NNT/NL : 02VEHT05P20/001ED184

THÈSE DE DOCTORAT

Soutenue à l'École Nationale des Sciences de L'Informatique (ENSI)
dans le cadre d'une cotutelle avec Aix-Marseille Université
le 24 Novembre 2023 par

Maha Mallek

Classification de relations d'un document textuel non structuré basée sur le contexte

Discipline
Informatique

École doctorale
Mathématiques et Informatique de Marseille Sciences et Technologies de l'Informatique, des COmmunicationsduDesignetdel'Environnement (STICODE)

Laboratoire de recherche
Laboratoire d'Informatique et Systèmes (LIS)
Unité de recherche LARIA

Composition du jury

Kais HADDAR Université de Sfax	Rapporteur
Lynda TAMINE-LECHANI Université Paul Sabatier	Rapporteuse
Antoine DOUCET La Rochelle Université	Examinateur
Henda BEN GHEZALA Université de la Manouba	Présidente du jury
Bernard ESPINASSE Aix Marseille Université	Directeur de thèse
Wided LEJOUAD CHAARI Université de la Manouba	Co-Directrice de thèse

Membres invités

Sébastien FOURNIER Aix Marseille Université	Encadrant
Ramzi GUETARI Université de Carthage	Co-encadrant

Affidavit

Je soussigné, Maha MALLEK, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Prof. Bernard ESPINASSE, la co-direction de Prof. Wided LEJOUAD CHAARI et le co-encadrement de Dr. Sebastien FOURNIER et Dr. Ramzi GUETARI dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte d'Aix-Marseille Université relative à la lutte contre le plagiat.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Tunis le 06 octobre 2022

Liste de publications et participation aux conférences

Liste des publications réalisées dans le cadre du projet de thèse :

1. **Maha Mallek**, Ramzi Guetari, Sébastien Fournier, Wided Lejouad Chaari et Bernard Espinasse, 'Relation Classification from Unstructured Text : a Systematic Literature Review' Knowledge and Information Systems, 2022. (première révision soumise le 17 octobre 2022, Impact Factor=2.9)
2. **Maha Mallek**, Sébastien Fournier, Ramzi Guetari, Bernard Espinasse et Wided Lejouad Chaari, 'An Unsupervised Approach for Precise Context Identification from Unstructured Text Documents', In Proceedings of 32nd IEEE International Conference on Tools with Artificial Intelligence, pp. 321-326, Baltimore, MD, USA, 2020. (Conférence classe B)
3. **Maha Mallek**, Ramzi Guetari, Sébastien Fournier, Bernard Espinasse et Wided Lejouad Chaari, 'Context–aware Relation Classification based on Deep Learning', In Proceedings of 34nd IEEE International Conference on Tools with Artificial Intelligence, 2022. (Conférence classe B)
4. **Maha Mallek**, Ramzi Guetari, Sébastien Fournier, Wided Lejouad Chaari et Bernard Espinasse, ' Accurate Context Extraction from Unstructured Text Based on Deep Learning', In Proceedings of IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, Niagara Falls, Canada / Virtual venue, 2022. (Conférence classe B)

Participation aux conférences et écoles d'été au cours de la période de thèse :

1. Participation à la conférence "32th International Conference on Tools with Artificial Intelligence (ICTAI)", 09-11 November 2020, conférence virtuelle
2. Participation à la conférence "32nd International Conference on Advanced Information Systems Engineering (CAISE 2020)", 08-12 juin 2020, conférence virtuelle
3. Assurer une formation à " l'Ecole d'été sur les sciences de données" durant la session " Intelligence artificielle", 23-25 août 2019, Palais des sciences, Mounastir, Tunisie.

Résumé

Cette thèse qui s'inscrit dans le domaine de l'extraction d'information, vise à améliorer la classification des relations sémantiques identifiées à partir de documents textuels non structurés. Plusieurs méthodes ont été proposées et mis en œuvre dans des systèmes spécifiques classant essentiellement les relations selon un certain nombre de types prédéfinis. Leur classement des relations est basé sur des aspects syntaxique du texte, sans prendre notamment en compte la sémantique du *contexte* du texte. Face à ces constats, le principal défi de cette thèse est la prise en compte du contexte associé au document dans la classification des relations qui en sont extraites.

Nous proposons tout d'abord *une méthode d'extraction du contexte d'un document textuel* et son implémentation avec deux variantes. Cette méthode repose sur l'identification des mots les plus importants caractérisant le contenu du document. Ces mots représentatifs sont ensuite utilisés pour définir la phrase du document donnant une brève idée de son contenu. Pour cela, le document est parcouru pour identifier la phrase qui contient le maximum des mots-clés ou leurs synonymes. Cette phrase est ensuite associée aux différents mots clés pour spécifier le contexte du document défini par son étiquette. Une première variante de la méthode réalise une extraction d'étiquette selon une approche *extractive*. Une seconde variante, améliorant la précédente, réalise cette extraction d'étiquette de façon *générative* en utilisant un réseau de neurones récurrent LSTM permettant, à partir des différents mots-clés identifiés, de de définir la phrase cohérente représentative du contenu du document. Les résultats de l'évaluation montrent que la variante *générative* est la plus performante et de plus améliore la qualité de la définition du contexte extrait ceci par rapport aux autres systèmes existants.

Nous proposons ensuite *une méthode de classification des relations d'un document textuel selon des types prédéfinis et surtout selon le contexte associé au document*. Cette méthode utilise les contextes extraits par la méthode précédente pour réaliser un filtrage pour éliminer les relations qui ne sont pas significatives pour le contexte du document. Cette méthode permet ainsi d'obtenir un degré de « contextualisation » des relations. Les résultats d'évaluation, sur les corpus SemEval-2010 Task-8, et sur une collection de données contextuelles, nommé WikiContext que nous avons construit à cet effet, montrent que notre système de classification de relations surpasse les systèmes de l'état de l'art, démontrant ainsi la pertinence de prendre en compte le contexte dans ce processus de classification.

Mots clés : Identification de contexte, Classification de relations, Extraction d'information, Apprentissage prodond

Abstract

This thesis fits into the field of information extraction. It aims at improving the relation classification in unstructured textual documents. Several methods have been proposed and implemented in specific systems that essentially classify relations into a number of predefined types, essentially based on syntactic aspects of the text. These methods neglect the semantic aspects and can be misleading during classification. A pillar of semantics is to take into account the *context* in the classification of relations. This is precisely the subject we address in this thesis and to which we propose some methods.

We first propose *a method for extracting the context of a given textual document* and its implementation with two variants. This method is based on the identification of the most important words that can best characterize the content of the document. These representative words are then used for the specification of the sentence that gives a brief idea of the document content. For this, the document is scanned to identify the sentence that contains the maximum number of keywords or their synonyms. This sentence is then associated with the different keywords to build the final context of the document defined by its label using an *extractive* approach. A second variant of the method, improving the previous one, performs a *generative* label extraction of the context using a recurrent neural network LSTM which allows, from the different identified keywords, to model a coherent sentence that represents the main idea of the document. The evaluation results show that this generative variant achieves a good performance while improving the quality of the extracted context definition compared to other existing systems.

We then propose *a method for classifying the relations of a textual document according to predefined types and especially according to the context associated with the document.* This method uses the contexts extracted by the previous method and perform a filtering process to eliminate the relations which are not significant for the context of the document. This method allows to obtain a degree of "contextualization" of relations. The evaluation results, on the e SemEval–2010 Task–8, New York Times corpora and a contextual dataset, named WikiContext, which we have built for this purpose, show that our system outperforms the state–of–the–art relation classification systems, thus demonstrating the relevance of taking context into account in this classification process.

Keywords: context Identification, Relation Classification, Information Extraction, Deep learning

Remerciements

Il m'est très agréable de dédier cette page comme un témoin de reconnaissance à toutes les personnes qui m'ont soutenu et encadré pour réaliser ce travail.

J'exprime, en premier lieu, ma profonde gratitude et reconnaissance à mes deux directeurs de thèse, **Pr. Wided Lejouad Chaari** et **Pr. Bernard Espinasse**, de m'avoir encadré et suivi tout au long de mes travaux de recherche. Je suis extrêmement reconnaissante, à vos encouragements, vos suggestions et vos orientations. Merci infiniment pour votre temps précieux, votre collaboration et votre attention m'ont largement appris pour l'élaboration de ce travail

Je tiens à remercier spécialement mon co-encadrant **Dr. Ramzi Guetari**, qui fut le premier à me faire découvrir le domaine de la recherche. Je désire exprimer, monsieur, mes remerciements les plus sincères pour votre générosité, votre patience tout au long de mes travaux de recherche.

Je tiens à présenter mes remerciements à mon co-encadrant de thèse,**Dr. Sébastien Fournier** pour votre soutien, encouragement et rigueur scientifique qui m'ont permis de mener à bien ce travail.

Je remercie également **Pr. Kais Haddar**, ainsi que **Pr. Lynda Tamine**, de l'honneur qu'ils m'ont fait en acceptant d'être rapporteurs de cette thèse.

Je remercie **Pr. Henda Ben Ghezala** de l'honneur qu'elle m'a accordé en acceptant d'examiner ma thèse et de présider le jury de ma soutenance.

Je tiens à remercier aussi **Pr. Antoine Doucet** pour l'honneur qu'il m'a accordé en acceptant d'examiner mon travail.

J'adresse mes plus vifs remerciements à **Mr. Nabil Morgham** pour votre coopération et votre aide surtout dans les moments difficiles.

Je voudrais profiter de cette occassion pour exprimer mes plus profonds remerciements à toute l'équipe de LIS pour leur accueil chaleureux. Je tiens également à témoigner toute ma reconnaissance à tous les membres de l'Ecole Nationale des Sciences de l'Informatique et toute l'équipe du laboratoire LARIA (**L**aboratoire de **R**echerche en **I**ntelligence **A**rtificielle).

J'adresse toutes les expressions de remerciement à mes chers amis et collègues qui m'ont apporté leur soutien moral et intellectuel tout au long de ma démarche. Je remercie particulièrement **Nour El Houda Ben Chaabane**, **Ibtissem Daoudi**, **Emna Hosni**, **Zayneb Gharsalla** et **Rym Alouane**.

Un grand merci à mon père **Abdelaziz** et ma mère **Raoudha**, pour leur amour, leurs conseils, leurs prières ainsi que leur soutien inconditionnel, à la fois moral et financier, qui m'a permis de réaliser les études que je voulais et par conséquent cette thèse de doctorat. Mes remerciemenst vont aussi à mon frère **Mejdi** et sa femme **Amal** pour leur appui continu et leurs mots encourageants.

Je remercie mon cher mari **Monem** pour son soutien quotidien indéfectible et son enthousiasme contagieux à l'égard de mes travaux comme de la vie en général. Ces remerciements ne peuvent s'achever, sans une pensée pour mon coeur et mon adorable petit prince **Hedi**. J'espère que tu sera fier de moi comme je le suis.

Je remercie spécialement ma belle soeur **Wafa** qui a toujours été là pour moi. Votre encouragement a été d'une grande aide.

Je remercie de tout cœur mes beaux-parents **Rachid** et **Mounira** aini que mes belles-soeurs **Neila**, **Imen** et **Mariem** pour leur soutien inconditionnel.

J'aimerais exprimer ma gratitude à tous les chercheurs et spécialistes, trop nombreux pour les citer, qui ont pris le temps de discuter de mon sujet. Chacun de ces échanges m'a aidé à faire avancer mon analyse.

Table des matières

Table des figures

Liste des tableaux

Introduction Générale

Contexte de la thèse

Notre travail s'inscrit dans le domaine très actif du *traitement automatique des langues naturelles* (TALN) et plus précisément dans le domaine de *l'extraction d'information* (EI). Naturellement, l'intérêt porté aujourd'hui à l'extraction d'information est le fruit de nombreux travaux qui cherchent à améliorer les techniques utilisées dans ce domaine. De nos jours, et grâce au Web, l'utilisation des technologies de l'information a entraîné une explosion du volume de données échangées entre particuliers, entreprises, etc. et a même conduit à un changement profond dans tous les aspects économiques, politiques et sociaux. Cette surcharge continue en croissance constante du volume de données hétérogènes entraîne une plus grande complexité dans la recherche et l'extraction d'information.

Face à cette énorme quantité de données bruitées, l'extraction de l'information pertinente devient un problème critique. L'un des grands défis du XXIème siècle « connecté » est la gestion de grands volumes de données complexes qui sont souvent faiblement structurées (HOLZINGER 2012, HOLZINGER 2011), voire pas du tout structurées (HOLZINGER, STOCKER, OFNER et al. 2013). Les données non structurées concernent le texte brut écrit en langue naturelle, les images, le son et la vidéo.

Notre problématique concerne essentiellement les données textuelles non structurées dont le volume augmente de manière constante et spectaculaire d'une année à l'autre. Ces données données non structurées peuvent véhiculer des informations essentielles et nécessiter un effort d'extraction de l'utilisateur. Cette extraction reste du domaine du possible tant que la quantité de ces données reste raisonnable. Cependant, la capacité humaine devient impuissante dès lors qu'il s'agisse de traiter un flux croissant et continu de données comme celui qui transite à travers les médias modernes (indexation, appariement, recherche etc.). Cette extraction est couteuse en termes de temps et de ressources, le taux d'erreur est élevé et les données doivent être constamment mises à jour. Le challenge consiste alors à rendre cette extraction d'information (semi)-automatique, tout en étant précise et pertinente.

Afin de produire un système (semi)-automatique capable de cette extraction d'information pour la classifier et éventuellement en extraire des connaissances pertinentes, une reconnaissance d'entités nommées et l'extraction de relations entre ces entités sont nécessaires. L'extraction de relations permet de découvrir des liaisons sémantiques entre des entités nommées dans un texte brut, sans poser d'hypothèses à priori sur ces relations. Les relations sont l'un des intérêts majeurs pour la quasi-totalité des applications de traitement automatique des langues naturelles. L'extraction des

relations ainsi que leur classification constituent en conséquence un enjeu majeur pour l'extraction d'information, voire l'acquisition de connaissances à partir de textes.

Problématique de la thèse

La problématique de ce travail de recherche consiste à identifier l'information pertinente à partir d'un texte écrit en langue naturelle, ceci par rapport à un contexte donné. Les données échangées à travers différents types de médias véhiculent des informations vitales ne pouvant être acquises qu'après une lecture et une analyse par l'utilisateur. Des estimations indiquent que seulement 15% des données publiées sur le Web sont structurées. Les 85% restants sont semi-structurés ou non structurés (HILBERT 2016).

Les *données structurées* sont généralement présentées sous forme de tableaux où l'association colonne/ligne représente une relation exprimant une certaine sémantique. Le traitement des données structurées est facilement compréhensible et l'automatisation de ce traitement ne présente aucune difficulté particulière.

Les *données non struturées* sont disponibles sous la forme de sources de données complexes, telles que les journaux Web, les courriers électroniques et les données de médias sociaux. Le traitement automatique des données non structurées est difficile et nécessite des techniques beaucoup plus élaborées que celles requises pour le traitement de données structurées. Puisque les relations permettent de faciliter la compréhension et l'automatisation du traitement de l'information, la question qui se pose est « y-a-t-il un moyen d'identifier automatiquement les relations dans des données brutes afin de faciliter la compréhension et le traitement de l'information associée ? ».

L'extraction et la classification des relations d'un document textuel sont des tâches cruciales, car situées en amont d'autres tâches comme l'expansion sémantique en recherche d'information ou encore l'extraction de relations pour la construction de ressources sémantiques. Afin de classer des relations automatiquement, plusieurs approches ont été expérimentées avec des résultats plus ou moins convaincants. Ainsi divers systèmes ont été proposés, utilisant généralement des méthodes supervisées (BACH et BADASKAR 2007), qui peuvent être basées (i) sur des règles conçues par l'homme, sur l'apprentissage relationnel logique comme la programmation logique inductive (LIMA, ESPINASSE et FREITAS 2019), sur l'apprentissage automatique statistique utilisant des méthodes basées sur les caractéristiques ou les noyaux (ZELENKO, AONE et RICHARDELLA 2003a) ou, plus récemment, sur l'apprentissage profond. De nombreux travaux en classification de relations, concernent l'utilisation de ce type l'apprentissage incluant les réseaux de neurones convolutionnels (CNN) (P. QIN, W. XU et J. GUO 2016), les réseaux de neurones récurrents (RNN) (Runyan ZHANG, MENG, Y. ZHOU et al. 2018), et les réseaux de mémoire à long terme (LSTM) Y. XU, MOU, G. LI et al. 2015. Plusieurs autres auteurs, comme (DHIWAR et DEWANGAN 2016), ont proposé des réseaux neuronaux profonds (DNN) pour classer les relations.

Quelles que soient les méthodes utilisées, ces systèmes classent les relations en un

certain nombre de types prédéfinis, essentiellement basés sur les aspects syntaxiques du texte. Ces méthodes négligent les aspects sémantiques et peuvent être trompeuses lors de la classification.

Ce travail avance l'hypothèse que l'extraction et la classification des relations entre entités nommées d'un document textuel sera plus pertinente si l'on prend en compte le contexte associé à ce document, ce contexte cristallisant ainsi une certaine sémantique. En effet, Considérons les deux phrases suivantes : « *Donald Trump est un candidat de l'élection présidentielle américaine de 2016* » et « *Le chauffard est un candidat au suicide* ».

Dans ces deux phrases, la même relation « *est un candidat* » exprime deux choses différentes, c'est- à-dire deux contextes différents, la première relation exprime le contexte « *élection présidentielle américaine de 2016* », et la seconde exprime le contexte « *Sécurité routière* ». C'est l'une des raisons pour lesquelles le contexte est un facteur important dans le processus d'extraction et de classification des relations. La question qui se pose alors est alors la suivante : comment trouver des moyens pour extraire et classer des relations pertinentes entre deux entités par rapport à un contexte d'une manière non-supervisée ou faiblement semi-supervisée ?

Principaux objectifs de la thèse

L'Extraction d'Information est désormais un sujet de recherche important dans le domaine de la fouille de textes ou « Text Mining ». Elle connaît ces dernières années un intérêt grandissant car elle répond à un besoin devenu incontournable dans la société de l'information. En terme applicatif, notre recherche vise à améliorer plusieurs processus notamment :

- **Le recueil de renseignements fiables par la découverte de termes et de concepts présents dans des documents :** prenons à titre d'exemple, le domaine de l'économie, les entreprises ont en permanence besoin d'informations fiables et pertinentes sur les marchés ainsi que sur leurs concurrents afin d'élaborer les stratégies leur permettant d'améliorer leurs résultats et de gagner des parts de marché.
- **La détection de variables statistiques à partir de textes :** prenant l'exemple d'un forum politique dans le Web qui porte sur l'élection présidentielle dans un pays donné. Les utilisateurs de ce média social peuvent y accéder pour exprimer leurs opinions. L'extraction et la classification des relations selon le contexte de ce forum amène à discerner et comprendre les opinions des gens. Ce type d'analyse permet d'extraire les statistiques des votes et ainsi permet au public d'avoir une idée précise sur les tendances de vote.
- **L'extraction de l'information pertinente à partir du Web :** En effet, pour effectuer une recherche sur le Web, les systèmes doivent extraire des données provenant de sources multiples. Toutefois, pour avoir l'information souhaitée, le système doit passer par deux étapes : l'indexation du corpus de documents et l'interrogation de la base documentaire construite par l'appariement requête-document. Cette méthode ne considère pas le sujet de la source et son contexte.

Cependant, Si on arrive à extraire les relations pertinentes par rapport à un contexte précis, il est évident qu'un tel traitement en amont de l'indexation dans des moteurs de recherche permettrait d'améliorer largement la qualité des résultats retournés lors d'une recherche en y associant un contexte, ce qui permettrait de diminuer les Faux Positis (FP) et Faux Négatifs (FN) en même temps.

- **La structuration de textes :** L'idée est de réduire les données contenues dans les documents en une structure plus facilement utilisable que le texte brut. À notre avis, le meilleur moyen de résumer ou de synthétiser des informations statistiques contenues dans des documents texte non structurés consiste tout d'abord à créer des données structurées (telles que des tableaux) contenant des données statistiques compatibles avec celles contenues dans le texte non structuré. La structuration des données dans un tableau est alors un outil efficace pour communiquer les résultats de sondages ou d'enquêtes statistiques par exemple. Cela permet, entre autres, une lecture plus rapide des informations, l'indépendance de la langue (même de la culture), une meilleure capture de l'attention du public, etc.

Contributions

Dans cette thèse, pour un document textuel donné, notre recherche s'intéresse à la classification des relations qui en sont extraites selon le *type* et surtout le *contexte* associé à ce document. Il s'agira tout d'abord d'extraire le contexte d'un document textuel non structuré. Les relations de ce document étant extraites il s'agira de les classer selon ce contexte tout en appliquant un processus de filtrage pour éliminer les relations qui ne sont pas significatives pour le contexte du document. Dans cette recherche, nos contributions peuvent être résumées ainsi :

Contribution 1 : une nouvelle classification des méthodes d'extraction et de classification des relations
L'étude bibliographique sur le domaine d'extraction et de classification des relations, effectuée au début de la thèse, a été exploitée pour proposer une nouvelle classification des méthodes proposées dans ce domaine. Cette classification sera d'un grand intérêt aux chercheurs pour obtenir un aperçu de la plupart des techniques d'extraction et de classification des relations proposées avec leurs avantages et leurs inconvénients, leurs performances relatives. Cette première contribution fait l'objet d'un article en cours de révision au journal Knowledge and Information Systems (KAIS) :
***M. Mallek**, R. Guetari, S. Fournier, W. Lejouad Chaari, B. Espinasse. Relation Classification from Unstructured Text : a Systematic Literature Review, Knowledge and Information Systems, 2022. (Première révision soumise le 17 octobre 2022)*

Contribution 2 : une méthode d'extraction du contexte d'un document textuel
Une deuxième contribution porte sur la proposition d'une méthode d'extraction du

contexte d'un document textuel donné et son implémentation avec deux variantes. Cette proposition repose sur l'identification des mots-clés les plus importants qui peuvent mieux caractériser le contenu du document. Ces mots représentatifs sont ensuite utilisés pour la spécification de la phrase qui donne une brève idée du contenu de document. Pour ce faire, le document est parcouru pour identifier la phrase qui contient le maximum des mots-clés ou leurs synonymes. Cette phrase représente le contexte final du document. La première variante de la méthode utilise une extraction du contexte extractive, et a fait l'objet d'un article publié à la conférence ICTAI 2020 : ***M. Mallek**, S. Fournier, R. Guetari, B. Espinasse and W. Lejouad Chaari, An Unsupervised Approach for Precise Context Identification from Unstructured Text Documents, ICTAI 2020 – 32th International Conference on Tools with Artificial Intelligence, November 09-11, 2020*

Une seconde variante de la méthode d'extraction du contexte proposée, améliorant la précédente, utilise une extraction de contexte *générative* en intégrant un réseau de neurones récurrent LSTM qui permet, à partir des différents mots-clés identifiés, de modéliser une phrase cohérente représentant l'idée principale du document. Les expérimentations montrent que cette variante *générative* est plus performante que la variante extractive précédente ainsi que les autres méthodes de la littérature en termes de mesure ROUGE, et de plus améliore la qualité du contexte extrait. Cette variante *générative* de la méthode a fait l'objet d'un article publié à la conférence WI-IAT 2022 : ***M. Mallek**, R. Guetari, S. Fournier, W. Lejouad Chaari, B. Espinasse. Accurate Context Extraction from Unstructured Text Based on Deep Learning, 21st IEEE/WIC/ACM WI-IAT Conference, WI-IAT 2022, November 17-20, 2022, Niagara Falls, Canada.*

Contribution 3 : une méthode de classification des relations selon le type et le contexte

Notre troisième contribution est la proposition d'une méthode de classification des relations d'un document textuel selon des types prédéfinis et surtout selon le contexte associé au document. Cette méthode utilise le contexte extraits du document par la méthode précédente pour réaliser un filtrage pour éliminer les relations qui ne sont pas significatives pour le contexte du document. Cette méthode permet ainsi d'obtenir un degré de « contextualisation » des relations. Les résultats d'évaluation, sur le corpus SemEval-2010 Task-8 et le corpus WikiContext, que nous avons construit à cet effet, montrent que le système implémentant notre méthode surpasse les systèmes de classification de relations de l'état de l'art, démontrant ainsi la pertinence de prendre en compte le contexte dans ce processus de classification. Cette contribution a fait l'objet d'un article publié à la conférence ICTAI 2022 :
***M. Mallek**, R. Guetari, S. Fournier, W. Lejouad Chaari, B. Espinasse. Context–aware Relation Classification based on Deep Learning, 34th International Conference on Tool with Artificial Intelligence, ICTAI 2022 , 31 Oct.-2 Nov. 2022*

Organisation du manuscrit

Ce manuscrit s'articule autour de quatre chapitres : les deux premiers chapitres constituent un état de l'art, et les deux autres développent nos contributions.

Chapitre 1 : De l'extraction à la classification de relations d'un document textuel

Ce chapitre introduit tout d'abord les principaux concepts de l'extraction d'information puis présente les différentes méthodes liées à l'extraction et surtout à la classification des relations. La classification de relations étant le cœur de notre recherche, nous développons, en complément de la présentation de ces méthodes associées, une revue systématique de la littérature ciblée sur la classification de relations. Enfin, nous concluons ce chapitre par un bilan sur les méthodes de classification de relations, en évoquant quelques perspectives.

Chapitre 2 : Extraction du contexte d'un document textuel

Ce chapitre introduit tout d'abord la notion de contexte et présente les travaux connexes qui ont abordé des aspects similaires à ceux liés à la notion de « contexte » et qui ont proposé des méthodes d'extraction de contenu à partir de documents textuels. En complément de la présentation de ces méthodes, nous réalisons une revue systématique de la littérature ciblée sur l'extraction de notions similaires à celle de « contexte », en adoptant la même méthodologie que celle utilisée dans le chapitre précédent pour la classification de relations. Enfin, nous concluons ce chapitre en présentant les limites majeures de ces travaux existants et en dessinant les grands axes d'une nouvelle méthode d'extraction de contexte permettant de dépasser ces limites.
Chapitre 3 : Une méthode d'extraction du contexte d'un document texte

Le troisième chapitre porte sur la présentation de notre deuxième contribution, à savoir une méthode d'extraction du contexte d'un document textuel avec ses deux variantes liées à une extraction de contexte soit *extractive*, soit *générative*. Ce chapitre défini tout d'abord les concepts fondamentaux utilisés dans notre approche notamment la notion de contexte. Puis, nous détaillons les différentes étapes de notre méthode d'extraction du contexte ainsi que son implémentation logicielle. Ensuite, nous procédons à des tests de comparaison des résultats de notre approche avec d'autres solutions existantes de l'état de l'art. Finalement, les différentes interfaces de notre application Web réalisée « EASYContext » ont été présentées.

Chapitre 4 : Une méthode de classification des relations d'un document textuel selon le type et le contexte

Le quatrième chapitre présente notre troisième contribution consistant en une méthode de classification des relations d'un document textuel selon le type et le contexte,

ainsi que sa mise en œuvre. Cette méthode a été développée en utilisant plusieurs techniques et algorithmes de l'intelligence artificielle et de la fouille de données. Elle utilise les contextes extraits par la méthode d'extraction de contextes précédente pour réaliser un filtrage pour éliminer les relations qui ne sont pas significatives pour le contexte du document. La première section défini la notion de relation, la deuxième présente la méthode proposée de classification des relations selon le type et le contexte en détaillant ses différentes étapes ainsi que leur implémentation, et la troisième section présente les tests réalisés pour l'évaluation de notre méthode et la comparaison des résultats obtenus avec ceux d'autres systèmes de l'état de l'art. La dernière section présente un exemple d'application de l'approche pour la visualisation automatique des données statistiques extraites du texte.

Ce manuscrit se termine par une conclusion générale qui rappelle les contributions apportées, leurs limites, et présente un certain nombre de perspectives.

1. De l'extraction à la classification de relations d'un document textuel

Sommaire

1.1. Introduction

De nos jours, l'explosion quantitative des données numériques a donné naissance à un grand phénomène, appelé BIG DATA. Littéralement, ce terme désigne les méga données, de grosses données ou aussi des données massives. Toute cette information constituée par des millions d'internautes représente un volume de contenu colossal disponible et accessible sur la toile.

Cependant le problème rencontré ne porte pas tant sur la gestion de ce gigantesque volume ou son interrogation, que celui du traitement d'une masse de données principalement non structurées.

Plusieurs méthodes et techniques permettant d'extraire des informations intéressantes d'un document ont été proposées, elles relèvent du domaine de l'extraction d'information (EI).

Ce premier chapitre de la thèse vise tout d'abord à introduire les principaux concepts clé liés au domaine de l'EI et présenter les différentes techniques liées à l'extraction et surtout la classification des relations. A cet effet, nous consacrons la première section à définir l'extraction d'information et à évoquer quelques applications représentatives dans le domaine de l'EI. La deuxième section présente en détail les différentes méthodes utilisées pour *l'extraction* ou *identification* de relations tandis que la section quatre présente les différentes méthodes de *classification* de relations. La classification de relations étant le coeur de notre recherche, nous développons dans la section cinq, en complément de la présentation des techniques associées présentées dans la section précédente, une revue systématique de la littérature ciblée sur la classification de relations. Enfin, nous concluons ce chapitre par un bilan sur les méthodes de classification de relations, leur limites et en évoquant quelques perspectives.

1.2. Extraction d'information

L'Extraction d'Information (CAFARELLA, BANKO et Oren ETZIONI 2011, MCCALLUM 2005) ou EI (en anglais, Information Extraction ou IE) est désormais un sujet de recherche important dans le domaine du Traitement Automatique des Langues Naturelles. Elle connaît ces dernières années un intérêt grandissant car elle répond à un besoin devenu incontournable dans la société de l'information.

L'EI est un processus qui consiste à extraire automatiquement des informations structurées de documents non structurés et/ou semi-structurés lisibles par machine. Dans la plupart des cas, cette activité concerne le traitement de textes en langage humain au moyen du traitement du langage naturel (NLP). Semblable à un système de recherche d'information (RI), un système d'extraction d'information répond au besoin d'information d'un utilisateur. Alors qu'un système de RI identifie un sous-ensemble de documents dans une grande base de données textuelles, un système d'EI identifie un sous-ensemble d'informations dans un document. Ce sous-ensemble d'informations représente généralement un résumé ou l'essentiel du contenu du document. Traditionnellement on distingue deux tâches principales en IE : *l'identification d'entités nommées* (Named Entity Recognition - NER) et sur *l'extraction de relations entre ces entités* (Relation Extraction – RE). Nous verrons par la suite que l'on distingue *l'identification* de relations de la *classification* de ces dernières.

1.2.1. Identification des entités nommées

Une sous-tâche très importante dans l'IE est de trouver et de classer les entités dans le texte. Les entités sont généralement des groupes nominaux. La forme la plus populaire d'entités est celle des entités nommées. Les entités nommées sont des éléments provenant de données textuelles appartenant à des catégories prédéfinies telles que les noms de personnes, les organisations, les emplacements, les valeurs monétaires, les pourcentages, popularisés par les compétitions MUC (CHINCHOR 1998, R. GUPTA, DIWAN et SARAWAGI 2007), ACE (DODDINGTON, MITCHELL, PRZYBOCKI et al. 2004) et CoNLL (SANG et DE MEULDER 2003).

Aujourd'hui, le terme d'entité est élargi pour inclure également des génériques comme les noms de maladies, les noms de protéines, les titres d'articles et les noms de revues. Le compétitions ACE pour l'extraction de relations entre entités dans des textes en de textes en langue naturelle répertorie plus de 100 types d'entités différents. La reconnaissance des entités nommées est une sous-tâche d'extraction d'informations qui cherche à identifier ces entités. Il s'agit d'annoter les différentes entités nommées d'un document afin d'identifier les relations sémantiques.

1.2.2. Extraction des relations

Une relation est définie par la cooccurrence de deux ou plusieurs éléments dans une phrase. Dans les systèmes d'extraction d'information, ces éléments peuvent être soit des entités nommées (HASEGAWA, SEKINE et GRISHMAN 2004), c'est-à-dire des éléments de la phrase qui ont été identifiés comme des noms propres tels que des noms de personnes, de lieux, d'organisations, etc., soit des syntagmes nominaux (ROZENFELD et FELDMAN 2006a), c'est-à-dire des groupes de mots contenant un nom ou un pronom comme noyau. Les entités nommées sont souvent privilégiées pour la classification des relations, car elles permettent une meilleure séparation entre les différents types de relations possibles. Cependant, l'utilisation de syntagmes nominaux peut permettre d'augmenter le nombre de candidats potentiels pour chaque relation, car elle permet de couvrir un éventail plus large de termes et de formulations qui peuvent être pertinents pour une relation donnée. Nous nous intéressons dans notre cas aux relations entre entités (syntagmes nominaux).

Plus formellement, les relations extraites des textes, que l'on devrait en toute rigueur appeler instances de relations, sont caractérisées par deux grandes catégories d'information : un couple d'entités (e1, e2) et la partie de la phrase se situant entre les deux entités. Ces catégories permettent tout à la fois de les définir et de fournir les éléments nécessaires à leur regroupement.

Le processus *d'extraction de relations* se décompose, comme l'illustre la Figure 1.1, autour de deux grandes étapes : *l'identification* des relations et leur *classification*.

L'identification des relations vise à extraire du texte les relations sémantiques qui existent généralement entre deux ou plusieurs entités, tandis que la *classification* des relations consiste à associer une étiquette à une relation extraite afin d'identifier la classe à laquelle cette relation appartient.

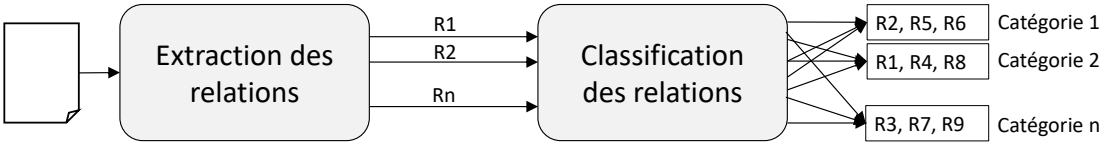


FIGURE 1.1. – Processus d'extraction des relations

1.2.3. Applications de l'Extraction d'information

L'extraction d'information est utile dans différentes applications. Nous présentons un sous-ensemble de ces applications classées selon qu'elles sont destinées aux entreprises, aux particuliers, aux scientifiques ou au Web.

— **Applications destinées aux Entreprises :** Toute entreprise orientée client collecte de nombreuses formes de données non structurées provenant de l'interaction avec le client. Pour une gestion efficace, ces données doivent être étroitement intégrées aux bases de données structurées et aux ontologies commerciales de l'entreprise. Ceci a donné lieu à de nombreux problèmes d'extraction intéressants tels que l'identification de noms et d'attributs de produits à partir de courriels de clients, la liaison de courriels de clients à une transaction spécifique dans une base de données de ventes (BHIDE, A. GUPTA, R. GUPTA et al. 2007, CHAKARAVARTHY, H. GUPTA, ROY et al. 2006), l'extraction de noms et d'adresses de commerçants à partir de factures de ventes (G. ZHU, BETHEA et KRISHNA 2007), l'extraction d'humeurs de clients à partir de transcriptions de conversations téléphoniques (JANSCHE et ABNEY 2002), et l'extraction de paires de valeurs d'attributs de produits à partir de descriptions textuelles de produits (GHANI, PROBST, Yan LIU et al. 2006).

— **Applications destinées aux particuliers :** Les systèmes de gestion des informations personnelles (PIM) cherchent à organiser les données personnelles telles que les documents, les courriers électroniques, les projets et les personnes dans un format structuré et interconnecté (Y. CAI, DONG, HALEVY et al. 2005, CHAKRABARTI, MIRCHANDANI et NANDI 2005, **16**). Le succès de ces systèmes dépendra de la capacité d'extraire automatiquement la structure des sources non structurées existantes, principalement basées sur des fichiers. Ainsi, par exemple, nous devrions être en mesure d'extraire automatiquement d'un fichier PowerPoint l'auteur d'une conférence et de lier cette personne au présentateur d'une conférence annoncée dans un courriel. Les courriels, en particulier, ont servi pour de nombreuses tâches d'extraction, telles la déduction de types de demandes dans des centres de service (COHEN, MINKOV et TOMASIC 2005).

— **Applications destinées aux scientifiques :** Par exemple, le domaine de la bioinformatique a élargi le champ des extractions précédentes d'entités nommées, aux objets bio- logiques tels que les protéines et les gènes. Un problème central est l'extraction des noms de protéines et de leurs interactions à partir de de documents tels que Pubmed (BLASCHKE et al. 2004, BUNESCU, GE, KATE et al. 2005, PLAKE, SCHIEMANN, PANKALLA et al. 2006). Comme la forme des entités

telles que les noms de gènes et de protéines est très différente de celle des entités nommées classiques telles que les personnes et les entreprises, cette tâche a permis d'élargir les techniques utilisées pour l'extraction d'information.

— **Applications destinées au Web :** Il existe plusieurs sites Web qui stockent des opinions non modérées sur toute une série de sujets, notamment des produits, des livres, des films, des personnes et de la musique. Beaucoup de ces opinions sont sous forme de texte non structuré, cachées derrière des blogs, des messages de groupes de discussion, des sites d'évaluation, etc. La valeur de ces avis peut être considérablement accrue s'ils sont organisés selon des champs structurés. Par exemple, pour les produits, il pourrait être utile de trouver, pour chaque caractéristique du produit, la polarité prédominante de l'opinion (Bing LIU, M. HU et J. CHENG 2005, POPESCU et Orena ETZIONI 2007).

Dans les sections suivantes, nous nous intéresseront essentiellement aux différentes méthodes utilisées dans *l'extraction (identification)* des relations, puis dans la *classification* de relations.

1.3. Méthodes d'identification des relations

Il existe plusieurs approches de méthodes d'extraction des relations. On distinguera les approches symboliques, statistiques et hybrides, comme l'illustre la figures 1.2.

1.3.1. Approches symboliques

Les systèmes symboliques, ou systèmes à base de règles, reposent sur un ensemble de règles ou de patrons : une règle est en général, dans ce contexte, assimilable à un patron contextuel, le plus souvent composé de mots et d'autres attributs issus de traitements linguistiques. Les règles peuvent être écrites manuellement par des experts humains ou alors acquises automatiquement notamment par apprentissage symbolique.

Les premiers systèmes développés pour l'extraction d'information étaient des systèmes symboliques à base de règles symboliques écrites par des experts humains. Ces systèmes étaient précis mais demandaient un effort significatif des experts humains pour définir ces règles et les maintenir, et présentaient des limitations dans leur couverture . Parmi les approches utilisées pour l'extraction des relations en utilisant des modèles symboliques, on peut citer (1) les approches basées sur les « patterns » et (2) les approches basées sur la programmation logique inductive :

1. **Méthodes basées sur les « patterns » :** Parmi Les méthodes basées sur les « patterns », on peut citer « DIPRE » et « SnowBall ». Brin (BRIN 1998) a proposé le premier algorithme de ce type nommé « DIPRE » (Dual Iterative Pattern Relation Expansion). Le système « DIPRE » permet d'extraire des relations entre les titres de livres et leurs auteurs. Il s'agit d'utiliser des instances de relations sous la forme de couples d'entités pour retrouver des occurrences de relations contenues dans des documents. Ces occurrences servent à générer des patrons

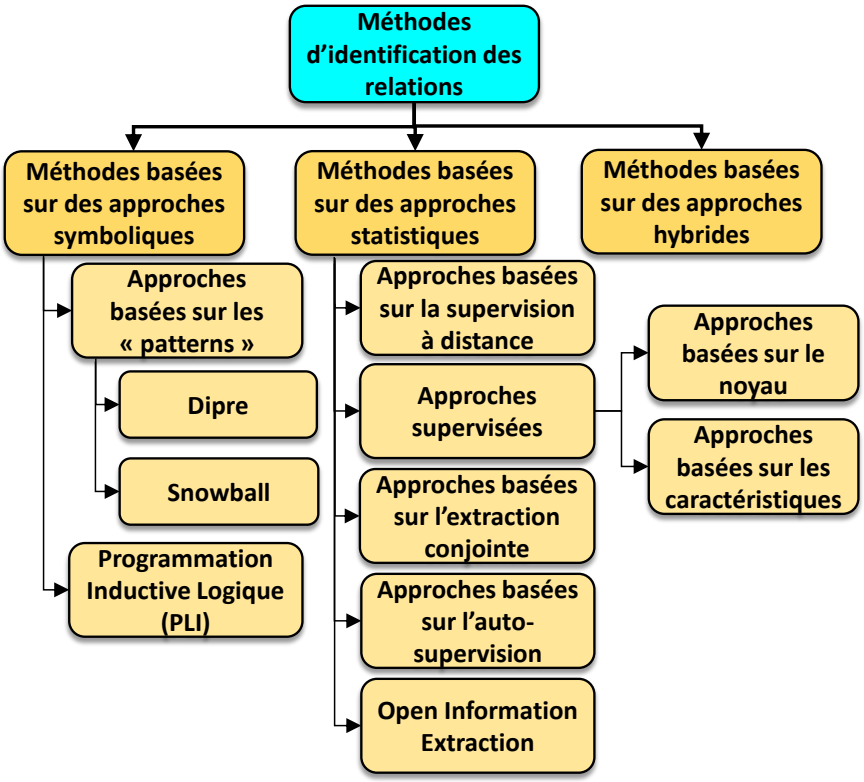


FIGURE 1.2. – Méthodes utilisées pour l'identification des relations

représentatifs des relations. Par la suite les patrons sont appliqués afin de trouver de nouvelles instances de relations. Ces nouvelles instances sont réintroduites dans le système pour une nouvelle itération. Le processus se termine lorsque le nombre d'instances renvoyé par le système est supérieur à une limite prédéfinie. A la manière de « DIPRE », (ZELENKO, AONE et RICHARDELLA 2003b) propose un système d'extraction de relations nommé « SnowBall » qui s'appuie sur un ensemble d'exemples de relations sous forme de couples d'entités. La distinction entre « DIPRE » et « SnowBall » se fait déjà au niveau de la génération des patrons de relations. Dans « SnowBall » les patrons sont aussi représentés par des tuples contenant les parties (prefix, middle, suffix), mais cette fois les patrons sont générés de façon itérative. Les mots contenus dans chacune de ces parties prefix, middle, suffix sont pondérés à chaque itération par le nombre d'occurrences du mot dans la partie concernée. Dans « SnowBall », le regroupement se fait en prenant en compte les parties préfixes et suffixes. Un score de similarité est attribué à chaque paire d'occurrences de relations. En pratique, le calcul de similarité se base sur le produit scalaire des vecteurs de composantes (préfixe, contexte, suffixe). Chaque groupe d'occurrences sert ensuite à la génération d'un patron et un score de confiance est associé à chaque patron généré. Pour ce qui concerne l'extraction de nouvelles relations, « SnowBall » commence par rechercher des occurrences potentielles de relations, ces occurrences sont transformées sous

forme de « tuples » et ensuite comparées aux patrons préalablement générés. Lors de la comparaison des occurrences avec les patrons, les occurrences ayant un score de similarité inférieur à un seuil donné sont éliminées. Enfin, parmi les occurrences conservées, celles qui sont issues de patrons ayant un score de confiance élevé et issues d'un grand nombre de patrons sont conservées.

2. **Programmation Logique Inductive :** D'autres auteurs utilisent « La programmation logique inductive » comme une méthode technique symbolique pour extraire par apprentissage symbolique des entités nommées et des relations. La programmation logique inductive permette de trouver des modèles à partir de données stockées dans des structures de don- nées complexes. Ces modèles sont utilisés pour classer de nouveaux exemples en positif ou en négatif. Comme Lima et al (LIMA, ESPINASSE et FREITAS 2015, LIMA, ESPINASSE et FREITAS 2018) présentent « OntoILPER », un système d'extraction d'information utilisant des ontologies et la programmation logique inductive une technique d'apprentissage symbolique, pour induire des règles d'extraction symboliques interprétables par des humains. « OntoILPER » utilise l'ontologie du domaine et profite d'un espace d'hypothèse relationnelle pour représenter des exemples dont la structure est pertinente pour l'extraction d'information. De plus, « OntoILPER » permet l'exploitation de l'ontologie du domaine et d'autres connaissances de base sous la forme de fonctionnalités relationnelles.

1.3.2. Méthodes basées sur des approches statistiques

Ces méthodes peuvent être divisées en cinq catégories : (1) Approches supervisées, (2) approches basées sur la supervision distante (Distantly supervised approaches), (3) l'extraction conjointe des entités et des relations (Joint Extraction), (4) approches basées sur l'auto-supervision (Self-supervised approaches) et (5) Open Information Extraction,

1. **Les approches supervisées :** Ces approches sont souvent séparées en deux catégories : les *méthodes basées sur les caractéristiques (Feature-based)* et les *méthodes basées sur les noyaux (Kernel-based)*.

   Les *méthodes basées sur les caractéristiques* pour l'extraction de relation utilisent des attributs explicitement syntaxiques et sémantiques. Ces caractéristiques extraites servent pour décider si les entités d'une phrase sont liées ou non. Les « caractéristiques » syntaxiques extraites d'une phrase sont obtenues en utilisant les techniques de traitement automatique de langues naturelles (NLP). Elles comprennent les différentes entités, leurs types, la séquence de mots entre les entités, leur nombre ainsi que le chemin dans l'arbre d'analyse contenant les deux entités. Les caractéristiques sémantiques comprennent le chemin entre les deux entités dans l'analyse de dépendance. Dans ce contexte, Kambhatla (KAMBHATLA 2004) a utilisé des modèles d'entropie maximale pour combiner diverses caractéristiques lexicales, syntaxiques et sémantiques déri-

vées du texte pour extraire des relations sémantiques. En effet, il a formé un classifieur d'entropie maximale avec 49 classes : deux pour chaque sous-type de relation (ACE 2003 a 24 sous-types de relation et chaque sous-type donne naissance à 2 classes en considérant l'ordre des arguments de relation) et une classe NONE pour le cas où les deux mentionnent ne sont pas liés. S'appuyant sur le travail de Kambhatla et al. (KAMBHATLA, FREITAG, MCCALLUM et al. 2004) a exploré quelques caractéristiques supplémentaires pour améliorer davantage les performances de l'extraction de relations. Zhou et al. (GuoDong ZHOU, SU, Jie ZHANG et al. 2005) ont incorporé les informations de segmentation de la phrase de base pour améliorer l'efficacité de l'extraction des relations tout en utilisant des caractéristiques d'arbre d'analyse. Ils ont obtenu de meilleures performances que le système de Kambhatla. Certaines caractéristiques plus intéressantes sont décrites par Nguyen et al. (D. P. NGUYEN, MATSUO et ISHIZUKA 2007) qui ont utilisé SVM pour identifier les relations entre les entités Wikipédia. Ils ont créé de façon semi-automatique une liste de mots-clés fournissant des indices pour chaque type de relation. D'autres auteurs s'intéressent à l'extraction des relations dans le domaine biomédical. Gu et al. (GU, QIAN et Guodong ZHOU 2016) décrit un système basé sur l'apprentissage automatique d'extraction des relations de maladies induites par des produits chimiques sur BioCreative-V Track-3b. Il qui utilise des caractéristiques lexicales simples mais efficaces. Des paires de mentions de produits chimiques et de maladies sont d'abord construites comme des instances de relation pour le training, puis ils ont fusionné les résultats de la classification pour acquérir les relations finales entre les produits chimiques et les maladies.

Les méthodes basées sur les noyaux (ROZENFELD et FELDMAN 2006b) sont un moyen d'explorer les « caractéristiques » structurelles en calculant la similarité entre deux entités à l'aide d'une fonction de noyau (MOONEY et BUNESCU 2005). En effet, les approches à base de noyau (Kernel-based) proposent de traiter l'extraction de relations comme un problème de calcul de similarité entre un ensemble annoté d'occurrences de relations (l'ensemble est composé d'exemples positifs et négatifs de la relation visée) et une nouvelle occurrence de la relation. L'idée est de comparer la nouvelle occurrence de relation à celles qui sont annotées comme positives ou négatives et de déterminer ainsi de quelle catégorie la nouvelle occurrence est plus proche. Le calcul de la similarité est effectué en s'appuyant sur une fonction noyau, qui permet d'appliquer des transformations au vecteur de features qui représente l'occurrence de relation afin de le projeter dans un espace de grande dimension. Les transformations appliquées par le noyau reviennent à appliquer un produit scalaire dans un espace de grande dimension.
Lorsque l'apprentissage du modèle est terminé, l'étiquetage d'une nouvelle occurrence revient à déterminer si sa distance vis-à-vis des exemples positifs est supérieure à celle vis-à-vis des exemples négatifs. Le problème, dans ce cas, ne se situe plus seulement au niveau de la sélection des features (comme dans les

approches Feature-based) mais aussi dans le choix de la fonction noyau.
L'intérêt des *méthodes à base de noyaux* est qu'elles peuvent être appliquées directement sur des structures complexes (des graphes ou des arbres) contrairement aux méthodes Feature-based. Par exemple, Zhao et Grishman (S. ZHAO et GRISHMAN 2005a) ont présenté une approche d'extraction de relations qui combine des informations provenant de trois niveaux différents de traitement de la NLP : la tokenisation, l'analyse syntaxique des phrases et l'analyse des dépendances. Les fonctions individuelles du noyau sont conçues pour capturer chaque source d'informations. Ensuite, des noyaux composites sont développés pour combiner ces noyaux individuels afin que les erreurs de traitement se produisant à un niveau puissent être surmontées par des informations provenant d'autres niveaux.

2. **Approches basées sur la supervision distante :** Ce type d'approche est une technique d'apprentissage qui ne requiert pas de données manuellement annotées, mais requiert plutôt une base de connaissances ainsi qu'un corpus de textes. Dans les méthodes basées sur la supervision distante, on utilise une base de données existante, telle que Freebase ou une base de données spécifique à un domaine, pour collecter des exemples de la relation à extraire. On utilise ensuite ces exemples pour générer automatiquement nos données d'entraîne- ment. Par exemple, Freebase contient le fait que Barack Obama et Michelle Obama sont mariés. Nous prenons ce fait étiquetons ensuite chaque paire de « Barack Obama » et de « Michelle Obama » qui apparaissent dans la même phrase comme un exemple positif pour notre relation conjugale. De cette façon, on peut facilement générer une grande quantité de données d'entraînement.
La supervision à distance combine les avantages des deux paradigmessupervisé et non supervisé en utilisant à la fois des données étiquetées et non étiquetées. Elle utilise des données étiquetées pour entraîner un classifieur probabiliste qui peut prédire les étiquettes de classification pour de nouvelles données. Ensuite, elle utilise des données non étiquetées pour identifier des motifs et des relations dans les données qui peuvent être utilisés pour améliorer la performance du classifieur. En utilisant ces deux types de données ensemble, la supervision à distance peut être très performante pour l'apprentissage de modèles de classification complexes.
Wang et al.(Q. WANG, P. ZHANG, Yang LIU et al. 2021) ont proposé une méthode d'extraction de relations distante qui utilise un graphe sémantique dynamique pour modéliser les relations entre les entités. Leur méthode utilise également un mécanisme de pondération adaptative pour mieux gérer les bruits de supervision distante.Li et al. (Xin LI, Yang LIU, P. ZHANG et al. 2021) ont proposé une méthode qui utilise un réseau à double attention pour mieux capturer les relations entre les entités et les phrases qui les mentionnent.Leur méthode utilise également un mécanisme de pondération adaptative pour mieux gérer les bruits de supervision distante.

3. **Approches basées sur l'extraction conjointe des entités et de relations :** Toutes les techniques d'ER expliquées dans la section précédente supposent que les connaissances sur les limites et les types de mentions d'entités sont connues à l'avance. Si de telles connaissances ne sont pas disponibles, pour utiliser ces techniques d'extraction de relations, il faut d'abord appliquer une technique d'extraction mentionnée par l'entité. Une fois que les mentions d'entité et leurs types d'entité sont identifiés, les techniques d'ER peuvent être appliquées. Une telle méthode « pipeline » est sujette à la propagation d'erreurs de la première phase (extraction des mentions d'entités) à la deuxième phase (extraction des relations). Pour éviter cette propagation d'erreurs, il existe une ligne de recherche qui modélise ou extrait conjointement des entités et des relations. Yu et Lam (X. YU et LAM 2010) ont proposé un framework basé sur un modèle graphique probabiliste non dirigé pour effectuer conjointement les tâches d'identification d'entité et d'extraction de relations. Singh et al. (S. SINGH, RIEDEL, MARTIN et al. 2013) est la première approche qui modélise même des co-références conjointement avec des entités et des relations d'entité. Ils ont proposé un modèle graphique commun non orienté qui représente les diverses dépendances entre ces trois tâches.
   Contrairement à la plupart des autres approches pour l'extraction de relations où la modélisation est au niveau de la phrase, dans cette approche, un modèle capture toutes les mentions d'entité dans un document ainsi que les relations et les coréférences entre elles. La majorité des approches qui extraient conjointement des entités et des relations font état d'une amélioration significative par rapport aux approches précédentes. L'extraction conjointe améliore non seulement les performances de l'extraction de relations, mais s'avère également efficace pour l'extraction d'entités. Parce que, contrairement aux méthodes de pipeline, le modèle conjoint facilite l'utilisation des informations de relations pour l'extraction d'entités.

4. **Approches basées sur l'auto-supervision (self-supervised) :** Ces approches sont basées sur une technique d'apprentissage relativement récente où les données sont étiquetées de manière autonome (ou automatique). Il s'agit toujours d'un apprentissage supervisé, mais les ensembles de données n'ont pas besoin d'être étiquetés manuellement par un humain, mais ils peuvent par exemple être étiquetés en trouvant et en exploitant les relations (ou corrélations) entre les différents signaux d'entrée. Un avantage de l'apprentissage auto-supervisé est qu'il peut être plus facilement effectué en ligne (étant donné que les données peuvent être collectées et étiquetées sans intervention humaine), où les modèles peuvent être mis à jour ou formé complètement à partir de zéro. Par conséquent, l'apprentissage auto-supervisé devrait également être bien adapté à l'évolution des environnements, des données et, en général, des situations.
   Wu et al. (Siyuan WU, Y. WU, Sheng WANG et al. 2020) ont proposé une méthode d'extraction de relations auto-supervisée basée sur l'attention multi-têtes et la raffinerie d'étiquettes. La méthode utilise un objectif d'entraînement de recons-

truction de phrases pour apprendre des représentations sémantiques des entités et des contextes, et utilise ensuite une couche d'attention multi-têtes pour capturer les relations contextuelles entre les entités. Enfin, une technique de raffinage des étiquettes est utilisée pour améliorer la qualité des annotations faibles obtenues par l'auto-étiquetage. Zhang et al.(C. ZHANG, Chen WANG, Y. YU et al. 2021) ont proposé une approche d'extraction de relations auto-supervisée basée sur l'apprentissage multi-tâches avec des modèles de langage pré-entraînés. En utilisant à la fois des données étiquetées et non étiquetées, ils ont entraîné un modèle à prédire les relations dans les données non étiquetées tout en effectuant une classification de relations dans les données étiquetées. En utilisant une telle approche, ils ont obtenu de meilleures performances en termes de précision et de rappel par rapport à (Siyuan WU, Y. WU, Sheng WANG et al. 2020).

5. **Open Information Extraction :** L'extraction de relations traditionnelles se concentre sur un ensemble de relations précises et prédéfinies. Une intervention humaine importante est généralement requise pour concevoir des règles d'extraction ou pour créer des données d'entraînement étiquetées. Par conséquent, il est difficile de faire fonctionner de tels systèmes dans un domaine différent. Pour surmonter ces limites, le paradigme de l'extraction d'information ouverte (Open IE) a été proposé pour la première fois par Banko et al. (Oren ETZIONI, BANKO, SODERLAND et al. 2008a). Les systèmes Open Information extraction découvrent automatiquement les relations d'intérêt possibles en utilisant le corpus de texte sans aucune intervention humaine. Ainsi, aucun effort supplémentaire n'est requis pour passer à un autre domaine. Banko et al. ont proposé d'utiliser le classifieur de séquences auto-supervisé O-CRF basé sur le champ aléatoire conditionnel au lieu du classifieur Naive Bayes utilisé dans TextRunner et ont observé de meilleures performances. Une autre amélioration de TextRunner a été suggérée par Wu et Weld (F. WU et WELD 2010) dans leur système Open Extractor (WOE) basé sur Wikipédia. Ils ont utilisé les infoboxes de Wikipédia pour générer avec plus de précision des données d'entraînement pour le module d'apprentissage auto-supervisé.

1.3.3. Approches hybrides

Plus récemment, certains des autres travaux sont basés sur la combinaison de deux ou plusieurs modèles pour effectuer la tâche d'extraction de relations. Dans ce cadre, (ALFONSECA, FILIPPOVA, DELORT et al. 2012) combine entre une approche basée sur les patrons (patterns) et la supervision distante. Il présente une nouvelle stratégie basée sur des règles pour l'extraction de relations sémantiques qui tire parti de l'analyse syntaxique partielle afin de simplifier les structures linguistiques contenant des instances de relations sémantiques. Il propose également une stratégie de supervision à distance qui extrait automatiquement des modèles lexico-syntaxiques génériques au moyen de ressources semi-structurées telles que les infoboxes Wikipedia. Ces modèles génériques sont ensuite transformés en règles d'extraction qui sont utilisées pour

mettre à jour une grammaire de dépendance partielle. De même, (TABA et CASELI 2014) combine entre une méthode symbolique et une méthode basée sur la supervision distante. Il présente une approche basée sur des règles faiblement supervisée pour l'extraction de relations qui effectue une analyse partielle des dépendances afin de simplifier la structure linguistique d'une phrase. Cette simplification nous permet d'appliquer des règles d'extraction sémantique génériques, obtenues avec une stratégie de surveillance distante qui tire parti des ressources semi-structurées. Les règles sont ajoutées à une grammaire de dépendance partielle, qui est compilée dans un analyseur capable d'extraire des instances des relations souhaitées. (ROCKTÄSCHEL, S. SINGH et RIEDEL 2015) propose une approche hybride d'extraction de relations basée sur un ensemble de caractéristiques (feature based method) et sur des règles (symbolic method). Il utilise ensuite différents algorithmes de classification tels que les classificateurs SVM, Naïve Bayes et Decision Tree pour la classification des relations. Il sélectionne sept caractéristiques avec la technique d'analyse peu profonde basée sur des règles.

1.4. Méthodes de classification de relations

Dans cette section, nous présentons différentes méthodes qui ont été utilisées pour la classification des relations (Figure 1.3). Nous distinguons trois grandes catégories de classification : (1) les méthodes supervisées, (2) les méthodes non-supervisées et (3) les méthodes basées sur l'apprentissage par renforcement.

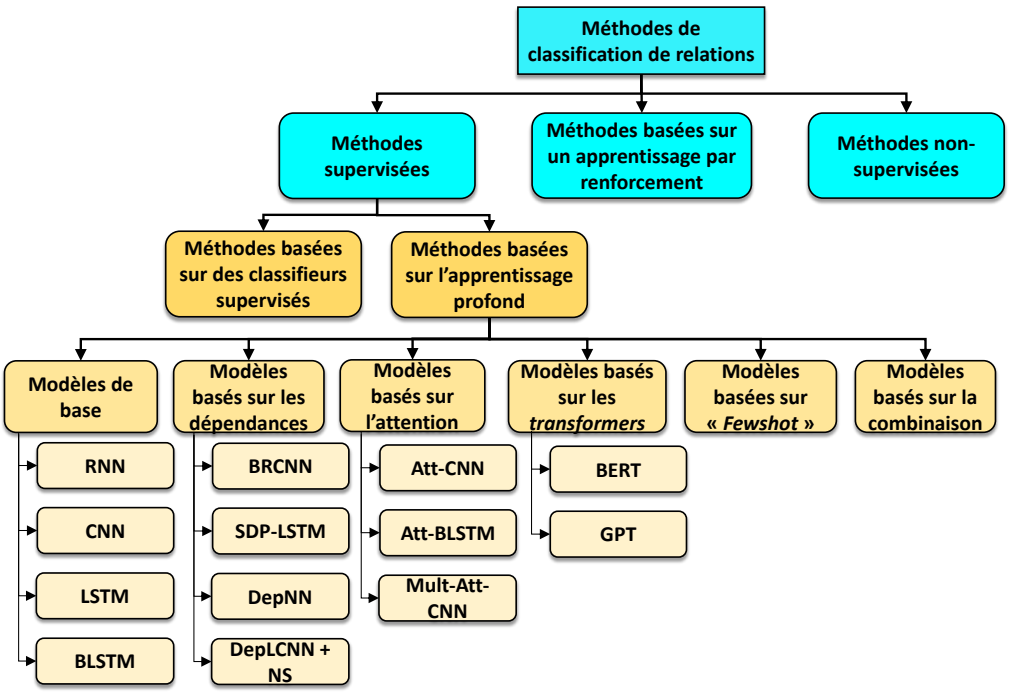


FIGURE 1.3. – Méthodes utilisées pour la classification des relations

1.4.1. Méthodes supervisées

Les méthodes supervisées pour la classification des relations peuvent être divisées en deux sous-catégories principales : (1) les méthodes basées sur des classifieurs supervisés et (2) les méthodes basées sur l'apprentissage profond. Les méthodes basées sur des classifieurs supervisés sont généralement utilisées pour la classification de relations spécifiques et nécessitent des données d'entraînement étiquetées pour chaque relation. D'autre part, les méthodes basées sur l'apprentissage profond sont souvent utilisées pour la classification de relations plus générales et peuvent être entraînées sur de grandes quantités de données non étiquetées.

1.4.1.1. Classification de relations par classifieurs supervisés

Les méthodes de classification supervisées utilisent des classifieurs entraînés sur un ensemble de données annotées manuellement pour prédire la relation entre deux entités dans une phrase. Le processus d'entraînement implique la présentation d'un grand nombre de paires d'entités préalablement annotées avec leur relation respective, au classifieur. Le classifieur utilise ces données pour apprendre les caractéristiques des relations et les patterns les plus fréquemment observés dans les données d'entraînement.

Les méthodes basées sur des classifieurs supervisés sont souvent efficaces lorsqu'il y a un grand nombre de données d'entraînement annotées, ce qui permet au classifieur d'apprendre les caractéristiques importantes pour la classification de manière précise. Cependant, leur efficacité peut diminuer lorsque le nombre de données d'entraînement est limité ou lorsque les caractéristiques importantes pour la classification ne sont pas évidentes.

Rink et al. (RINK et HARABAGIU 2010) ont proposé une approche pour la classification des relations sémantiques en combinant des ressources lexicales et sémantiques. L'approche utilise un ensemble de caractéristiques lexicales telles que les patrons lexicaux, les dépendances syntaxiques et les informations morphologiques, ainsi qu'un ensemble de caractéristiques sémantiques telles que les synsets de WordNet et les caractéristiques de cooccurrence basées sur le corpus. Ensuite, un classifieur supervisé, en l'occurrence un SVM, est utilisé pour classer les relations sémantiques en fonction de ces caractéristiques. Hendrickx et al. (HENDRICKX, S. N. KIM, KOZAREVA et al. 2019) utilisent un SVM avec des caractéristiques lexicales et de dépendance syntaxique pour classer les relations sémantiques entre les entités. Il améliore le travail de (RINK et HARABAGIU 2010) en utilisant une combinaison de ressources lexicales et de connaissances sémantiques supplémentaires.

1.4.1.2. Classification de relations par apprentissage profond

Les modèles d'apprentissage profond sont basés sur les réseaux de neurones artificiels (Artificial Neural Networks - ANN). Les ANN sont une forme d'algorithmes d'apprentissage automatique qui tentent d'imiter le comportement des neurones humains pendant le processus de traitement de l'information. Ils comprennent un grand

nombre d'unités de traitement interconnectées, appelées « neurones », qui collaborent pour traiter les données. Dans la plupart des cas, un réseau de neurones artificiels est organisé en couches. On trouve généralement trois couches dans ce réseau : la couche d'entrée, la couche de sortie et la couche cachée. L'apprentissage profond utilise des réseaux de neurones artificiels comportant de nombreuses couches cachées. Plusieurs modèles d'apprentissage profond sont utilisés pour classer les relations Ces modèles peuvent être classés en six sous-modèles : (1) les modèles de base, (2) les modèles basés sur les dépendances, (3) les modèles basés sur l'attention, (4)les modèles basés sur les transformers, (5) les modèles basés sur la combinaison de deux ou plusieurs réseaux de neurones, et (6) les modèles basés sur « Few-Shot ».

1. **Classification des relations par modèles de base :** L'apprentissage automatique permet de construire un algorithme de résolution complexe en adaptant les paramètres à partir de données d'apprentissage. Cela permet de construire des modèles de résolution sans avoir besoin de règles de résolution explicites ou de schémas de calcul comme c'est le cas dans le génie logiciel traditionnel. L'idée de supprimer complètement les schémas algorithmiques classiques de la résolution de problèmes complexes et d'utiliser uniquement l'idée d'apprentissage à partir des données est ce que l'on appelle les modèles basiques.Nous allons trouver différentes méthodes de classification utilisant un certain type de réseau de neurones basiques, notamment (a) les réseaux de neurones convolutifs (CNN), (b) les réseaux de neurones récurrents (RNN), (c) les réseaux récurrents à mémoire court et long terme (LSTM) et (d) les réseaux à mémoire à long et court terme bidirectionnelle (BLSTM).
   a) **Classification de relations par réseau de neurone convolutifs (CNN) :** Les réseaux de neurones convolutifs sont une sous-catégorie de réseaux de neurones. Ils possèdent donc toutes les caractéristiques des ANN. Bien que les CNN soient principalement conçus pour traiter les images, leur efficacité a également été démontrée dans d'autres applications telles que la classification de relations. Zeng et al. ont proposé de ne pas extraire les caractéristiques des systèmes de traitement du langage naturel, mais d'utiliser le modèle CNN pour extraire à la fois les caractéristiques de niveau lexical (Lexical level features) et les caractéristiques de niveau de la phrase (sentence level features). Les caractéristiques de niveau lexical sont extraites en fonction des noms donnés (les deux entités nommes, les mots qui viennent à gauche et à droite des deux entités et les hypernymes de wordnet des deux entités). Pendant ce temps, les caractéristiques de niveau de phrase sont apprises en ajoutant des caractéristiques WPE (WF (Feature Word) + PF (Feature position)) au vecteur du mot. En effet, la théorie des hypothèses de distribution indique que les mots qui apparaissent dans le même contexte ont tendance à avoir des significations similaires. Pour capturer cette caractéristique, le WF combine la représentation vectorielle d'un mot et les représentations vectorielles des mots dans son contexte. De plus, il n'est pas possible de capturer les

caractéristiques de structure uniquement via WF. Il est nécessaire de spécifier les paires d'entités dans la phrase. À cet effet, PF est proposé pour la classification des relations. Ensuite le WF et le PF de chaque mot sont concaténées pour former l'entrée pour la couche convolutionnelle. Enfin, toutes les entités sont entrées dans la couche softmax pour prédire la relation entre les deux entités. Cependant, afin d'obtenir des résultats meilleurs, (D. ZENG, K. LIU, LAI et al. 2014) utilisent toujours certaines caractéristiques extraites de ressources lexicales telles que WordNet. (SANTOS, B. XIANG et B. ZHOU 2015) propose un nouveau réseau neuronal convolutif (CNN), nommé (CR-CNN). Le réseau propose une nouvelle fonction de perte de classement par paire qui permet de réduire facilement l'impact des classes artificielles. En utilisant CRCNN, et sans avoir besoin de caractéristiques extraites de ressources lexicales, le CR-CNN est plus efficace que le CNN suivi d'un classificateur softmax (D. ZENG, K. LIU, LAI et al. 2014). En effet, Le CR-CNN utilise une méthode de classement par paire, tandis que (D. ZENG, K. LIU, LAI et al. 2014) applique une classification multiclasse en utilisant la fonction softmax au sommet du CNN. Pengda Qin et al. (P. QIN, W. XU et J. GUO 2016) proposent un réseau de neurones convolutif plus approprié pour la classification de relations en utilisant la fonction ETF (Entity Tag Feature). Cette fonction permet d'indiquer les informations de position de l'entité annotée, ce qui est plus simple mais plus efficace que la fonction de position proposée dans (D. ZENG, K. LIU, LAI et al. 2014) et utilisée dans (SANTOS, B. XIANG et B. ZHOU 2015). Cependant, ces méthodes souffrent souvent de sous-séquences ou de clauses non pertinentes, en particulier lorsque les sujets et les objets sont plus éloignés. Par exemple, dans la phrase « [Emmanual Macron]e1, qui est marié depuis octobre 2007 à Brigitte Trogneux, est un membre du [parti socialiste]e2 », la clause « qui » est utilisée pour modifier le sujet e1, mais elle n'est pas liée à la Relation de Member-Collection entre « Emmanual Macron » et « Parti socialiste ». L'intégration de ces informations dans le modèle nuira aux performances d'extraction. Pour cette raison, (D. JIN, DERNONCOURT, SERGEEVA et al. 2018a) a proposé d'apprendre une représentation de relation plus robuste à partir d'un modèle de réseau de neurones à convolution qui fonctionne sur le chemin de dépendance simple entre les sujets et les objets, ce qui caractérise naturellement la relation entre deux nominaux et évite les effets négatifs d'autres morceaux ou clauses non pertinents. De plus, les auteurs de (D. JIN, DERNONCOURT, SERGEEVA et al. 2018a) ont introduit une stratégie d'échantillonnage négative dans les modèles CNN pour aborder la directionnalité de la relation, c'est-à-dire affecter correctement le sujet et l'objet dans une relation. Dans l'exemple de Macron ci-dessus, (Emmanual Macron , Parti socialiste) détient la relation Member-collection, tandis que (Parti socialiste, Emmanual Macron) non. Cette stratégie permet d'exploiter le chemin de dépendance pour apprendre les affectations des sujets et des objets. Les résultats expérimentaux montrent que la méthode d'échantillonnage négatif améliore considérablement les performances.

b) **Classification de relations par réseau neuronaux récurrent (RNN) :** La plupart des approches d'apprentissage basées sur CNN actuelles pour l'apprentissage relationnel et la classification sont des modèles statiques et sont potentiellement faibles, en particulier lors de l'apprentissage de modèles de relations à longue distance. Par exemple, le modèle CNN ne peut apprendre que des modèles locaux, et il est donc difficile de traiter des modèles qui sont en dehors de la fenêtre du filtre convolutionnel. Comparé aux modèles CNN, RNN est un modèle temporel et est particulièrement bon pour modéliser des données séquentielles. En effet, un RNN reçoit des informations d'entrée et un contexte et produit un résultat. Ce résultat est renvoyé au réseau en tant que nouveau contexte. Ce nouveau contexte est combiné à une nouvelle entrée, et utilisé par le RNN pour produire une nouvelle sortie. Ce processus est répété autant de fois que nécessaire pour produire un résultat final. Grâce à leur mémoire interne, les RNN sont capables de « se souvenir » d'éléments importants concernant les données qu'ils ont reçues. Des éléments importants concernant les données qu'ils ont reçues, ce qui leur permet de de prédire très précisément ce qui va suivre. Les RNN sont des réseaux neuronaux puissants et robustes et sont très prometteuses puisqu'ils sont les seuls à posséder une mémoire interne.

   Un des premiers travaux qui a utilisé un RNN a été proposé par Socher et al. (SOCHER, HUVAL, MANNING et al. 2012). Ce modèle nommé MVRNN est capable de capturer non seulement les différentes caractéristiques des mots mais aussi la signification des mots adjacents. Zhang et al. (D. ZHANG et Dong WANG 2015) ont constaté que la distance entre deux entités dans une phrase était généralement longue, ils ont donc proposé le modèle RNN pour s'attaquer au problème de l'apprentissage de modèle à longue distance. Le modèle MV-RNN (SOCHER, HUVAL, MANNING et al. 2012) est basé sur NN récursif tandis le modèle proposé par (D. ZHANG et Dong WANG 2015) est basé sur NN récurrent qui est un modèle temporel. Xu et al. (Y. XU, R. JIA, MOU et al. 2016) ont noté que les réseaux de neurones existants pour la classification des relations ont généralement des architectures peu profondes et ont donc conçu un RNN profond (DRNN) avec augmentation des données pour résoudre ce problème.

c) **Classification de relations par réseau récurrent à mémoire court et long terme (LSTM) :** Les RNNs « classiques » ne sont capables de mémoriser que le passé dit proche, et commencent à « oublier » au bout d'une cinquantaine d'itérations environ. Ce transfert d'information à double sens rend leur entrainement beaucoup plus compliqué. Pour cette raison, des méthodes efficaces ont été mises au point comme les LSTM (Long Short Term Memory). Les LSTM étendent leur mémoire. Cette dernière peut être vue comme une cellule gated, où gated signifie que la cellule décide de « se souvenir » ou d'ignorer des données, en fonction de l'importance qu'elle attribue à l'infor-

mation. L'attribution de l'importance se fait par le biais de poids, qui sont également appris par l'algorithme. Le LSTM a montré son mérite dans la capture de relations à longue distance dans la classification des relations.

d) **Classification de relations par réseau à mémoire à long et court terme bidirectionnelle (BLSTM) :** Le modèle LSTM est uniquement capable d'utiliser un contexte précédent. Cependant, dans ces applications, le contexte futur peut être très utile par rapport à un moment donné. C'est le cas, par exemple, dans la tâche de traitement des séquences de mots. Pour cette raison, le LSTM bidirectionnel (BLSTM) a été introduit par (S. ZHANG, D. ZHENG, Xinchen HU et al. 2015). Il s'agit d'une extension des LSTM traditionnels qui peut améliorer les performances du modèle pour les problèmes de classification de relations. Ce type d'architecture peut être entraîné dans les deux directions, avant et arrière. En effet, le BLSTM offre une capacité d'apprentissage supplémentaire, car la couche de sortie reçoit simultanément les informations des instances passées (vers l'arrière) et futures (vers l'avant), ce qui permet d'obtenir de meilleures performances.

Zhang et al ont proposé des réseaux de mémoire bidirectionnels à long terme (BLSTM) pour modéliser la phrase avec des informations complètes et séquentielles sur tous les mots. Ils ont utilisé également des caractéristiques extraites des ressources lexicales telles que WordNet et de NLP tels que l'analyseur de dépendances et les identificateurs d'entités nommées (NER).

2. **Classification de relations basées sur les dépendances :** Les modèles basées sur les dépendances peuvent obtenir de meilleurs résultats. Dans ces modèles, les caractéristiques syntaxiques sont tirées de l'analyseur syntaxique de dépendance. Cet analyseur est une caractéristique commune utilisée dans les tâches d'extraction de relations. Le résultat de l'analyseur syntaxique est un arbre de dépendance où les sommets représentent les mots et les bords représentent les relations syntaxiques entre ces mots. Les arbres d'analyse syntaxique des dépendances révèlent les dépendances non locales au sein des phrases, c'est-à-dire entre des mots très éloignés les uns des autres dans une phrase. Les analyseurs syntaxiques de dépendance sont donc utilisés pour capturer les dépendances à longue distance entre deux noms. L'arbre d'analyse syntaxique des dépendances contient la structure grammaticale d'une phrase comme le sujet, l'objet, etc. et devient donc une caractéristique importante dans l'extraction de relations. Parmi les modèles basés sur les dépendances les plus utilisés, nous citons (a) réseau de neurones récurrent à longue mémoire à court terme selon le chemin de dépendance le plus court, (b) réseau de neurones convolutionnel récurrent bidirectionnel, (c) réseau de neurones basé sur les dépendances et (d) CNN avec échantillonnage négatif simple.

a) **Classification de relations par réseau de neurones récurrent à longue mémoire à court terme selon le chemin de dépendance le plus court (SDP-LSTM) :** Xu et al. (Y. XU, MOU, G. LI et al. 2015) proposent une nouvelle

architecture LSTM avec le chemin de dépendance le plus court (SDP) pour classer les relations entre les entités dans une phrase. Cette architecture présente plusieurs caractéristiques : (a) le chemin de dépendance le plus court permet de conserver les informations les plus pertinentes, tout en éliminant les mots non pertinents dans la phrase, (b) les réseaux LSTM permettent une intégration efficace des informations provenant de sources hétérogènes sur les chemins de dépendance, (c) une stratégie d'abandon régule le réseau neuronal pour éviter le sur-ajustement. Le processus se déroule comme suit : la phrase d'entrée est analysée dans un arbre de dépendance, puis le chemin de dépendance le plus court est extrait en entrée. Quatre éléments d'information supplémentaires tels que le Wordembedding, les balises de parties du discours, les relations grammaticales et les hypernoms WorldNet sont pris en compte dans ce processus. Deux réseaux neuronaux récurrents sont utilisés pour capturer la directionnalité des relations. Les unités LSTM sont utilisées pour propager efficacement les informations le long du chemin de dépendance le plus court.

b) **Classification de relations par réseau de neurones convolutionnel récurrent bidirectionnel (BRCNN) :** Dans (R. CAI, Xiaodong ZHANG et Houfeng WANG 2016a), Cai et al. proposent une nouvelle architecture appelée « BRCNN » combinant des réseaux neuronaux convolutifs et des réseaux neuronaux récurrents avec des unités LSTM. Cette architecture est bidirectionnelle et utilise les informations sur les relations de dépendance dans le chemin de dépendance le plus court (SDP). Elle permet d'apprendre des représentations relationnelles avec des informations directionnelles le long du SDP en avant et en arrière en même temps. Le long du SDP, deux LSTM sont appliqués pour apprendre les représentations cachées des mots et les relations de dépendance respectivement. Les caractéristiques locales des représentations cachées de chaque paire de mots voisins, ainsi que les relations de dépendance entre eux, sont capturées en utilisant une couche de convolution. Ensuite, les informations sont rassemblées à partir des caractéristiques locales du SDP et du SDP inverse en utilisant une couche de mise en commun maximale. Enfin, la couche de sortie softmax est utilisée pour la classification. Le BRCNN surpasse les différentes méthodes de l'état de l'art et présente une efficacité supérieure pour la classification des relations.

c) **Classification de relations par réseau de neurones basé sur les dépendances (DepNN) :** Le chemin de dépendance augmenté (ADP) est une nouvelle structure proposée par Liu et al (Yang LIU, F. WEI, S. LI et al. 2015). Cette structure est composée du plus court chemin de dépendance entre deux entités et des sous-arbres liés au plus court chemin. Liu et al (Yang LIU, F. WEI, S. LI et al. 2015) développent des réseaux neuronaux basés sur la dépendance (DepNN) où un réseau neuronal récurrent (RNN) est conçu pour modéliser les sous-arbres, et un réseau neuronal convolutif (CNN) est appliqué pour

capturer les caractéristiques les plus importantes sur le chemin le plus court. Le DepNN tire parti des réseaux de neurones convolutifs et récurrents. Afin de connecter les modèles RNN et CNN, chaque mot du chemin le plus court est combiné avec une représentation générée à partir de son sous-arbre. Le chemin de dépendance augmenté est ainsi représenté comme un vecteur sémantique continu, qui peut ensuite être utilisé pour classer les relations.

d) **CNN avec échantillonnage négatif simple (DepLCNN+NS) :** Les auteurs de (K. XU, Y. FENG, Songfang HUANG et al. 2015) proposent un réseau neuronal convolutif afin d'apprendre des représentations de relations plus robustes à partir des chemins de dépendance les plus courts. Ce réseau considère le chemin de dépendance le plus court entre deux entités qui décrit la relation. Le plus court chemin de dépendance du sujet à l'objet représente l'entrée du modèle. Ce chemin est traité par une couche de table de consultation. En cherchant dans la matrice d'intégration, chaque mot et chaque étiquette dans le chemin de dépendance est transformé en un vecteur. Les caractéristiques locales capturées autour de chaque nœud sont combinées en un vecteur de caractéristiques global. Enfin, ce vecteur de caractéristiques est transmis à une couche softmax pour la classification des relations. Les auteurs de (K. XU, Y. FENG, Songfang HUANG et al. 2015) ont également proposé une méthode simple d'échantillonnage négatif pour aider à effectuer des affectations correctes des sujets et des objets dans une relation.

3. **Classification de relations par modèles basés sur l'attention :** Au cours des quatre dernières années, certains modèles basés sur l'attention ont été proposés. Le mécanisme d'attention attribue des poids différents à chaque mot d'une phrase et rend l'apprentissage du modèle de réseau de neurones plus flexible. En NLP, le mécanisme d'attention a d'abord été utilisé par Bahdanau et al. (BAHDANAU, CHO et BENGIO 2014) pour l'alignement du texte dans la traduction automatique. Il a rapidement attiré l'attention et a été largement utilisé pour de nombreuses autres tâches.
   Les réseaux neuronaux utilisent le mécanisme d'attention pour tenter d'imiter le cerveau humain, en particulier le processus de vision, pour traiter les données. Au lieu de se concentrer sur la totalité des données d'entrée, l'accent est mis sur les sous-ensembles pertinents tout en minimisant le travail consacré aux sous-ensembles de données jugés non pertinents. Les modèles d'attention les plus utilises sont (a) CNN basé sur l'attention, (b) BLTM basé sur l'attention et (c) Mult-Att-CNN.

   a) **Classification de relations par CNN basé sur l'attention (Att-CNN) :** Le modèle CNN basé sur l'attention utilise les informations d'intégration des mots, les étiquettes d'intégration des parties du discours et les informations d'intégration de la position pour la tâche de classification des relations. Un mécanisme d'attention pour sélectionner les mots pertinents en relation avec

les entités est proposé dans (N. LI, Hui ZHANG et Yong CHEN 2018). Le modèle d'attention est composé de modèles hétérogènes qui sont une phrase et deux entités. Le mécanisme d'attention aux mots est utilisé pour modéliser quantitativement la pertinence contextuelle des mots par rapport aux entités. Le poids de chaque mot dans la phrase est calculé en intégrant chaque mot et entité de la phrase dans un perceptron multicouche (MLP). (N. LI, Hui ZHANG et Yong CHEN 2018) applique une nouvelle forme de mécanisme d'attention qui repose sur le chemin de dépendance le plus court entre les entités pour la classification des relations. Il propose sur une architecture CNN qui combine le mécanisme d'attention avec SDP pour extraire les mots qui ont un effet décisif sur la classification. De plus, une nouvelle fonction objective est utilisée pour réduire les interférences de la classe artificielle qui améliore l'effet. De même, (L. WANG, Z. CAO, DE MELO et al. 2016) utilise une architecture CNN mais qui s'appuie sur un nouveau mécanisme d'attention à plusieurs niveaux. Cette architecture permet de détecter des indices plus subtils malgré la structure hétérogène des phrases d'entrée, ce qui lui permet de détecter automatiquement les parties les plus pertinentes pour une classification donnée. Cette architecture permet un apprentissage sans devoir recourir à des connaissances externes telles que les structures de dépendance ou des connaissances extraites des systèmes NLP

b) **Classification de relations par BLSTM basé sur l'attention (Att-BLSTM) :** Les réseaux BLSTM basés sur l'attention bidirectionnelle, sont conçus pour capturer des informations sémantiques importantes à n'importe quelle position dans la phrase. Ce modèle contient essentiellement cinq composants : Couche d'entrée, couche d'incorporation, couche LSTM, couche d'attention et couche de sortie.
Zhou et al. (P. ZHOU, Wenhao SHI, Jinjun TIAN et al. 2016) ont conçu le modèle Att-BLSTM. La contribution de ce modèle est d'utiliser BLSTM avec un mécanisme d'attention, qui peut automatiquement se concentrer sur les mots qui ont un effet décisif sur a classification, pour capturer les informations sémantiques les plus importantes dans une phrase, sans utiliser de connaissances supplémentaires et de systèmes de NLP. Alfattni et al. (ALFATTNI, PEEK et NENADIC 2021a) décrit un BLSTM avec un mécanisme d'attention, qui peut se concentrer automatiquement sur les mots qui ont un effet décisif sur la classification, pour capturer les informations sémantiques les plus importantes dans une phrase, sans utiliser de connaissances supplémentaires et de systèmes de traitement automatique du langage.

c) **Classification de relations par Mult-Att-CNN :** Un mécanisme d'attention à plusieurs niveaux est proposé dans (L. WANG, Z. CAO, DE MELO et al. 2016). Cette attention multi-niveaux du CNN permet un apprentissage de bout en bout à partir de données étiquetées spécifiques à la tâche, sans utiliser de connaissances externes telles que des structures de dépendance explicites. Le mécanisme d'attention multi-niveaux consiste à appliquer l'attention sur

plusieurs couches. L'attention multi-niveaux est utilisée pour capturer à la fois l'attention spécifique à l'entité et l'attention de mise en commun spécifique à la relation. Cela lui permet de détecter des indices plus subtils malgré la structure hétérogène des phrases en entrée, ce qui lui permet d'apprendre automatiquement quelles parties sont pertinentes pour une classification donnée.

4. **Modèles basés sur les transformers :**
Les méthodes basées sur les Transformers sont une famille de modèles d'apprentissage profond largement utilisés dans le traitement automatique du langage naturel. Les Transformers ont été introduits en 2017 par Vaswani et al. pour la traduction automatique et ont depuis été utilisés dans de nombreux autres domaines, notamment la génération de texte, la compréhension de texte, la traduction automatique, la classification de texte et bien plus encore.
Ces modèles ont la capacité de traiter des séquences de texte de manière simultanée, en utilisant des mécanismes d'attention pour capturer les interactions entre les différents éléments de la séquence. Les modèles basés sur les Transformers sont capables de capturer des informations sémantiques complexes à partir du contexte global d'un texte et de produire des représentations vectorielles riches pour chaque mot ou entité nommée.
Les modèles basés sur les Transformers ont connu un grand succès ces dernières années, en grande partie grâce à des modèles pré-entraînés tels que BERT (Bidirectional Encoder Representations from Transformers) et GPT (Generative Pre-trained Transformer).

   a) **BERT (Bidirectional Encoder Representations from Transformers) :** BERT (DEVLIN, CHANG, K. LEE et al. 2018a) est un modèle de traitement du langage naturel basé sur les Transformers, développé par Google en 2018. BERT a été pré-entraîné sur de vastes ensembles de données non étiquetées, en utilisant une approche de pré-entraînement masqué, où le modèle doit prédire les mots manquants dans une phrase. Après le pré-entraînement, BERT peut être finement ajusté pour des tâches spécifiques en utilisant des ensembles de données étiquetées. BERT a atteint des résultats remarquables sur plusieurs tâches de traitement automatique du langage naturel, notamment la classification de relations.

      Plusieurs travaux ont utilisé BERT pour la classification de relations. Hu et al. (L. HU, Luhao ZHANG, C. SHI et al. 2019) ont proposé un modèle qui combine une méthode d'apprentissage par étiquetage distant avec l'utilisation de BERT pour la classification de relations. Les auteurs ont également introduit une technique d'embedding d'étiquettes conjointes pour améliorer les performances du modèle.Yu et al. (Y. YAO, D. YE, Peng LI et al. 2019) ont proposé une tâche de classification de relations à l'échelle des documents et ont introduit un nouvel ensemble de données appelé DocRED pour évaluer

la performance de différents modèles de classification de relations, y compris un modèle basé sur BERT avec un schéma d'encodage d'entités et un mécanisme d'attention basé sur l'attention de tête multiple.Les résultats ont montré que leur modèle basé sur BERT a obtenu des performances significativement meilleures que celles obtenues par (L. HU, Luhao ZHANG, C. SHI et al. 2019).
Zhang et al. (Xiaoya ZHANG, X. HAN, J. MA et al. 2020) ont proposé une méthode de classification de relations qui utilise l'apprentissage multi-tâches pour améliorer la performance de l'extraction d'entités et de relations. Les auteurs proposent un modèle basé sur BERT qui effectue simultanément l'extraction d'entités et de relations, en utilisant un mécanisme de partage de poids pour optimiser les deux tâches de manière conjointe. Les auteurs concluent que l'utilisation de l'apprentissage multi-tâches avec BERT est une approche prometteuse pour l'extraction d'entités et de relations, et qu'elle peut être utilisée dans de nombreuses applications pratiques telles que l'analyse de sentiments, la classification de documents, et l'extraction d'informations.

b) **GPT (Generative Pre-trained Transformer) :** GPT (RADFORD, NARASIMHAN, SALIMANS et al. 2018) est un modèle de traitement du langage naturel basé sur les Transformers, développé par OpenAI en 2018. GPT a été pré-entraîné sur de vastes ensembles de données non étiquetées, en utilisant une approche de prédiction de mots suivants, où le modèle doit prédire le mot suivant dans une phrase donnée le contexte précédent. Après le pré-entraînement, GPT peut être finement ajusté pour des tâches spécifiques en utilisant des ensembles de données étiquetées.
GPT a atteint des résultats remarquables sur plusieurs tâches de traitement automatique du langage naturel, notamment pour la classification de relations.Li et al. (Peng LI, X. QIU et X. HUANG 2019) ont proposé une méthode qui utilise GPT pour extraire des représentations de phrases pour la classification de relations. Ils ont montré que leur méthode surpassait les méthodes basées sur les réseaux de neurones à rétropropagation classiques. Plus récemment, Chen et al. (T. CHEN, JI, Z. GUO et al. 2020) ont proposé une méthode de classification de relations qui utilise une combinaison de GPT et de la propagation de graphes pour extraire des informations de relations dans un corpus de texte non structure. Ces travaux montrent que GPT peut être efficace pour la classification de relations en utilisant une approche différente de celle de BERT, en se concentrant sur la compréhension du contexte global plutôt que sur des informations spécifiques d'entités.

5. **Modèles basés sur la combinaison de deux ou plusieurs réseaux de neurones :** Plus récemment, certains des autres travaux sont basés sur la combinaison de deux ou plusieurs réseaux de neurones pour effectuer la tâche de classification des relations. Dans ce cadre, des réseaux neurones (DepNN) ont été développés pour résoudre le problème de classification des relations. Ils combinent entre

les réseaux de neurones ecurrents et les réseaux de neurones convolutifs afin de tirer les avantages des deux réseaux. Zhang et al (Ye ZHANG, L. YANG, Zhiyuan LIN et al. 2016) présentent une nouvelle approche DepNN qui combine entre un RNN et un CNN. Ils proposent une nouvelle structure, appelée chemin de dépendance augmentée (ADP), qui est composée du chemin de dépendance le plus court entre deux entités et des sous arbres attachés au chemin le plus court. Le RNN est conçu pour modéliser les sous arbres vu qu'il est bon pour spécifier les structures hiérarchiques. Le CNN est conçu pour capturer les caractéristiques les plus importantes sur le chemin le plus court.
D'autres auteurs ont combiné entre un réseau BLSTM avec un RNN ou un CNN. Li et al (F. LI, Meishan ZHANG, G. FU et al. 2016) proposent un modèle (BLSTM-RNN) qui est modélisé à partir des représentations d'entités et de contextes tirées des LSTM-RNN. En effet, la motivation de ce modèle est que la relation entre deux entités cibles peut être représentée par les deux entités et les trois contextes qui les entourent. Des fonctions de « pooling » standard sont, ensuite, appliquées sur les représentations de mots de chaque partie pour obtenir cinq représentations correspondant aux cinq parties. Enfin, ils sont concaténés et introduits dans une couche softmax pour la classification des relations. De même, (T. XU, DU, C. FU et al. 2018) a utilisé un BLSTM mais combiner avec un CNN pour la classification des relations qui a donné une meilleure performance par rapport à la combinaison avec un RNN (F. LI, Meishan ZHANG, G. FU et al. 2016). Wang et al (P. WANG, Z. XIE et Junfeng HU 2017) proposent de combiner le CNN mais avec SDP-BLSTM en ajoutant les chemins de dépendances. Ils ont utilisé un CNN pour construire les caractéristiques locales. Puis, un réseau neuronal SDP-BLSTM est appliqué pour produire la représentation vectorielle de taille fixe finale de l'instance de relation.

Plus récemment, d'autres auteurs ont essayé d'ajouter un mécanisme d'attention avec la combinaison de deux ou plusieurs réseaux de neurones afin d'améliorer les performances. Guo et al (X. GUO, Hui ZHANG, H. YANG et al. 2019a) ont proposé un nouveau modèle Att-RCNN. Ce modèle utilise une combinaison des deux types de NN pour capturer les caractéristiques. En effet, il utilise un RNN pour extraire des représentations contextuelles de niveau supérieur de mots et un CNN pour obtenir des caractéristiques de phrase pour la tâche de classification de relation. De plus, afin d'améliorer les performances du modèle, ils ont appliqué des mécanismes d'attention à deux niveaux pour capturer des caractéristiques plus sensibles, pertinentes et précieuses pour la tâche de classification des relations. Aussi, Shen et al (S. SHEN, WEN, L. ZHOU et al. 2018) ont proposé un nouveau mécanisme d'attention personnalisé pour améliorer les performances. De plus, ils ont combiné entre un RNN et deux CNN. Cependant, les CNNs et le RNN sont architecturés en un seul réseau profond au lieu d'un réseau parallèle pour obtenir les avantages des deux réseaux.

6. **Méthodes basées sur l'apprentissage « Few-Shot » :** Les données annotées ont toujours été un défi pour l'apprentissage supervisé et représentent un facteur

qui détermine les coûts des ressources. Afin de réduire les coûts d'analyse des données, nous pouvons utiliser d'autres méthodes, telles que les méthodes d'apprentissage « Few-Shot ».
Les modèles « Few-Shot » pour la classification de relations sont particulièrement pertinents dans les scénarios où les ensembles de données étiquetées sont coûteux à annoter ou limités en termes de quantité de données disponibles, tels que les langues moins courantes ou les domaines de spécialisation. Ces modèles sont également utiles dans les cas où de nouvelles relations doivent être détectées et classées sans qu'il soit nécessaire de recueillir un grand nombre d'exemples étiquetés.
De nombreux travaux ont exploré l'utilisation de modèles « Few-Shot » pour la classification de relations. Parmi ces travaux, Han et al. (X. HAN, H. ZHU, P. YU, Ziyang LIU et al. 2018) qui ont apporté une contribution importante à la recherche en apprentissage few-shot en proposant un ensemble de données de classification de relations de grande envergure « FewRel » et en proposant des méthodes de référence pour la classification de relations few-shot, y compris une méthode d'apprentissage par prototypage (prototype-based learning) qui utilise des représentations de relation pondérées pour construire des prototypes de relation. Liu et al. (Baitao LIU, Jingxuan ZHU, Xiaoyan LI et al. 2019) ont proposé une méthode de classification de relations few-shot basée sur l'apprentissage multitâche. La méthode combine plusieurs tâches de classification de relations liées dans un modèle d'apprentissage commun pour tirer parti des similarités entre les tâches et améliorer les performances en few-shot.
Tian et al.(F. TIAN, Lei ZHANG, Wei HUANG et al. 2021) ont proposé une méthode de classification de relations few-shot basée sur plusieurs hypergraphes pour mieux modéliser les relations complexes et les hiérarchies entre les entités et les relations. Leur méthode utilise des représentations d'entités et de relations pour construire des hypergraphes de différentes granularités, et utilise des techniques de propagation de graphes pour intégrer l'information de différents hypergraphes. Les expériences montrent que leur méthode améliore les performances de classification de relations few-shot par rapport à (X. HAN, H. ZHU, P. YU, Ziyang LIU et al. 2018) et (Baitao LIU, Jingxuan ZHU, Xiaoyan LI et al. 2019)

Ces travaux ont montré que les modèles « Few-Shot » peuvent être efficaces pour la classification de relations dans des scénarios où les données étiquetées sont limitées.

1.4.2. Classification de relations avec apprentissage par renforcement

Un autre type de modèle de classification des relations est l'apprentissage par renforcement (RL). L'apprentissage par renforcement est une technique populaire

d'apprentissage automatique qui excelle dans la résolution de problèmes dans des environnements dynamiques et adaptatifs. Il s'agit d'une forme d'apprentissage supervisé dans laquelle seules des informations partielles sont fournies. Dans la tâche de classification des relations, les chercheurs ont utilisé une méthode de sélection dure afin de filtrer les noises dans l'ensemble d'apprentissage. Ils ont conclu que traiter les phrases incorrectement étiquetées avec des poids d'attention n'était pas une bonne idée, et que ces phrases devaient être traitées avec une décision dure. Pour cette raison, ils ont utilisé RL pour prendre des décisions.

Feng et al. (J. FENG, M. HUANG, Li ZHAO et al. 2018) ont proposé un nouveau modèle de classification des relations basé sur la RL. Ce modèle vise à sélectionner les phrases correctes à partir de données bruitées pour une meilleure classification des relations. Il se compose de deux modules clés : le sélecteur d'instance et le classificateur de relations. Le sélecteur d'instance vise à sélectionner des phrases correctes à partir de données bruitées, et le classificateur de relation prédit le niveau de la phrase et fournit des récompenses au sélecteur d'instance. Au cours du processus d'apprentissage, ces deux modules interagissent l'un avec l'autre. Comme Feng et al. (J. FENG, M. HUANG, Li ZHAO et al. 2018), Qin et al. (P. QIN, W. XU et W. Y. WANG 2018) explorent une stratégie d'apprentissage par renforcement profond pour apprendre un sélecteur d'instance. Les probabilités de prédiction sont utilisées pour déterminer la récompense dans (J. FENG, M. HUANG, Li ZHAO et al. 2018) . Cependant, la récompense de Qin et al. (P. QIN, W. XU et W. Y. WANG 2018) est intuitivement reflétée dans le changement de performance du classificateur de relations.

Contrairement aux travaux précédents, Yang et al. (K. YANG, L. HE, X. DAI et al. 2019) proposent RCEND, un nouveau cadre pour améliorer la classification des relations en exploitant les données bruitées. Avec RL, Yang et al. (K. YANG, L. HE, X. DAI et al. 2019) utilisent un discriminateur pour séparer les données bruitées en données correctement étiquetées et en données incorrectement étiquetées. Le modèle a ensuite traité les données qui avaient été étiquetées comme des données non étiquetées et a utilisé une méthode d'apprentissage semi-supervisée pour les exploiter.

1.4.3. Méthodes non supervisées

Les méthodes non supervisées visent à réduire la quantité d'exemples annotés nécessaire pour l'extraction des relations. Dans ce cas précis, l'objectif est de ne pas en utiliser du tout. Pour la classification des relations, des approches à base de clusters ont été proposées. L'idée générale des approches de regroupement (ou clustering) est de regrouper les exemples non annotés en fonction des caractéristiques qu'ils partagent. Les groupes ainsi constitués sont appelés clusters. Les méthodes de clustering se distinguent par : (i) la manière de calculer la similarité (ou distance) entre les éléments à regrouper, (ii) la manière de construire les clusters, (iii) la manière de définir le nombre de clusters à construire. Une des premières approches pour l'extraction des relations d'une manière complètement non supervisées a été proposée par Fader et al. (FADER, SODERLAND et Oren ETZIONI 2011b) . Ils nécessitent uniquement un marqueur NER pour identifier les entités nommées dans le texte afin que le système

se concentre uniquement sur les mentions d'entités nommées. Wang et al. (Yanjun WANG, X. YAO et T. LIU 2014) ont proposé quelques améliorations dans l'approche de clustering de base de Fader et al. (FADER, SODERLAND et Oren ETZIONI 2011b) . Ils ont développé une méthode de sélection de fonctionnalités non supervisée afin de supprimer les mots bruités non informatifs. Une autre approche similaire pour l'extraction de relations non supervisées des textes a été proposée par Zeng et al. (X. ZENG, Yankai LIU, Z. CHEN et al. 2019) . Ces derniers propose une méthode de clustering basée sur la similarité de contexte, en se concentrant sur les contextes lexicaux qui entourent les entités dans le texte.

1.4.4. Conclusion

La classification de relations est une tâche importante dans le traitement automatique du langage naturel qui a attiré beaucoup d'attention ces dernières années. Les méthodes basées sur le deep learning ont prouvé leur efficacité pour cette tâche, avec des performances supérieures à celles des méthodes non supervisées et des méthodes basées sur le renforcement. Les modèles de deep learning peuvent apprendre des représentations de haute qualité à partir de grandes quantités de données étiquetées, ce qui leur permet de capturer des modèles complexes dans les données. Ces représentations peuvent ensuite être utilisées pour effectuer des tâches spécifiques telles que la classification de relations. Pour cette raison, nous nous concentrons dans la section suivante sur les travaux de deep learning.L'organisation hiérarchique des différentes méthodes existantes pour la classification des relations peut être très utile pour mener une revue systématique de la littérature.

En utilisant cette hiérarchie pour notre état de l'art Systematic Literature Review, nous pouvons nous assurer que notre revue est exhaustive et complète, en couvrant toutes les méthodes pertinentes pour la classification des relations. En outre, cela nous permettra de fournir des recommandations utiles pour les chercheurs et les praticiens qui travaillent dans ce domaine, et d'orienter la recherche future dans des directions qui peuvent améliorer les performances de la classification des relations. En somme, cette hiérarchie peut apporter une structure et une méthodologie rigoureuse à notre revue, qui peut être utile pour la communauté de recherche en général.

1.5. Revue systématique de la littérature sur la classification de relations

Avec l'acquis du temps qui passe, le domaine de la recherche évolue et, particulièrement, un nombre croissant de recherches se rapportent au domaine de la classification des relations. Cela peut être remarqué par le nombre de documents publiés dans les conférences et/ou les revues, et même avec la création de nouvelles conférences axées sur ce domaine particulier. Cependant, la réalisation d'un état de l'art lié à la classification des relations constitue un travail complexe et nécessite une préparation optimale. La méthode classique de réalisation de l'état de l'art, connue sous le nom

de revue narrative, utilise des méthodes informelles, non systématiques et subjectives pour rechercher, collecter et interpréter les informations. C'est une revue qui utilise des méthodes informelles, non systématiques et subjectives pour rechercher, collecter et interpréter les informations.

Aussi en complement de la présentation des différentes méthodes de classification de relations présenté dans la section précédente, il nous a semblé pertinent d'effectuer une revue systématique de la littérature (Systematic Literature Review en anglais - SLR) (KITCHENHAM et CHARTERS 2007) centrée sur la littérature de travaux portant sur la classification de relations.

L'utilisation d'une SLR présente de nombreux avantages, notamment en offrant une approche rigoureuse et méthodique pour identifier et synthétiser les preuves de manière systématique. En utilisant une SLR, nous pouvons minimiser le risque de biais dans la sélection des papiers et améliorer la qualité globale de leur synthèse.

En adoptant cette méthode, nous avons utilisé le protocole présenté dans la figure 1.4 pour la réalisation des revues systématiques et des méta-analyses. Ce protocole et constitué, principalement, de trois phases : la planification, la conduite et finalement l'établissement de rapports. Chaque phase est associée à plusieurs étapes qui lui sont associées. La phase de planification proposée par cette méthodologie présente les objectifs poursuivis par la revue ainsi que l'identification des questions de recherche. La phase de conduite comprend la stratégie de recherche utilisée pour récupérer les articles pertinents, l'approche d'extraction des données utilisée pour répondre aux questions de recherche et enfin la synthèse des données. Au stade de la stratégie de recherche, nous spécifions les termes de recherche, les critères de sélection des articles et les règles d'évaluation de la qualité qui ont été utilisées pour filtrer les articles recherchés. Enfin, La phase d'établissement de rapports qui concerne principalement l'interprétation des résultats et présente les conclusions de la revue. Nous expliquons dans ce qui suit les différentes phases citées précédemment.

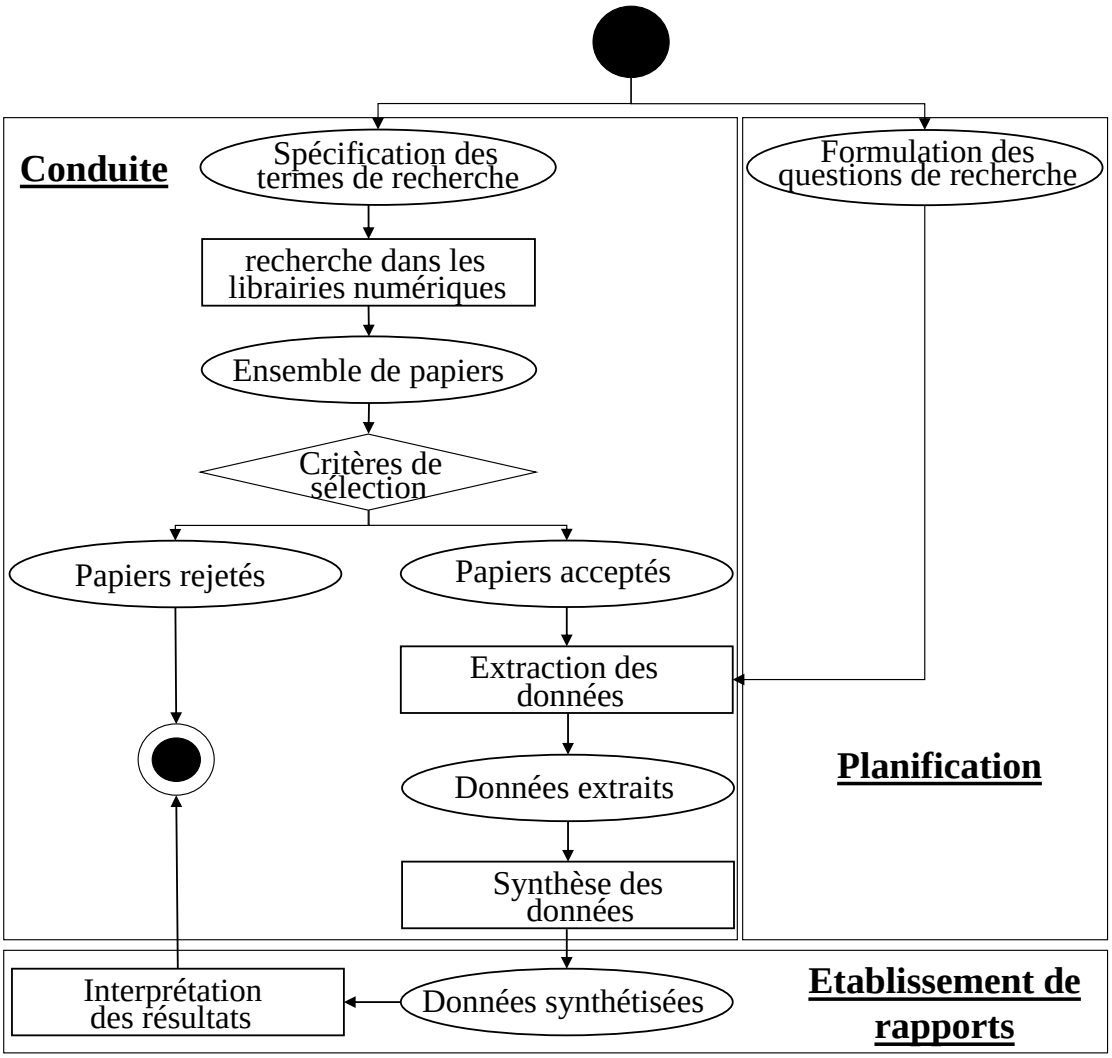


FIGURE 1.4. – Protocole utilisé pour la réalisation des revues systématiques

1.5.1. Planification : Identification des questions de recherche

Les questions de recherche sont une composante essentielle de toute revue systématique de la littérature (SLR) car elles permettent de définir l'objectif de la revue et d'orienter la recherche des études pertinentes. Les questions de recherche doivent être clairement formulées et précises pour permettre une recherche efficace des études pertinentes, tout en étant suffisamment larges pour permettre une analyse globale des résultats.

les questions de recherche dans une SLR peuvent être à la fois génériques et spécifiques.Les questions de recherche génériques telles que celles qui portent sur les informations bibliographiques, l'évolution du nombre et des types de papiers, et les domaines de recherche abordés sont d'une grande importance dans une SLR. Ces

questions permettent d'obtenir une vue d'ensemble du domaine de classification des relations et de mieux comprendre l'état de la recherche existante. La connaissance de l'évolution du nombre et des types de papiers peut aider à identifier les zones de recherche les plus actives et les plus productives.

Les questions de recherche spécifiques peuvent également être très importantes pour approfondir la compréhension du domaine de classification des relations et pour identifier les lacunes dans l'état de l'art actuel. Par exemple, les questions sur les modèles dédiés à la classification des relations, les corpus utilisés pour évaluer les modèles, les métriques d'évaluation les plus appropriées, et les performances des systèmes de classification des relations. Les réponses à ces questions peuvent aider à identifier les approches les plus performantes pour la classification des relations, les lacunes dans les méthodes et à identifier les directions futures pour la recherche dans ce domaine.

En raison de l'importance de comprendre à la fois le contexte général du domaine de classification des relations et les lacunes spécifiques, nous avons choisi de poser les questions de recherche suivantes, qui incluent à la fois des questions génériques et spécifiques :

- QR1 :Quelles sont les informations bibliographiques des études existantes ?
  - Comment le nombre de papiers a-t-il évolué au fil des années ?
  - : Quels sont les types de papiers (les revues, les conférences ou les workshops) utilisés ?
- QR2 : Quels sont les domaines de recherche qui ont été abordés dans la classification des relations ?
- QR3 : Quels sont les modèles d'apprentissage automatique dédiés la classification des relations ?
- QR4 : Quels sont les corpus qui ont été utilisés pour évaluer les approches et les modèles pour la classification des relations ?
- QR5 : Quels sont les métriques d'évaluation les plus appropriés pour évaluer la classification des relations ?
- QR6 : Quelles sont les performances des systèmes de classification des relations en termes des métriques d'évaluation utilisées ?

1.5.2. Conduite

1. **Stratégie de recherche**
   a) **Termes de recherche :** Afin de trouver presque toutes les publications pertinentes, nous établissons un ensemble de termes de recherche qui ont un lien avec notre sujet d'intérêt. Ainsi, trois termes de recherche sont utilisés pour le domaine de la classification des relations : "Relation classification" OR "Classifying relation" OR "Classification of relation".
   b) **Bibliothèques numériques :** Après avoir fixé nos termes de recherche, une recherche automatisée a été effectuée dans quatre bibliothèques numériques (tableau 1.1) pour trouver des publications de recherche pertinentes :

SpringerLink, IEEE Xplore, Google Scholar et ACM Digital Library. Ces bibliothèques ont été sélectionnées car elles offrent un accès facile et permettent de récupérer le texte intégral des articles.

TABLEAU 1.1. – Bibliothèques numériques choisies

Bibliothèques numériques	URL
SpringerLink	http://link.springer.com
IEEE Xplore	http://ieeexplore.ieee.org
Google Scholar	http://scholar.google.com
ACM Digital Library	http://dl.acm.org

c) **Critères d'inclusion et d'exclusion :** Pour filtrer davantage les papiers les plus pertinents, plusieurs critères de sélection ont été appliqués (KITCHENHAM et CHARTERS 2007). Ces critères nous permettent soit de retenir les papiers pertinents soit de les éliminer car ils répondent à un ou des critère(s) d'exclusion. L'ensemble de critères d'inclusion/exclusion que nous avons adoptés sont définis dans le tableau 1.2

TABLEAU 1.2. – Critères d'inclusion et d'exclusion

Type I : Critères d'inclusion et d'exclusion
Articles liés aux questions de recherche (QR1 à QR6)
Papiers complets (Full papers)
Articles publiés entre 2005 et 2021
Articles rédigés en anglais uniquement
Type II : Critères d'exclusion
Articles non liés aux questions de recherche (QR1 à QR6)
Posters
Articles publiés avant 2005 ou après 2021
Articles rédigés dans des langues non anglaises
Les articles dupliqués

La recherche s'est concentrée sur les articles dans le domaine de la classification des relations. Par conséquent, le premier critère était d'inclure les articles publiés entre 2005 et 2021. Ensuite, le deuxième critère était de se concentrer sur les articles écrits en anglais. Après cela, les articles redondants ont été supprimés. Pour le reste des articles, l'objectif de l'article devait être lié aux différentes questions de recherche (QRs).

d) **Règles d'évaluation de la qualité :**Règles d'évaluation de la qualité : L'application des règles d'évaluation de la qualité (QAR) était la dernière étape utilisée pour identifier la liste finale des articles inclus dans cette thèse. Les QAR ont été appliquées pour évaluer la qualité des articles de recherche en fonction

des questions de recherche établies. Six QAR ont été utilisées pour évaluer la qualité des articles, à savoir :
— QAR 1 : Le document est-il bien organisé ?
— QAR 2 : Les objectifs de recherche sont-ils clairement identifiés dans le document ?
— QAR 3 : Le document inclut-il des expériences pratiques ?
— QAR 4 : L'ensemble de données utilisé est-il clairement identifié ?
— QAR 5 : Les résultats des expériences menées sont-ils clairement identifiés et rapportés ?

L'étape d'évaluation de la qualité était basée sur (KITCHENHAM et CHARTERS 2007). Le score de chaque QAR a été défini comme suit : lorsque l'article répond entièrement à la question, il reçoit un score de 1 ; il reçoit un score de 0,5 s'il répond partiellement à la question ; il reçoit zéro s'il ne répond pas du tout à la question. Après la notation, la note globale de chaque article est calculée par la somme des notes des cinq RAQ. Un score de 3 ou plus signifie que l'article a été inclus dans cette revue. Si la note globale est inférieure à 3, l'article est rejeté.

Dans notre cas, cette évaluation a été effectuée par moi-même et ma collègue qui travaille sur la classification des relations. Pour ce faire, nous avons suivi les mêmes critères d'évaluation de qualité définis et convenus précédemment. Nous avons également travaillé de manière indépendante, sans influencer l'autre, pour garantir une évaluation objective. Les résultats de l'évaluation de qualité ont été comparés et les différences entre nous ont été résolues par des discussions pour garantir une évaluation fiable et objective. En fin de compte, une évaluation objective et fiable des règles de qualité est essentielle pour garantir la validité et la fiabilité de la revue.

e) **Phase de recherche :** Les termes de recherche mentionnés précédemment ont été utilisées pour récupérer les articles dans les bibliothèques numériques spécifiées. Au début, le nombre d'articles était de 8257. La première sélection s'est faite par titre et le nombre a été réduit à 996 articles. Ensuite, deux examinateurs ont vérifié de manière anonyme si les articles abordaient une ou plusieurs des questions de recherche. Pour ce faire, les articles ont été importés dans une application web appelée Rayyan (INSTITUTE Accessed : 2 April 2023).
Rayyan est une application web de gestion de revue systématique de la littérature qui permet aux chercheurs de collaborer sur la sélection des articles pertinents pour leur étude. Ell5e est conçue pour aider les chercheurs à organiser et à trier les articles en utilisant un système de vote collaboratif.
Les articles sont téléchargés dans Rayyan et chaque article est examiné par moi et ma collègue indépendants pour déterminer s'il répond aux questions de recherche. Pour ce faire, nous votons pour chaque article en utilisant l'une des trois options : « inclus », « exclu » ou « peut-être ». Si un article reçoit deux

votes « inclus », il est automatiquement inclus dans la revue. Si un article reçoit deux votes « exclu », il est automatiquement exclu. Si un article reçoit un vote « inclus » et un vote « exclu », il est discuté pour prendre une décision finale.
Après l'application des règles d'évaluation de la qualité, seuls 130 ont été identifiés comme pertinents pour la revue de la littérature. Le tableaux 1.3 présente un résumé des résultats de la recherche et des différentes bibliothèques numériques sur lesquelles ils ont été recherchés.
Comme indiqué précédemment, pour identifier les papiers portant sur la classification des relations, nous avons utilisé quatre bibliothèques différentes. La majorité des papiers, soit 85, ont été identifiés à partir de la bibliothèque Google Scholar. En outre, nous avons identifié 20 papiers à partir de la bibliothèque de l'Institute of Electrical and Electronics Engineers (IEEE). De même, nous avons trouvé 14 papiers pertinents à partir de la bibliothèque Springer. Enfin, nous avons identifié 11 papiers pertinents à partir de la bibliothèque de l'Association for Computing Machinery (ACM). En utilisant ces différentes bibliothèques, nous avons été en mesure de rassembler une collection diversifiée de papiers sur la classification des relations qui nous a permis de réaliser notre recherche de manière exhaustive. La bibliothèque Google Scholar est une ressource incontournable pour les chercheurs en informatique et en linguistique computationnelle. Sur les 85 papiers extraits de cette bibliothèque, 66 proviennent de la librairie ACL Anthology. Cette bibliothèque est une collection en ligne de papiers scientifiques en traitement automatique du langage naturel et en linguistique computationnelle.

Parmi les 66 papiers, 17 ont été publiés dans « Annual Meeting of the Association for Computational Linguistics », une conférence de renommée internationale dans le domaine de la linguistique computationnelle. D'autre part, 11 ont été publiés dans la conférence « Conference on Empirical Methods in Natural Language Processing », une autre conférence importante pour les chercheurs en traitement automatique du langage naturel. En outre, 9 papiers ont été publiés dans le cadre de l'International « Workshop on Semantic Evaluation », une conférence qui se concentre sur l'évaluation sémantique et la disambiguïsation dans le traitement du langage naturel. Enfin, 6 papiers ont été publiés dans le cadre de l'International « Conference on Computational Linguistics », qui rassemble des chercheurs et des professionnels du monde entier pour présenter leurs travaux dans divers aspects de la linguistique computationnelle.

2. **Extraction des données :** A ce stade, la liste finale des articles a été utilisée pour extraire les informations nécessaires pour répondre à l'ensemble des questions de recherche. Les informations extraites de chaque article sont les suivantes : le titre de l'article, l'année de publication, le type de publication, le nom de l'auteur, la date de publication, les différentes méthodes utilisées, les types de bases de données, les métriques d'évaluation qui ont été utilisées pour évaluer

TABLEAU 1.3. – Résumé des résultats de recherche de classification des relations

Bibliothèque numérique	Sélection primaire	Sélection par titre	Application des règles d'évaluation de la qualité	% du total des articles pertinents
Google Scholar	7040	591	85	65.5%
IEEE Xplore	104	77	20	15.5%
SpringerLink	804	170	14	10.5%
ACM Digital Library	309	158	11	8.5
Total	8257	996	130	100%

les résultats dans chaque article et enfin une comparaison des différents travaux relatifs à la tâche de classification des relations.

3. **Synthèse des données :** Les informations extraites pour les questions de recherche QR1 à QR4 ont été présentées sous forme de données quantitatives qui ont été utilisées pour développer une comparaison statistique entre les différents résultats pour chaque question de recherche. Ces statistiques élaborées ont permis de découvrir certains modèles ainsi que les directions dans lesquelles la recherche a été menée. En ce qui concerne QR5 et QR6, les données extraites étant qualitatives, une comparaison descriptive a été effectuée pour faire une synthèse sur les différentes approches proposées et identifier les différentes directions futures.

1.5.3. Résultats obtenus

Les résultats obtenus par cette revue systématique de la littérature (SRL) sont des analyses statistiques représentées par des graphiques et des tableaux concernant divers travaux étudiés portant sur la classification de relations. Pour plus de facilité chacun de ces travaux est cité dans ces analyses avec sa référence bibliographique notés AXX (par exemple A05). Les références bibliographiques complètes de ces travaux sont présentées dans l'annexe A de ce manuscrit.

1. **Quelles sont les informations bibliographiques des études existantes ?**
   — Comment le nombre de papiers a-t-il évolué au fil des années ? La figure 1.5 illustre la distribution des papiers de classification des relations par année et type de publication. Les colonnes verticales présentent le nombre d'articles de recherche publiés par année. Comme nous pouvons le voir sur la figure, le nombre de papiers dans le domaine de la classification des relations a augmenté depuis 2016.
   — : Quels sont les types de papiers (les journaux, les conférences ou les workshops) utilisés ?

   Les 130 articles qui ont été identifiés comme pertinents peuvent être divi-

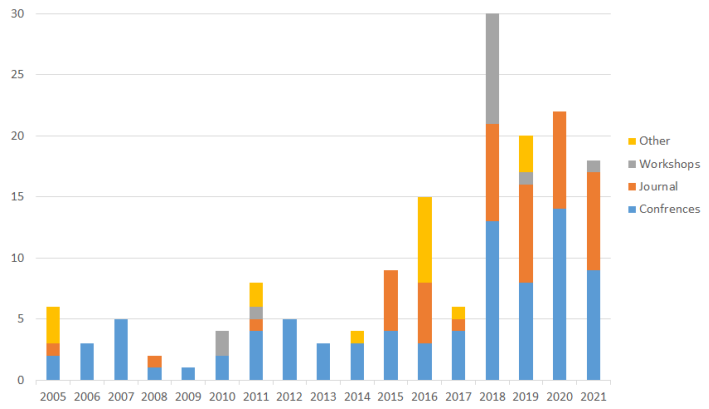


FIGURE 1.5. – Répartition des papiers de classification des relations par année et type de publication

sés en quatre types principaux : papiers de conférence, papiers de journal, papiers de workshops et autres. La figure 1.6 montre la répartition des articles entre ces principaux types. La majorité des articles utilisés dans l'analyse documentaire ont été identifiés comme des articles de conférence (63%). Les 37% restants se répartissent entre les articles de journaux, les articles de workshops et autres, avec respectivement 17.5%, 9.5% et 10%.

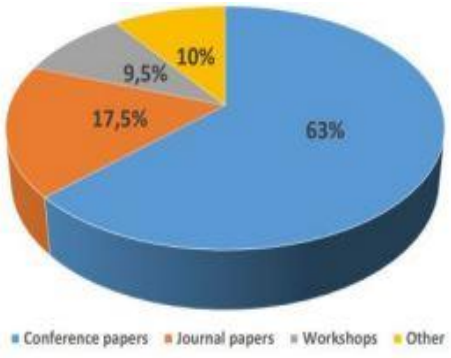


FIGURE 1.6. – Pourcentage de papiers pour chaque type

2. **Quels sont les domaines de recherche qui ont été abordés dans la classification des relations ?**
   Parmi les 130 articles, différents types de classification des relations ont été identifiés, la classification selon des types prédéfinis (par exemple, cause-effet), la classification des relations temporelles, la classification des relations dans les articles scientifiques (USER, RESULT, MODEL, PART-WHOLE, TOPIC and COMPARISON) et la classification des relations lexicales (synonyme, hyperonyme). Le pourcentage d'articles dans chaque type est présenté dans le tableau 1.4. La majorité des articles se situent dans le domaine de la classification des relations

selon des types prédéfinis (74,5%), suivi par environ 11% dans le domaine de la classification des relations temporelles, 8,5% dans le domaine de la classification des relations dans les articles scientifiques et 4% pour la classification des relations lexicales. En outre, 2% des articles ont été classés comme autres. Étant donné qu'un pourcentage énorme d'articles relève du domaine de la classification des relations selon des types prédéfini, une analyse plus approfondie a été effectuée sur ces articles.

3. **RQ3 : Quels sont les modèles d'apprentissage automatique dédiés la classification des relations?** Les modèles utilisés pour classer les relations varient entre l'apprentissage profond, les méthodes non supervisées, l'apprentissage par renforcement et le « Few-Shot ». 50% des papiers utilisent des modèles basés sur l'apprentissage profond. 11,5% des papiers utilisent des modèles non supervisées. Seulement, 2,5% des papiers utilisent l'apprentissage par renforcement. Les méthodes basées sur l'apprentissage « Few-Shot » sont utilisées par environ 17,5% des papiers. 18,5% des papiers utilisent d'autres modèles tels que le méta-apprentissage. Le tableau 1.5 fournit plus de détails.
4. **QR4 : Quels sont les corpus qui ont été utilisés pour évaluer les approches et les modèles pour la classification des relations?** Pour tester les différentes approches proposées, plusieurs « collections de données » ont été utilisées dans les différents papiers. Certaines sont privées, tandis que la majorité des collections étaient publiques et disponibles sur le Web. Le tableau 1.6 présente les différentes collections utilisées pour classer les relations.

   Ces collections comprennent TimeBank Dense, SemEval-2010 Task 8, TempEval3, ACE 2003-2004, TACRED, NYT et « Fewrel ». Comme montre le tableau 1.7, la majorité des travaux utilisent « SemEval 2010 Task8 » pour classer les relations selon des types prédéfinis afin de tester leurs approches. « SemEval 2018 » a été utilisé pour la classification des relations dans les articles scientifiques. En ce qui concerne la classification des relations temporelles, deux collections de données ont été utilisés : « TimeBank Dense » et « TempEval3 ».

TABLEAU 1.4. – Différents domaines de classification de relations identifiés dans les papiers

Domaine de classification des relations	Papiers de conférences	Papiers de journal	Workshops	Autres	Total of papers	%
Classification des relations selon des types prédéfinis	A1, A2, A3, A4, A5, A6, A7, A8, A9, A10, A11, A12, A13, A14, A15, A16, A17, A18, A19, A20, A21, A22, A23, A24, A25, A26, A27, A28, A29, A30, A31, A32, A33, A34, A35, A36, A37, A38, A39, A40, A41, A42, A43, A44, A45, A46, A47, A48, A49, A50, A51, A52, A53, A54, A55, A56, A57, A58, A59, A60, A61, A62	A63, A64, A65, A66, A67, A68, A69, A70, A71, A72, A73, A74, A75, A76, A77, A78, A79, A80	A81, A82, A83, A84, A85	A86, A87, A88, A89, A90, A91, A92, A93, A94, A95, A96, A97	97	74.5%
Classification des relations temporelles (BEFORE, IBEFORE, Begins, Ends, SIMULTENOUS, INCLUDES)	A98, A99, A100, A101, A102, A103, A104, A105, A106, A107	A108, A109, A110		A111	14	11%
Classification des relations dans des articles scientifiques (USE, RESULT, MODEL, PART-WHOLE, TOPIC and COMPARISON)	A112, A113, A114	A115	A116, A117, A118, A119, A120, A121, A122		11	8.5%
Classification des relations lexicales (Hypernym, Hyponym, Antonym, etc)	A123, A124, A125, A126, A127				5	4%
others	A128, A129	A130			3	2%
Total	82	23	12	13	130	100%

TABLEAU 1.5. – Modèles utilisés pour la classification des relations

Méthodes de classification	Modèles utilisés		Papiers associés	nombre de papiers	% de papiers
Méthodes d'apprentissage profond	Modèles basiques	RNN	[A10], [A12], [A124], [A64]	32	24.5%
		CNN	[A63], [A123], [A14], [A9], [A118], [A112], [A119], [A120], [A122], [A113], [A101], [A17], [A71], [A27], [A114], [A110], [A54]		
		LSTM	[A1], [A24], [A53]		
		BLSTM	[A2], [A1], [A65], [A99]		
		Autres	[A62], [A103], [A105]		
	Modèles basés sur les dépendances	BRCNN	[A86], [A7]	11	8.5%
		SDP-LSTM	[A3], [A111]		
		DepNN	[A16], [A69], [A15]		
		DepLCNN + NS	[A13]		
		Autres	[A28], [A5], [A30]		
	Modèles basés sur l'attention	Att-CNN	[A19]	11	8.5%
		Att-BLSTM	[A8], [A11], [A70], [A68]		
		MultAtt-CNN	[A44]		
		Autres	[A20], [A21], [A67], [A31], [A55]		
	Combination	BRCNN+ 2CNN	[A6]	11	8.5%
		DepBLSTM	[A100]		
		LSTM+ CNN	[A125]		
		BLSTM+ CNN	[A26]		
		BLSTM+ Bert	[A107]		
		CNN+B-GRU	[A72]		
		C-LSTM	[A117]		
		BLSTM + GCN	[A57]		

		Att-2CNN+ LSTM Att-RCNN CNN+SDP-LSTM	[A4] [A66] [A25]		
Modèles basés sur l'apprentissage par renforcement			[A39], [A92], [A40]	3	2.5%
Méthodes non supervisées			[A45], [A75], [A48], [A74], [A95], [A46], [A47], [A49], [A84], [A50], [A51], [A76], [A128], [A98], [A108]	15	11.5%
Modèles basés sur la classification « Few-Shot »			[A33], [A34], [A35], [A36], [A91], [A37], [A73], [A38], [A52], [A96], [A78], [A79], [A58], [A59], [A60], [A61], [A41], [A42], [A43], [A93], [A94], [A44]	23	17.5%
Autres			[A115], [A32], [A89], [A126], [A90], [A29], [A127], [A18], [A22], [A23], [A82], [A102], [A83], [A106], [A77], [A56], [A85], [A80], [A88], [A116], [A121], [A81], [A129], [A104]	24	18.5%
Total				130	100%

TABLEAU 1.6. – Différentes collections de test utilisées dans le domaine de classification des relations

Dataset	Number of classes
TimeBank Dense	6
SemEval-2010 Task 8	9
TempEval3	13
ACE	24
TACRED	42
NYT	57
FewRel	100

TABLEAU 1.7. – Les collections de test utilisées dans les papiers

Domaine de classification	Collection de test utilisées	Nombre de papiers
Classification des relations selon des types prédéfinis	SemEval 2010 Task 8	58
	ACE	19
	FewRel	15
	KBP37	9
	Autre	9
Classification des relations dans des articles scientifiques	SemEval 2018	8
	Autre	3
Classification des relations temporelles	TimeBankDense	9
	TempEval 3	5
Classification des relations lexicales	Root09/Bless/ K2H+N/Evalluation	5

5. **QR5 : Quels sont les métriques d'évaluation les plus appropriés pour évaluer la classification des relations ?** Plusieurs métriques d'évaluation ont été utilisées dans les documents de recherche pour évaluer la performance globale du système développé. Le tableau 1.8 illustre les différentes techniques identifiées, ainsi que le pourcentage d'articles ayant utilisé chaque technique. Comme on peut le constater, 72% des articles ont utilisé le score F1 pour évaluer les performances de leur système, tandis que 13,5% ont utilisé l'accuracy. D'autre part, 8,5% des articles ont utilisé la précision et 3% Mean Average precision. 3% des articles ont été classés dans la catégorie « autres », les techniques utilisées par chacun des articles ayant obtenu un score inférieur à 1

TABLEAU 1.8. – Métrique d'évaluation utilisées

Métrique d'évaluation	Nombre de papiers	% de papiers
F1-score	94	72%
Precision	11	8.5%
Accuracy	17	13.5%
Mean Average precision	4	3%
Autres	4	3%

6. **RQ6 : Quelles sont les performances des systèmes de classification des relations en termes des métriques d'évaluation utilisées ?**
Nous nous sommes principalement concentrés sur les articles traitant de la classification des relations selon les neuf types prédéfinis (74,5%). En effet, une analyse plus approfondie a été réalisée sur ces articles afin d'identifier encore plus de détails sur les différentes approches présentées. La majorité des travaux utilisent « SemEval 2010 Task8 » comme corpus pour tester leurs approches. Pour cette raison, dans cette section, nous présenterons les résultats d'évaluation de quelques articles testés sur cette collection de données.

Les expériences menées à l'aide du corpus SemEval-2010 sont présentées dans le tableau 1.8. Les modèles CNN et RNN sont les premiers réseaux de neurones utilisés dans la tâche de classification de relations, et il a été démontré que l'apprentissage profond peut effectivement améliorer l'efficacité de la classification. Cependant, le tableau 1.8 montre que les modèles CNN (D. ZENG, K. LIU, LAI et al. 2014) et RNN (D. ZHANG et Dong WANG 2015) de base ne sont pas satisfaisants car ils n'ont pas de structure spécifique pour la classification des relations : leurs scores sont respectivement de 82,7% et 79,6%. Le modèle MVRNN (SOCHER, HUVAL, MANNING et al. 2012) construit un vecteur et une matrice dans chaque nœud de l'arbre, et son score est de 82,4%.

DepNN (Yang LIU, F. WEI, S. LI et al. 2015) et le SDP-LSTM (Y. XU, MOU, G. LI et al. 2015) sont tous deux basés sur SDP, et ils ont obtenu des résultats similaires, 83,6% et 83,7%, respectivement. Les résultats de DepNN et SDP-LSTM sont seulement 1% plus élevés que ceux de CNN et RNN, ce qui peut être dû au fait que SDP ne considère pas la phrase entière. Le CR-CNN (SANTOS, B. XIANG et B. ZHOU 2015) utilisant la phrase entière, l'intégration des mots et le WPE surpasse tous les résultats précédemment rapportés et atteint un nouveau F1 de 84.1. Il s'agit d'un résultat remarquable car il n'utilise pas de caractéristiques compliquées qui dépendent de ressources lexicales externes telles que WordNet et des outils de TAL. Le modèle BLSTM (S. ZHANG, D. ZHENG, Xinchen HU et al. 2015) applique des informations supplémentaires telles que WPE, POS, NER, WNSYN et DEP au LSTM bidirectionnel par rapport à DepNN et SDP-LSTM et obtient un meilleur score de 84,3%.

Comme le SDP-LSTM, le depLCNN utilise un réseau neuronal convolutionnel pour modéliser le chemin de dépendance et obtient de meilleurs résultats que le DepNN. Mais la stratégie d'échantillonnage négatif proposée dans (K. XU, Y. FENG, Songfang HUANG et al. 2015) est efficace pour modéliser les chemins de dépendance, ce qui permet au depLCNN d'atteindre une performance impressionnante de 85,6%. Comme le DepNN, le modèle BRCNN utilise le CNN pour modéliser le SDP. La différence est que BRCNN (R. CAI, Xiaodong ZHANG et Houfeng WANG 2016a) utilise le SDP dans différentes directions d'entrée, ce qui augmente l'efficacité de la classification à 86,3%.

Comme montre le tableau 1.8, La combinaison de deux ou plusieurs réseaux neuronaux peut améliorer les résultats en tirant parti des RNN et des CNN. Par

exemple, le CNN-SDP-LSTM (P. WANG, Z. XIE et Junfeng HU 2017) a obtenu un score F1 de 84,7%. De même, le réseau BLSTM-CNN (T. XU, DU, C. FU et al. 2018) a atteint un score de 85,9%.

Les expériences montrent que les modèles basés sur l'attention sont plus performants que les modèles basés sur le CNN et le RNN. Par exemple, Att-RCNN (X. GUO, Hui ZHANG, H. YANG et al. 2019a) atteint un score supérieur de près de 4% à celui de CNN et de près de 3% à celui de SDP-LSTM. De plus, par rapport au modèle CR-CNN, bien que les fonctions objectives de CR-CNN et Att-RCNN soient similaires, Att-RCNN améliore le score F1 de 2,5%. Dans les mêmes conditions, Att-RCNN obtient un score F1 supérieur de 1,2% à celui de BRCNN, malgré le fait qu'il lui ressemble beaucoup. De même, les modèles Att-BLSTM (Runyan ZHANG, MENG, Y. ZHOU et al. 2018) et EAtt-BiGRU (P. QIN, W. XU et J. GUO 2017) tentent d'appliquer des mécanismes d'attention à la classification relationnelle. Ils ont atteint respectivement 86,3% et 84,7%, ce qui montre que le mécanisme d'attention peut encore améliorer l'efficacité du réseau neuronal. Le modèle MultiAtt-CNN a également obtenu un score F1 élevé (88,0%). Le modèle AttPoolingCNN utilise principalement une nouvelle fonction de perte sur le CNN.

Contrairement à ces modèles, SSL-KAS-MuBiGRU (L. LI, Jiabing WANG, Jichang LI et al. 2019) améliore la capacité d'extraction du mécanisme d'attention et combine les mots-clés et la phrase originale pour renforcer la structure du réseau neuronal, ce qui améliore les performances (88,1%) malgré l'utilisPation d'un seul mécanisme d'attention. Comme montre le tableau 1.8, La combinaison du CNN et du RNN en ajoutant un mécanisme d'attention (S. SHEN, WEN, L. ZHOU et al. 2018) est la meilleure avec un score F1 égal à 89,3%.
Récemment, le modèle de langage pré-entraîné BERT a été à l'origine d'avancées considérables dans diverses tâches de NLP. La première étude systématique de l'application du modèle BERT pré-entraîné (R-BERT) Shanchan WU et Yifan HE 2019a à l'extraction de relations a été rapportée. Le modèle proposé intègre des informations provenant des entités cibles et ajoute des symboles spéciaux pour marquer la position des paires d'entités afin de mettre en évidence les entités cibles. Il surpasse les modèles proposés précédemment avec un score F1 de 89,25 %. Soares et al. SOARES, FITZGERALD, J. LING et al. 2019a ont étudié les effets des différents modes d'entrée et de sortie du modèle BERT pré-entraîné sur les résultats de l'extraction de relations et ont obtenu le meilleur score F1 avec 89,5%.
En résumé, sur la base des expériences, le tableau 1.9 montre que la combinaison de réseaux avec un mécanisme d'attention et le modèle BERT sont plus appropriés pour la tâche de classification des relations.

TABLEAU 1.9. – Résultats de performance de quelques papiers sur le corpus SemEval 2010 Task 8 dataset

Modèles utilisés		Papiers associés	F1 Score
Modèles de base	MVRNN	[A64]	82.4%
	RNN	[A10]	79.6%
	DRNN	[A124]	86.1%
	CNN	[A63]	84.8%
		[A9]	82.7%
		[A64]	82.32%
	BLSTM	[A2]	84.3%
	CR-CNN	[A14]	84.1%
Modèles basés sur les dépendances	BRCNN	[A86]	86.3%
	SDP-LSTM	[A3]	83.7%
	DepLCNN + NS	[A13]	85.6%
Modèles basés sur l'attention	Att-CNN	[A19]	85.3%
	Att-BLSTM	[A8]	83.7%
		[A11]	84.3%
		[A70]	86.3%
	MultAtt-CNN	[A87]	88.0%
	SSL-KAS-MuBiGRU	[A67]	88.1%
	Att-BiGRU	[A20]	84.7%
		[A55]	85.3%
Modèles basés sur la combinaison	BLSTM-CNN	[A62]	85.9%
	BLSTM-RNN	[A12]	83.1%
	DepN(RNN-CNN)	[A16]	83.6%
	CNN-SDP-BLSTM	[A25]	84.7%
	Att-RCNN	[A66]	86.6%
	Att-CNN-LSTM	**[A4]**	**89.3%**
	GCN+BLSTM	[A57]	85.7%
Modèles basés sur les transformers	Bert	[A85]	89.41%
	Bert+maxpooling	[A116]	89.95%
	R-BERT	[A88]	89.25%
	D-Bert	**[A81]**	**90.1%**
	Bert+MTB	[A129]	89.5%

1.6. Conclusion

Dans ce chapitre, après avoir introduit le domaine de l'extraction d'information et ses applications, nous avons fait un tour d'horizon de méthodes d'extraction de

relations, à savoir les méthodes d'*identification* de relations et les méthodes de *classification* de relations.

Nous nous sommes plus particulièrement intéressé aux méthodes existantes de *classification* de relations, leurs limitations, et en évoquant quelques perspectives, en esquissant ce que pourrait être notre contribution à la classification de relation dans un document textuel.

L'étude bibliographique systématique effectuée en classification des relations montre que la majorité des travaux existants visent à classer les relations selon neuf types prédéfinis (Cause-Effet, Contenu-Contenant, Entité-Destination, etc.) en tenant compte uniquement de l'aspect syntaxique du texte. Ces différentes méthodes de classification de relations négligent un aspect important qui est la sémantique, et en particulier le contexte associé au document. Une relation identifiée peut être pertinente ou non en fonction du contexte. Ainsi, considérons les deux phrases suivantes :
(1) La cuisinière à gaz est dans la cuisine.
(2) Jean est de bonne humeur.
Si l'on se fie aux approches existantes de classification des relations, les deux relations exprimées par les deux phrases ci-dessus sont classées dans le même type prédéfini « Content- Container ». Ces approches ne tiennent pas compte du contexte qui dépend de la nature des entités. Alors que la première phrase exprime une relation « Content-Container », la seconde exprime un état d'esprit qui, hormis la syn- taxe, n'a rien à voir avec la relation « Content-Container ». Le contexte est tout à fait essentiel puisqu'il permet de préciser la nature du contenu d'un document ou ce que représente une relation.

En nous appuyant sur la dernière limitation, il ne nous semble pas exister de méthode ou d'approche permettant la classification des relations selon un contexte précis.

Afin de combler cette lacune, nous proposons dans le chapitre 4 une nouvelle approche qui consiste alors à classer les relations en fonction de leurs types tout en tenant compte du contexte. Elle permet d'obtenir un degré de « contextualisation » des relations. L'ajout de capacités sémantiques et d'une contextualisation plus précise aux relations extraites de documents texte non structurés pourrait être intégré à différentes applications comme la prise de décision. En particulier, dans le cadre des moteurs de recherche, la classification des relations en fonction d'un contexte précis améliore l'efficacité de la recherche d'informations.

Ce travail de cette thèse étant focalisé sur la classification des relations de documents textuels selon leur contenu, le chapitre suivant cernera tout d'abord cette notion de « contexte » permettant de prendre en compte ce contenu, ainsi que sur les méthodes d'extraction de notions voisines liées au contenu d'un document.

2. Extraction de contenu d'un document textuel, notion de contexte

Sommaire

2.1. Introduction

La plupart des documents textuels que nous utilisons sont électroniques, qu'il s'agisse par exemple de résultats de recherche (articles scientifiques), de rapports de médecins (rapports cliniques), de rapports de police (enquêtes), d'articles de presse ou d'enquêtes, etc. Les documents électroniques sont des versions numériques des documents papier qui peuvent être créées, modifiées, visualisées et partagées sur un ordinateur ou un appareil mobile. Le traitement automatique de ces documents est une tâche dont la complexité peut varier considérablement notamment selon leur niveau de structuration.

De façon générale un document électronique est caractérisé par des métadonnées qui décrivent ou identifient le document lui-même. Ces métadonnées permettant une meilleure utilisation de son contenu. Certaines métadonnées sont attribuées automatiquement et sont liées stockage du document (taille, date de création, date de dernière modification), d'autres liées à leur contenu, plus sémantiques, sont principalement fixées par ses auteurs (nom du document, titre, mots-clés, résumé, etc.). Ces informations peuvent être utilisées pour organiser, rechercher et récupérer des documents électroniques.

Le traitement automatique de documents électroniques textuels non-structurés doit de plus en plus prendre en compte le contenu du document. Cette prise en compte s'avère très complexe, et d'autres métadonnées peuvent alors être fort utiles, notamment celle de ***« contexte »*** d'un document, que nous proposons dans cette recherche.

Dans un premier temps, le ***« contexte »*** d'un document peut être vu comme une métadonnée permettant une certaine synthétisation de son contenu. Ce contexte permettra notamment de mieux classer et retrouver un document donné dans un ensemble conséquent de documents, et de mieux exploiter son contenu, notamment dans le classement des relations qu'il contient, auquel ce travail de thèse s'intéresse plus particulièrement.

Dans ce chapitre nous essayons dans la section 2.2 de cerner la notion de ***« contexte »*** d'un document textuel non structuré, notion pas encore formellement définie. Dans les sections 2.3 à 2.7 nous nous intéressons à des méthodes d'extraction de contenu basées sur des notions voisines de celle du contexte mais mieux formalisées, à savoir des méthodes d'extraction de mots-clés, de résumés, de titres et de thèmes. Dans la section 2.8, en complément des sections précédentes, nous présentons une revue systématique de la littérature ciblée sur l'extraction de ces notions voisines du contexte, en adoptant la méthodologie déjà utilisée dans le chapitre précédent pour la classification de relations. Enfin, dans la section 2.9 nous concluons par un bilan sur les méthodes d'extraction de contenu actuelles, leurs limites et nous esquissons les grandes lignes d'une nouvelle méthode d'extraction de contenu basée sur une notion de contexte formalisée permettant de combler certaines de ces limites, méthode qui sera développée dans le chapitre suivant.

2.2. Notion de « contexte » d'un document

La prise en compte du ***« contexte »*** d'un document est une pratique courante depuis des siècles. Baruch Spinoza (1632-1677) disait qu'il faut regrouper les différentes idées et concepts d'un livre en thèmes principaux, simplifiant ainsi l'organisation et la classification des informations contenues dans le document (SPINOZA 1974). En considérant le ***« contexte »*** d'un document comme une métadonnée, on obtient une synthèse concise et précise du contenu, ce qui permet de classer efficacement les documents pour faciliter leur recherche et leur utilisation.

Ainsi le ***« contexte »*** d'un document textuel apparaît comme une métadonnée associée au document constituée d'un ensemble d'éléments qui éclaire sur son contenu et par la même, renvoie aux notions de *thème*, *titre*, *résumé*, *sujet* et *mots-clés* pouvant être associés à ce document. Cependant cette notion de ***« contexte »*** est plus puissante en fournissant une « synthèse » plus élaborée du contenu du document. A ce stade, la question qui se pose alors est la suivante : En quoi cette nouvelle notion de ***« contexte »*** dans un document diffère-t-elle des notions plus familières et mieux formalisées telles que le *thème*, le *titre*, le *résumé*, le *sujet* et les *mots-clés*?

Tout d'abord il faut distinguer le contexte du thème d'un document. Considérons les deux phrases suivantes :

— « L'élection présidentielle indienne de 2022 sera la 16ème élection présidentielle organisée en Inde » (P1)

— « Les instituts de sondage prévoient que le président Emmanuel Macron a remporté l'élection française de 2022. » (P2)

Ces deux phrases peuvent être classées dans le même *thème « politique »*. Cependant, le *thème* identifié n'est pas assez précis pour caractériser au mieux le *contexte* réel de chaque phrase. En effet, la première concerne le *contexte« élection présidentielle indienne de 2022 »*, alors que la seconde fait référence au *contexte « l'élection présidentielle française de 2022 »*. Ainsi le *thème* associé à un document, est trop général, il ne suffit pas à caractériser son *contexte*.

Bien que le *titre* ou le *résumé* d'un document puissent sembler de bons candidats pour représenter son contexte, il est important de prendre en compte certaines limitations. La qualité du *titre* et du *résumé* dépend étroitement de la compétence et du style d'écriture de l'auteur. En effet, l'auteur peut avoir une vision subjective de son propre travail, ce qui peut affecter la qualité de ces éléments.De plus, certains documents peuvent avoir des *titres* très similaires et pourtant traiter de contenus très différents. Et par là même avoir des *contextes* différents.

Quant à la notion de *sujet*, qui apparaît proche de la notion de *contexte*, elle semble être une généralisation associée à un corpus de documents et non, comme pour le *contexte*, être lié à un document particulier.

Concernant les *mots-clés*, qu'ils soient définis par les auteurs ou extraits automatiquement, ils sont très souvent les métadonnées qui représentent le mieux le contenu d'un texte. Cependant, l'ensemble des mots extraits d'un document ne permet pas de synthétiser son contenu, ce que doit permettre la notion de *contexte*.

En conclusion, le *titre*, le *thème*, le *résumé*, le *sujet* ainsi que les *mots-clés* sont

des notions générant des métadonnées d'un document, qui ne permettent pas de cerner le contenu d'un document de manière satisfaisante pour sa classification et son traitement.

Nous proposons une nouvelle métadonnée caractérisant le contenu d'un document, basée sur la notion de *« contexte »*, défini comme un groupe structuré de mots extraits du document permettant de caractériser le plus précisément possible son contenu,« ce dont parle le document ».

Dans la section suivante nous allons présenter un ensemble de méthodes utilisées dans l'extraction de contenu d'un document textuel permettant de définir des métadonnées basées sur les notions voisines du *contexte*, comme celles de *mots-clés*, de *titre*, de *résumé* et de *thème*, notions mieux formalisées que celle de *« contexte »*.

2.3. Méthodes d'extraction de contenu d'un document textuel centrées sur les mots-clés, les résumés, les titres et les thèmes

La notion de ***« contexte »*** est évoquée à différents niveaux du traitement de l'information linguistique sans pour autant avoir été toujours clairement définie. Son rôle dans la compréhension des textes a été bien mis en évidence par de nombreux travaux en intelligence artificielle. Néanmoins, il n'existe pas à notre connaissance de méthode d'extraction de contexte d'un document basée explicitement sur le concept de ***« contexte »***.

Cependant, nous avons vu précédemment que la notion de ***« contexte »*** d'un document est voisine des notions de *titre*, *thème*, *résumé* et *mots-clés* d'un document, pour lesquelles plusieurs méthodes d'extractions ont déjà été proposées. Aussi dans les sections suivantes nous présentons différentes méthodes d'extraction des *mots-clés*, de *résumé automatique*, de *titre* et de *thème*, comme l'illustre la Figure 2.1. .

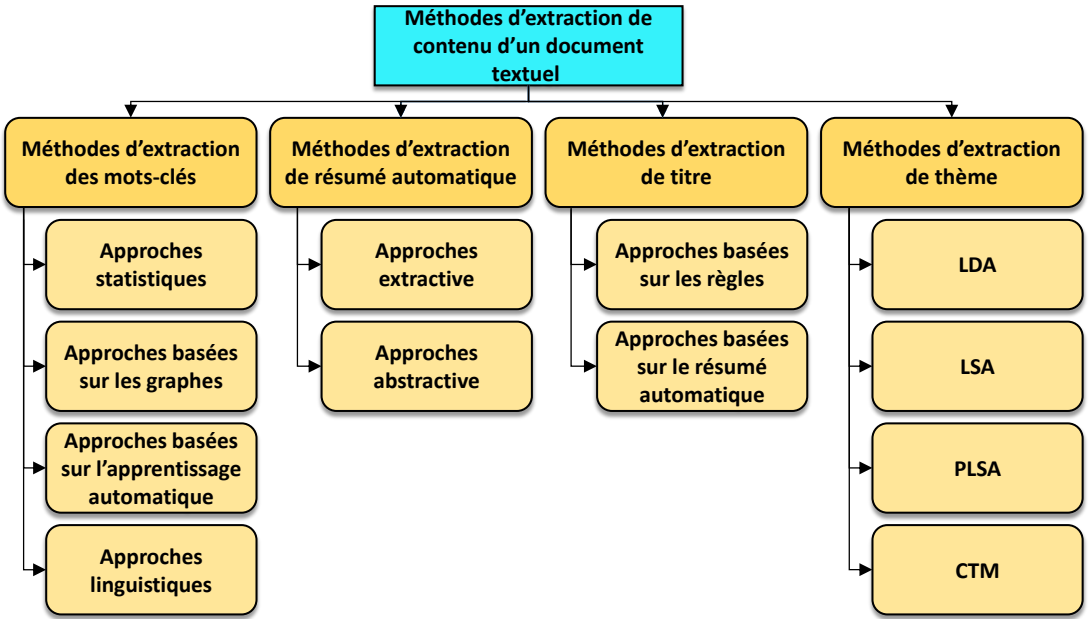


FIGURE 2.1. – Méthodes utilisées pour l'extraction de contexte

2.4. Méthodes d'extraction des mots clés

Les *mots-clés* jouent un rôle crucial dans la recherche de données dans une base de données ou un moteur de recherche, et sont donc des éléments essentiels d'un document. Dans le domaine du traitement du langage naturel, il existe des techniques pour identifier automatiquement les mots les plus pertinents d'un document.

L'extraction automatique de *mots-clés* est un processus de traitement du langage naturel qui consiste à identifier les termes les plus importants d'un texte en utilisant des algorithmes et des modèles. L'avantage de cette méthode est qu'elle élimine la subjectivité inhérente à l'extraction manuelle, permettant ainsi de traiter de grands volumes de données de manière rapide et efficace. De plus, cette technique peut réduire ou même éliminer les incohérences qui peuvent survenir lors de l'analyse manuelle des textes.

Il existe différentes méthodes d'extraction automatique de *mots-clés*, dont les plus connues sont celles basées sur les approches statistiques, celles basées sur les approches linguistiques, les approches basées sur les graphes et enfin les approches basées sur l'apprentissage automatique.

2.4.1. Approches statistiques

L'utilisation de statistiques est l'une des méthodes les plus simples pour identifier les *mots-clés* d'un document et peuvent être efficaces en termes de calcul. Ces approches ne nécessitent pas de données d'entraînement pour extraire les *mots-clés*. Leur idée de base est de trouver le score des mots présents dans le document en utilisant

différents types de statistiques calculées sur un seul ou plusieurs documents. Une fois que les scores sont calculés, les mots sont ordonnés en fonction de leurs poids et les n premiers sont identifiés comme *mots-clés* du document. Parmi les méthodes statistiques les plus simples on peut citer la fréquence des mots (*Word Frequency*), la collocation des mots et la cooccurrence *(Word Collocations and Co–occurrences*). Toutefois, il existe également des méthodes plus sophistiquées, telle que *TF-IDF*.

- ***Fréquence des mots (Word Frequency) :*** (BAAYEN 2001) consiste à identifier les mots les plus fréquemment utilisés dans un texte ou un corpus de textes. Cette méthode s'appuie sur le principe que les mots les plus fréquents dans un texte sont souvent ceux qui reflètent le mieux le contenu et le thème de ce texte. Pour extraire les mots clés par fréquence des mots, on commence par identifier tous les mots du texte, puis on compte le nombre d'occurrences de chaque mot. Les mots les plus fréquents sont ensuite sélectionnés comme mots clés potentiels. Cette méthode est simple et rapide à mettre en œuvre, cependant elle peut être biaisée par les mots les plus courants qui ne sont pas toujours pertinents pour l'analyse du contenu. De plus, Cette méthode considère les documents comme un simple « sac de mots », ignorant ainsi les aspects cruciaux liés au sens, à la structure, à la grammaire et à la séquence des mots.
- ***Collocation des mots et Cooccurrence (Word Collocations and Co–occurrences) :*** (NOVOA GREEN 1992) La collocation des mots et la cooccurrence sont des méthodes utiles pour l'extraction de mots clés dans un texte ou un corpus de textes.Pour extraire des mots clés à partir des collocations, on commence par identifier les paires ou les groupes de mots qui ont une fréquence de cooccurrence significative dans le texte. On peut ensuite utiliser ces combinaisons de mots pour générer des termes clés et des expressions à partir desquels on peut déduire les thèmes et les sujets principaux abordés dans le texte. Quant à la cooccurrence, elle peut également être utilisée pour extraire des mots clés en identifiant les mots qui apparaissent fréquemment ensemble dans le texte ou le corpus. Les mots les plus fréquemment associés peuvent être extraits comme mots clés potentiels, qui peuvent ensuite être utilisés pour résumer le contenu et le thème du texte. En utilisant à la fois la collocation des mots et la cooccurrence, il est possible de générer une liste de mots clés qui reflètent avec précision le contenu et les sujets principaux du texte ou du corpus de textes.
- ***TF-IDF (Term Frequency, Inverse Document Frequency) :*** (SALTON et BUCKLEY 1988) est une formule qui mesure l'importance d'un mot contenu dans un document, relativement à un corpus. L'importance d'un mot est un facteur calculé en fonction de la fréquence d'un mot contenu dans un document (term frequency(TF)) et de la fréquence inverse par rapport aux documents (inverse document frequency (IDF)). La multiplication de ces deux fréquences donne le score TF-IDF d'un mot dans un document. Plus le score est élevé, plus le mot est pertinent pour le document. Le poids d'un mot exprimé selon le TF-IDF est calculé par la formule suivante.

$$W_{ij} = TF_{ij} IDF_i = TF_{ij} log_2(\frac{N}{DF_i}) \tag{2.1}$$

Avec :

$TF_{ij} = \frac{Nombred'occurrencedumotidansledocumentj}{Nombretotaldemot}$
DF_i : représente le nombre de documents qui contiennent le mot i
$IDF_i = log_2(\frac{N}{DF_i})$
N : représente le nombre total des documents dans le corpus

Les algorithmes TD-IDF ont plusieurs applications dans l'apprentissage automatique. Lorsqu'il s'agit d'extraction de mots-clés, cette métrique peut aider à identifier les mots les plus pertinents d'un document (ceux qui ont les scores les plus élevés) et à les considérer comme des mots-clés. Généralement, un mot qui apparaît dans un seul document mais qui n'apparaît pas dans les autres peut être très important pour comprendre le contenu de ce document.

2.4.2. Approches basées sur les graphes

Les méthodes basées sur les graphes génèrent un graphe de mots liés à partir des documents. Ces méthodes utilisent un algorithme de classement comme PageRank qui prend en compte la structure du graphe pour calculer l'importance des sommets. Ce calcul s'appuie non seulement sur les informations locales spécifiques au sommet, mais aussi sur les informations globales calculées récursivement à partir du graphe entier. L'une des méthodes basées sur les graphes les plus connues est TextRank.

TextRank (MIHALCEA et TARAU 2004) est basée sur le calcul du score d'importance des sommets en utilisant le principe de vote ou de recommandation entre deux sommets et inspiré de l'algorithme Pagerank (BRIN et PAGE 1998). TextRank utilise une représentation efficace d'un document, elle peut aussi être utilisée pour faire des résumés automatiques d'un document. L'algorithme d'extraction du mot clé TextRank se déroule comme suit :

— Premièrement, le texte est segmenté et annoté avec une partie de discours (part of speech qui sert à identifier le type de chaque unité lexicale du texte : nom, verbe, adjectif, etc.), une étape de prétraitement est nécessaire pour permettre l'application de filtres syntaxiques (qui sélectionnent uniquement les unités lexicales d'une certaine partie de discours, par exemple, on ajoute au graphe seulement les noms et les verbes, seulement les noms et les adjectifs, etc.). Pour éviter une croissance excessive de la taille du graphe en ajoutant toutes les combinaisons possibles de séquences composées de plus d'une unité lexicale (n-grammes), on considère seulement des mots simples comme candidats pour l'addition au graphe, avec des mots-clés multi-mots reconstruits dans la phase de post-traitement.
— Ensuite, les unités lexicales qui passent le filtre syntaxique sont ajoutées au graphe, et une arête est ajoutée entre les unités lexicales qui co-occurrent dans

une fenêtre de N mots. Une fois le graphe construit (graphe non pondéré non orienté), le score associé à chaque sommet est fixé à une valeur initiale de 1, la formule (1) et l'algorithme de classement sont appliqués sur le graphe pour plusieurs itérations jusqu'à ce qu'il converge. Généralement pour 20 - 30 itérations, à un seuil de 0,0001.

— Une fois un score final est obtenu pour chaque sommet dans le graphe, les sommets sont triés et les meilleurs (qui ont les poids les plus élevés) T sommets dans le classement sont conservés pour le post-traitement. T peut-être défini à n'importe quelle valeur fixe, généralement de 5 à 20 mots-clés, on utilise dans ce travail une méthode plus flexible, qui décide le nombre de mots-clés en fonction de la taille du texte. T est fixé à un tiers du nombre de sommets dans le graphe.

— Pendant le post-traitement, toutes les unités lexicales sélectionnées comme mots-clés potentiels par l'algorithme TextRank sont marquées dans le texte, et les séquences de mots-clés adjacents sont regroupées en un mot-clé multi-mots. Par exemple, dans le texte « Matlab code for plotting ambiguity functions », si à la fois les mots « Matlab » et « code » sont sélectionnés comme mots-clés potentiels par TextRank, puisqu'ils sont adjacents, ils sont réduits en un seul mot-clé « Matlab code ».

2.4.3. Approches linguistiques

Les approches linguistiques pour l'extraction des mots clés se concentrent sur l'analyse des caractéristiques linguistiques des textes. Ces approches utilisent souvent des techniques d'analyse lexicale, syntaxique et sémantique pour extraire les termes clés.

L'analyse lexicale se concentre sur les propriétés des mots individuels, comme leur fréquence et leur position dans le texte. Les approches basées sur l'analyse lexicale peuvent inclure l'utilisation de méthodes telles que la fréquence de termes et la mesure de la spécificité du terme. L'analyse syntaxique se concentre sur les relations entre les mots dans le texte, en utilisant des techniques telles que l'analyse de dépendance et l'analyse en constituants. Cette approche peut aider à identifier les phrases clés et les relations entre les termes.L'analyse sémantique se concentre sur le sens des mots et les relations sémantiques entre eux. Les approches basées sur l'analyse sémantique peuvent inclure l'utilisation de techniques telles que l'analyse de similarité sémantique et l'analyse de cooccurrence de concepts.

Ces approches linguistiques sont généralement plus complexes que les approches statistiques, car elles impliquent souvent des connaissances linguistiques et des règles spécifiques pour identifier les termes clés dans les textes. Cependant, elles peuvent également fournir des résultats plus précis et plus complets en identifiant les relations sémantiques et les contextes d'utilisation des termes clés.

Hulth et al. (HULTH 2003) ont proposé une méthode d'extraction de mots clés qui se base sur l'intégration de connaissances linguistiques avancées. La contribution majeure de ce travail est d'apporter des améliorations significatives aux méthodes d'extraction de mots clés en utilisant des informations linguistiques plus avancées,

telles que les segments constitués de phrases nominales (NP-chunks), plutôt que de se fier uniquement à la fréquence des termes ou aux n-grammes. La méthode proposée par Hulth utilise un ensemble de règles syntaxiques pour identifier les NP-chunks dans les textes, qui sont ensuite considérés comme des candidats potentiels pour les mots clés. Ensuite, un score de pertinence est calculé pour chaque candidat en se basant sur diverses caractéristiques telles que la position dans le document, la longueur, le nombre de co-occurrences avec d'autres termes, etc. Rose et al. (ROSE, ENGEL, CRAMER et al. 2010) ont proposé une méthode d'extraction de mots clés qui utilise des techniques linguistiques avancées pour identifier les termes les plus importants dans un document. Contrairement aux approches statistiques qui se basent sur des critères tels que la fréquence des mots, l'approche proposée dans ce papier prend en compte la structure syntaxique des phrases et les relations entre les mots pour identifier les expressions nominales qui sont les plus pertinentes.

2.4.4. Approches basées sur l'apprentissage automatique

Les méthodes d'extraction de mots clés basées sur l'apprentissage automatique utilisent des algorithmes d'apprentissage pour entraîner un modèle à identifier les mots clés dans un texte. Le modèle est généralement entraîné sur un ensemble de données étiquetées où chaque document est annoté avec les mots clés pertinents. Ces méthodes peuvent utiliser une variété de caractéristiques pour représenter les mots, telles que la fréquence de co-occurrence des mots, les caractéristiques lexicales et syntaxiques, les traits sémantiques et contextuels, etc.

Parmi les méthodes les plus populaires, on peut citer l'algorithme de classification Bayésienne, les modèles de Markov cachés (HMM), les réseaux de neurones, les forêts aléatoires, et les algorithmes de clustering. Les modèles d'apprentissage automatique ont l'avantage de pouvoir généraliser les règles d'extraction de mots clés à partir des données d'entraînement pour identifier les mots clés dans des textes inconnus. Cependant, ces méthodes nécessitent souvent une grande quantité de données annotées pour atteindre une précision élevée,

Sarkar et al. (SARKAR, NASIPURI et S. GHOSE 2012) ont comparé trois algorithmes d'apprentissage automatique (arbres de décision, Bayes naïf et réseaux de neurones artificiels) pour l'extraction de mots clés à partir de documents. Les auteurs ont utilisé deux ensembles de données différents pour évaluer la performance de chaque algorithme en termes de précision, de rappel et de F1-score. Les résultats ont montré que les réseaux de neurones artificiels ont obtenu les meilleurs résultats en termes de précision et de F1-score, tandis que les arbres de décision ont obtenu les meilleurs résultats en termes de rappel. Les auteurs ont conclu que l'utilisation de l'apprentissage automatique pour l'extraction de mots clés peut améliorer considérablement la précision et le rappel par rapport aux méthodes statistiques traditionnelles. Jo et al. (JO et J.-H. LEE 2015) ont proposé une méthode de représentation sémantique des documents basée sur des réseaux de croyance profonde qui permettent de capturer les relations complexes entre les mots dans le document. Ils utilisent ensuite cette représentation pour identifier les mots clés en utilisant une approche de classification

multi-étiquettes.

2.5. Méthodes d'extraction de résumé

Le résumé automatique de texte est le processus de compresser automatiquement le texte d'entrée en une version courte, tout en préservant son contenu d'information et sa signification globale. . Il existe deux grands types de méthodes de résumé automatique de texte : les méthodes « extractives » et les méthodes « abstractives ». Un résumé obtenu par une méthode extractive résulte de la concaténation de phrases extraites du document source. Ces phrases extraites sont considérées comme des phrases clés du document source. La nouvelle séquence de phrases résultant de cette concaténation représente le résumé du texte. En revanche, un résumé construit par une méthode abstractive est composé de phrases qui n'appartiennent pas au texte source et sont construites par un modèle intelligent basé sur l'apprentissage automatique.

2.5.1. Méthodes Extractives

Le résumé automatique extractif est une technique de résumé qui consiste à extraire les phrases les plus pertinentes d'un texte pour les assembler en un résumé cohérent. les systèmes basés sur des méthodes extractives ne prennent pas beaucoup de temps. De plus, comme le résumé final est obtenu en sélectionnant un certain nombre de phrases extraites du texte source, il est normalement bien écrit et grammaticalement correct, mais aussi dépend au document original. On trouve plusieurs travaux qui utilisent des techniques ou des méthodes qu'on peut classer en trois classes ou familles : les méthodes basées sur les graphes, les méthodes basées sur les l'apprentissage automatique, et les méthodes basées sur les statistiques.

2.5.1.1. Méthodes basées sur les graphes

Les méthodes basées sur les graphes modélisent un document sous forme d'un graphe où les sommets représentent les unités textuelles à traiter (mots, termes, phrases, etc.) et les arcs / arêtes représentent les relations entre les unités textuelles (occurrence, cooccurrence, similarité sémantique, chevauchement de contenu, etc.). Parmi les méthodes à base de graphe, nous pouvons citer : (1) REG et (2) TextRank.

1. **REG :** REG (TORRES-MORENO et RAMIREZ 2010)(REsumeur à base de Graphes) est une approche à base de graphes pour aborder, dans le traitement du langage naturel, le résumé automatique de document. L'algorithme modélise un document sous forme de graphe, pour obtenir des phrases pondérées, puis extraire celles qui ont les scores les plus élevés et les concaténer afin de construire le résumé. REG est basé sur deux phases principales : (a) une phase de représentation des documents, réalisée à travers une représentation vectorielle indépendante de la langue et (b) une phase de pondération des phrases, réalisée à travers un algorithme d'optimisation glouton.

a) Prétraitement et représentation vectorielle : dans cette phase, les documents sont prétraités en utilisant des algorithmes de normalisation et de lemmatisation (BARANES et SAGOT 2014). La normalisation est le processus de canonisation des mots de sorte que les correspondances se produisent avec les mêmes mots qui ont d'autres formes malgré les différences superficielles dans les séquences de caractères. Par exemple, si on cherche le mot « Tunisien », on peut espérer d'également faire correspondance aux documents qui contiennent les mots « Tunisiens », « Tunisienne » et « Tunisiennes ». La lemmatisation (BARANES et SAGOT 2014) est une tâche qui a pour but d'extraire, pour chaque forme que peut prendre un mot (nom, verbe, pluriel, singulier, etc.), sa forme canonique (son lemme)) pour éliminer les mots qui n'ont pas d'influence au sens générale de la phrase ou du document ce qui emmène à réduire la dimensionnalité.

b) Solution gloutonne : dans cette phase, on va créer un graphe en se basant sur le modèle vectoriel de représentation des documents décrit dans la phase précédente, où les sommets S du graphe représentent les phrases et A l'ensemble d'arêtes qui représentent les relations entre les phrases. On crée une arête entre deux sommets si les phrases correspondantes ont au moins un mot en commun. Une matrice d'adjacence est construite par la suite à partir de la matrice S[phrases x mots] en suivant cet enchaînement : Si l'élément Si, k = 1 de la matrice S (c'est-à-dire le mot k est présent dans la phrase i), on vérifie dans la colonne k (si le même mot existe dans la phrase j) et quand un élément Sj, k=1 on met 1 dans la case Ai, j de la matrice d'adjacence A, ce qui signifie que les phrases i et j partagent le mot k. Sinon on met 0, ce qui signifie que les phrases i et j ne partagent pas le mot k. Pour sélectionner les phrases pertinentes pour le résumé, les auteurs ont trouvé qu'il faut chercher une variante du problème de l'arbre de poids maximum, où les poids sont sur les sommets, pas sur les arêtes. Ils ont ainsi construit un algorithme inspiré de l'algorithme de Kruskal (GOULD 2012).

2. **TextRank :** L'autre application TextRank (WONGCHAISUWAT 2019) consiste à extraire les phrases pertinentes pour un résumé automatique. Le problème d'extraction des phrases est similaire à celui de l'extraction des mots-clés, puisque les deux applications visent à identifier les séquences les plus représentatives pour un texte donné. Dans cette application, les unités de texte candidates sont des phrases entières, et donc un sommet est ajouté au graphe pour chaque phrase dans le texte. Un lien entre deux sommets est ajouté dans le graphe s'il y a une relation de similarité entre ces phrases correspondantes, où la similarité entre deux phrases est mesurée en fonction de leur chevauchement de contenu. Le chevauchement de deux phrases peut être déterminé comme le nombre de tokens (unités lexicales) communs entre les représentations lexicales des deux phrases.

2.5.1.2. Méthodes basées sur les statistiques

Les méthodes statistiques sont basées essentiellement sur des formules mathématiques. Dans le contexte de leur utilisation dans le résumé automatique de texte, ils servent généralement à fournir des mots/phrases pondérés qui indiquent leurs importances dans le document afin de décider par la suite s'ils/elles vont être inclus(es) dans le résumé ou non. Nous ne présenterons ici que deux modèles significatifs, (1) TF-IDF et (2)OKAPI BM-25.

1. **TF-IDF :** Il s'agit d'accorder un poids à chaque mot du texte afin de mesurer le degré d'importance de ce dernier dans un document donné. L'idée de base de cette méthode se manifeste par le fait qu'un mot qui se répète souvent dans un document et pas dans d'autres signifie que ce mot est important pour ce document. Par conséquent, les phrases qui contiennent plus de mots importants sont importantes et vont par la suite être incluses dans le résumé.
2. **Okapi BM-25 :** Le modèle Okapi BM-25 (MINING 2006a) est vu comme un tf-idf qui prend mieux en compte la longueur des documents. De plus, il suit le même principe qui est le calcul du poids de chaque mot du texte, par la suite le calcul du poids pour chaque phrase est fait afin d'extraire les phrases les plus importantes du texte qui vont représenter son résumé.

   Sa définition est donnée par la formule suivante :

$$W_{ij} = TF_{BM25}(i,j) \times IDF_{BM25}(i)$$

$$W_{ij} = \frac{TF_{ij}(k_1+1)}{TF_{ij} + (k_1(1-b) + b(\frac{dl_j}{avgdl}))} \times log_2(\frac{N - df_i + 0.5}{dF_i + 0.5}) \tag{2.2}$$

   Avec :
   tf_ij représente la fréquence d'apparition du mot i dans le document j
   df_i représente le nombre de documents contenant le mot i
   N représente le nombre total des documents du corpus
   dl_i représente le nombre de mots dans le document j
   avgdl représente le nombre moyen des mots dans un document du corpus

   les valeurs de b et k_1 sont déterminées à partir de l'expérimentation, les valeurs b = 0.75 et k_1 = 2 donnent les meilleurs résultats ; ce sont des constantes.

   Pour montrer le fonctionnement de la méthode Okapi BM-25, on va présenter son application sur un exemple. Comme Okapi BM-25 est vu comme un TF-IDF qui prend mieux en compte la longueur des documents, on va suivre l'exemple rencontré dans tf-idf. On va calculer le poids du mot « élection » dans un document spécifique. On suppose que ce document contient 500 mots, avec une occurrence du mot élection de 15. Le corpus de document contient 3000 documents, le mot élection apparaît dans 20 d'entre eux. Le nombre d'occurrence maximale de ce mot dans ces 20 documents est 30. Le nombre moyen des mots

dans un document du corpus est 600. On va calculer maintenant le poids du mot élection à partir de ces données :
$tf(élection) = \frac{15}{17} = 0.88$

Okapi BM-25(élection) = $\frac{0.88(2+1)}{0.88+2((1-0.75)+0.75(\frac{500}{600})} \times log2(\frac{3000-20+0.5}{20+0.5} = 7.21$

2.5.1.3. Méthodes basées sur l'apprentissage automatique

Le domaine d'apprentissage automatique est concerné par la question de savoir comment construire des programmes ou des logiciels pour l'ordinateur qui s'améliorent automatiquement avec l'expérience. Un apprentissage automatique est tout programme informatique qui améliore ses performances à une tâche donnée avec l'expérience. On ne s'intéresse dans ce travail qu'aux applications de traitement des données, précisément aux applications de résumé automatique de texte.

On va décrire dans cette partie les deux méthodes principales à partir des méthodes utilisées dans le résumé automatique de texte : (1) les arbres de décision et (2) les classifieurs Bayésiens.

1. **Arbre de décision :** Un arbre de décision (SUTHAHARAN 2016) est une structure récursive simple pour exprimer un processus de classification séquentielle dans lequel un cas, décrit par un ensemble d'attributs, est attribué à l'un d'un ensemble disjoint de classes. Chaque feuille de l'arbre indique une classe. Un nœud intérieur désigne un test sur un ou plusieurs attribut(s) avec un arbre de décision secondaire pour chaque résultat possible du test.

   Les arbres de décision classifient les instances en les triant de la racine de l'arbre à un nœud feuille. Chaque nœud de l'arbre spécifie un test d'un attribut de l'instance et chaque branche descendant de ce nœud correspond à l'une des valeurs possibles de cet attribut. Une instance est classée en commençant par le nœud racine de l'arbre, en testant l'attribut spécifié par ce nœud, puis en descendant la branche arborescente correspondante à la valeur de l'attribut dans l'exemple donné. Ce processus est ensuite répété pour le sous-arbre enraciné sur le nouveau nœud. Les arbres de décision peuvent également être re-représentés en tant qu'un ensemble de règles if-then pour améliorer la lisibilité humaine.

   Dans (R. KUMAR, SURI et CHAUHAN 2005), les auteurs ont utilisé l'arbre de décision afin de sélectionner les phrases pertinentes qui servent à construire le résumé d'un document. Mais d'abord, ils ont décrit 23 caractéristiques (features) afin de les utiliser pour choisir les phrases les plus représentatives du document. Parmi ces caractéristiques, nous pouvons citer : numéro de paragraphe, fréquence moyenne de mots, nombre de mots de titre, nombre de mots bonus, et bien d'autres.

   Un arbre de décision est généré en trouvant une caractéristique qui donne le gain d'information maximum. Un nœud est ensuite généré avec un ensemble de règles correspondant à la caractéristique. Ce processus est répété successivement pour les autres caractéristiques jusqu'à ce qu'aucun gain d'information

ne soit disponible. Lors du test, un modèle est comparé à plusieurs reprises au nœud de l'arbre de décision en commençant par la racine et en suivant le chemin approprié en fonction de la caractéristique jusqu'à atteindre le nœud feuille. Ce modèle est alors supposé appartenir à la classe que ce nœud feuille représente. Dans ce cas, le modèle représente chacune des phrases du document, le nœud feuille possède deux valeurs possibles ; la phrase est incluse dans le résumé ou la phrase n'est pas incluse dans le résumé.

2. **Naïve Bayésien :** Le classifieur Bayésien (KUPIEC, PEDERSEN et Francine CHEN 1995) est un classifieur statistique, qui classifie une instance j en déterminant la probabilité qu'elle appartient à une classe Ci. Il est basé sur le théorème de Bayes. Les classifieurs Bayésiens naïfs supposent que l'effet d'une valeur d'attribut sur une classe donnée est indépendant des valeurs des autres attributs. Cette hypothèse est appelée indépendance conditionnelle de classe. Il est fait pour simplifier le calcul impliqué, et dans ce sens, est considéré comme « naïf ». La formule suivante représente la définition du théorème de Bayes (KANTARDZIC 2011, MINING 2006b) :

$$\mathbb{P}(H|X) = \frac{\mathbb{P}(X|H)\mathbb{P}(H)}{\mathbb{P}(X)} \tag{2.3}$$

Le classifieur Bayésien naïf est basé sur le théorème de Bayes, qui fonctionne comme suit :

— Soit T un ensemble d'exemples, chacun avec son étiquette de classe. Il y a k classes, $C_1, C_2, \ldots, C_k$. Chaque exemple est représenté avec un vecteur à n dimensions, $X = x_1, x_2, \ldots, x_n$ décrivant respectivement n valeurs mesurés des n attributs $A_1, A_2, \ldots, A_n$.

— Étant donné un exemple X, le classifieur prédit que X appartient à la classe ayant la probabilité à posteriori la plus élevée, conditionnée par X. L'exemple X est prévu qu'il appartient à la classe C_i si et seulement si
$\mathbb{P}(C_i|X) > \mathbb{P}(C_j|X)$ pour $1 \leq j \leq m; i \neq j$.
On trouve donc la classe qui maximise $\mathbb{P}(C_i|X)$. La classe C_i pour laquelle $\mathbb{P}(C_i|X)$ est maximisée s'appelle hypothèse postérieure maximale. Par le théorème de Bayes, on obtient
$\mathbb{P}(C_i|X) = \frac{\mathbb{P}(X|C_i)\mathbb{P}(C_i)}{P(X)}$
Avec $\mathbb{P}(C_i|X)$ = probabilité de l'instance X étant dans la classe C_i
$\mathbb{P}(X|Ci)$ = probabilité de génération de l'instance X étant la classe C_i
$\mathbb{P}(Ci)$ = probabilité d'occurrence de la classe C_i
$\mathbb{P}(X)$ = probabilité d'occurrence de l'instance X

— Comme $\mathbb{P}(X)$ est la même pour toute les classes, seulement $\mathbb{P}(X|C_i)\mathbb{P}(C_i)$ doit être maximisée. Si les probabilités à priori des classes, $\mathbb{P}(C_i)$, ne sont pas connues, alors on suppose que les classes sont également probables, de façon que $\mathbb{P}(C_1) = P(C_2) = \ldots = \mathbb{P}(Ck)$, et donc on maximise $\mathbb{P}(X|C_i)$. Sinon, on maximise $\mathbb{P}(X|C_i)P(C_i)$
Les probabilités à priori des classes peuvent être estimées par la formule suivante :

$\mathbb{P}(C_i) = fréquence(C_i, T)/|T|$

— Etant donné un ensemble de données avec plusieurs attributs, le calcul de $\mathbb{P}(X|C_i)$ devient coûteux, afin de réduire le calcul dans l'évaluation de $\mathbb{P}(X)(X|C_i)P(C_i)$, l'hypothèse naïve de l'indépendance conditionnelle de classe est faite. Cela suppose que les valeurs des attributs soient conditionnellement indépendantes les unes des autres. Mathématiquement, cela veut dire $\mathbb{P}(X|C_i) = \prod_{k=1}^{n} \mathbb{P}(X_k|C_i)$
Les probabilités $\mathbb{P}(x_1|C_1), \mathbb{P}(x_2|C_2), \ldots, \mathbb{P}(x_n|C_i)$ peuvent êtres estimées à partir de l'ensemble de données. On rappelle que x_k fait référence à la valeur de l'attribut A_k pour l'exemple X.

— Afin de prédire la classe de X, $\mathbb{P}(X|C_i)\mathbb{P}(C_i)$ est évaluée pour chaque classe C_i. Le classifieur prédit que la classe de X est C_i si et seulement si c'est la classe qui maximise $\mathbb{P}(X|Ci)\mathbb{P}(Ci)$. Après avoir décrit le mode de fonctionnement du classifieur Bayésien naïf ainsi que sa formule, on arrive maintenant au stade de son utilisation dans le résumé automatique de texte.

2.5.2. Méthodes abstractives

Un résumé abstractif est une méthode de résumé automatique qui vise à produire un résumé en créant de nouvelles phrases qui récapitulent les informations les plus importantes du texte d'origine. Contrairement au résumé extractif qui sélectionne des phrases existantes, le résumé abstractif peut générer des phrases qui ne figurent pas dans le texte original. Pour ce faire, les modèles de résumé abstractif utilisent des techniques d'apprentissage profond, telles que les réseaux de neurones récurrents, pour créer un modèle de langue capable de prédire la probabilité d'une phrase donnée étant la suite logique d'une autre phrase. Ensuite, le modèle utilise cette probabilité pour générer de nouvelles phrases qui résument le contenu du texte original. Cependant, le résumé abstractif est encore un défi en raison de la difficulté de produire des phrases précises et cohérentes tout en conservant le sens et le style du texte d'origine. Nallapati, et al. (NALLAPATI, B. ZHOU, GULCEHRE et al. 2016) ont proposé une méthode d'extraction de résumé abstrait à l'aide de réseaux de neurones récurrents (RNN) basés sur une architecture d'encodeur-décodeur. Les auteurs proposent une approche qui permet de générer des résumés abstraits en prenant en compte la sémantique des phrases du texte source, plutôt que de simplement sélectionner des phrases existantes. Ils proposent également l'utilisation de techniques avancées telles que l'attention contextuelle et le décodage par échantillonnage pour améliorer la qualité des résumés générés.

See et al. (SEE, P. J. LIU et MANNING 2017) ont apporté une amélioration significative à l'approche de résumé abstrait utilisant les réseaux de neurones séquence à séquence en introduisant une méthode de pointage-génération (pointer-generator) qui combine les avantages des approches extractive et abstractive. Cette méthode permet de pointer vers des mots importants du texte source à inclure dans le résumé plutôt que de simplement générer de nouveaux mots. Cette technique permet de conserver la précision de l'information tout en générant des résumés plus fluides et lisibles. Le

papier a montré des résultats améliorés par rapport à la méthode précédente, en particulier pour les textes longs et complexes.

Plus récemment, Zhang et al. (Haoyu ZHANG, Jianjun XU et Ji WANG 2019) ont proposé un nouveau cadre de génération de texte en utilisant un processus de pré-entraînement basé sur BERT pour améliorer la qualité de la génération de résumés abstratifs de texte. Le modèle proposé utilise un encodage basé sur BERT pour représenter le contexte du texte source et une combinaison de décodeurs basés sur Transformers pour générer le résumé. Le modèle utilise également une méthode de raffinement en deux étapes pour améliorer la qualité du résumé en combinant l'information du décodeur et de BERT. Les résultats expérimentaux ont montré que la méthode proposée améliore significativement les performances de la génération de résumés abstratifs par rapport à (NALLAPATI, B. ZHOU, GULCEHRE et al. 2016) et (SEE, P. J. LIU et MANNING 2017)

2.6. Méthodes d'extraction de titre

Les titres de documents sont des éléments essentiels pour permettre aux lecteurs de comprendre rapidement de quoi traite un texte. Ils doivent être concis, clairs et donner une idée précise du contenu du document. La génération automatique de titres est un processus qui consiste à créer un titre à partir du contenu textuel. Cette approche permet de résumer le contenu d'un document de manière succincte et d'en faciliter la compréhension pour les lecteurs. Cependant, la génération automatique de titres présente des défis particuliers en raison de la quantité d'informations à synthétiser en une seule phrase. Les approches utilisées pour la génération automatique de titres peuvent varier, telles que les approches basées sur les règles, les approches basées sur les métadonnées et les approches à base de résumé automatique

2.6.1. Approches basées sur les règles

Les approches basées sur les règles pour la génération automatique de titres utilisent des règles de linguistique et de logique pour extraire les informations les plus importantes du texte et les combiner en un titre concis. Ces règles peuvent inclure la suppression de mots vides, la détection des verbes d'action, la reconnaissance des noms propres et l'identification des phrases avec une structure syntaxique particulière. Les règles sont souvent définies manuellement par des experts en linguistique ou en traitement automatique du langage naturel. Bien que ces approches puissent être efficaces pour extraire des titres à partir de documents simples et bien structurés, elles peuvent être limitées dans leur capacité à traiter des textes complexes ou ambigus. De plus, l'ajout de nouvelles règles pour gérer de nouveaux types de textes peut nécessiter un effort important et une expertise linguistique.

2.6.2. Approches basées sur le résumé automatique

L'extraction automatique des titres basées sur le résumé automatique consiste à extraire le titre d'un document à partir de son résumé. Cette méthode est basée sur l'hypothèse que le résumé contient les informations les plus importantes du document et qu'il peut être utilisé pour générer un titre informatif et concis.

Pour extraire automatiquement le titre, des techniques de résumé automatique sont utilisées pour réduire le texte à ses parties les plus importantes. Ensuite, des règles ou des algorithmes sont appliqués pour sélectionner les phrases les plus représentatives ou les plus informatives du résumé et les combiner en un titre. Cette méthode peut être plus facile à mettre en œuvre que la génération de titres à partir de zéro, car elle peut être basée sur des techniques de résumé automatique déjà existantes.

Cependant, l'extraction automatique de titres basée sur le résumé automatique peut avoir des limites car les résumés automatiques ne contiennent pas toujours les informations les plus importantes ou les plus représentatives du document. De plus, certains documents peuvent avoir plusieurs sujets principaux, ce qui peut rendre difficile la sélection d'un titre approprié à partir du résumé.

2.7. Méthodes d'extraction de « Topics »

L'extraction de « Topics » dans le traitement automatique du langage naturel est une technique qui attribue un thème à un corpus donné sur la base des mots présents. La tâche d'extraction de « Topics » est importante, car dans ce monde plein de données, il est devenu de plus en plus important de catégoriser les documents. Par exemple, si une entreprise reçoit des centaines de critiques, il est important pour elle de savoir quelles catégories de critiques sont les plus importantes et vice versa. Parmi les différentes méthodes utilisées pour l'extraction des « Topics », citons LSA, PLSA, LDA et CTM.

2.7.1. Analyse Sémantique Latente(LSA)

L'analyse sémantique latente LSA (J. ZHAO, Xinguang LI et Xia LI 2016, A. R. MISHRA, PANCHAL et P. KUMAR 2019) a pour but, à partir d'un ensemble de documents, par exemple des pages web, d'établir automatiquement des relations entre les termes contenus dans ces documents, les documents eux-mêmes et des "concepts" associés aux termes. Elle est notamment utilisée pour :

— Établir des similitudes entre des termes (recherche des synonymes).
— Associer des documents à des "concepts" à partir de l'analyse de leurs termes et donc établir une éventuelle proximité sémantique entre eux.
— Associer un concept à une requête de recherche en analysant ses termes.

L'analyse sémantique latente se base sur une matrice mathématique à deux dimensions. Autrement dit, elle utilise un tableau avec des lignes contenant les termes utilisés dans les différents documents (une colonne par document). Les cellules du tableau contiennent les occurrences des différents termes dans chaque document. Ce tableau est ensuite utilisé pour réaliser des associations entre les documents et

des concepts (à partir des termes), et donc de relier les documents entre eux sur le plan sémantique (une forme de proximité thématique). Nous réalisons pour cela des opérations mathématiques sur la matrice, dans l'ordre suivant :

— Extraction des termes les plus informatifs.
— Réduction (au sens matriciel) du tableau en utilisant uniquement les valeurs singulières (celles qui caractérisent les documents).
— Calcul de la proximité sémantique entre les documents, à partir des similarités entre les mots.

2.7.2. Analyse Sémantique Latente Probabiliste (PLSA)

Analyse sémantique latente probabiliste (PLSA) (Yan CHEN, Yang YANG, Huisan ZHANG et al. 2012,BABU, ANNAVARAPU et MOHAPATRA 2019), également connu sous l'indexation sémantique latente probabiliste est une technique statistique pour l'analyse de deux modes et des données de co-occurrence. En effet, nous pouvons déduire une représentation de faible dimension des variables observées en fonction de leur affinité pour certaines variables cachées, tout comme dans l'analyse sémantique latente, à partir de laquelle PLSA a évolué. Avec le modèle (PLSA), chaque document d'une collection D est représenté par une distribution de probabilité sur les K valeurs de la variable thématique latente A = 1, ..., K, où chaque valeur de correspond à une distribution de probabilité sur l'ensemble des mots de la collection.

2.7.3. L'allocation de Dirichlet latente(LDA)

L'allocation de Dirichlet latente (LDA) (S. LEE, J. LEE, C.-Y. PARK et al. 2013a, Qihua LIU 2015a) est un algorithme d'apprentissage non supervisé qui tente de décrire un ensemble d'observations sous la forme d'une combinaison de catégories distinctes. Le modèle LDA est plus couramment utilisé pour découvrir un certain nombre de rubriques partagées par les documents au sein d'un corpus de texte (ce nombre est spécifié par l'utilisateur). Ici, chaque observation est un document, les fonctions sont la présence (ou nombre d'occurrences) de chaque mot, et les catégories sont les rubriques. Étant donné que la méthode n'est pas supervisée, les rubriques ne sont pas spécifiées à l'avance et leur alignement avec la façon dont les humains peuvent naturellement classer les documents n'est pas garanti. Les rubriques sont apprises sous la forme d'une distribution de probabilité sur les mots rencontrés dans chaque document. Chaque document est à son tour décrit comme un mélange de rubriques. Les contenus exacts de deux documents aux combinaisons de rubriques similaires ne seront pas identiques. Mais surtout, nous pouvons supposer que ces documents utilisent plus fréquemment un sous-ensemble partagé de mots qu'un document issu d'une combinaison de rubriques différentes. Cela permet au modèle LDA de découvrir ces nouveaux groupes de mots et de les utiliser pour former des rubriques.

2.7.4. Modèle des Topics Corrélés(CTM)

LDA ne peut pas modéliser les corrélations entre les Topics. Par exemple, le Topic « génétique » est plus susceptible de ressembler à « maladie » qu'à « astronaute ». Le LDA ne parvient pas à décrire cette corrélation des Topics. Le modèle des Topics corrélés (CTM) (« Rayyan Systems Inc. Available online : https ://www.rayyan.ai/ (accessed on 1 August 2022) » s. d.)est qui est une extension de LDA peut modéliser les corrélations entre les Topics. CTM fournit une représentation graphique des relations entre les Topics, tandis que la LDA impose une indépendance mutuelle entre les Topics. Dans ce modèle, ils ont utilisé un algorithme « meanfield variational » pour former une distribution factorisée des variables latentes, paramétrée par des variables libres qui sont appelées les paramètres variationnels. Ces paramètres sont choisis de telle sorte que l'écart K-L entre la partie postérieure.iorer les performances. De plus, ils ont combiné entre un RNN et deux CNN. Cependant, les CNNs et le RNN sont architecturés en un seul réseau profond au lieu d'un réseau parallèle pour obtenir les avantages des deux réseaux.

2.8. Revue systématique de la littérature sur l'extraction de contexte

La méthode utilisée pour mener cette étude bibliographique est celle décrite dans le premier chapitre : revue systématique de la littérature. D'abord, les questions de recherche sont décrites, suivies par la stratégie de recherche. Les critères d'inclusion sont ensuite énoncés. Ensuite, le processus d'évaluation de la qualité est présenté. Finalement, une phase d'établissement de rapports est réalisée pour présenter des conclusions sur les travaux existants.

2.8.1. Planification : Identification des questions de recherche

La première étape dans La première étape de cette revue systématique consiste à l'identification des questions de recherche. Nous utilisons les six questions de recherche suivantes :

- QR1 : Quels sont les types de papiers (revues, conférences ou ateliers) inclus dans l'étude bibliographique ?
- QR2 : Quels sont les différents domaines d'extraction du contexte identifiés dans les papiers ?
- QR3 : Quelles sont les méthodes proposées pour l'extraction du contexte ?
- QR4 : Quelles sont les collections de test qui ont été utilisées pour évaluer les approches et les modèles pour l'extraction du contexte ?
- QR5 : Quels sont les métriques d'évaluation les plus appropriés pour évaluer l'extraction du contexte ?
- QR6 : Quelles sont les limites des différents travaux liés à l'extraction du contexte

2.8.2. Conduite

1. **Termes de recherche :** La recherche spécifique dans des bibliothèques numérique a été effectuée en utilisant des termes de recherche avec des mots-clés comme suivants : "Relation classification" OR "Classifying relation" OR "Classification of relation". Dans le cas du domaine d'extraction du contexte, les termes retenus peuvent être exprimés comme suit : "Title extraction" OR "Subject extraction" OR "Context extraction" OR "Topic extraction" OR "keyphrase extraction".
2. **Bibliothèques numériques :** Le processus de recherche a été mené en recherchant des publications de recherche pertinentes à travers les quatre bibliothèques présentées dans 1.9.1 dans le premier chapitre : SpringerLink, IEEE Xplore, Google Scholar et ACM Digital Library.
3. **Critères d'inclusion et d'exclusion :** Des critères d'inclusion et d'exclusion ont été appliqués pour filtrer les articles. La recherche se concentrait sur les papiers récents dans le domaine d'extraction des relations. Par conséquent, le premier critère était d'inclure les papiers publiés entre 2010 et 2021. Le deuxième critère consistait à filtrer les papiers en fonction de la langue des textes examinés. L'accent a été mis sur les papiers dans lesquels la langue examinée était l'anglais. Le dernier critère consiste à garder les papiers qui répondent aux différentes questions de recherche. Chaque critère a été appliqué séparément. Tous les articles qui répondaient aux autres critères ont été conservés dans l'ensemble de données.
4. **Règles d'évaluation de la qualité :** Le processus d'évaluation de la qualité était basé sur les questions de qualité prédéfinies suivantes :
   — Les objectifs de la recherche sont-ils clairement énoncés ?
   — L'article fournit-il de nouvelles techniques ou contributions pour le domaine d'extraction de contexte ?
   — L'article mentionne-t-il des défis en matière d'extraction de contexte ?
   — L'article fournit-il des réponses aux questions de recherche formulées ?
   Chaque réponse positive à l'une de ces questions d'évaluation de la qualité valait un point pour le score de qualité. Tous les articles ont été évalués en fonction des questions de qualité et des questions de recherche associées. Comme expliquer dans la section 1.9.1 dans le premier chapitre. L'article de recherche était sélectionné s'il avait un score de qualité supérieur ou égal à trois.
5. **Phase de recherche :** Les termes de recherche énumérés précédemment ont été utilisés pour récupérer les papiers des bibliothèques numériques spécifiées. Au début, le nombre de papiers était de 8086 pour le domaine d'extraction de contexte. La première sélection s'est faite par titre et le nombre a été réduit à 935 papiers. Ensuite, deux réviseurs ont vérifié de manière anonyme si les papiers répondaient à une ou plusieurs des questions de recherche. Pour ce faire, les papiers ont été importés dans une application web appelée Rayyan (« Rayyan Systems Inc. Available online : https ://www.rayyan.ai/ (accessed on 1 August 2022) » s. d.). Cette application web permet aux auteurs de revues systématiques

TABLEAU 2.1. – Résumé des résultats de recherche de l'extraction de contexte

Bibliothèque numérique	Sélection primaire	Sélection par titre	Application des règles d'évaluation de la qualité	% du total des articles pertinents
Google Scholar	5429	320	28	35%
IEEE Xplore	513	251	19	26%
SpringerLink	1205	214	13	18%
ACM Digital Library	939	150	15	21%
Total	8086	935	75	100%

de la littérature de collaborer en votant sur les papiers en fonction des questions de recherche. Trois options de vote peuvent être utilisées dans l'application Rayyan, à savoir « exclure », « inclure » et « peut-être » :

— Les papiers qui ont reçu deux votes « inclure » ou un vote « peut-être » et un vote « inclure » ont été retenus
— Les papiers ayant reçu deux votes « exclure » ou un vote « peut-être » et un vote « exclure » ont été éliminés de l'ensemble des papiers.
— Les papiers ayant reçu deux votes « peut-être » ou un vote « inclure » et un vote « exclure » ont été résolus par discussion. Dans ces cas, le vote décisif sur l'inclusion ou l'exclusion de l'article est pris par le troisième réviseur.

Lors de cette étape, seulement 211 papiers ont été inclus dans la revue. Après application des règles d'évaluation de la qualité, 75 papiers été identifiés comme pertinents pour la revue systématique. L'organigramme PRISMA montrant un rapport des résultats obtenus à chaque phase de la présente revue systématique de la littérature sur l'extraction du contexte est illustré dans la Figure 2.2. De même, le tableau 2.1 présente un résumé des résultats de recherche et des différentes bibliothèques numériques sur lesquelles ils ont été recherchés.

2.8.3. Résultats obtenus

Les informations extraites pour les questions de re cherche QR1 à QR5 sont des analyses statistiques représentées par des graphiques et des tableaux concernant divers travaux étudiés portant sur l'extraction du contexte. Ces statistiques élaborées ont permis de découvrir certains modèles ainsi que les directions dans lesquelles la recherche a été menée. En ce qui concerne QR6 une comparaison descriptive a été effectuée pour faire une synthèse sur les différentes approches proposées et identifier les différentes directions futures. Pour plus de facilité chacun de ces travaux est cité dans ces analyses avec sa référence bibliographique notés BXX (par exemple B15). Les références bibliographiques complètes de ces travaux sont présentées dans l'annexe B de ce manuscrit.

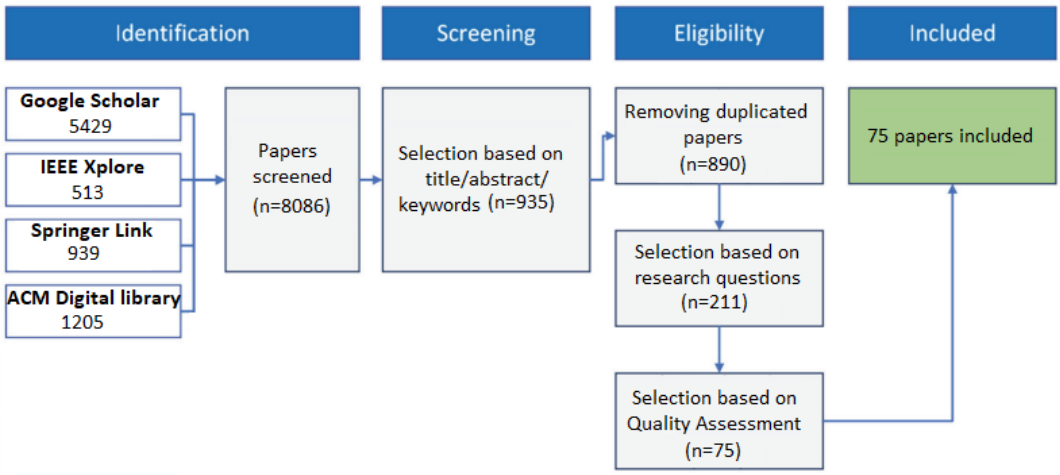


FIGURE 2.2. – Etapes de la revue selon le protocole PRISMA

1. **Question de recherche1 : Quels sont les types de papiers (journal, conférence ou workshop) inclus dans l'étude bibliographique ?**
   Les 75 articles qui ont été ont été identifiés comme pertinents se répartissent en troix principaux types différents : les articles de conférence, les articles de revues et les articles d'atelier (Workshops). La figure 2.3 présente la répartition des articles entre ces principaux types. La majorité des articles utilisés dans l'étude bibliographique ont été identifiés comme des articles de conférence 57%. Les 43% restants ont été répartis entre des articles de revues et des articles d'ateliers à 41.5% et 1.5% respectivement.

FIGURE 2.3. – Pourcentage de papiers dans chaque type

2. **Question de recherche2 : Quels sont les différents domaines d'extraction du contexte identifiés dans les papiers?**
   Parmi les 75 papiers, différents domaines d'extraction du contexte ont été identifiés, notamment : l'extraction du titre à partir des métadonnées du document (Marge, police, etc...), l'extraction des sujets d'actualités à travers les micro-Blogs, Extraction du topic du document, Extraction d'un résumé du document et Catégorisation des commentaires, avis sur les Micro-blogs. Le pourcentage d'articles dans chacun de ces domaines est indiqué dans le tableau 2.2. La majorité des papiers sont inclus dans le domaine d'extraction des topics à 36%, suivis par environ 20% dans le domaine d'extraction des sujets d'actualité à travers les microblogs.
3. **Question de recherche3 : Quelles sont les méthodes proposées pour l'extraction du contexte?**
   Les modèles utilisés pour l'extraction du contexte varient entre méthodes basées sur l'extraction des mots clés, méthodes basées sur le résumé automatique, Méthodes basées sur les métadonnées et méthodes basées sur les techniques du « Topic Modeling ». 33.5% des documents mentionnés utilisent des modèles à base de Topic Modeling. En ce qui concerne les modèles à base de mots clés, 24% des articles utilisent ce modèle pour l'extraction du contexte.18.5% des papiers utilisent les métadonnées pour extraire le contexte à partir d'un document. Seulement 5.5% des papiers utilisent le résumé pour l'extraction du contexte et 18.5% utilisent autres méthodes. Le tableau 2.3 fournit plus de détail.

TABLEAU 2.2. – Résumé des résultats de recherche de l'extraction de contexte

Domaine d'extraction du contexte	Papiers associés	Total des articles	Pourcentage
Extraction du titre à partir des métadonnées de document (police gras, marge, etc. . .).	[B27] [B34] [B42] [B56] [B60] [B65] [B66] [B71] [B20] [B16] [B50] [B63]	12	16%
Extraction des sujets d'actualités à travers les microbloggings	[B2] [B7] [B8] [B9] [B13] [B17] [B22] [B25] [B26] [B28] [B31] [B35] [B39] [B59] [B69]	15	20%
Extraction du résumé automatique du document	[B3] [B14] [B19] [B46]	4	5,5%
Catégorisation des commentaires, avis sur le microblogging	[B1] [B5] [B15] [B21] [B30]	5	6.5%
Extraction des topics	[B4] [B6] [B10] [B11] [B12] [B18] [B33] [B36] [B37] [B38] [B40] [B44] [B45] [B55] [B64] [B67] [B72] [B23] [B51] [B52] [B53] [B48] [B49] [B57] [B73] [B74] [B75]	27	36%
Autres	[B29] [B41] [B43] [B68] [B54] [B24] [B32] [B47] [B61] [B62] [B70] [B58]	12	16%
Total	75	75	100%

TABLEAU 2.3. – Différentes méthodes utilisées pour l'extraction du contexte

Différents Modèles	Articles associés à chaque modèle	Nombre d'articles pour chaque modèle	% d'articles pour chaque modèle
Méthodes basées sur les mots clés	[B8] [B30] [B1] [B12] [B13] [B15] [B17] [B18] [B21] [B22] [B26] [B38] [B14] [B9] [B53] [B62] [B69] [B70]	18	26
Méthodes basées sur les techniques du topic modeling	[B2] [B4] [B6] [B10] [B25] [B29] [B31] [B33] [B35] [B36] [B40] [B41] [B43] [B45] [B48] [B49] [B51] [B57] [B61] [B68] [B23] [B32] [B73] [B74] [B75]	25	24%
Méthodes basées sur le résumé automatique	[B3] [B19] [B44] [B46]	4	5,5 %
Méthodes basées sur les titres	[B55] [B64] [B67] [B72] [B20] [B65] [B39] [B28] [B47] [B50] [B52] [B56] [B63] [B24]	14	18.5%
Autres	[B27] [B34] [B42] [B5] [B7] [B11] [B37] [B58] [B59] [B54] [B16] [B60] [B66] [B71]	14	18.5%
Total	75	75	100%

4. **Question de recherche 4 : Quelles sont les collections de test qui ont été utilisées pour évaluer les approches et les modèles pour l'extraction du contexte ?** Pour tester les différentes approches proposées, plusieurs les collections de test ont été utilisées dans les papiers. Certaines étaient privées tandis que dans la majorité des papiers étaient publiques et disponibles sur le Web. Ces collections sont présentées dans le tableau 2.4, la majorité des travaux utilisent 20 New groups, New York times and BBC News pour tester leurs approches. « Twitter » et « Facebook » ont été utilisées pour l'extraction des contextes des microbloggings. Concernant les autres travaux à savoir le topic modeling et l'extraction du contexte à partir des mots clés, plusieurs collections de test ont été utilisées à savoir « NIPS dataset » et « Wikepédia ». Finalement, plusieurs auteurs ont créé leur collections manuellement.

TABLEAU 2.4. – Les collections de test utilisées pour évaluer les différentes approches proposées

Types de collection de test	Différentes collections de test utilisées	Nombre de papiers
Articles de presse	BBC News	8
	New York Times	9
	20newsgroup	5
	Autres	3
Microblogging	Facebook	6
	Titter	12
	Autres	3
Autres	Nips dataset	4
	Word Similarity	4
	Wikipedia	3
	Autres	3
collection de test créés par les auteurs	Manuelle	15

5. **Question de recherche5 : Quels sont les métriques d'évaluation les plus appropriés pour évaluer l'extraction du contexte ?**
Plusieurs techniques d'évaluation ont été utilisées dans les articles de recherches pour évaluer la performance globale du système développé. Le tableau 2.5 illustre les différentes techniques identifiées, ainsi que le pourcentage d'articles qui ont utilisé chaque technique. Comme on peut le voir, la majorité des articles (26.5%) ont utilisé « F-mesure » pour évaluer les performances de leur système. D'un autre côté, La métrique « Accuracy » a été fortement utilisé dans la majorité des évaluations après le F1-Score avec un pourcentage de 24%. La précision accompagnée du rappel a été utilisée comme mesure d'évaluation dans 7 documents (9.5%). L'évaluation qualitative a été utiliséé avec un pourcentage de 8%. La précision, la perplexité ainsi que l'average ont été utilisés dans 4 évaluations respectivement soit 5.5% chacune. La pureté et Normolized Mutual information ont été rarement utilisées dans la totalité des documents avec approximativement 2.5% chacune. En revanche, 10.5% des articles ont été classés comme « autres », car les techniques utilisées par chacun des articles ont obtenu moins de 1%.

TABLEAU 2.5. – Les métriques d'évaluation utilisées pour tester les différentes approches proposées

Méthodes d'évaluation	Documents associés	Nombre d'articles	Pourcentage
Accuracy	[B25] [B28] [B29] [B30] [B34] [B36] [B37] [B39] [B48] [B53] [B68] [B24] [B42] [B65] [B64] [B73] [B74] [B75]	18	24 %
F mesure	[B1] [B3] [B4] [B5] [B10] [B11] [B16] [B21] [B38] [B40] [B45] [B50] [B55] [B62] [B67] [B20] [B14] [B51] [B6] [B72]	20	26.5%
Precision Recall	[B2] [B13] [B18] [B41] [B69][B70] [B26]	7	9.5%
Purity	[B15] [B33]	2	2.5%
Precision	[B43] [B44] [B57][B71]	4	5.5%
Perplexity	[B31] [B32] [B52] [B35]	4	5.5%
Average	[B56] [B60][B19] [B54]	4	5.5%
Normolized Mutual information	[B47][B23]	2	2.5%
Evaluation Qualitative	[B7] [B9] [B49] [B66] [B58] [B59]	6	8%
Autres mesures	[B8] [B12] [B17] [B46] [B22][B27] [B61] [B63]	8	10.5%
Total	75	75	100%

6. **Question de recherche6 : Quelles sont les limites des différents travaux liés à l'extraction du contexte ?**
   Nous avons présenté un aperçu global sur les techniques adoptés pour l'extraction du contexte dans la partie précédente. Divers auteurs ont avancé diverses hypothèses pour extraire le contexte de documents textuels non structurés. Quatre approches principales pour l'extraction du contexte ont été identifiées : (a) Les approches basées sur le « Topic Modeling », (b) les approches basées sur les mots-clés, (c) les approches basées sur les méta-données et (d) les approches basées sur les résumés automatiques du texte

   a) **Approches basées sur le « Thème » :** De nombreuses recherches ont été menées sur le « Topic Modeling » et les approches basées sur l'extraction de mots-clés. Dans les approches de « Topic Modeling », le sujet représente généralement soit un ensemble de mots clés décrivant le contexte, soit un seul mot qui catégorise le sujet du texte, par exemple (santé, musique, politique, etc.). Dans [13], les auteurs ont proposé une approche pour extraire le thème général d'un document. Par exemple, si l'on considère les deux phrases sui-

vantes :

« Emanuel Macron a obtenu 66,1%des voix à l'élection présidentielle française de 2017 »
« Barack Obama a obtenu 52,9% des voix »

Selon l'approche présentée dans (MALLEK, GUETARI, ETTEYEB et al. 2017), les deux phrases ci-dessus sont classées dans le même contexte prédéfini « politique ». Cependant, le sujet identifié n'est pas assez précis pour caractériser au mieux le contexte réel de chacune des deux phrases. En effet, la première phrase concerne le contexte de l'élection présidentielle française de 2017, tandis que la seconde se rapporte à l'élection présidentielle américaine de 2012. Il nous semble évident qu'un seul mot-clé est loin de pouvoir caractériser de manière concise le contexte d'un texte.

b) **Approches basées sur les mots clés :** La littérature sur les modèles basés sur les mots-clés montre une variété d'approches. En effet, de nombreux travaux adoptent ce type de modèle afin d'extraire les sujets d'actualités à partir des blogs et des news. Récemment, plusieurs auteurs (RANGU, CHATTERJEE et VALLURU 2017,ZHONG, Yuefeng LI et S.-T. WU 2010,COLLOBERT, WESTON, BOTTOU et al. 2011) ont proposé de nouveaux modèles pour trouver des sujets d'actualité. En fait, ils analysent les nouvelles sur le web et renvoient les mots avec la plus grande fréquence pendant une période de temps prédéfinie. C'est une très bonne approche, sauf qu'elle est dédiée aux données du web et des microblogs uniquement. En outre, les auteurs de (BLACKBURN et BOS 2005, GUETARI et MALLEK 2017) ont proposé une approche pour automatiser le processus d'extraction du sujet et du titre d'un document unique en utilisant des techniques basées sur des mots clés. Cependant, le sujet extrait peut ne pas être pertinent. En effet, si l'on prend comme exemple un document dont le titre est « Europe Gets Tough on Facebook », le résultat de l'extraction du sujet est « Facebook Industry Company Data ». Cet exemple montre que l'ensemble des mots extraits du document ne permet pas de construire une phrase ou que la phrase obtenue est dépourvue de tout sens.

c) **Approches basées sur les titres :** De grands efforts ont été consacrés à l'étude de l'extraction des titres, considérant qu'ils représentent l'idée principale d'un document. Ces chercheurs se répartissent en deux catégories : Les approches pour les documents PDF et les approches pour les pages web HTML. Ces approches sont basées sur le style appliqué au document (taille de la police, alignement, marge, etc.) et sur certaines métadonnées pour obtenir cette phrase clé et ignorer la sémantique du contenu. (NIBOONKIT, KRATHU et PADUNGWEANG 2017) ont développé une heuristique simple basée sur des règles, qui prend en compte les informations de style (taille de la police) pour identifier le titre d'un PDF. Pour ce faire, ils ont appliqué des règles empiriques simples reflétant les pratiques habituelles lors de la présentation d'un texte. Parmi les règles utilisées, on peut citer les plus habituelles telles

que « les titres sont généralement situés dans les parties supérieures des premières pages », « les titres sont généralement dans les plus grandes tailles de police », etc... Concernant les approches pour les documents HTML, ces approches sont basées sur des éléments (balises) dans l'en-tête et le corps du document, etc. pour extraire le titre. Parmi les balises les plus utilisées dans ce sens, on trouve par exemple (où n 1, 2, 3, 4, 5, 6). Dans (CROUCH, BERG, SALVETTI et al. 2014), les auteurs ont proposé un schéma général permettant d'apprendre les titres des textes en fonction des informations de style. Les approches traitant des documents PDF et HTML souffrent de plusieurs lacunes et incohérences. En effet, les métadonnées sont généralement saisies par les auteurs des documents et sont donc subjectives. Les styles et les règles sur lesquels reposent ces méthodes d'extraction des titres ne sont pas toujours fiables, d'autant plus que les styles peuvent être modifiés par les auteurs. Si les documents PDF souffrent d'un manque de structure, la structure des documents HTML est également douteuse et manque de fiabilité. En effet, si l'on considère les balises <H1>,<H2>,...,<H6>, rien dans les règles de rédaction d'un document HTML n'exige l'utilisation de <H1> avant <H2> et de <H2> avant <H3> etc. En outre, ces méthodes ne fonctionnent pas du tout lorsque le titre n'est pas mentionné dans le texte.

d) **Approches basées sur le résumé automatique :** On trouve très peu de publications dans la littérature qui traitent l'extraction du contexte basée sur le résumé automatique du texte. L'étude de cas (HONG et ZHEN 2012a) présente une approche qui utilise l'occurrence des mots afin de résumer et d'extraire le sujet principal d'un document textuel. Cette approche est efficace pour extraire automatiquement le résumé d'un texte mais pas nécessairement son contexte. Un résumé donne une idée générale du contenu d'un document et n'offre pas nécessairement son contexte.

2.9. Bilan

Le ***« contexte »*** d'un document, s'apparente à une métadonnée associée à ce document et constituée d'un ensemble d'éléments qui permettent de mieux cerner son contenu. La notion de ***« contexte »*** est voisine de notions (ou métadonnées) liées au contenu du document comme celles de thème, titre, résumé, sujet et mots-clés associés au document. Le ***« contexte »*** devant fournir une synthèse plus pertinente du contenu du document.

Les méthodes d'extraction du contenu de document basées sur l'extraction de métadonnées comme *les mots-clés*, le *résumé*, le *titre* ou le *thème* présentées dans ce chapitre souffrent toutes du même problème : elles manquent de précision et de pertinence dans l'identification de ce dont parle un document, ce qui dégrade considérablement toutes les utilisations qui peuvent être faites dans des tâches de classification, de recherche d'information, de compréhension du contenu d'un document, etc.

Assimiler un ***« contexte »*** d'un document à un *« thème »* n'est pas satisfaisant. En effet un *thème* reste une généralisation du sujet traité par un corpus de documents et non le ***« contexte »*** précis lié à un document particulier. Si nous considérons par exemple deux documents différents traitant du *thème* des élections présidentielles. Il s'agit bien de deux documents traitant du même *thème*, mais ce thème reste vague et manque de détails. Le niveau de précision d'un tel sujet doit inclure le pays, l'année et, selon la législation et le pays, peut aller jusqu'au tour (comme c'est le cas en France). Par conséquent, ces documents peuvent faire référence à deux élections présidentielles différentes.

De même, assimiler un ***« contexte »*** d'un document à un *« résumé »*, écrit par les auteurs ou extrait de façon automatique du document, n'est pas satisfaisant. En effet le ***« contexte »*** serait composé d'une phrase clé de ce *résumé*, phrase très difficile à déterminer. De plus, si le *résumé* est écrit par ses auteurs, même si une telle phrase existe, la qualité du résultat est intimement liée à la qualité de l'écriture et à la subjectivité de ces auteurs.

Le *« titre »* pourrait être raisonnablement considéré comme un bon candidat pour représenter le ***« contexte »*** du document, mais il souffre de plusieurs limites. En effet la qualité du *titre* dépend de l'imagination des auteurs et de leurs compétences littéraires et est souvent sujette à une certaine subjectivité.

Ainsi, ce qui est commun et récurrent à toutes ces métadonnées d'un document donné, ce sont les *mots-clés* que l'on trouve dans le *titre* et dans le *résumé*. Les *mots-clés*, qu'ils soient définis par les auteurs ou extraits automatiquement, sont des métadonnées qui représentent souvent le mieux le contenu d'un document. Cependant, en s'appuyant sur les *mots-clés*, le ***« contexte »*** dépend du choix d'un ensemble de mots-clés, et la qualité de ce choix n'est pas garantie.

2.10. Conclusion

Afin d'améliorer l'extraction d'informations pertinentes d'un document basé sur son contenu, nous avons cerné dans ce chapitre la notion de ***« contexte »*** d'un document textuel non structuré, notion à notre connaissance pas vraiment formalisée, assimilable à une nouvelle métadonnée du document. Nous nous sommes ensuite intéressés à des méthodes existantes pour l'extraction de métadonnées du document aussi associées à son contenu, basées sur les notions voisines du ***« contexte »***, comme celles de *mots-clés*, de *titre*, de *résumé* et de *thème*, notions mieux formalisées.

Après avoir présenté une revue systématique de la littérature axée sur l'extraction de métadonnées associées au contenu d'un document, nous avons présenté une synthèse générale des différentes approches existantes qui nous a permis de prendre connaissance des différentes perceptions des travaux proposés et de récapituler les limitations de ces travaux afin de nous positionner par rapport à l'existant. Nous avons constaté qu'il n'y avait pas de méthode permettant l'extraction d'une métadonnée pertinente et précise associée au contenu d'un document.

Dans le chapitre suivant, après avoir formalisé le concept de ***« contexte »*** consti-

tuant une nouvelle métadonnée d'un document, nous allons proposer une méthode d'extraction de ce ***« contexte »***. Cette méthode, explicitement basée sur le concept de ***« contexte »***, effectue à la fois l'extraction ou l'identification du ***« contexte »***du document et sa classification dans une base de contextes. Cette méthode produit une métadonnée***« contexte »*** associée au contenu du document, plus précise et plus pertinente que celles obtenues par les autres méthodes d'extraction de contenu étudiées précédemment.

3. Une méthode d'extraction du contexte d'un document textuel

Sommaire

3.1. Introduction

Dans le chapitre précédent nous avons constaté qu'il n'y avait pas de méthode permettant l'extraction d'un contexte pertinent et précis d'un document, tout au plus des méthodes d'extraction de mots-clés, de résumé et de topics, notions mieux formalisées que celle de contexte. Dans ce chapitre nous présentons la méthode d'extraction de ***« contexte »*** d'un document que nous proposons explicitement basée sur le concept de ***« contexte »***.

Pour un document donné, cette méthode effectue à la fois l'extraction ou l'identification de son ***« contexte »*** et son stockage dans la base de contextes. Elle permet de construire un contexte compréhensible par un humain associé au document, et enfin produit pour un document donné une métadonnée plus précise et plus pertinente sur le contenu du document.

Nous commençons dans la section 3.2. par définir formellement les notions sur lesquelles est fondée cette méthode, notamment celle de ***« contexte »***. Nous présentons ensuite dans la section 3.3 une vue d'ensemble de notre méthode d'extraction de contexte qui repose sur deux phases principales : (i) *la phase d'extraction des mots-clés du document* et (ii) *la phase d'extraction de contexte du document*. Ces phases dont détaillées dans les sections suivantes (section 3.4 et 3.5). Nous avons identifié deux approches différentes de l'extraction de contexte : l'une extractive et l'autre générative, donnant lieu ainsi à deux variantes de la méthode. Ensuite, dans la section 3.7 nous évaluons par expérimentation notre méthode d'extraction du contexte selon un protocole bien défini, avant de conclure.

3.2. Fondements de la méthode proposée

Comme nous l'avons déjà évoqué, le ***« contexte »*** d'un document peut être vue comme une métadonnée permettant une certaine synthétisation de son contenu. Ce ***« contexte »*** permettra notamment de mieux classer et retrouver un document donné dans un ensemble conséquent de documents, et de mieux exploiter son contenu, notamment dans le classement des relations qu'il contient, auquel ce travail de thèse s'intéresse plus particulièrement.

De façon générale nous définirons le ***« contexte »*** d'un document comme un groupe de mots permettant de caractériser « ce dont parle un document ». Un même contexte peut caractériser une collection de documents textuels (corpus) traitant de contenus voisins.

Comme nous l'avons déjà dit, la notion de ***« contexte »*** d'un document n'est pas clairement définie. D'autres notions sont mieux définies comme celles de *mots-clés*, *résumé*, *titre* et *thème*. Dans le cadre de l'extraction automatiques de ces dernières notions, nous avons dans le chapitre précédent étudiées diverses méthodes et techniques d'extraction de contenu notamment basées sur l'extraction de *mots-clés*, de *résumés*, de *titres* et de *thèmes*.

Cependant, toutes ces méthodes manquent de précision dans leur extraction de

contenu d'un document, ce qui pénalise leur mise en œuvre notamment dans la classification, la recherche d'information, la compréhension du contenu du document. Aussi nous proposons une nouvelle méthode d'extraction du ***« contexte »*** d'un document explicitement basée sur le concept de ***« contexte »***.

Les sections suivantes présentent les fondements de cette méthode d'extraction de contexte d'un document.

3.2.1. Définition d'un document

Pour notre recherche, un document sera défini ainsi :

Un document « D » est un texte écrit en langage naturelle. Il est défini par un contenu « C » et des métadonnées associées qui permettent de décrire et préciser son contenu. Ces métadonnées comprennent un ensemble de mots-clés « Mc » et un contexte « Ctx », soit :

$$D = \{C, Mc, Ctx\} \tag{3.1}$$

3.2.2. Définition des mots-clés « Mc » d'un document

Les mots-clés sont très souvent les métadonnées qui représentent le mieux le contenu d'un texte et sa thématique. Les mots-clés peuvent être fournis par les auteurs d'un texte. Selon la manière dont les mots-clés ont été sélectionnés, ils peuvent souffrir d'un certain nombre de lacunes. En effet, les mots-clés fournis par un auteur peuvent être trop généraux ou vagues, ce qui peut ne pas aider à trouver des informations précises dans le texte. En outre, Si les mots-clés ne couvrent pas toutes les facettes importantes du sujet abordé dans le texte, cela peut rendre l'extraction d'un contexte précis à partir de mots-clés difficile. C'est pour ces raisons pour lesquelles nous extrayons automatiquement ces mots-clés. Pour notre part, l'ensemble de mots-clés d'un document sera défini comme suit :

Définition 1 *Les mots-clés « Mc » d'un document « D » est un ensemble de mots extraits du document « D » et fournissant des informations pertinentes sur son contenu, c'est à dire qui aident à la compréhension de sa thématique, soit :*

$$Mc = \{Mc_i \ ; \ i = 1..n \ \ et \ \ Mc_i \in D\} \tag{3.2}$$

Où : n est le nombre de mots-clés à extraire (ce nombre sera défini dans la section 3.4.)

3.2.3. Définition du contexte « Ctx » d'un document

Il est courant que deux documents partagent les mêmes mots-clés mais qui parlent de deux choses différentes. Par conséquent, assembler des mots-clés en une phrase en utilisant des mots vides pour les lier peut sembler similaire pour deux documents, mais en réalité, les phrases décrivent des contextes différents. Pour cette raison, les

mots-clés ne peuvent pas être compris sans leur contexte spécifique et une phase d'extraction du contexte est si importante. En effet, une phrase construite à partir des mots-clés peut fournir un contexte clair et précis, permettant ainsi une compréhension de quoi parle un document.

Dans le cadre d'un texte, le ***« contexte »*** désigne l'ensemble des éléments qui entourent un mot, une expression, une phrase, ou toute autre unité linguistique, et qui peuvent aider à clarifier son sens. De manière informelle, on peut dire que le ***« contexte »*** capture « ce dont parle un document". Nous définissons la notion de contexte ainsi :

Définition 2 *Le contexte « Ctx » d'un document « D » est présenté comme une phrase représentant ce dont parle le document. Cette phrase est construite à partir des mots-clés « Mc » associés au document « D ». Ces mots sont liés par des mots de liaison « Ml » pour créer une phrase qui permet de transmettre une idée claire et précise. Les mots de liaison peuvent inclure des mots vides ou des verbes, soit :*

$$Ctx = f(Mc, Ml) \tag{3.3}$$

Où : f :C'et la fonction qui permet de construire une phrase d'une manière cohérente à partir des mots-clés et les mots de liaison.

Exemple : ***Mc*** = { candidate, Presidential election, French, vote, intentions }

- ***Ctx1*** = The candidate's performance in the debates swayed voting intentions of many during the French presidential election.
- ***Ctx2*** = Candidates' voting intentions of the French presidential election.

Nous constatons que l'utilisation d'une même liste de mots-clés peut donner lieu à deux phrases totalement différentes en termes de sens. Ce phénomène est susceptible de se produire pour deux documents qui parlent de deux choses différentes, mais utilisant des mots-clés similaires.

3.3. Présentation générale de la méthode d'extraction du contexte proposée

Pour l'extraction de contexte d'un document textuel non structuré, nous proposons une approche fondée sur le Traitement Automatique de Langage Naturel (TALN). Nous présentons dans la Figure 3.1 la méthode proposée pour extraire le contexte associé à un texte non structuré. La méthode d'extraction du contexte proposée est organisée en deux phases distinctes : la phase d'extraction de mots-clés du document et la phase d'extraction du contexte du document :

- Phase d'extraction de mots-clés (section 3.4). Cette phase permet de déterminer les mots qui fournissent des informations pertinentes sur le contenu du document en se basant sur deux types de mots-clés : ceux qui sont extraits d'un

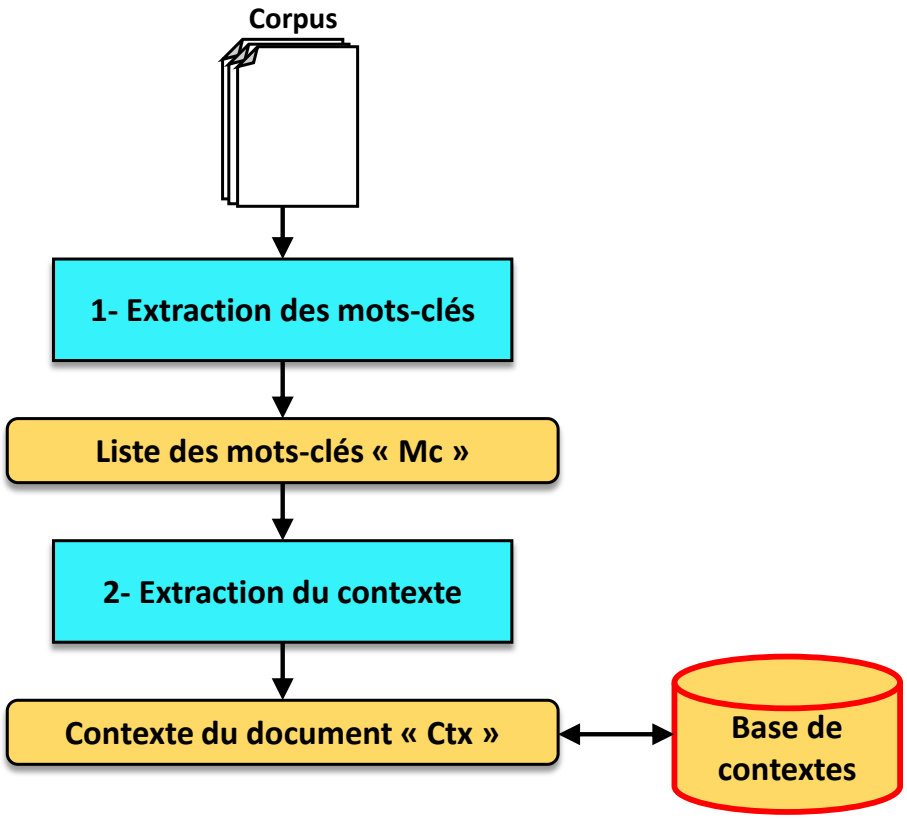
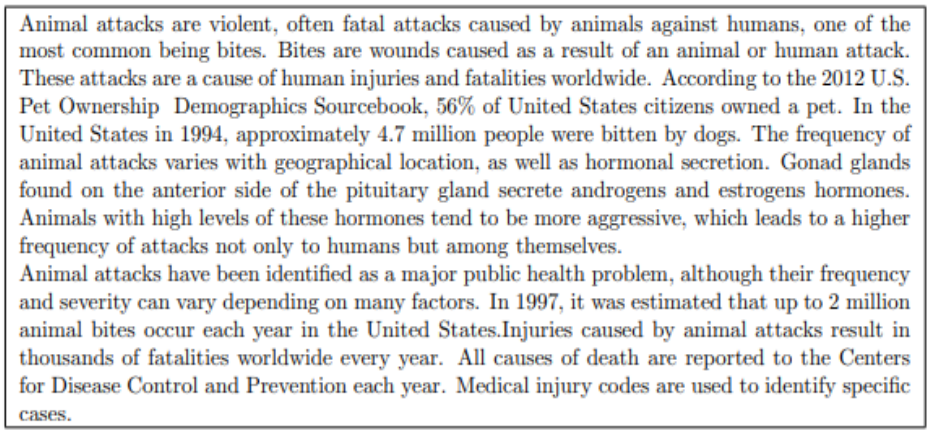


FIGURE 3.1. – Vue d'ensemble de la méthode d'extraction du contexte proposée

corpus de documents et ceux qui sont extraits d'un seul document.

— Phase d'extraction du contexte (section 3.5). Il est possible que deux documents présentent des mots-clés en commun alors qu'ils traitent de contextes différents. En effet, les mots-clés sont des termes qui renvoient à des concepts spécifiques, mais ils ne prennent leur signification que dans le contexte dans lequel ils sont utilisés. Ainsi, si l'on assemble ces mots-clés en une phrase, en utilisant des mots vides pour les lier, il est possible d'obtenir deux phrases qui semblent similaires, mais qui décrivent en réalité des contextes différents. Il est donc important de prendre en compte le contexte global du document pour comprendre la signification des mots-clés et éviter toute confusion. Pour cette raison, nous passons en deuxième lieu à une phase d'identification du contexte qui permet de construire, à partir des mots-clés sélectionnés, une phrase qui représente ce dont parle le document. Pour se faire, deux approches différentes sont appliquées : l'une extractive et l'autre générative. L'approche extractive permet de chercher, à partir du document, une phrase qui contient le maximum des mots-clés. Tandis que l'approche abstractive permet de construire cette phrase à partir des mots-clés du document en se basant sur la modélisation du langage à l'aide de réseaux neuronaux récurrents et en particulier du modèle LSTM (Long Short-Term Memory).

Dans les sections suivantes, nous détaillons chacune des deux phases de la méthode proposée. Pour des raisons pédagogiques, tout au long des sections, nous utilisons un même document pour illustrer pas à pas la mise en œuvre de notre méthode d'extraction de contexte. Ce document, présenté à la figure 3.2, est un article intitulé « The Public Health Impact of Animal Attacks » qui traite l'impact des attaques animales sur la santé publique.



Animal attacks are violent, often fatal attacks caused by animals against humans, one of the most common being bites. Bites are wounds caused as a result of an animal or human attack. These attacks are a cause of human injuries and fatalities worldwide. According to the 2012 U.S. Pet Ownership Demographics Sourcebook, 56% of United States citizens owned a pet. In the United States in 1994, approximately 4.7 million people were bitten by dogs. The frequency of animal attacks varies with geographical location, as well as hormonal secretion. Gonad glands found on the anterior side of the pituitary gland secrete androgens and estrogens hormones. Animals with high levels of these hormones tend to be more aggressive, which leads to a higher frequency of attacks not only to humans but among themselves.
Animal attacks have been identified as a major public health problem, although their frequency and severity can vary depending on many factors. In 1997, it was estimated that up to 2 million animal bites occur each year in the United States.Injuries caused by animal attacks result in thousands of fatalities worldwide every year. All causes of death are reported to the Centers for Disease Control and Prevention each year. Medical injury codes are used to identify specific cases.

FIGURE 3.2. – Document de la collection de données WikiContext qui concerne le contexte des attaques d'animaux

Notons enfin que les contextes extraits d'un corpus de documents par la méthode sont stockés dans une base de contextes (BdC) qui sera présentée plus loin (section 3.6). Lors de la phase d'Extraction de contexte d'un nouveau document, deux cas peuvent se présenter. Soit le contexte extrait du document existe déjà dans la base de contextes, ce contexte sera alors associé au document, soit il n'existe pas et il sera alors rajouté à la base de contextes.

3.4. Phase d'extraction des mots-clés

L'extraction de mots-clés est un processus essentiel dans l'analyse de texte, car elle permet de résumer le contenu d'un document en quelques mots pertinents, appelé « mots-clés ». Les mots-clés extraits en utilisant des techniques existantes ne sont pas forcément les mots-clés les plus pertinents pour générer, par la suite, une « phrase » définissant le contexte du document, et donnant une idée générale du contenu du document et être en mesure de transmettre l'essentiel de celui-ci en quelques mots.

Notons que la pertinence se réfère à la qualité d'être approprié par rapport à un contexte ou une situation particulière. Ainsi, dans cette thèse, un mot-clé est considéré comme pertinent s'il est en rapport avec le contenu du document et qu'il contribue à le décrire de manière précise et concise.

Pour cette raison, il est important de mettre en place une approche différente de l'extraction de mots-clés traditionnelle permettant de sélectionner les concepts les

plus intéressants pour le document. Cette approche doit prendre en compte la sémantique et la signification des mots pour sélectionner les mots-clés les plus pertinents qui permettent de construire cette phrase représentative du document.

La phase d'extraction des mots-clés de notre méthode se compose de trois étapes principales : (1) extraction des mots-clés basée sur un seul document, (2) extraction des mots-clés basée sur un corpus de documents, et enfin (3) la sélection des mots-clés pertinents, comme illustré dans la figure 3.3.

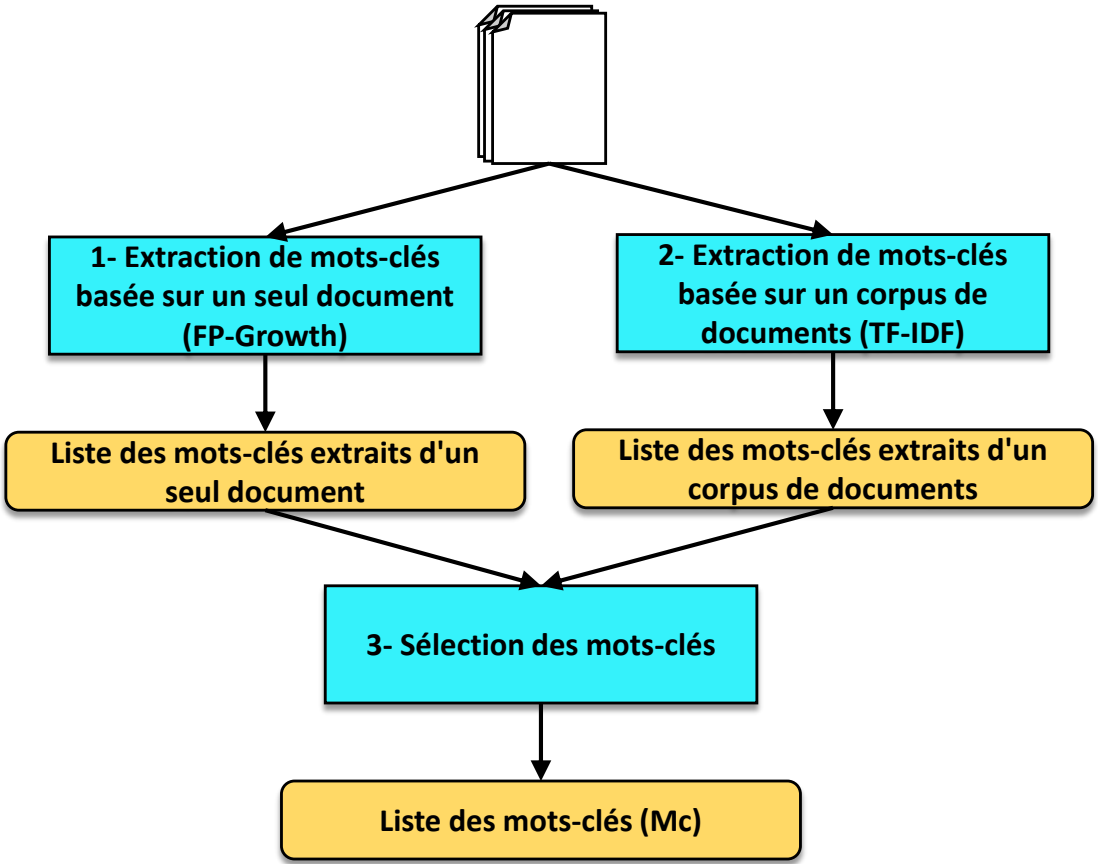


FIGURE 3.3. – Processus général de la phase d'extraction des mots-clés

Afin d'extraire les mots-clés les plus pertinents pour un document « D », il est possible de combiner plusieurs techniques. Les méthodes basées sur les statistiques peuvent éliminer les mots qui sont peu fréquents dans le document, mais qui peuvent néanmoins être importants pour comprendre le contexte du document. Pour cette raison, nous nous sommes basés dans une première étape à extraire tous les motifs fréquents du document, y compris les mots qui ne sont pas très fréquents mais qui peuvent être importants pour comprendre les relations sémantiques entre les concepts. Pour ce faire, nous avons appliqué l'algorithme FP-growth (section 3.4.1) sur le document « D ». L'utilisation de FP-growth peut nous donner une vue complète des mots-clés présents dans le document, ce qui peut être utile pour comprendre le contenu et la signification du document.

L'extraction de mots-clés à partir d'un seul document est une méthode courante pour extraire des informations importantes à partir d'un seul document. Cependant, cette méthode peut parfois être insuffisante pour appréhender pleinement le contenu du document. Elle ne permet pas toujours d'identifier les mots-clés les plus pertinents dans le domaine traité dans le document. Pour cette raison, il est souvent bénéfique d'extraire des mots-clés à partir d'un corpus de documents, qui permettent d'obtenir une vue d'ensemble plus large des termes clés dans le domaine. Pour cette raison, il est souvent préférable de combiner l'extraction de mots-clés à partir d'un seul document et à partir d'un corpus. Cette approche permet de tirer parti des avantages des deux méthodes et de fournir une compréhension plus complète du contenu du document et du domaine.

Afin d'extraire les mots-clés à partir d'un corpus, nous avons appliqué une approche statistique. Ce corpus contient plusieurs documents dans des contextes différents y compris le document « D ». Nous avons utilisé l'algorithme TF-IDF (section 3.4.2) qui ne considère l'importance d'un mot pour un document que par sa rareté dans d'autres documents.

Les termes extraits par FP-Growth ne sont pas tous nécessairement pertinents pour le document. Certains termes peuvent apparaître fréquemment simplement parce qu'ils sont utilisés fréquemment dans la langue ou dans le domaine traité par le document, sans nécessairement être pertinents pour le contenu spécifique du document. Pour cette raison, le calcul de poids des mots est nécessaire dans une troisième phase. Pour ce faire, nous utilisons l'Universal Sentence Encoder (USE) pour calculer les poids des mots (section 3.4.3). Les deux avantages d'utiliser USE pour calculer les poids des mots, sont la prise en compte la similarité sémantique entre les mots et l'identification de mots qui ont des significations similaires, même s'ils sont formulés différemment. Par conséquent, l'utilisation de l'USE pour générer les poids des mots améliore la précision de l'extraction des mots-clés en prenant en compte la signification sémantique des mots.

Pour calculer la pondération d'un mot, nous calculons la similarité sémantique de chaque mot-clé extrait par FP-Growth avec chaque mot-clé généré par TF-IDF en utilisant l'USE. La pondération d'un mot-clé peut ensuite être calculée en prenant la moyenne des similarités avec tous les mots-clés générés par TF-IDF. L'ensemble des mots-clés est trié par ordre décroissant de score pour sélectionner les mots les plus pertinents (le nombre de mots-clés sera défini dans la section 3.4.3). Ces mots-clés peuvent permettre de mieux comprendre le contexte et servent, par la suite, pour la construction du contexte du document.

3.4.1. Extraction de mots-clés basée sur un seul document

Dans le but d'extraire des mots-clés à partir d'un seul document, nous avons opté pour l'algorithme FP-growth (BORGELT 2005). Nous avons choisi FP-growth plutôt que les algorithmes Eclat (ZAKI 2002) et A priori (HEGLAND 2007) en raison de ses avantages significatifs dans les situations où la taille des données est importante. En effet, FP-growth utilise une structure de données compacte qui permet d'éviter de générer

tous les sous-ensembles possibles, ce qui le rend plus efficace en termes de temps de traitement et de consommation de mémoire. De plus, FP-growth est capable de gérer des ensembles d'éléments avec des occurrences faibles, ce qui permet également une analyse plus précise et plus complète du contexte.

FP-growth est un algorithme d'exploration de données qui permet d'extraire un ensemble de mots fréquents à partir d'un grand ensemble de transactions (dans notre cas des phrases). Dans notre approche, nous commençons par un prétraitement de nos documents afin d'éliminer tous les mots vides du texte. Cette étape nous permet de réduire la dimensionnalité. Il faut ensuite segmenter le texte sous forme de phrases. Chaque phrase est découpée en un ensemble de mots. Ensuite, les fréquences des termes sont extraites du document pour créer une liste de termes fréquents. L'arbre FP-Growth est ensuite construit à partir de la liste de termes fréquents, permettant ainsi d'extraire les ensembles de motifs fréquents à partir de l'arbre. Les ensembles de motifs fréquents sont des groupes de termes qui apparaissent ensemble plus souvent que prévu. Enfin, pour extraire les mots-clés à partir des ensembles de motifs fréquents, des techniques de traitement du langage naturel telles que la lemmatisation et la reconnaissance d'entités nommées peuvent être utilisées. Les mots-clés extraits peuvent aider à comprendre le contenu du document.

Si on applique FP-growth sur le document présenté dans la figure 3.2, on obtient les mots-clés fréquents extraits de l'arbre FP-growth suivants : animal attacks, injuries, fatalities, bites, geographical location, wounds, United States, frequency, hormonal secretion, pet ownership, dog bites, aggressive behavior, public health, problem, specific cases.

3.4.2. Extraction de mots-clés basée sur un corpus de documents

Dans cette étape, nous considérons un ensemble de documents contenant le document D (corpus de documents) pour identifier les mots-clés du document. Pour ce faire, nous avons testé des méthodes basées sur les statistiques. Parmi ces méthodes, nous citons fréquence des mots (PARSING 2009), le TF-IDF (SALTON et BUCKLEY 1988) et Likey (PAUKKERI et HONKELA 2010). Les résultats des expériences, présentées dans la section 3.7, montrent que l'algorithme TF-IDF donne de meilleurs résultats que les autres méthodes.

L'un des principaux avantages de l'utilisation de TF-IDF pour l'extraction de mots-clés est qu'elle permet de donner plus de poids aux termes qui sont importants pour un document particulier, tout en réduisant le poids des termes communs à tous les documents du corpus. Ainsi, les termes qui sont fréquents dans un document donné, mais qui ne sont pas très fréquents dans l'ensemble du corpus, obtiennent un poids plus élevé, tandis que les termes qui sont fréquents dans tous les documents, comme les mots outils, obtiennent un poids plus faible. Cette méthode peut également être appliquée à des corpus de documents de tailles variables et peut être adaptée pour tenir compte de la longueur des documents.

En appliquant TF-IDF à l'exemple illustré dans la figure 3.2, les mots-clés extraits

du texte annotés par leur poids (calculé en utilisant la formule de TF-IDF) sont les suivants : Attacks : 0.1384, Bites : 0.1384, Injuries : 0.1317, Estrogens hormones : 0.1121, Estrogens hormones : 0.1121, United States : 0.1121, Public health : 0.1055, Hormonal secretion : 0.1055, Gonad glands : 0.1055.

3.4.3. Sélection des mots-clés pertinents

Pour choisir les mots les plus pertinents à partir d'un ensemble de mots extraits par FP-Growth, il est nécessaire de calculer la pondération de chaque mot. Cette pondération est déterminée en prenant en compte la fréquence des mots dans le document et leur importance relative par rapport à l'ensemble des mots extraits. Cependant, il est important de noter que la fréquence ne suffit pas à elle seule pour déterminer la pertinence d'un mot. En effet, certains mots peuvent apparaître fréquemment dans le document simplement parce qu'ils sont couramment utilisés dans le langage ou le domaine traité par le document, sans pour autant être pertinents pour le contenu spécifique du document. Pour résoudre ce problème, une approche plus sophistiquée est nécessaire. Le calcul de la pondération des mots est nécessaire dans une troisième phase pour choisir les mots qui ont un poids élevé. Cette pondération doit prendre en compte le contexte et la similarité sémantique pour améliorer la précision de l'extraction des mots-clés.

Le calcul des poids des mots générés par FP-growth est basé sur la similarité sémantique entre ces mots et les mots extraits en utilisant TF-IDF. Il existe plusieurs méthodes et modèles pour calculer la similarité entre deux mots comme la distance de Jaccard (JACCARD 1901), la similarité de Levenshtein (LEVENSHTEIN et al. 1966), la distance de Hamming (HAMMING 1950), les modèles Word2Vec (MIKOLOV, K. CHEN, CORRADO et al. 2013) et l'Universal Sentence Encoder (USE) (**c9)**). Universal Sentence Encoder (USE) a l'avantage de prendre en compte la sémantique des phrases et des mots dans leur contexte, ce qui permet d'obtenir des résultats plus précis et adaptés à la tâche de l'extraction de mots-clés. De plus, USE est pré-entraîné sur de larges corpus de texte, ce qui lui permet de capturer des nuances subtiles de la langue et d'identifier des relations sémantiques entre des mots qui pourraient passer inaperçues pour les autres méthodes. Pour cela, l'USE est utilisé.

Lorsque l'USE est utilisé pour calculer la pondération des mots, il génère une matrice qui représente les vecteurs de chaque mot. Cette matrice est basée sur une représentation vectorielle de chaque mot, qui est créée en utilisant un réseau de neurones entraîné sur des données textuelles massives.

Plus précisément, le réseau de neurones du USE prend en entrée une séquence de mots et génère un vecteur qui représente cette séquence. Pour chaque mot de la séquence, le réseau de neurones utilise un modèle de langage préalablement entraîné pour produire une représentation vectorielle du mot. Ces représentations sont ensuite agrégées pour former un vecteur qui représente l'ensemble de la séquence.

Pour calculer la pondération d'un mot extrait par FP-Growth, le vecteur « u » correspondant est extrait de la matrice USE. De même, le vecteur « v » correspondant à un mot-clé généré par TF-IDF est extrait de la matrice USE. La similarité entre le vecteur

d'un mot extrait par FP-Growth et le vecteur de chaque mot-clé généré par TF-IDF est ensuite calculée en utilisant la mesure de similarité cosinus, représentée par la formule suivante :

$$Similarity(u,v) = \frac{u \cdot v}{\|u\| \times \|v\|} \tag{3.4}$$

Cette mesure de similarité est une mesure standard dans le traitement du langage naturel qui mesure la similitude entre deux vecteurs. En calculant la similarité cosinus pour chaque paire de vecteurs, on obtient une matrice de similarité qui mesure la similarité sémantique entre chaque mot extrait par FP-Growth et chaque mot-clé généré par TF-IDF, représentée par la matrice de taille N x M, où N est le nombre de mots-clés extraits par TF-IDF et M le nombre de mots extraits par FP-Growth. Les valeurs de la matrice représentent les similarités cosinus entre les vecteurs de chaque paire de mots-clés.

En utilisant la matrice de similarité, on peut calculer la pondération de chaque mot extrait par FP-Growth en prenant la moyenne des similarités avec tous les mots-clés générés par TF-IDF, représentée par la formule suivante :

$$Weight(\text{mot}_i) = \frac{1}{n} \sum_{j=1}^{n} \text{similarity}(\mathbf{u}_i, \mathbf{v}_j) \tag{3.5}$$

où ui correspond au vecteur du moti généré par FPgrowth, vj correspond au vecteur du motj généré par TF-IDF et n est le nombre total de mots-clés générés par TF-IDF.

Enfin, l'ensemble des mots-clés est trié par ordre décroissant de score pour sélectionner les mots les plus pertinents. Ces mots-clés sont utiles pour mieux comprendre le contexte du document et servent à la construction de celui-ci.

Dans le cadre de cette méthode, nous avons choisi de fixer initialement le nombre de mots- clés pertinents à 5 qui servent par la suite pour résumer efficacement le contenu du document en une phrase. Ce nombre a été choisi sur la base d'expériences et de considérations pratiques effectuées. Parmi les mots-clés restants, ceux qui ont un poids proche du cinquième mot-clé sont également retenus. La proximité des scores est calculée en fonction d'un hyper paramètre β. Tous les mots-clés Mc_t tels que |Poids(Mc_5) - Poids(Mc_t)| $\leq \beta$ sont retenus. Mc_5 représente le cinquième mot-clé et Mc_t est un mot-clé quelconque parmi ceux qui n'ont pas été conservés. L'hyper paramètre sera définit par les expérimentations.

Après l'application de cette phase à l'exemple illustré dans la figure 3.2, on obtient les poids des mots suivants : animal attacks 0.320, injuries 0.333, fatalities 0.306, bites 0.310, problem 0.286, public health 0.286, dog bites 0.267, aggressive behavior 0.243, hormonal secretion 0.232, United States 0.219, pet ownership 0.213, wounds 0.189, frequency 0.183, specific cases 0.124, and geographical location 0.099. En choisissant les mots avec les poids les plus élevés, les mots-clés les plus pertinents pour le contexte du document sont les suivants : animal attacks, injuries, fatalities, bites, problem et public health. Ces mots sont plus précis et pertinents par rapport aux mots générés par TF-IDF seul ou FP-Growth, ce qui permet d'améliorer la qualité de l'extraction des mots-clés.

3.5. Phase d'extraction du contexte du document

Après avoir extraire les mots-clés du document, la question la plus importante qui se pose est : comment construire le contexte à partir de ces mots ? Dans cette section, nous avons identifié deux approches pour l'identification de la phrase représentant le contexte : l'une *extractive* et l'autre *générative*. Cette phrase est obtenue à partir des mots-clés du document.

3.5.1. Approche *extractive* pour l'extraction de contexte

La première étape de l'approche extractive d'extraction de contexte (Figure**??**) est l'identification de la phrase qui résume brièvement le contenu du texte. Pour ce faire, une analyse minutieuse du document est effectuée afin d'identifier la phrase qui contient le maximum de mots-clés extraits dans la phase précédente. Cependant, cette première phrase identifiée peut contenir des informations non pertinentes pour le contexte visé. À cet effet, deux cas de figure peuvent être distingués :

— Si la phrase contient des structures grammaticales complexes : Dans ce cas il faut passer par un processus de simplification. Ce processus peut être effectué en trois étapes. Tout d'abord, il est important d'identifier la proposition principale en utilisant des outils linguistiques tels que l'analyse syntaxique. Cette proposition principale contient l'information principale de la phrase et peut exister seule. Ensuite, il faut éliminer les phrases subordonnées en cherchant les mots subordonnants tels que « while », « although », « because », « if », etc. Les phrases subordonnées peuvent être éliminées pour simplifier la phrase. Cependant, si la phrase subordonnée contient des mots-clés, elle sera conservée dans le processus d'extraction du contexte afin de s'assurer que toutes les informations importantes sont incluses. Enfin, après avoir éliminé les phrases subordonnées, il est important d'identifier et d'éliminer les mots inutiles tels que « very », « realy », « quite », etc. Ces mots couramment utilisés mais qui peuvent être considérés comme des mots inutiles car ils ne contribuent pas toujours de manière significative à la compréhension de la phrase. La phrase obtenue sera considérée comme le contexte final du document

— Si la phrase ne contient ni de phrases subordonnées ni de mots inutiles, alors elle peut être considérée comme le contexte final du document.

Après avoir appliqué l'approche extractive sur le document présenté dans la figure 3.2, la première phrase extraite est « Animal attacks are identified as a public health problem, although their frequency and severity can vary depending on many factors.". Cette phrase contient le maximum de mots-clés, mais nous constatons qu'elle contient également deux phrases subordonnées : « as a public health problem" et « although their frequency and severity can vary depending on many factors". La première phrase étant pertinente car elle contient des mots-clés, elle sera gardée dans le contexte final tandis que la deuxième phrase sera éliminée. Ainsi, le contexte final sera :« Animal attacks are identified as a public health problem". Cette phrase finale représente une version plus claire et concise de l'information principale du document, facilement

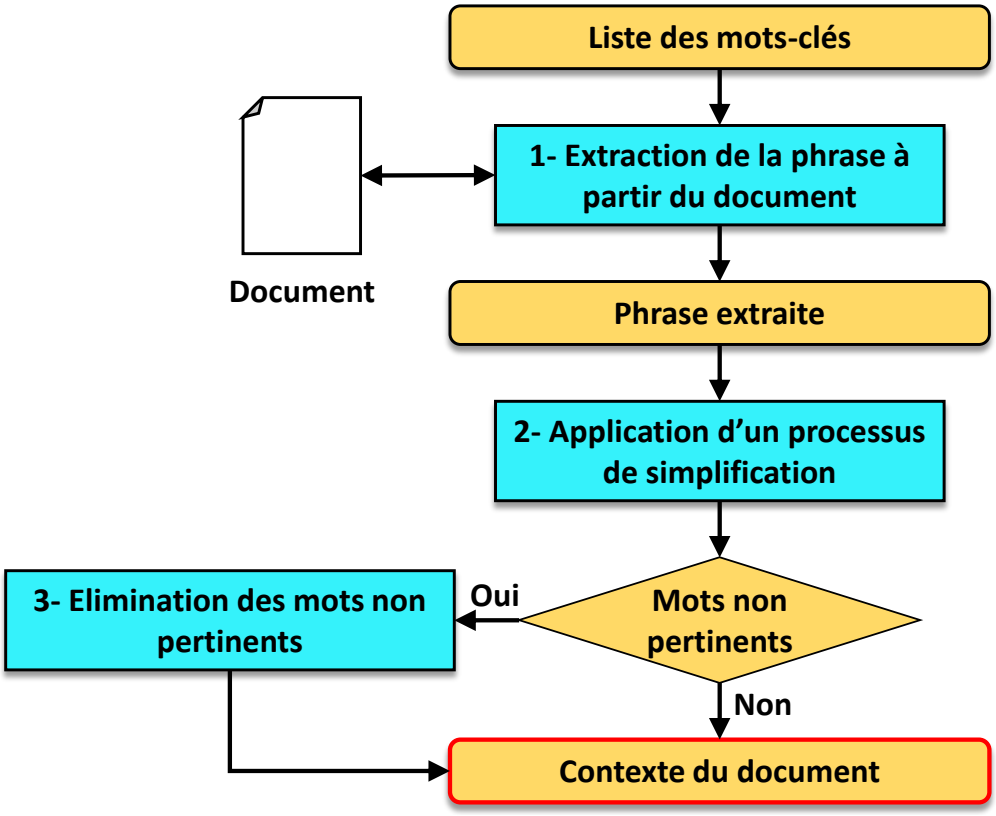


FIGURE 3.4. – : Processus de l'approche *extractive* d'extraction de contexte

compréhensible pour le lecteur.

Cette approche peut être améliorée, principalement au niveau de la qualité de la phrase générée. En effet, cette approche ne peut pas générer un contexte de très haute qualité. Sa performance dépend fortement des différentes phrases incluses dans le document. Ce dernier peut ne pas inclure une phrase avec la majorité des mots-clés extraits. Pour cette raison, l'intégration des réseaux de neurones peuvent améliorer la phase de génération des phrases et ainsi la qualité du contexte final.

Dans la section suivante, nous proposons une nouvelle approche générative basée sur l'apprentissage profond pour l'extraction de contexte.

3.5.2. Approche *générative* pour l'extraction de contexte

Notre approche générative pour l'extraction du contexte est basée sur la modélisation du langage. Cette approche nécessite trois phases principales : (1) la phase de construction de notre modèle LSTM et son entraînement, (2) la phase de construction de la chaine de Markov et (3) la phase de modélisation du contexte de document.

3.5.2.1. Construction de notre modèle LSTM et son entrainement

— **Structure d'une couche LSTM :** Une couche LSTM (Long Short-Term Memory) est une couche récurrente utilisée dans les réseaux de neurones pour traiter des séquences de données, telles que des séquences de mots dans un texte. En entrée, la couche LSTM reçoit un mot (ou un vecteur représentant un mot) ainsi que l'état caché et l'état de la cellule de la couche LSTM provenant de la précédente itération de la séquence. Ces états cachés et de cellules sont des vecteurs qui stockent des informations sur les mots précédents de la séquence.
La structure d'une cellule LSTM est présentée dans la figure 3.5, qui est composée d'une porte d'entrée, d'une porte d'oubli, d'une porte de sortie et de l'état de la cellule. En effet, à l'intérieur de la couche LSTM, plusieurs calculs sont effectués pour traiter l'entrée et mettre à jour l'état caché et l'état de la cellule. Ces calculs comprennent des portes (gate) qui permettent de contrôler l'information qui est stockée et oubliée dans la cellule, ainsi que des fonctions d'activation qui transforment l'entrée et l'état précédent pour produire la sortie de la couche. Ces portes sont calculées à l'aide de fonctions d'activation telles que la fonction

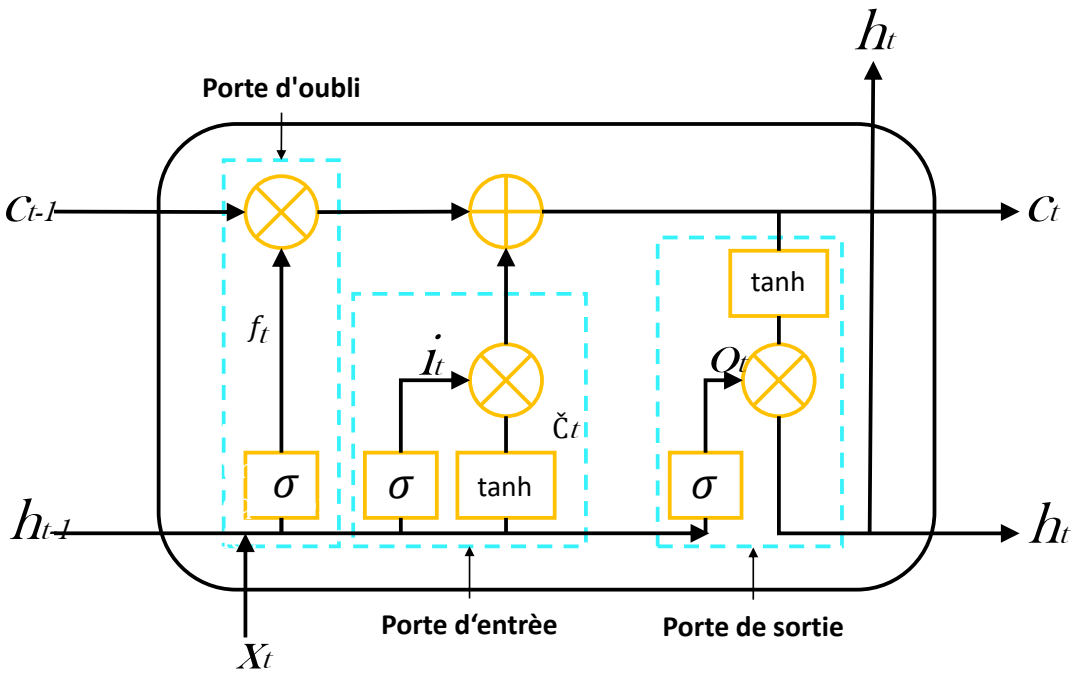


FIGURE 3.5. – Structure d'une cellule LSTM

sigmoïde, qui transforme les entrées en valeurs entre 0 et 1. Ensuite, l'entrée actuelle et l'état caché précédent sont transformés en un vecteur d'activation intermédiaire qui est stocké dans la cellule LSTM en utilisant la fonction d'activation tanh, qui produit des valeurs entre -1 et 1. Enfin, les portes sont utilisées pour combiner l'activation intermédiaire avec l'état de la cellule précédent pour

produire l'état de la cellule actuel. Cet état de cellule est ensuite combiné avec la sortie produite par la porte de sortie pour produire l'état caché de l'unité LSTM. Finalement, la sortie de la couche LSTM est le mot suivant dans la séquence, représenté par un vecteur de probabilités indiquant la probabilité de chaque mot dans le vocabulaire d'être le mot suivant.

— **Construction de notre modèle LSTM pour la prédiction du mot prochain et son entrainement :** Notre modèle de prédiction de mots suivant basé sur « LSTM », construit dans cette thèse, est illustré par la figure **??** Ce modèle se compose principalement d'une couche d'entrée, d'une couche cachée et d'une couche de sortie. La couche d'entrée correspond à la couche « embedding » qui permet de convertir les mots-clés en vecteurs de mots denses de faible dimension. La couche cachée est constituée de deux couches « LSTM » et de deux couches « dropout ». En effet, chaque couche « LSTM » doit être accompagnée d'une couche « dropout ». Une telle couche permet d'éviter le sur-apprentissage. Enfin, pour calculer le score de chaque mot généré prédit par le modèle, une couche « dense » est utilisée avec « Softmax » comme fonction d'activation.
Dans notre cas, nous entrainons notre modèle pour faire des prédictions précises et d'effectuer la tâche de génération du contexte de document. Pour notre modèle « LSTM », nous avons préparé le corpus « WikiContext » qui contient 7592 documents dans des contextes différents. Ce corpus est introduit dans nos modèles « LSTM » pour l'entraînement. Initialement, nous avons défini le nombre d'époques à 1000 pour l'entraînement de notre modèle. Cependant, nous avons utilisé une fonction prédéfinie de Python appelée 'Early Stopping' qui permet de surveiller la performance du modèle sur un ensemble de validation et de stopper l'entraînement automatiquement lorsque la performance ne s'améliore plus pendant un certain nombre d'époques consécutives. Cela permet d'éviter un sur-apprentissage (overfitting) du modèle et d'obtenir un modèle plus généralisable.

3.5.2.2. Construction de la table de Markov

Dans notre approche, la table de Markov est utilisée pour compléter les prédictions du modèle LSTM. En effet, bien que le modèle LSTM soit capable de générer une séquence de mots cohérente, il est parfois nécessaire de faire appel à d'autres techniques pour améliorer la qualité des prédictions. C'est là que la table de Markov intervient.

Une table de Markov est une représentation de la probabilité de transition entre les différents états possibles d'un système, où chaque état représente un mot possible dans la séquence de mots que nous essayons de prédire. Cette table contient des informations sur la probabilité qu'un mot donné suive un autre mot, en fonction de l'historique des mots précédents.

En pratique, la table de Markov est souvent représentée sous forme d'un tableau bidimensionnel où les colonnes et les lignes représentent les mots possibles, et chaque entrée du tableau représente la probabilité de transition entre deux mots. Cette probabilité est calculée en fonction de la fréquence de la transition entre les deux mots

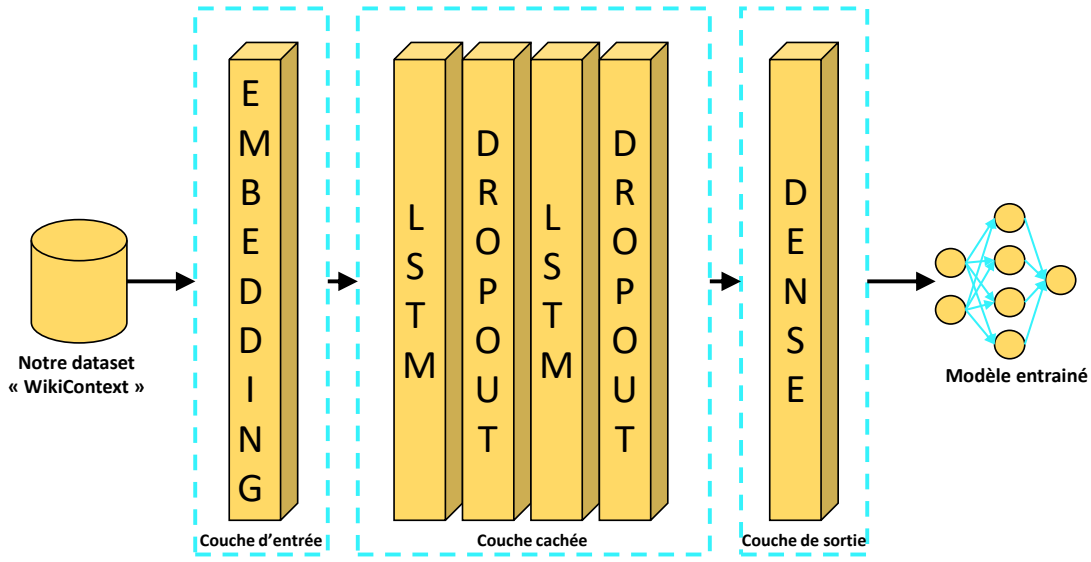


FIGURE 3.6. – : Modèle LSTM utilisé pour générer le contexte

dans l'ensemble de données d'entraînement. Plus précisément, la probabilité de transition d'un mot X à un mot Y est calculée en divisant le nombre de fois où le mot Y suit le mot X par le nombre total de transitions de mot dans l'ensemble de données d'entraînement.

Plus précisément, la table de Markov est utilisée dans trois situations. Tout d'abord, elle est utilisée pour décider quel est le premier mot, parmi les mots-clés généré dans l'étape précédente, à anticiper comme entrée dans le modèle LSTM. Ensuite, la table de Markov est également utilisée pour prédire le mot suivant dans la séquence, lorsque le modèle LSTM génère un ensemble de mots possibles. Dans ce cas, la table de Markov peut être utilisée pour déterminer quelle est la meilleure prédiction parmi l'ensemble des mots générés par le modèle LSTM, en fonction du contexte et des probabilités de transition entre les différents mots.

Enfin, la table de Markov peut également être utilisée pour insérer des mots de liaison ou des mots vides entre les mots générés par le modèle LSTM, de manière à améliorer la cohérence et la fluidité de la séquence de mots, dans le cas ou notre modèle ne génère pas un mot parmi la liste des mots-clés extraits. Nous utilisons notre corpus « WikiContext » pour construire la table de Markov.

3.5.2.3. Modélisation du contexte du document

La figure**??** ainsi que l'algorithme 1 montrent notre processus pour la modélisation du contexte du document. Le processus de modélisation du contexte du document implique l'utilisation de notre table de Markov et de notre modèle LSTM pour prédire la suite de la phrase en cours de génération. Tout d'abord, nous utilisons notre table de Markov pour sélectionner le mot-clé qui a la plus forte probabilité d'être le premier de la phrase. Une fois que nous avons sélectionné le mot-clé initial, nous utilisons

notre modèle LSTM pour prédire le mot suivant en utilisant le mot initial comme partie du préfixe pour la prochaine entrée du modèle.

Le modèle LSTM peut prédire une liste de mots suivants, et le choix du mot approprié doit être effectué. Pour prendre cette décision, nous comparons chaque mot prédit par LSTM avec les mots-clés extraits du document pour déterminer lequel est le plus similaire. Si un mot prédit par LSTM a une forte similarité avec les mots-clés du document, il est sélectionné comme prochain mot.

Cependant, si le modèle LSTM ne génère pas un mot qui a une forte similarité avec les mots-clés du document, nous utilisons notre table de Markov pour prédire le mot suivant. Dans ce cas, le choix du mot approprié est basé sur la probabilité définie par la table de Markov. Ce processus est répété jusqu'à ce que la liste des mots-clés extraits soit vide. En d'autres termes, nous générons une phrase en utilisant tous les mots-clés extraits.

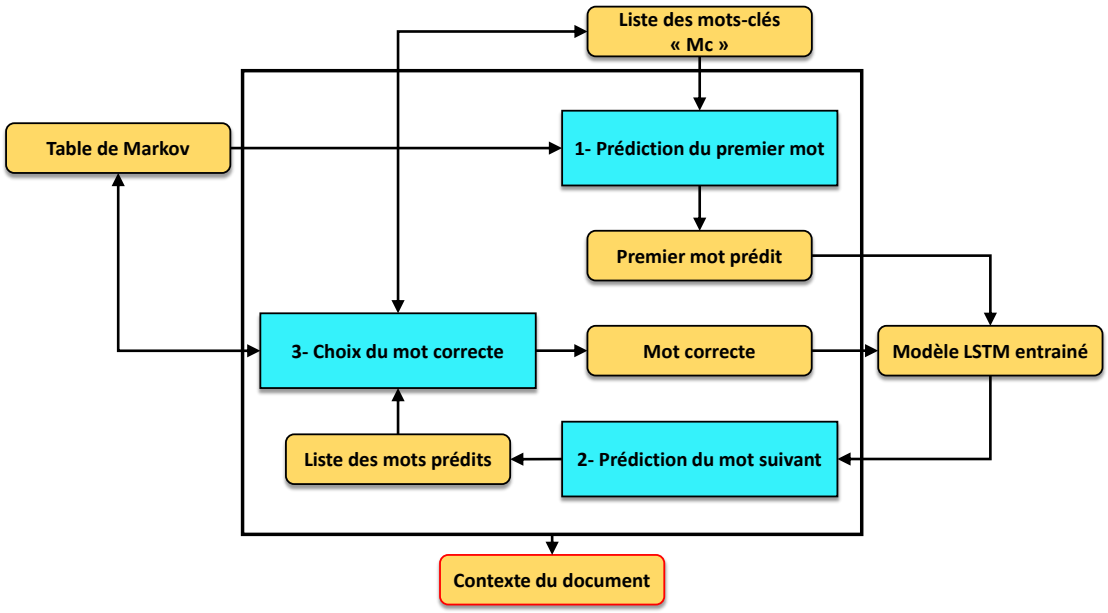


FIGURE 3.7. – : Approche *générative* d'extraction de contexte

Algorithm 1 Algorithme de modélisation de contexte du document

input : Liste-motsclés
input : TableMarkov
output : Context
function ModélisationContext($Liste-motsclés, TableMarkov$)
$Context \leftarrow$ ""
/*Prédiction du premier mot*/
for $mot \in List-motsclés$ **do**
    $premiermot \leftarrow PredirePremierMot()$
end for
$Context \leftarrow premiermot$
/*Suppression du premier mot prédit de la liste des mots-clés*/
$Liste-motsclés \leftarrow supprimermot(premiermot)$
/*Prédiction du mot suivant*/
repeat
    $Liste-MotPrédit \leftarrow ProchainePrédiction(Context)$
    $Motprochain \leftarrow ChoixMotProchain(Liste-MotPrédit)$
    $Liste-motsclés \leftarrow supprimermot(Motprochain)$
    $Context \leftarrow Context + Motprochain$
until $Liste-motsclés = \varnothing$
return $Context$
End function

Avec :

— **PredirePremierMot() :** est une méthode qui utilise les probabilités fournies par la table de Markov, de chaque mot se trouvant au début d'une phrase afin de prédire le premier mot.
— **supprimermot() :** est une méthode qui permet de supprimer un mot donné de la liste de mots-clés.
— **ProchainePrédiction() :** est une méthode qui utilise un modèle LSTM entraîné pour prédire le mot suivant

Une fois que la méthode générative a été appliquée à l'exemple présenté dans la figure 3.2, la phrase construite sera la suivante : Animal attacks and bites are a public health problem that cause injuries and fatalities.

3.6. Stockage du contexte et du document associé dans une base de contextes « BdC »

L'association entre un document et son contexte correspondant, à travers une base de contextes (BdC), est cruciale pour faciliter l'extraction de l'information. Comme l'illustre la figure**??**, la BdC est composée de trois colonnes : la colonne *'contextes'* contient les différents contextes, la colonne *'id'* contient les ID correspondants et la

colonne *'documents'* contient les *URLs* des documents associés à chaque contexte.

id	contextes	documents
1	French Presidential Elections of 2017	https://en.wikipedia.org/wiki/2017_French_presidential_election
2	FIFA World Cup 2018	https://en.wikipedia.org/wiki/2022_FIFA_World_Cup
3	Education in the USA	https://en.wikipedia.org/wiki/Education_in_the_United_States
4	Natural Language Processing	https://en.wikipedia.org/wiki/Natural_language_processing

FIGURE 3.8. – : Organisation de la base de contextes (BdC)

Ce stockage conjoint du contexte et du document associé dans la base de contextes BdC est d'une grande utilité :

— Tout d'abord, cela permet de vérifier rapidement si un document a déjà été traité précédemment, évitant ainsi de répéter tout le processus d'extraction du contexte. On peut simplement annoter le document par le contexte adéquat de la base.

— Ensuite, si le contexte généré n'est pas assez compréhensible ou si la phrase générée ne correspond pas à nos attentes, nous pouvons consulter la base de contextes pour voir si le document appartient à l'un des contextes stockés. Dans ce cas, nous pouvons annoter le document avec le contexte correspondant, plutôt que d'utiliser la phrase générée par notre modèle définissant le contexte extrait par la méthode.

Le processus de stockage du nouveau contexte et du document associé dans BdC est présenté dans la figure3.9. Le processus commence par la vérification dans la BdC si le document a été traité auparavant ou non. Si c'est le cas, on n'a pas besoin d'extraire le contexte à nouveau. Au lieu de cela, on peut simplement annoter le document avec le contexte stocké dans la BdC. Si le document n'a pas encore été traité, on procède à l'extraction du contexte en utilisant notre méthode générative décrite précédemment.

Une fois que le contexte a été extrait, on procède au stockage du contexte dans la BdC. Nous vérifions si le contexte extrait existe déjà dans la base en effectuant une comparaison avec les contextes de la BdC. Si le contexte existe déjà dans la base, on met à jour la BdC en ajoutant simplement le document associé au contexte existant.

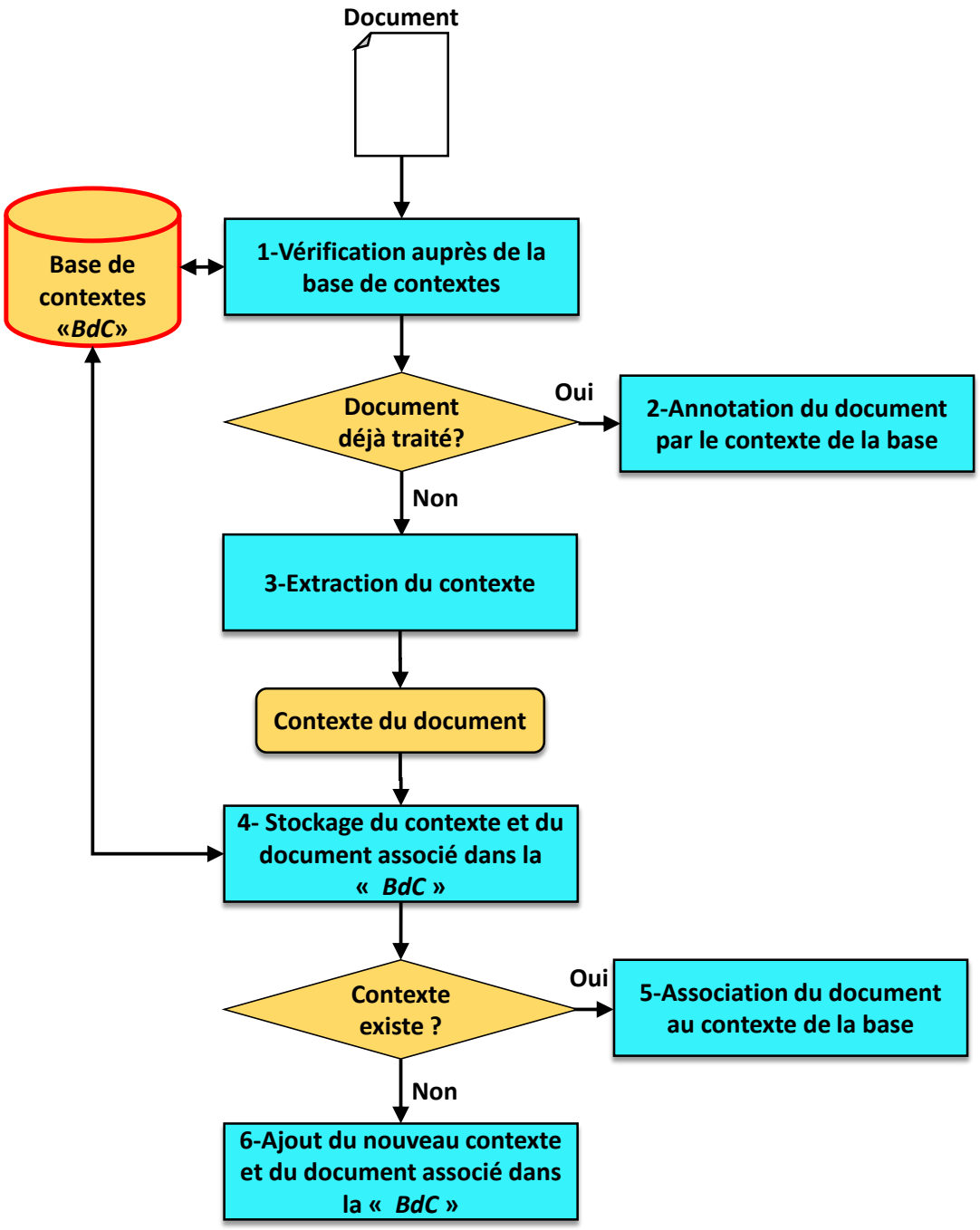


FIGURE 3.9. – : Processus de stockage du contexte et du document associé dans une base de contextes

Dans le cas où le contexte extrait est nouveau et n'existe pas encore dans la base, nous ajoutons le nouveau contexte à la BdC avec le document associé. Cela permet de continuellement enrichir la BdC avec de nouveaux contextes et de nouveaux documents pour améliorer les performances du système. En utilisant cette approche, nous pouvons garantir que chaque document est annoté avec le contexte approprié

sans avoir besoin de répéter le processus d'extraction de contexte pour les documents déjà traités. De plus, en stockant les contextes et les documents associés dans la même BdC, on peut rapidement rechercher et associer de nouveaux documents à des contextes existants.

Nous devons maintenant aborder la question de la comparaison entre le contexte extrait et les contextes existants dans la base. Il est crucial de mesurer la similarité entre ces contextes afin de déterminer si le contexte est déjà présent dans la base ou s'il s'agit d'un nouveau contexte.

Pour effectuer la comparaison entre le contexte extrait et les contextes de la base de contextes, nous avons testé plusieurs techniques de calcul de similarité à savoir la similarité cosinus, le clustering, LSTM et BERT. Les expériences présentées dans la section 3.8.3.1 nous ont permis d'utiliser l'algorithme Bert qui a obtenu les meilleurs résultats par rapport aux autres. En effet, comparé aux modèles LSTM, BERT a l'avantage de capturer les dépendances à long terme entre les mots, ce qui est important pour comprendre le contexte global d'une phrase.

En ce qui concerne le clustering et la similarité cosinus, ces méthodes sont basées sur une approche de représentation vectorielle des phrases. Cependant, ces méthodes peuvent manquer de précision pour la mesure de la similarité sémantique entre les phrases, car elles ne tiennent pas compte du contexte global dans lequel les phrases sont utilisées. BERT, en revanche, est capable de saisir des informations contextuelles plus précises, ce qui peut conduire à une mesure de similarité plus précise et plus adaptée à l'objectif du calcul de similarité entre deux contextes.

3.6.1. Structure et pré-entraînement du modèle BERT :

« BERT » ou encore « Bidirectional Encoder Representations from Transformers » est un modèle de représentation de textes écrit en langage naturel. « BERT » partage la même architecture qu'un encodeur de transformateurs et est largement pré-entraîné sur des données textuelles brutes et non étiquetées à l'aide d'un objectif d'apprentissage auto-supervisé (self-supervised), avant d'être affiné par un apprentissage par transfert (fine-tuning) pour résoudre d'autres tâches. La représentation faite par « BERT » à la particularité d'être contextuelle. En effet, ce modèle permet de traiter chaque mot de texte d'entrée dans le contexte complet de tous les mots avant et après. C'est-à-dire qu'un mot n'est pas représenté de façon statique comme dans un « embedding » classique mais en fonction du sens du mot dans le contexte du texte. Comme montre la figure 3.10, le modèle « BERT » se compose principalement des deux parties : (1) « BERT tokenizer » et (2) « BERT encoder ».

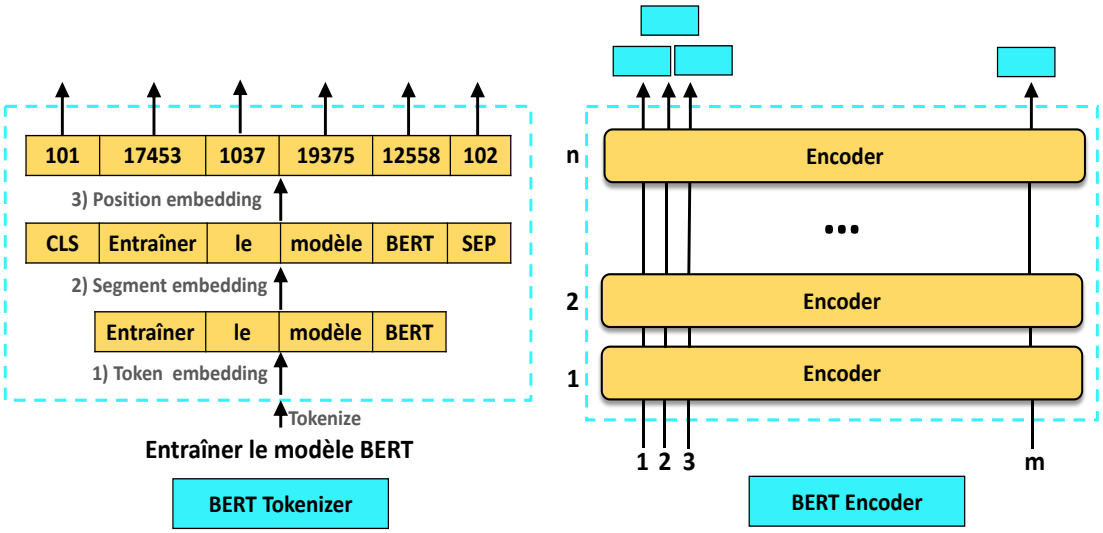


FIGURE 3.10. – : Les différents composants du modèle BERT

Comme montre la figure 3.10, la représentation en entrée de chaque mot est obtenue par « BERT tokenizer » en additionnant les trois tâches « token embedding », « segment embedding » et « position embedding ». Le « token embedding » permet de décomposer en éléments individuels qui représentent des mots ou des parties de mots. Le « segment embedding » permet d'insérer de « jetons spéciaux » : la séquence d'entrée de « BERT » commence par un jeton [CLS] et se termine par un jeton [SEP], indiquant le début/la fin d'une phrase. Finalement, dans la tâche « position embedding », des « additive embeddings » sont ajoutés à chaque élément de cette séquence, représentant la position de l'élément dans la séquence.

« BERT encoder » est basé sur les transformateurs. Il utilise des « n » encodeurs transformateurs bidirectionnels multicouches pour les représentations du langage. En fonction de la profondeur de l'architecture du modèle, deux types de modèles « BERT » sont introduits : « BERT Base » et « BERT Large ». Le modèle « BERT Base » utilise 12 couches de bloc de transformateurs avec une taille cachée de 768. D'un autre côté, « BERT Large » utilise 24 couches de transformateurs avec une taille cachée de 1024.

« BERT » se différencie par la façon dont il est pré-entraîné. Ce pré-entraînement est auto-supervisé c'est-à-dire qu'il ne nécessite pas de jeu de données labellisé. « BERT » est pré-entraîné sur un grand jeu de données constitué de textes des pages Wikipédia en anglais (2 500 millions de mots) ainsi qu'un ensemble de livres (800 millions de mots).

3.6.2. Modèle BERT utilisé pour le calcul de similarité entre deux contextes

La figure 3.11 montre le modèle BERT utilisé pour le calcul de similarité entre deux contextes.

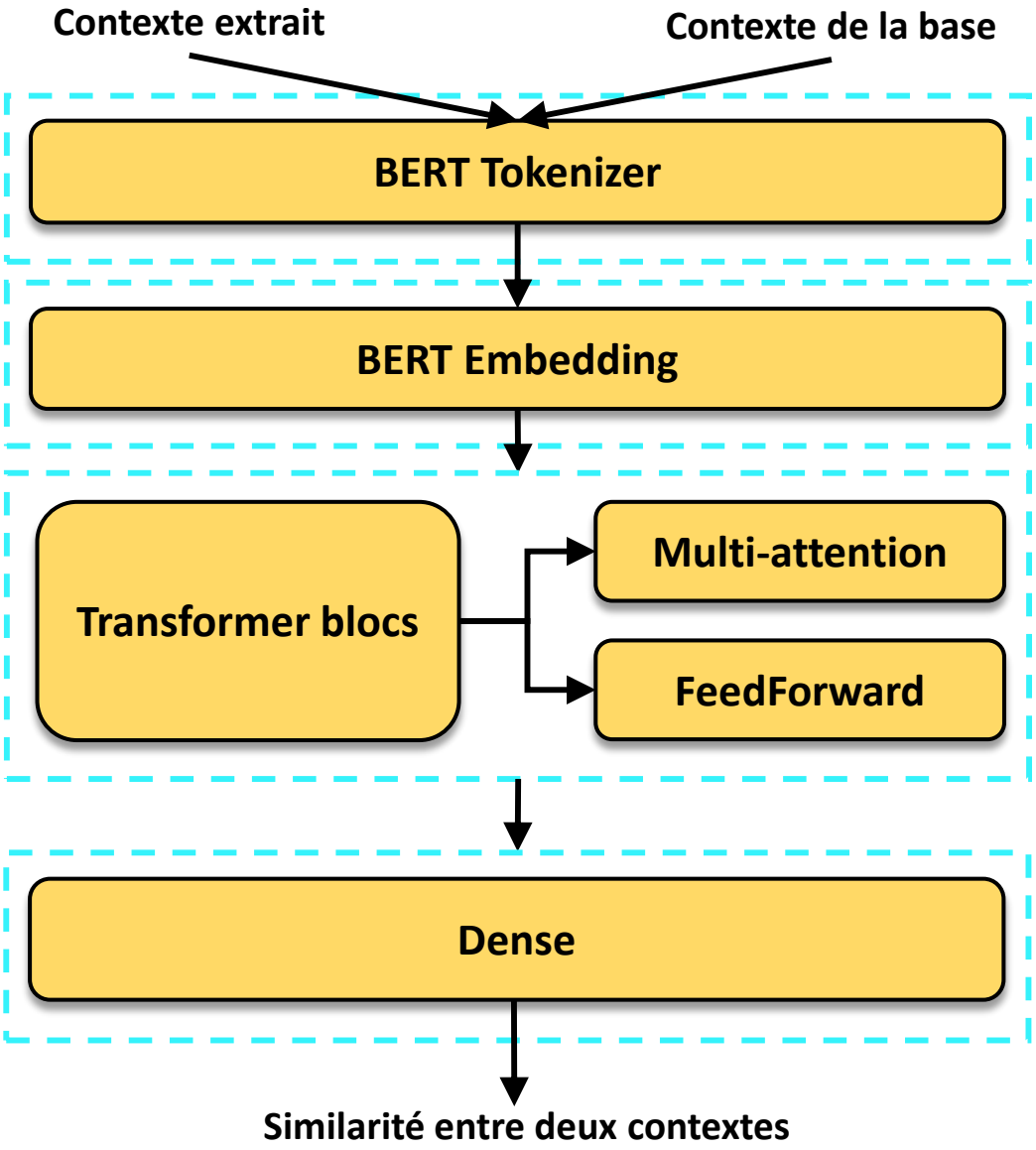


FIGURE 3.11. – : Modèle BERT utilisé pour le calcul de similarité entre deux contextes

Notre modèle de calcul de similarité entre deux contextes est composé de quatre couches où chaque couche prend comme paramètre le résultat de la couche précédente : La première couche correspond à la couche Tokinizer qui est responsable de la tokenisation des mots en sous-mots, ainsi que de l'ajout des tokens spéciaux [CLS] et [SEP]. Le token [CLS] est ajouté au début de la première phrase et un token [SEP] est ajouté entre les deux phrases qui sont concaténées en entrée du modèle Bert. La couche Embedding de BERT prend ensuite ces tokens et les transforme en vecteurs de mots. La troisième couche correspond à la couche des transformers. Cette couche contient la couche de multi-attention. Cette couche calcule la similarité entre tous les tokens des deux contextes en utilisant une matrice de poids de similarité. La similarité

est calculée en utilisant une attention multi-têtes pour capturer différentes relations entre les tokens et une couche de feedforward cachée. Cette couche prend en entrée les vecteurs de similarité calculés dans la couche précédente et les transforme en une représentation de dimension fixe. Finalement, la cinquième couche est la couche de sortie. Cette couche prend en entrée les vecteurs de la couche de feedforward cachée et calcule la similarité entre les deux phrases en utilisant une fonction de similarité, telle que la similarité cosinus.

Nous affinons notre modèle par une méthode de transfert Learning (« fine-tuning ») sur notre corpus « WikiContext » pour améliorer la précision de notre modèle.

3.7. Expérimentation et évaluation

Nous avons décrit dans les sections précédentes un ensemble de propositions décrivant les différentes phases d'une méthode d'extraction de contexte à partir d'un document texte non structuré. Comparée aux travaux existants dans le domaine, notre méthode se distingue par l'extraction d'un contexte précis permettant une extraction d'informations pertinentes du texte.

L'évaluation globale de notre méthode dépend étroitement de l'évaluation des résultats de chacune des trois phases :

— Evaluation des résultats de la phase « Extraction des mots-clés »
— Evaluation des résultats de la phase « Extraction de contexte »
— Evaluation des résultats de la phase « Stockage du contexte et du document associé dans une base de contextes « BdC »

Dans le but de valider notre méthode, notre démarche est d'effectuer une série d'expérimentations sur des corpus de données donnés dans le cadre d'un protocole expérimental. Dans les sections suivantes nous décrivons l'environnement expérimental ainsi que le protocole expérimental suivi.

3.7.1. Environnement expérimental

Dans cette section, nous présentons l'environnement matériel, le choix des technologies ainsi que l'environnement logiciel utilisés dans le développement de la méthode d'extraction de contexte. Pour la mise en œuvre de cette méthode, nous avons travaillé sur une machine dont la configuration est la suivante : CPU Intel Core i9-7960X, 2.8GHz, 16 Cores et 16G de RAM. Comme langages de programmation, nous avons choisie Python pour implémenter notre solution. Ce langage puissant nous permet de mettre en œuvre plus facilement des algorithmes d'apprentissage automatique et d'exploration de texte grâce à ses nombreuses bibliothèques disponibles. Parmi les bibliothèques utilisées, citons NLTK, Pandas, Gensim, Keras, Mlxtend et Markovify :

— ***NLTK :*** Nous avons choisi de travailler avec ***NLTK*** car elle nous fournit de nombreuses fonctionnalités que nous devons utiliser dans la phase de prétraitement des données pour nettoyer notre dataset afin d'extraire le contexte ainsi que les mots-clés.

— ***Pandas*** nous a permis de mieux visualiser et effectuer des opérations sur notre dataset. Elle nous a permis de transformer l'ensemble de données en un Dataframe.
— Nous avons bénéficié des fonctions de la bibliothèque ***Gensim*** dans la vectorisation de texte pour en recevoir des représentations numériques distribuées des caractéristiques des mots.
— ***Keras*** permet d'interagir avec les réseaux de neurones profonds et d'apprentissage automatique. Elle met à notre disposition des modules permettant de créer nos modèles.
— ***Mlxtend*** nous a permis l'implémentation de l'algorithme FP-Growth.
— ***Markovify*** nous a servi dans la construction de la table de Markov.

Concernant l'environnement logiciel, nous avons utilisé ***Google-Colab*** qui représente un service gratuit offert par Google. Il nous a permis d'effectuer l'entraînement de nos modèles de machine learning directement dans le cloud.

3.7.2. Protocole expérimental

Dans le domaine de la recherche, le protocole expérimental représente une étape essentielle dans la validation des résultats de recherche. Un protocole expérimental peut être défini comme une liste de tâches expérimentales organisées de façon temporelle permettant d'aboutir à des résultats exploitables. Le protocole que nous avons suivi est composé de trois tâches principales : *Pré-évaluation*, *Evaluation* et *Post-évaluation*, comme l'illustra la figure 3.12.

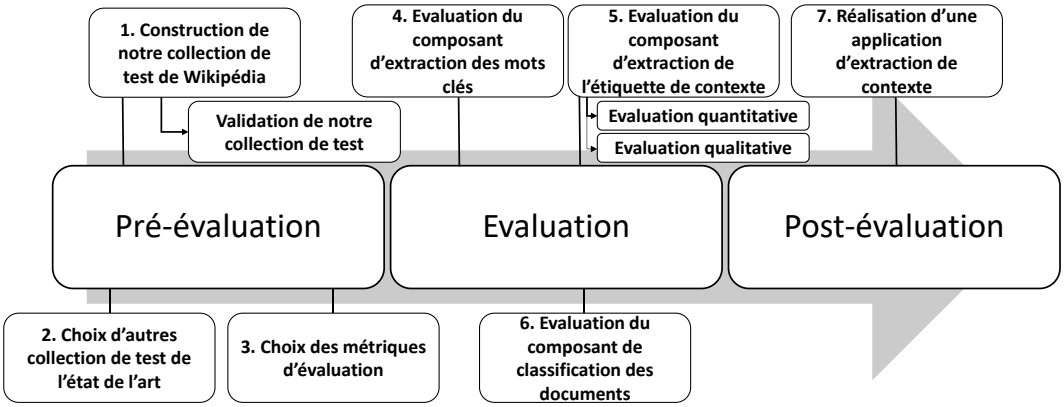


FIGURE 3.12. – : Protocole expérimental suivi lors du processus d'évaluation

3.7.3. Tâche de Pré-évaluation

1. **corpus de données :** Dans le but de valider l'approche présentée dans le chapitre 3, il est nécessaire d'effectuer une série d'expérimentations sur trois corpus de données.

— **WikiContext :** Afin de fournir des données riches et authentiques à la fois pour notre application Web et nos expérimentations, nous avons préparé manuellement une base de textes en anglais composée de 30 classes de contextes avec 20 articles Wikipédia pour chacune, ce qui constitue un ensemble de 600 textes. Cette collection va nous aider à l'expérimentation et le test de notre solution d'extraction de contexte. Ce choix est motivé par le fait que la collection préparée est basée sur des textes intégraux de Wikipédia qui parlent de plusieurs topics ce qui nous permettent d'avoir une évaluation efficace. Afin de valider ce corpus, nous avons suivi les étapes suivantes :
  - Récupération des données de Wikipédia en utilisant la bibliothèque python Scrapy 2.
  - Filtrage des textes récupérés (par exemple élimination des textes dupliqués).
  - Annotation manuelle du corpus par un titre ainsi que des mots-clés. Chaque texte a été annoté par 5 annotateurs pour valider ce corpus.
  - Classer les textes en 30 catégories différentes, couvrant ainsi 30 sujets. Chaque sujet contient 20 textes ce qui donne 600 documents en total.
  - Afin d'augmenter la taille de notre corpus WikiContext et pour améliorer la précision de notre approche, nous avons appliqué des méthodes d'augmentation des données. Pour ce faire, nous avons utilisé le framework « TextAttack » qui permet, à partir de chaque document, de générer jusqu'à 12 textes supplémentaires en effectuant certaines transformations sur le texte original telles que la reformulation ou le remplacement de certains mots par leurs synonymes. La figure 3.13 présente notre base qui contient 7592 documents, après avoir effectué l'augmentation des données.

    ⇒ **WikiContext** est utilisé pour effectuer une évaluation quantitative de trois processus : l'extraction de mots-clés, l'extraction de de contexte et le stockage des contextes ainsi que les documents associés.

FIGURE 3.13. – : Notre collection de données WikiContext

— **New York Times :** New York Times contient plus de 1,8 million d'articles écrits et publiés par le New York Times entre 1987 et 2007. Les métadonnées des articles sont fournies par la salle de rédaction du New York Times, le service d'indexation du New York Times et le personnel de production en ligne de « nytimes.com ».
— **BBC News :** Cette collection de données est composée de 2225 documents provenant du site Web d'actualités de la BBC correspondant à des articles dans cinq domaines d'actualité entre 2004 et 2005.

⇒ **NewYorkTimes** et **BBC News** sont utilisées pour évaluer qualitativement la phase d'extraction de contexte.

2. **Choix des métriques d'évaluation :** Dans ce travail, nous avons opté pour les deux métriques d'évaluation suivantes :
— **F1-score :** Afin d'évaluer quantitativement les performances de notre système, nous adoptons la métrique d'évaluation officielle, qui est basée sur le score F1 score. Cette métrique peut être interprétée comme une moyenne pondérée de la *précision* et du *rappel*.

La *précision* mesure la capacité du système à rejeter les mots-clés, le contexte ou les documents qui ne sont pas pertinents pour une requête. Nous définissons A comme le nombre d'attributions correctes, B le nombre d'attributions erronées. La *précision* est définie comme suit :

$$précision = \frac{A}{A+B} \tag{3.6}$$

Le *rappel* mesure la capacité du système à trouver tous les mots-clés, le

contexte ou les documents pertinents. Nous définissons A comme le nombre d'affectations correctes et C comme le nombre de positifs incorrectement prédits comme négatifs par le modèle. Le *rappel* est défini comme suit :

$$rappel = \frac{A}{A+C} \tag{3.7}$$

— **Mesure ROUGE :** Nous utilisons la meure ROUGE [63] pour évaluer qualitativement la performance du contexte généré par notre système. ROUGE, qui est une abréviation de Recall Oriented Understudy for Gisting Evaluation, comprend un ensemble de mesures utilisées pour l'évaluation de la synthèse automatique de texte et des traductions automatiques. Cette mesure permet d'évaluer la qualité du contexte en comptant les chevauchements unitaires entre le contexte généré par notre système et l'ensemble des contextes annotés par des humains dans les collections de test utilisées. Le contexte qui obtient le score ROUGE le plus élevé est considérée comme le plus similaire au contexte généré par l'homme. Dans notre évaluation, nous choisissons les mesures ROUGE-1, ROUGE-2, et ROUGE-L. ROUGE-1 mesure le chevauchement de l'unigramme dans le contexte de référence et le contexte candidat (généré par notre solution). ROUGE-2 fait référence au chevauchement des bigrammes entre le système et les contextes de référence (annotés par l'humain). ROUGE-L permet de calculer la plus longue séquence commune entre le contexte de référence et le contexte généré par notre modèle. Chaque phrase d'un contexte est considérée comme une séquence de mots. Par conséquence, deux contextes qui ont une séquence de mots commune plus longue se ressemblent d'avantage.

3.7.4. Tâche d'Evaluations : résultats obtenus et interprétations

Dans cette section, nous présentons les résultats d'évaluation des deux phases de notre approche d'extraction de contexte à partir d'un document texte non structuré : extraction des mots-clés et extraction de contexte. Finalement, nous évaluons le processus du stockage du contexte et du document associé dans une base de contextes « BdC ».

1. **Résultats de l'évaluation de la phase d'extraction des mots-clés :** La méthode d'évaluation de l'approche proposée d'extraction de mots-clés consiste en une évaluation quantitative sur notre collection de données « WikiContext ». En effet, pour chaque document, le résultat extrait de notre solution est évalué par rapport aux mots-clés de référence annotés dans notre corpus. Afin d'évaluer la faisabilité de l'approche proposée et de la valider, nous avons essayé de comparer les résultats de notre approche avec les résultats d'autres approches, telles que TF-IDF, Likey, TextRank, Rake, Yake et KeyBert en utilisant la mesure F1-score. Le tableau 3.1 ainsi que la figure 3.14 illustre les résultats de cette évaluation.

TABLEAU 3.1. – Performance de la tâche d'extraction des mots-clés sur la collection de données « WikiContexte »

Méthode utilisée pour l'extraction des mots-clés	**Précision(%)**	**Rappel(%)**	**F1–score**
TF–IDF	46.89	47.71	47.30
Likey	23.46	24.25	23.84
TextRank	15.48	20.94	17.80
Yake	27.17	27.25	27.21
Rake	37.25	37.25	37.26
KeyBERT	14.80	14.86	14.83
TF–IDF+Jiang and conrath (notre approche)	49.20	65.25	56.09
TF–IDF+USE (notre approche)	**65.80**	**73.20**	**69.30**

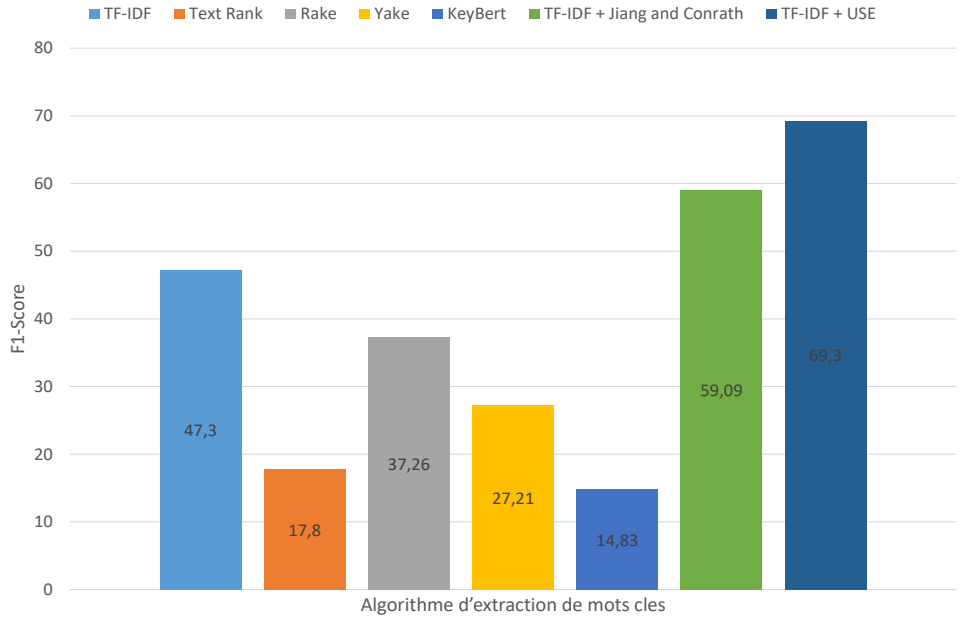


FIGURE 3.14. – : Performance de la tâche d'extraction des mots-clés sur la collection de données « WikiContexte »

Nous pouvons voir que notre approche d'extraction des mots-clés en utilisant Universal sentence Encoder surpasse significativement toutes les autres méthodes. En effet, le meilleur résultat en termes de F1-score est basé sur la combinaison de TF-IDF, FPgrowth et USE, et il atteint 69,30%, ce qui est supérieur à l'utilisation de la similarité de Jiang et conrath (56,09%) et l'utilisation seulement de TF-IDF (47.3%).

2. **Résultats de l'évaluation de la phase d'extraction de contexte :** Dans cette section, nous présentons une discussion, sur la phase d'extraction de contexte, sur la base des résultats de l'état de l'art et des résultats expérimentaux obtenus par notre méthode. Nous présentons dans ce qui suit, la préparation des corpus utilisés, les paramètres d'expérimentation du modèle utilisé pour l'extraction du contexte ainsi que l'évaluation de nos méthodes *extractive* et *générative*.
   — **Préparation des corpus :** Les corpus utilisés dans l'expérience pour évaluer la phase d'extraction de contexte sont WikiContext NewYorkTimes et BBC News. Ces corpus sont divisés en 80% pour l'entrainement et 20% pour le test. Pour WikiContext, sur les 7592 documents collectés dans 30 contextes, 6067 documents (80% de chaque contexte) ont été utilisés pour l'entraînement et les 1525 restants (20% de chaque contexte) pour les tests. Concernant BBC News, nous avons utilisé 1780 documents pour l'entrainement et 445 pour le test. Le corpus New York Times (1980-current) contient une quantité énorme de données (3 millions de documents), ce qui peut rendre le traitement et l'analyse de l'ensemble du corpus très coûteux en termes de temps et de ressources. Pour cette raison, il est courant de prendre des échantillons aléatoires de ce corpus pour effectuer des expérimentations. Ces échantillons peuvent être sélectionnés en fonction de divers critères tels que la période de temps, le contexte ou la taille du document. Cela permet de limiter la quantité de données à traiter et d'optimiser les ressources de calcul, tout en garantissant des résultats représentatifs de l'ensemble du corpus. Ainsi, dans notre approche nous avons pris un échantillon de 7500 documents dans des contextes différents pour effectuer l'entrainement et le test de notre modèle. De même, ce corpus a été divisé en 80 :20. Pour le prétraitement des différents corpus, nous avons organisé nos données dans un dataframe, en associant à chaque texte ses mots-clés et son contexte, ensuite nous avons procédé au nettoyage en éliminant tout caractère non-alphabétique ainsi que les mots sans importance (mots vides), obtenir la racine de chaque mot (stem) et supprimer les mots de longueur inférieure à 2.
   — **Paramètres des expériences :** Les valeurs des paramètres du modèle LSTM ont été choisies en se basant sur les expérimentations de l'état de l'art. En effet, les valeurs retenues sont considérées comme les meilleures selon les résultats de ces expérimentations. L'entrée de notre modèle LSTM est les vecteurs de mots, et la sortie est le vecteur de probabilité de prédiction. La fonction de perte est réduite par l'optimiseur « Adam » [20], et le modèle optimal est obtenu en actualisant constamment le poids itératif. Dans l'expérience, le nombre de nœuds de la couche cachée LSTM est fixé à 512, le taux d'apprentissage initial est fixé à 0,1. Ce taux ralenti considérablement l'apprentissage ce qui permettra au modèle de converger en douceur. La capacité de traitement d'échantillons (Batch size) est fixée à 256.
   — **Résultats d'évaluation :** La méthode d'évaluation de la solution d'extraction de contexte proposée consiste à évaluer *quantitativement* et *qualitativement* cette approche.

— **Evaluation quantitative :** Les contextes générés par nos deux solutions extractive et générative sont évalués par rapport aux contextes de référence annotés dans notre collection WikiContext afin de vérifier s'ils sont pertinents. Pour ce faire, nous avons utilisé F1-score pour effectuer une évaluation quantitative. Les résultats de cette évaluation sont présentés dans le tableau 3.2 : En se basant sur les données présentées dans le tableau 3.2, le meilleur résultat est donné par notre approche générative qui atteint 73%. Ce résultat est dû au fait que les méthodes génératives peuvent fournir des résultats plus précis que les méthodes extractives en prenant en compte le contexte sémantique et syntaxique des phrases. Contrairement aux méthodes extractives qui se concentrent uniquement sur les phrases qui contiennent les mots-clés, les méthodes génératives sont capables de générer une nouvelle phrase qui synthétise les informations pertinentes extraites de différentes parties du texte. En outre, les modèles LSTM sont capables de capturer les dépendances à long terme dans le texte, ce qui leur permet de fournir des résultats plus cohérents et précis que les modèles extractifs. Cela est particulièrement vrai lorsque le contexte est complexe et nécessite une compréhension approfondie de la structure du langage.

TABLEAU 3.2. – Performance de l'extraction du contexte en termes de F1-score sur la collection de données WikiContext

Approche utilisée pour l'extraction de contexte	**Méthode utilisée pour l'extraction des mots-clés**	**Méthode utilisée pour le calcul de la similarité**	**Précision**	**Rappel**	**F1–score**
Approche extractive	TF-IDF	Jiang and Conrath	49%	65%	56%
		USE	56%	70%	62%
Approche générative	TF-IDF	Jiang and Conrath	65%	73%	68%
		USE	**68%**	**78%**	**73%**

— **Evaluation qualitative :** Afin de comparer et d'analyser qualitativement les résultats de notre approche avec les méthodes existantes d'extraction de contexte, nous avons réalisé nos expériences sur les collections de données du New York Times et de la BBC News. Les tableaux 3.3 et 3.4 présentent les résultats d'extraction de contexte fournis par nos deux approches (extractive et générative) comparés aux résultats obtenus par (SAJID, JAN et SHAH 2017a) et (N. GUO, Yuan HE, C. YAN et al. 2016a), pour 5 documents sélectionnés aléatoirement à partir de New York Times et de

BBC News.

TABLEAU 3.3. – Résultats fournis par nos approches comparés aux résultats obtenus par (SAJID, JAN et SHAH 2017a) sur la collection de données New York Times

Titre fourni par New York Times	**Contexte généré par (SAJID, JAN et SHAH 2017a)**	**Contexte généré par notre approche extractive**	**Contexte généré par notre approche générative**
2015 Was Hottest Year in Historical Record, Scientists Say Record Year Heat	Record Year Heat	The Hottest Year in the Historical Record	**2015 Was the Hottest Year in Historical Record**
Apple Settles Legal Dispute With Nokia	Apple Nokia Company	Apple and Nokia Settled a Legal Dispute	**Apple Settles a Legal Dispute**
Atlantic Hurricane Season Is Expected to Be Busy	Storm Hurricane Season Forecast	Atlantic Hurricane Season	**Atlantic Hurricane Season Is Busy**
Britain Accuses Ghana Lawmakers of Visa Fraud	Visa Ghana Parliament Britain	British authorities accused Ghana's Parliament of Visa Fraud	**Britain Accuses Ghana's Parliament of Visa Fraud**
Trump Will Withdraw U-.S. From Paris Climate Agreement	Trump Agreement Paris President	The Paris Climate Accord	**Paris Climate Agreement**

TABLEAU 3.4. – Résultats fournis par nos approches comparés aux résultats obtenus par (N. GUO, Yuan HE, C. YAN et al. 2016a) sur la collection de données BBC News

Titre fourni par BBC News	**Contexte généré par (N. GUO, Yuan HE, C. YAN et al. 2016a)**	**Contexte généré par notre approche extractive**	**Contexte généré par notre approche générative**
India students caught 'cheating' in exams in Bihar	Society	Cheating in exams in the Indian state	**Cheating in exams in Bihar**
Glasgow School of Art : 'One of the great buildings'	Art	Glasgow school of art	**The great building of Glasgow School of Art**
Martin Guptill hits highest World Cup score in New Zealand victory	Sport	Martin Guptill made the highest score in World Cup	**Martin Guptill hits the highest score in New Zealand victory**
Possible fatty acid detected on Mars	Science	A fatty acid discovered on Mars	**fatty acid detected on Mars**
US winemakers reject arsenic claim	Society	California winemakers rejecting claims	**California winemakers rejecting arsenic claim**

Le tableau 3.3 montre que nos résultats d'extraction de contexte sont riches en contenu et faciles à comprendre par rapport aux résultats obtenus par (SAJID, JAN et SHAH 2017a).. Les mots associés à chaque contexte constituent une phrase cohérence qui transmet un titre significatif du document. Par exemple, si nous considérons le titre « Apple Settles Legal Dispute with Nokia" fourni par la collection de données New York Times, nous remarquons que nos deux modèles développent un résultat beaucoup plus compréhensible « Apple and Nokia Settled a Legal Dispute" et Apple Settles a Legal Dispute" qui couvrent le sujet du document. Cependant, le contexte obtenu par (SAJID, JAN et SHAH 2017a). (Apple Nokia Company) est dépourvu de toute signification.

Concernant les résultats obtenus par (N. GUO, Yuan HE, C. YAN et al. 2016a) sur de la collection de données BBC News (tableau 3.4), le document « India students caught 'cheating' in examins in Bihar » a pour contexte général « Société ». Cependant, le contenu de ce document concerne l'éducation en Inde et plus précisément la fraude aux examens en Bihar. De plus, les deux documents « India students caught 'cheating' in exams in Bihar » et « US winemakers reject arsenic claim » sont classés par le modèle proposé par (N. GUO, Yuan HE, C. YAN et al. 2016a) dans le même contexte « Société ». Le contexte identifié n'est pas assez précis pour caractériser au mieux le contexte réel. En effet, le premier document

concerne l'éducation en Inde, tandis que le second porte sur la santé et plus précisément sur le rejet d'arsenic dans l'environnement. Par rapport au contexte original du document, nos modèles sont capables d'extraire un contexte plus précis que (N. GUO, Yuan HE, C. YAN et al. 2016a).

La mesure rouge (ROUGE : Recall-Oriented Understudy for Gisting Evaluation) a été développée spécifiquement pour évaluer la qualité des textes en comparant le texte généré par le système avec une référence de texte humain. Elle prend en compte la similarité entre les phrases, les mots et les n-grammes présents dans les deux textes, et permet ainsi d'évaluer la qualité du texte généré. En utilisant la mesure rouge, on peut donc comparer les performances de l'approche extractive et de l'approche générative en termes de qualité du contexte produit. En effet, nous avons calculé les mesures ROUGE1 (R-1), ROUGE2 (R-2) et ROUGE-L (R-L) pour chaque contexte généré à partir d'un document de la collection de données New York Times. Puis nous avons calculé la moyenne de ces mesures pour avoir une évaluation sur toute la collection de données. Le tableau 3.5 résume les résultats obtenus sur les 5 textes sélectionnés aléatoirement et présentés dans 3.4.

TABLEAU 3.5. – Performance des contextes générés en termes de mesure ROUGE sur toute la collection de données New York Times

Algorithme utilise pour l'extraction du contexte	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
label extracted by (SAJID, JAN et SHAH 2017a)	'f' : 0.3, 'p' : 0.5, 'r' : 0.22	'f' : 0.18, 'p' : 0.33, 'r' : 0.12	'f' : 0.32, 'p' : 0.60, 'r' : 0.22
Approche extractive	'f' : 0.48, 'p' : 0.75, 'r' : 0.35	'f' : 0.21, 'p' : 0.55, 'r' : 0.13	'f' : 0.44, 'p' : 0.70, 'r' : 0.33
Approche générative	**'f' : 0.63, 'p' : 0.65, 'r' : 0.62**	**'f' : 0.27, 'p' : 0.58, 'r' : 0.18**	**'f' : 0.52, 'p' : 0.55, 'r' : 0.50**

En se basant sur les résultats présentés dans les deux tableaux 3.5 et 3.6, nous pouvons voir que nos approches extractive et générative surpassent la méthode proposée par (SAJID, JAN et SHAH 2017a). De plus, les valeurs ROUGE-1, ROUGE-2 et ROUGE-L de l'approche productive sont bien meilleures que les résultats de l'approche extractive et de (SAJID, JAN et SHAH 2017a) sur la collection de données New York Times. En conclusion, notre évaluation montre que notre approche atteint une bonne performance tout en améliorant la qualité du contexte extrait par rapport aux autres systèmes.

3. **Résultats de l'évaluation de la phase de stockage du contexte et du document associé dans une base de contextes « BdC » :** Dans cette section, nous présentons

TABLEAU 3.6. – Performance des contextes générés en termes de mesure ROUGE sur les cinq textes sélectionnés aléatoirement et présentés dans le tableau 3.4

Titre fourni par New York Times	**Mesure ROUGE du contexte généré par (SAJID, JAN et SHAH 2017a)**	**Mesure ROUGE du contexte généré par l'approche extractive**	**Mesure ROUGE du contexte généré par l'approche génerativee**
2015 Was Hottest Year in Historical Record, Scientists Say Record Year Heat	**R-1**'f' :0.42, 'p' :1.00, 'r' :0.27 **R-2**'f' :0.30, 'p' :1.00, 'r' :0.18 **R-L**'f' :0.42, 'p' :1.00, 'r' :0.27	**R-1**'f' :0.62, 'p' :1.00, 'r' :0.45 **R-2**'f' :0.39, 'p' :0.75, 'r' :0.27 **R-L**'f' :0.62, 'p' :1.00, 'r' :0.45	**R-1'f' :0.73, 'p' :0.87, 'r' :0.63** **R-2'f' :0.30, 'p' :0.37, 'r' :0.27** **R-L'f' :0.73, 'p' :0.87, 'r' :0.63**
Apple Settles Legal Dispute With Nokia	**R-1**'f' :0.44, 'p' :0.66, 'r' :0.33 **R-2**'f' :0.00, 'p' :0.00, 'r' :0.00 **R-L**'f' :0.44, 'p' :0.66, 'r' :0.33	**R-1**'f' :0.73, 'p' :0.31, 'r' :0.73 **R-2**'f' :0.87, 'p' :0.37, 'r' :0.87 **R-L**'f' :0.63, 'p' :0.27, 'r' :0.63	**R-1'f' : 0.73, 'p' :0.44, 'r' :0.73** **R-2'f' :0.80, 'p' :0.50, 'r' :0.80** **R-L'f' :0.66, 'p' :0.40, 'r' :0.66**
Atlantic Hurricane Season Is Expected to Be Busy	**R-1**'f' :0.33, 'p' :0.50, 'r' :0.25 **R-2**'f' :0.19, 'p' :0.33, 'r' :0.14 **R-L**'f' :0.33, 'p' :0.50, 'r' :0.25	**R-1** f' :0.54, 'p' :1.00, 'r' :0.37 **R-2**'f' :0.44, 'p' :1.00, 'r' :0.28 **R-L**'f' :0.54, 'p' :1.00, 'r' :0.37	**R-1'f' :0.76, 'p' :1.00, 'r' :0.62** **R-2'f' :0.54, 'p' :0.75, 'r' :0.42** **R-L'f' :0.76, 'p' :1.00, 'r' :0.62**
Britain Accuses Ghana Lawmakers of Visa Fraud	**R-1**'f' :0.54, 'p' :0.75, 'r' :0.42 **R-2**'f' :0.00, 'p' :0.00, 'r' :0.00 **R-L**'f' :0.18, 'p' :0.25, 'r' :0.14	**R-1** f' :0.39, 'p' :0.37, 'r' :0.42 **R-2**'f' :0.30, 'p' :0.28, 'r' :0.33 **R-L**'f' :0.39, 'p' :0.37, 'r' :0.42	**R-1'f' :0.71, 'p' :0.71, 'r' :0.71** **R-2'f' :0.49, 'p' :0.50, 'r' :0.50** **R-L'f' :0.71, 'p' :0.71, 'r' :0.71**
Trump Will Withdraw U-.S. From Paris Climate Agreement	**R-1**'f' :0.46, 'p' :0.75, 'r' :0.33 **R-2**'f' :0.00, 'p' :0.00, 'r' :0.00 **R-L**'f' :0.30, 'p' :0.50, 'r' :0.22	**R-1** f' :0.30, 'p' :0.50, 'r' :0.22 **R-2**'f' :0.18, 'p' :0.33, 'r' :0.12 **R-L**'f' :0.30, 'p' :0.50, 'r' :0.22	**R-1'f' :0.49, 'p' :1.00, 'r' :0.33** **R-2'f' :0.39, 'p' :1.00, 'r' :0.25** **R-L'f' :0.49, 'p' :1.00, 'r' :0.33**

la préparation du corpus utilisé, les paramètres d'expérimentation du modèle BERT utilisé pour le calcul de similarité entre deux contextes ainsi que les résultats obtenus.

— **Préparation du corpus :** Pour étudier l'efficacité du calcul de similarité entre deux contextes, nous avons utilisé notre corpus « WikiContext ». Pour cette tâche, ce corpus est divisé en 6067 documents pour l'apprentissage et le reste pour le test. Nous avons effectué de même le prétraitement des documents avant la phase d'entrainement.

— **Paramètres des expériences :** Nous effectuons un pré-entraînement supplémentaire avec BERT avec un « batch size » de 32, une longueur de séquence maximale de 128. Concernant la phase de « fine-tuning » de notre modèle BERT avec notre base « WikiContext », nous utilisons Adam comme fonction de perte et un taux d'apprentissage de 0.1 dans la phase d'entrainement. Nous fixons le nombre maximal d'époques à 20 et sauvegardons le meilleur modèle sur l'ensemble de validation pour le test.

— **Résultats de l'évaluation :** Dans cette étape, nous avons transformé chaque contexte ou mots-clés en un vecteur numérique de dimension fixe à l'aide des modèles prédéfinis de différentes librairies comme BERT, Universal sentence encoder (USE) et Doc2vec. Les résultats de l'expérimentation réalisée sur notre dataset « WikiContext » ont présentés dans le tableau 3.7.

Dans cette évaluation, nous avons tout d'abord nettoyé le document en éliminant tout caractère non-alphabétique ainsi que les mots vides. Ensuite, nous avons passé à la vectorisation du texte qui représente l'une des étapes cruciales après le nettoyage de données. En effet, celle-ci permet à la machine de comprendre un texte puisqu'elle n'est pas capable de s'entraîner que sur des valeurs numériques. Dans cette étape, nous avons transformé chaque contexte ou mots-clés en un vecteur numérique de dimension fixe à l'aide des modèles prédéfinis de différentes librairies comme le BERT, Universal sentence encoder (USE) et Doc2vec. Les résultats d'expérimentation réalisée sur notre dataset WikiContext sont présentés dans le tableau 3.7.

TABLEAU 3.7. – Évaluation des modèles de vectorisation

Mesure	Doc2vec(%)	USE(%)	BERT(%)
Précision	47	71	82
Rappel	52	87	96
F1-score	49	78	88

Nous observons que le meilleur résultat (88%) est donné par Bert. Nous avons donc choisi ce modèle comme un transformateur de texte en vecteur.

Afin d'évaluer le calcul de similarité entre deux contextes, Nous avons testé trois techniques similarité cosinus, LSTM et BERT. Les résultats de cette évaluation sont présentés dans le tableau 3.8.

TABLEAU 3.8. – Résultats obtenus par le processus de calcul de similarité entre deux contextes

Modèle utilisé	Précision	Rappel	F1-score
Similarité Cosinus	25%	23%	24%
LSTM	40%	36%	38%
BERT	83%	81%	82%

Selon le tableau 3.8, Nous concluons que le modèle LSTM n'est pas une bonne solution vu qu'il ne prend pas en compte le contexte du texte dans son apprentissage, nous avons donc eu recours aux « Transformers » (BERT dans notre cas). Nous pouvons voir que le modèle BERT pré-entraîné surpasse les autres modèles avec un score F1 de 82%.

L'application de l'algorithme d'optimisation itérative de premier ordre« Gradient Descent", nous a permis d'utiliser le seuil de 0,67 comme score minimum indiquant qu'un document est pertinent pour un contexte. Ce score aide notre modèle à bien distinguer entre un contexte connu et un nouveau contexte inconnu pour aboutir à une bonne précision.

3.7.5. Tâche de Post-évaluation : « EasyContext » une application pour l'extraction du contexte d'un document textuel

Afin de prouver l'intérêt pratique de notre approche, une application Web, nommée « EASYContext », a été développée mettant en œuvre notre méthode d'extraction de contexte. Notre application propose un nouveau paradigme pour l'extraction du contexte avec l'ajout du contexte qui le décrit avec précision. Dans cette section, nous présentons l'architecture de notre application, l'environnement expérimental ainsi que quelques interfaces et ses différentes pages.

1. **Architecture de l'application :** Pour la réalisation de notre application « EASYContext », l'architecture aux trois tiers a été choisie. Cette architecture est composée du tiers Client qui représente un navigateur web. Le tiers Serveur Web qui se compose de deux composants : le Front-end, y compris les formulaires d'inscriptions de connexions et de demande d'extraction de contexte et le Back-end reliant notre application à la base de données de contexte (Bdc) et aux algorithmes d'extraction de contexte. Finalement, le tiers Serveur Base de données contenant la base des contextes (Bdc).
2. **Environnement Expérimental :** Concernant la partie back-end de notre application, nous avons opté à l'utilisation du Framework Python Django. L'utilisation

de python comme langage de développement web nous a poussé à bien choisir Django avec ses librairies et modules qui facilitent l'introduction des algorithmes de texte mining. Pour la partie front-end nous avons opté à utiliser le Framework Angular avec sa version 9. Ce Framework nous a permet avec sa richesse de fonctionnalité de développer cette partie de notre application tout en faisant le contrôle des accès et de bien présenter l'ergonomie des interfaces utilisateurs. La communication entre la partie front-end et back-end se fait à travers les requêtes HTTP et le format d'échange de données JSON.

3. **Description de l'application :** Dans cette section, nous présentons les différentes interfaces de notre application Web « EASYContext ». La page d'accueil permet l'accès aux différentes fonctionnalités de notre application à savoir l'accès aux espaces d'authentification, la création d'un compte ainsi que l'interface d'extraction de contexte. La figure 3.15 représente l'interface dédiée à l'extraction de contexte. L'utilisateur a le choix de saisir un texte directement ou d'importer un fichier. Il clique par la suite, sur le bouton "Extract Context" pour afficher le contexte relatif à ce dernier ainsi que les mots-clés associés.

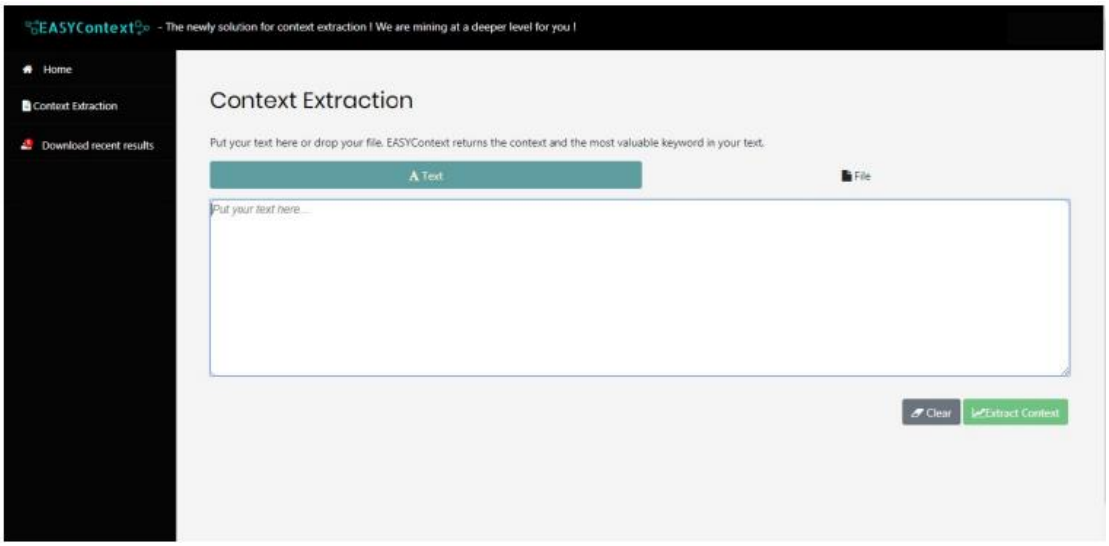


FIGURE 3.15. – Page d'extraction de contexte

La figure 3.16 présente un exemple d'extraction de contexte et de ses mots-clés relatifs à un texte intitulé « Natural language processing ».

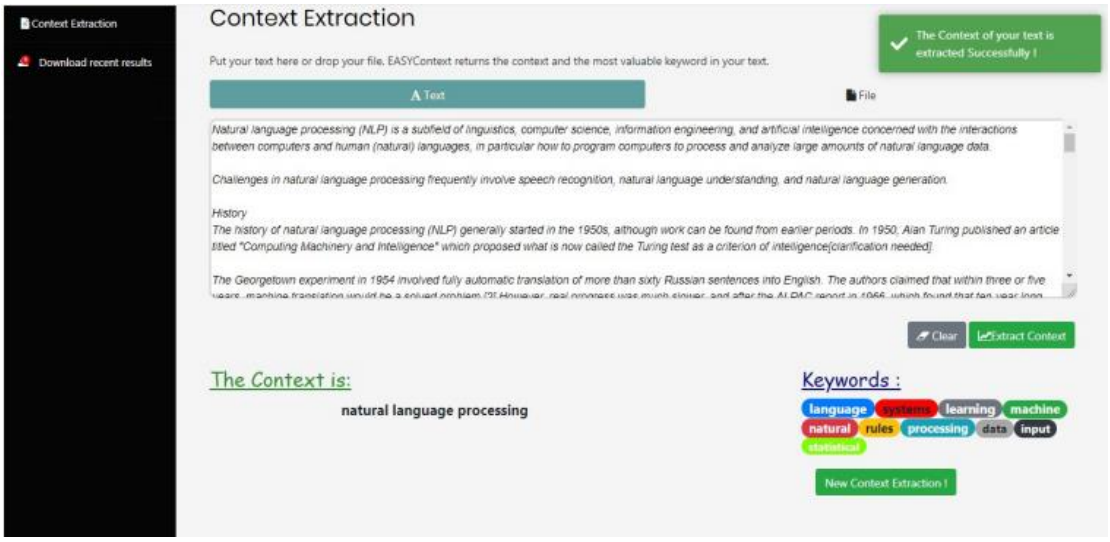


FIGURE 3.16. – Exemple d'extraction de contexte

La figure 3.17 montre une page qui offre aux utilisateurs de « EASYContext » de télécharger leurs résultats d'extraction de contexte des documents récents.

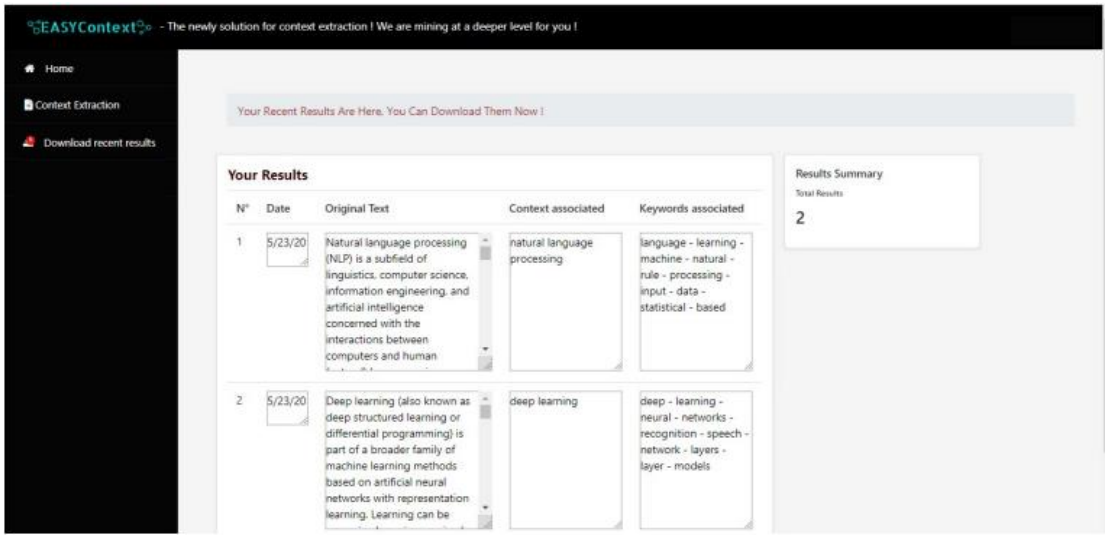


FIGURE 3.17. – Page de téléchargement des résultats récents

3.8. Conclusion

Dans ce chapitre, nous avons présenté une nouvelle méthode qui extrait automatiquement le contexte d'un document textuel donné. Cette méthode générative est basée sur une nouvelle approche d'extraction de mots-clés ainsi que sur la modélisation de langage en utilisant un LSTM. En utilisant ces techniques, la méthode permet

de produire une phrase cohérente qui représente le contexte du document de manière précise et pertinente par rapport aux systèmes d'extraction de contexte existants. Les résultats montrent que les contextes extraits sont quantitativement et qualitativement beaucoup plus précis que ceux identifiés par ces autres systèmes.

L'ajout d'une contextualisation plus précise des documents pourraient être intégrés à de nombreuses applications. En particulier, dans le cas des moteurs de recherche, la classification des documents en fonction d'un contexte précis doit améliorer l'efficacité de la recherche. Ainsi l'utilisation de contextes pertinents et clairement identifiés, et spécifiés devrait permettre d'obtenir des systèmes de recherche d'information plus performants et plus robustes.

Dans le chapitre suivant, nous proposons une nouvelle approche de classification des relations selon le type et le contexte en se basant sur l'approche d'extraction de contexte proposée.

4. Une méthode de classification des relations selon le type et le contexte

Sommaire

4.1. Introduction

Le chapitre précédent a présenté une nouvelle méthode d'extraction du ***« contexte »*** d'un document textuel, le ***« contexte »*** constituant une nouvelle métadonnée associée

au document permettant de cerner de façon précise et pertinente son contenu. Ce chapitre présente notre seconde contribution, à savoir une méthode de classification des relations d'un document textuel non structuré selon le *« type de relation »* et surtout selon le ***« contexte »*** associé à ce document. Cette méthode utilise le contexte extrait du document par la méthode présentée dans le chapitre précédent. La classification des relations d'un document selon son contexte conduit ainsi une certaine *« contextualisation »* des relations de ce document.

Dans ce chapitre, nous commençons dans la section 4.2 par rappeler la notion de relation. Dans la section 4.3 nous présentons les grandes étapes de la méthode proposée de classification des relations selon le type et le contexte. Ensuite, dans les sections 4.4, 4.5 et 4.6 nous détaillons chacune des grandes étapes de cette méthode, ainsi que les composants spécifiques les implémentant. Dans la section 4.7, nous procédons à diverses expérimentations permettant d'évaluer les résultats des composants principaux de la méthode, en les comparant avec les résultats obtenus par d'autres solutions existantes de l'état de l'art. Enfin, dans la section 4.8, nous présentons un exemple d'application de cette méthode relative à la visualisation automatique des données statistiques extraites d'un texte.

4.2. Rappel de la notion de relation

Dans la théorie des ensembles, une relation est un moyen de montrer une connexion ou une relation entre deux ensembles quelconques. Dans le domaine de traitement automatique de langue naturelle, une relation est une association significative entre deux ou plusieurs entités. Une entité est un mot ou un groupe nominal qui se réfère à un objet ou un concept tel que livre, compte bancaire, un lieu, une personne, etc. Une relation entre deux entités correspond à une relation *binaire*.

Exemple 1 : Soit la phrase « Donald Trump est membre du parti républicain » véhicule une *relation* « est membre de » entre les entités « Donald Trump » et « parti républicain ».

Une *relation sémantique* entre plus de deux entités correspond à un *événement*.

Exemple 2 : Soit la phrase « Joe Biden a été élu par l'électorat électeurs à la présidence des États-Unis le 20 décembre 2021 ». Dans cette phrase, nous avons la *relation ternaire* ou *l'événement* « est élu », entre les entités « Joe Biden », « Président des États-Unis » et « 20 décembre 2021 ».

La classification des relations consiste à classer une relation qui existe entre entités nominales dans un texte. Dans cette thèse, nous nous concentrons sur la classification des relations sémantiques entre deux paires d'entités (relations binaires). Étant donné une phrase *« P »* avec une paire annotée d'entités e1 et e2, la tâche est de classer laquelle des neuf types de relations suivantes : Cause-Effet, Instrument-Agence, Produit-Producteur, Contenu-Conteneur, Entité-Origine, Entité-Destination, Composant-Global, Membre-Collection, Sujet-Message, ou Autre si elle n'appartient à aucune des neuf relations annotées. Cette classification s'appelle une classification par *« type »*.

Comme mentionné dans le premier chapitre, la classification selon le *type* tient en compte uniquement de l'aspect *syntaxique* du texte. Elles négligent un aspect important qui est la *sémantique*, et en particulier le ***contexte*** associé au document qui sera extrait selon la méthode présentée dans le chapitre précédent. Aussi l'objectif de notre travail est d'annoter la relation non seulement par le *type* mais aussi par le ***contexte*** du document dont elle fait partie. Une relation sera définie comme suit :

Définition 3 *Une relation binaire « R » et une association significative entre deux entités. Elle est définie sous la forme d'un coupe c = (e_1, e_2) où e_1 et e_2 sont deux entités liées par un prédicat « P ». Le prédicat « P » représente la partie de la phrase se situant entre les deux entités. Cette relation est annotée par un « type » et un contexte « Ctx ». Ainsi, une relation est définie comme suit :*

$$R = \{(e_1, e_2), P, type, Ctx\} \tag{4.1}$$

4.3. Présentation générale de la méthode de classification des relations selon le type et le contexte

Dans cette section, nous présentons une méthode de classification des relations basée sur le type et le contexte. Comme l'illustre la figure 4.1, cette méthode comprend quatre étapes principales : (1) l'*extraction du contexte du contenu textuel*, (2) l'*identification des relations*, (3) l'*annotation des relations par le contexte*, et enfin, (4) la *classification des relations identifiées par le type*. La méthode de classification des relations procède ainsi :

- Etape 1 : *Extraction de contexte*. On commence par extraire le contexte du document en utilisant l'approche générative présentée dans le chapitre 3 précédent. Cette approche consiste, tout d'abord à identifier et extraire les différents mots-clés du document, puis générer à partir de ces mots-clés une phrase cohérente représentant l'idée principale du document (par un modèle LSTM).
- Etape 2 : *Identification des relations*. Pour cette étape qui sera développé dans la section suivante, une solution existante est utilisée (Stanford OpenIE).
- Etape 3 : *Annotation (et classification) des relations par le contexte*. L'extraction initiale est exécutée avec des critères d'extraction minimaux pour identifier toutes les relations. Il s'agira de sélectionner les relations significatives parmi toutes les relations extraites initialement. Pour ce faire, un filtrage sera fait pour éliminer les relations non valides afin d'annoter, et par là même classifier, uniquement les relations qui sont significatives pour le contexte retenu.
- Etape 4 : *Classification des relations selon le type*. Finalement on procède à la classification des relations en fonction de leur type. Cette classification utilisera un modèle neuronal de type Att-RCNN. A l'issu de ce processus les relations sont classées selon le contexte et leur type.

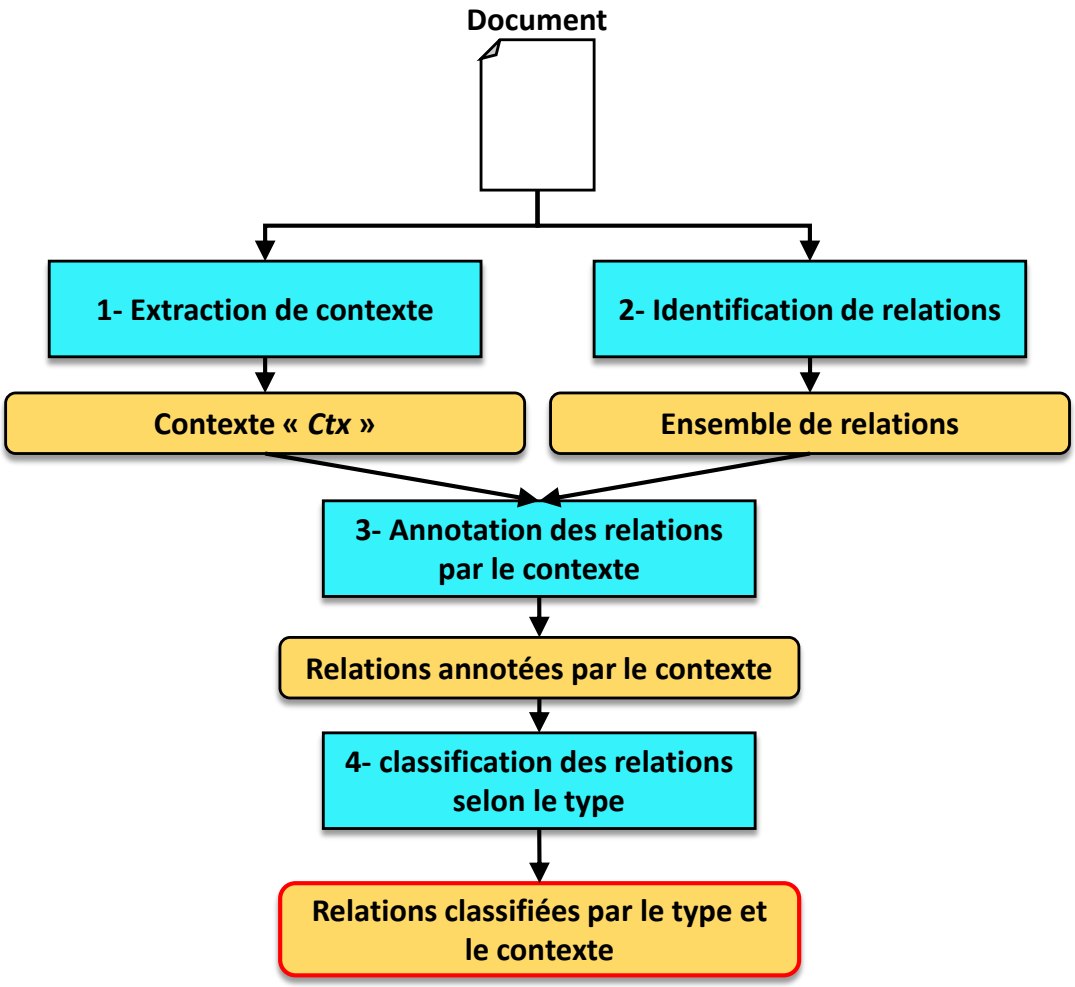


FIGURE 4.1. – Présentation générale de la méthode de classification des relations selon le type et le contexte

Dans les sections suivantes nous présentons en détail les différentes étapes de la méthode de classification de relations proposée, concernant notamment les composants logiciels les mettant en œuvre. Nous renvoyons le lecteur au chapitre précédent pour plus de détails sur la première étape d'*extraction de contexte du document*.

4.4. Etape d'identification des relations

L'extraction d'information (EI) transforme l'information non structurée exprimée dans un texte en langage naturel en une représentation structurée (Jurafsky et Martin, 2009) sous la forme de tuples relationnels composés d'un ensemble d'entités (Ei) et d'une phrase indiquant une relation sémantique P entre eux : (E1 ; P ; E2). Les approches traditionnelles de l'EI se concentrent sur la réponse à des requêtes bien définies sur un ensemble prédéfini de relations, sur de petits corpus homogènes. Pour

ce faire, elles prennent en entrée la relation cible ainsi que des patrons d'extraction ou des patrons appris à partir d'exemples d'entraînements étiquetés à la main. Par conséquent, le passage à un nouveau domaine nécessite de nommer les relations cibles et de définir manuellement de nouvelles règles d'extraction ou d'annoter de nouvelles données d'entraînement. Ainsi, ces systèmes reposent sur une implication humaine importante.

Afin de réduire l'effort manuel requis par les approches EI, (Oren ETZIONI, BANKO, SODERLAND et al. 2008b) ont introduit un nouveau paradigme d'extraction : Open IE. Contrairement aux méthodes d'IE traditionnelles, l'Open IE ne se limite pas à un petit ensemble de relations cibles connues à l'avance, mais extrait plutôt tous les types de relations. De cette façon, elle facilite la découverte indépendante du domaine des relations extraites du texte et s'adapte à de grands corpus hétérogènes tels que le Web.

Comme notre travail se focalise sur un contexte libre et indépendant d'un domaine particulier, il est indispensable d'utiliser une méthode basée sur l'*Open IE* afin d'extraire tous les types de relations présentes dans un texte donné. Plusieurs approches sur la tâche de l'IE ont été proposées depuis son introduction de TextRunner par Banko et al. (Oren ETZIONI, BANKO, SODERLAND et al. 2008b). Dans cette thèse, nous avons testé quatre approches OpenIE à savoir, Reverb (Oren ETZIONI, FADER, CHRISTENSEN et al. 2011), Ollie (SCHMITZ, SODERLAND, BART et al. 2012), ClausIE(DEL CORRO et GEMULLA 2013) et Stanford OpenIE (ANGELI, PREMKUMAR et MANNING 2015) afin d'identifier l'ensemble des relations présentent dans un document.

4.4.1. Aperçu des différentes approches proposées pour résoudre la tâche d'identification de relations basées sur l'Open IE

TextRunner (Oren ETZIONI, BANKO, SODERLAND et al. 2008b) a introduit le paradigme de l'EI OpenIE, une approche d'apprentissage auto-supervisée. Tout d'abord, à partir d'un petit échantillon de phrases de la Penn Treebank (MARCINKIEWICZ 1994), l'apprenant applique un analyseur syntaxique de dépendances pour identifier et étiqueter de manière heuristique un ensemble d'extractions comme les phrases POS et les phrases nominales NP. Ces données sont ensuite utilisées en entrée d'un classificateur Naive Bayes qui apprend un modèle de relations de confiance à l'aide de caractéristiques non lexicalisées de POS et de NP. La nature auto-supervisée atténue le besoin de données d'entraînements étiquetées manuellement, et les caractéristiques non-lexicalisées aident à s'adapter aux multitudes de relations trouvées sur le Web. Ce système a été évalué en utilisant un corpus de test de 9 millions de documents Web, obtenant 7,8 millions de tuples. Des humains ont évalué un ensemble de 400 tuples sélectionnés au hasard, dont 80,4% ont été considérés comme corrects.

Fader et al. (FADER, SODERLAND et Oren ETZIONI 2011a) décrivent le système ReVerb, le premier système qui utilise la stratégie heuristique, lançant une deuxième génération d'extracteurs Open IE. La conception du système est basée sur une heuristique simple qui identifie les verbes exprimant des relations en anglais. Il reçoit en

entrée des phrases étiquetées et découpées par POS. L'algorithme identifie d'abord les relations et obtient ensuite leurs arguments. Comme la méthode obtient un rappel élevé, mais une précision faible, ils établissent un seuil pour attribuer un score de confiance à chaque extraction. Dans (FADER, SODERLAND et Oren ETZIONI 2011a), ces auteurs rapportent que ReVerb a atteint une AUC (area under precision-recall curve) deux fois plus grande que celle de TextRunner.

Dans le but d'améliorer l'OpenIE en élargissant la portée syntaxique des phrases qui expriment des relations, Mausam et al. (SCHMITZ, SODERLAND, BART et al. 2012) présentent le système OLLIE, qui introduit la stratégie hybride. Le système est basé sur un apprentissage par bootstrap de patrons basés sur des chemins d'analyse de dépendance. OLLIE applique un ensemble de n-uplets de haute précision à partir de son système prédécesseur ReVerb, pour bootstrapper un vaste ensemble d'entraînements sur lequel il apprend un ensemble de patrons d'extraction à l'aide de l'analyses de dépendance. OLLIE comprend une étape d'analyse du contexte dans laquelle les informations contextuelles de la phrase d'entrée sont analysées pour étendre la représentation en sortie en ajoutant si nécessaire des modificateurs d'attribution et en augmentant ainsi la précision du système. Les auteurs rapportent que le système obtient une surface sous le rendement de précision 1,9 fois plus grande, si celle-ci est comparée à celles du système ReVerb.

L'extracteur ClausIE, présenté dans (DEL CORRO et GEMULLA 2013), utilise aussi une stratégie hybride. Il sépare la détection des clauses et des types de clauses de la formation des tuples de relations. Une clause est définie comme une partie d'une phrase qui exprime un élément d'information cohérent. En combinant l'analyse syntaxique des dépendances et la connaissance des propriétés des verbes, elle utilise une méthode de classification pour identifier les arguments (entités) d'une relation. Les auteurs rapportent des tests effectués sur trois ensembles de données différents. ClausIE a obtenu une précision et un rappel supérieurs à ceux des autres extracteurs. Plus précisément, ClausIE a produit 2,5 à 3,5 fois plus d'extractions correctes qu'OLLIE.

Stanford Open IE (ANGELI, PREMKUMAR et MANNING 2015) est une approche dans laquelle un classifieur est appris pour diviser une phrase en un ensemble d'énoncés plus logiques impliqués implicitement en parcourant de manière récursive son arbre de dépendance. Afin d'accroître l'utilité des propositions extraites pour les applications en aval, chaque clause indépendante est ensuite raccourcie au maximum en effectuant une inférence de logique naturelle.

4.4.2. Discussion

Malgré la simplicité et la rapidité d'exécution du système ReVerb, il ne couvre que les relations qui se situent entre les entités, dès que la relation se positionne avant la première entité ou après la deuxième entité dans une phrase ReVerb la néglige. En effet, au total, 65% des relations incorrectement extraites par le ReVerb représentent des cas dans lesquels l'identification de relation a échoué. Les deux problèmes les plus importants pour cet échec sont : les extractions incohérentes et les extractions non informatives.

Les deux outils OLLIE et ClausIE, ont pour but d'améliorer le problème principal des systèmes précédents : l'identification des entités des relations. Del Corro et Gemulla (DEL CORRO et GEMULLA 2013) affirment que la plupart des erreurs d'extraction sont dues à deux problèmes : les défaillances de l'analyseur syntaxique et l'incapacité à exprimer les relations dans les textes en relations binaires, c'est-à-dire que si les relations dans le texte n'impliquent pas exactement deux entités, les relations extraites s'avèrent incorrectes parce que les extracteurs se concentrent sur l'apprentissage des relations binaires. Un problème important dans les systèmes de pointe ClausIE et OLLIE est que, en visant à extraire le plus grand nombre de relations de la même phrase, ils perdent en précision. De plus, OLLIE est un système qui retourne très peu de relations extraites par rapport à ReVerb, cela est dû principalement à des erreurs d'analyse. ClausIE produit des extractions correctes 2,5 à 3,5 fois plus correctes qu'OLLIE. Cependant, ClausIE exploite les connaissances linguistiques relatives à la grammaire de la langue anglaise pour détecter les clauses d'une phrase. Par conséquent, il ne supporte que la langue anglaise.

Pour résoudre ces problèmes, Stanford OpenIE permet d'extraire des triplets de relations de domaine ouvert en divisant une phrase longue en clauses courtes et cohérentes, puis en recherchant les triplets de relations extrêmement simples qui sont justifiés pour chacune de ces clauses. Cela nous a permis d'avoir une meilleure connaissance du contexte de chaque extraction et de fournir des triplets informatifs. De plus Stanford OpenIE supporte plusieurs langues (Anglais, Français, Allemand, Chinois, Espagnol et l'Arabe). Pour toutes ces raisons, nous avons choisi Stanford OpenIE dans nos expérimentations.

4.5. Etape d'annotation des relations par le contexte

L'identification initiale des relations est effectuée avec des critères d'extraction minimaux pour récupérer toutes les relations. Par conséquent, l'accent sera mis sur la manière de sélectionner les relations significatives parmi toutes les relations initialement extraites. Par conséquent, un processus de filtrage visant à éliminer les relations non pertinentes est appliqué. La figure 4.2 illustre le processus d'annotation des relations par le contexte, qui se divise en trois étapes principales : (1) *Elimination des relations redondantes*, (2) *filtrage des relations en fonction du contexte* et (3) *Annotation des relations avec le contexte correspondant.*

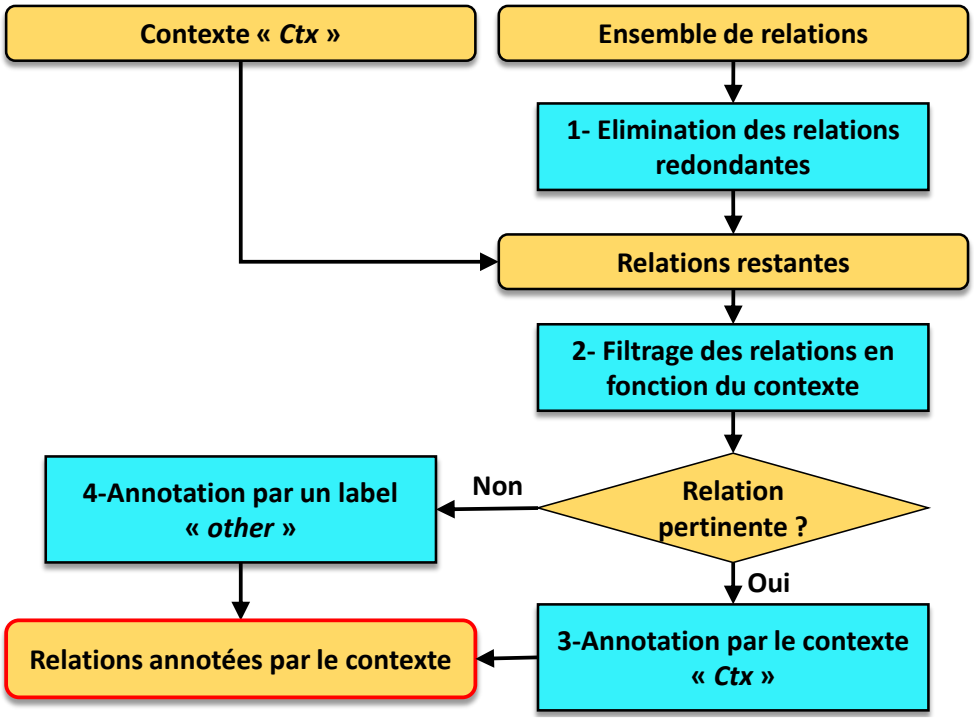


FIGURE 4.2. – Processus d'annotation des relations par le contexte

4.5.1. Elimination des relations redondantes

Ce filtrage commence par une élimination des relations redondantes en identifiant et en regroupant les relations similaires et en ne conservant qu'une seule relation représentative pour chaque groupe. Au début, un appariement est effectué entre les entités et le prédicat des différentes relations extraites (E1, P et E2) en utilisant le dictionnaire WordNet [130]. WordNet est utilisé en raison de sa capacité à identifier et les termes similaires (synonymes) en groupes. WordNet permet ainsi de réduire la redondance dans les relations en identifiant et en regroupant les relations similaires, ne conservant ainsi qu'une seule relation représentative pour chaque groupe. Grâce à cela, l'utilisation de WordNet pour la détection de synonymes permet d'obtenir des résultats plus précis et pertinents, en évitant les doublons.

Dans le cas où le prédicat ainsi que les deux entités de deux relations sont identiques ou des synonymes, ces dernières sont regroupées dans une même classe appelée « classe de relation ». Enfin, à partir de chaque classe, une seule relation est arbitrairement choisie pour représenter toutes les autres relations.

A titre d'exemple, si nous considérons les deux relations suivantes :
R1= ('Emmanual Macron', 'de la France'), 'est président'
R2= ('Macron', 'de la France'), 'est dirigeant

Ces deux relations sont considérées comme similaires grâce à WordNet (est président et est dirigeant sont des synonymes)

4.5.2. Filtrage des relations en fonction du contexte

L'objectif de ce filtrage est de ne conserver que les relations qui sont significatives pour le contexte du document. Par exemple, si nous considérons le paragraphe suivant (§1) :

(§1) : « L'élection présidentielle française a eu lieu les 10 et 24 avril 2022. Emmanuel Macron, qui est marié à Brigitte Trogneux, a battu Marine Le Pen au second tour de la présidentielle et effectuera un nouveau mandat à la présidence de la France »

L'application du processus de génération du contexte au paragraphe (§1) permet d'identifier le contexte suivant : « élection présidentielle française de 2022 ». La relation « Emmanual Macron est marié à Brigitte Trogneux » est considérée comme non pertinente pour ce contexte.

Afin de déterminer si la relation est pertinente ou non pour le contexte du document, le processus commence par calculer la similarité sémantique entre chaque paire de relations identifiées dans un document. Pour ce faire, nous avons évalué trois techniques différentes : USE (Yinfei YANG, CER, AHMAD et al. 2019), BERT-paire (DEVLIN, CHANG, K. LEE et al. 2018b) et BERT-paire avec apprentissage par transfert. Les expérimentations présentées dans la section 4.7.4 montrent que le modèle BERT-paire en appliquant le fine tuning donne de meilleurs résultats.

Une fois la similarité entre chaque paire de relation est calculée, nous stockons les scores de similarité dans une matrice. Enfin, pour calculer la similarité de chaque relation avec l'ensemble des relations du document, nous prenons la moyenne des scores de similarité de cette phrase avec chaque autre phrase du texte. Cela permet d'avoir une mesure de la similarité sémantique globale de la relation avec le document entier. Cette mesure peut être utilisée pour classer les relations en fonction de leur similarité sémantique avec le texte pour prédire si la relation est pertinente ou non pour le contexte du document. Afin de dire si la relation est pertinente pour le contexte ou non, nous avons choisi un seuil qui sera défini par nos expérimentations (section 4.7.4).

L'architecture du modèle BERT-paire utilisée pour calculer la similarité sémantique entre deux relations est présentée dans la figure 4.3. Dans ce modèle, deux relations sont entrées en parallèle dans le même modèle BERT, chacune étant tokenisées en utilisant un tokenizer BERT. Chaque séquence de tokens est ensuite encodée en vecteurs de représentation sémantique à l'aide de la couche d'embedding. Les embeddings de chaque relation sont ensuite traités en parallèle par deux ensembles de couches de transformers, chacun comprenant plusieurs couches de transformation avec attention multi-tête, qui calcule les représentations de chaque token en prenant en compte les autres tokens dans la séquence, et d'autres couches feedforward qui ajoute une non-linéarité à la représentation. Ensuite, des couches de pooling sont appliquées à chaque ensemble de sorties de la dernière couche des transformers. Ces couches de pooling convertissent les sorties de chaque couche de transformer en un vecteur de caractéristiques fixe pour chaque phrase, qui est ensuite concaténé. La couche de concaténation prend les deux vecteurs en entrée et les joint en un seul vecteur en les plaçant côte à côte. Cette fonction de concaténation simple permet de conserver

toutes les informations des deux vecteurs dans un seul vecteur, ce qui permet à la couche dense de prendre en compte les deux vecteurs simultanément pour calculer le score de similarité entre les deux relations.

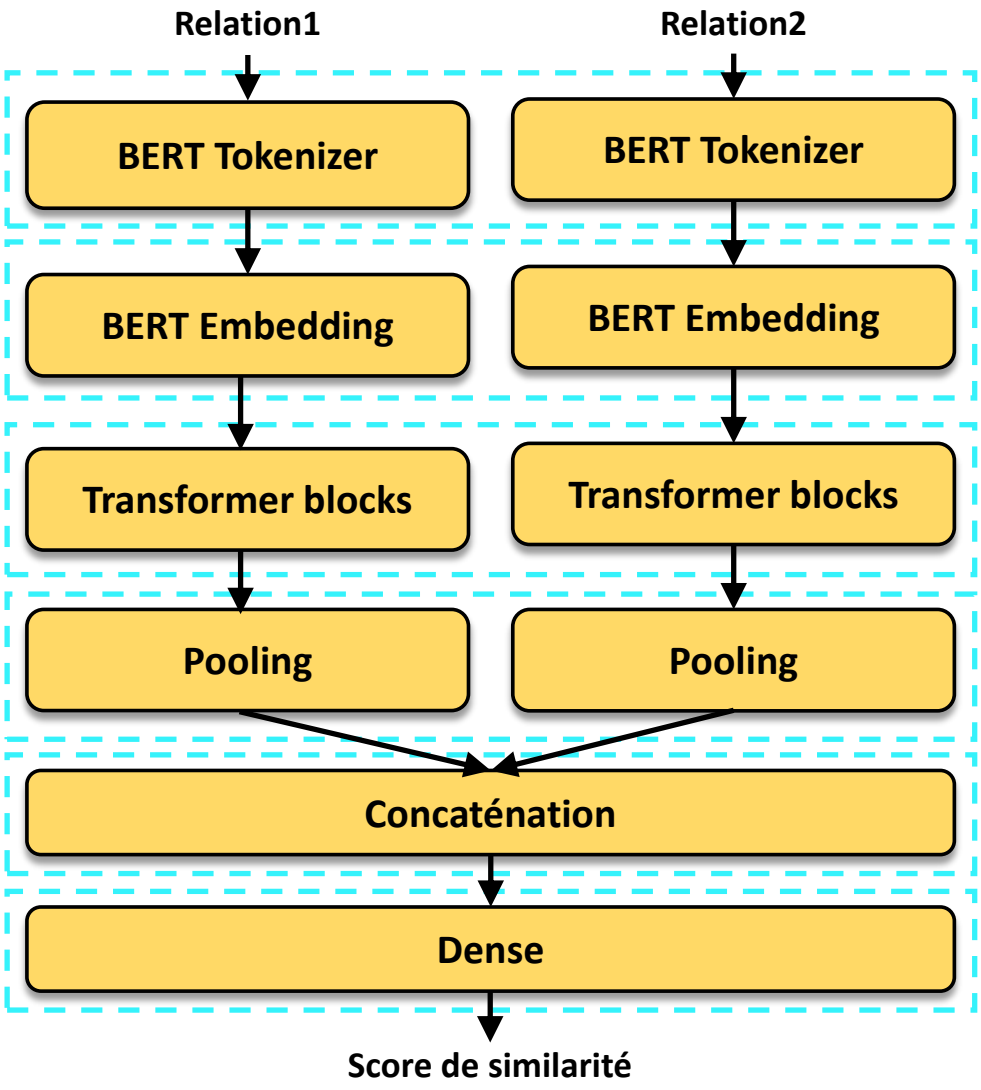


FIGURE 4.3. – L'architecture du modèle BERT-paire utilisée

4.5.3. Annotation des relations avec le contexte correspondant

Une fois les scores de similarité calculés, deux cas peuvent se présenter :

— Si la relation est pertinente pour le contexte du document, dans ce cas, elle sera annotée avec le contexte généré par le processus 1.

— Dans le cas contraire, lorsque la relation est considérée comme non pertinente, elle sera annotée avec le label « autre ».

Considérons le paragraphe (§1), les trois relations ('L'élection présidentielle française', 'a eu lieu', ' les 10 et 24 avril 2022'), ('Emmanuel Macron', 'a battu', 'Marine Le Pen') et ('Emmanuel Macron', 'effectuera', 'un nouveau mandat à la présidence de la France') sont annotées par le contexte 'élection présidentielle française de 2022'. Par contre, la relation ('Emmanuel Macron', 'est marié à', 'Brigitte Trogneux') est annotée par le label 'autre'.

4.6. Classification des relations selon le type

4.6.1. Classification selon le type

La classification des relations vise à reconnaître la relation sémantique entre deux entités dans le texte en se basant sur neuf types prédéfinis : Cause-Effet, Instrument-Agence, Produit-Producteur, Contenu-Conteneur, Entité-Origine, Entité-Destination, Composant-Global, Membre-Collection, Message-Topic, ou Autre si elle n'appartient à aucune des 9 relations annotées. Prenons à titre d'exemple la phrase suivante : *« fifty* $[essayes]_{e1}$ *collected in this* $[volume]_{e2}$ *testify to most of the prominent themes from Professor Quispel's scholarly career »*, Où les indices e_1 et e_2 désignent la première et la seconde entité. L'objectif de la classification des relations est d'identifier le type de la relation entre « $essayes$ » et « $volume$ », qui dans cette relation est « $Membre-Collection$ ».

4.6.2. Choix du modèle

Comme présenté dans le chapitre 1, la tâche de classification des relations selon des types prédéfinis à partir de données textuelles ont suscité un énorme intérêt au cours de ces dernières années. De nombreux chercheurs ont proposé diverses approches pour résoudre ce problème essentiellement des approches basées sur l'apprentissage profond. Les réseaux neuronaux profonds ont récemment démontré leur capacité à apprendre automatiquement des caractéristiques à partir d'ensembles de données.

Les modèles CNN et RNN sont les premiers réseaux neuronaux utilisés dans la tâche de classification des relations et qui ont démontré que l'apprentissage profond peut effectivement améliorer l'efficacité de la classification. Cependant, les modèles CNN (D. ZENG, K. LIU, LAI et al. 2014) et (D. ZHANG et Dong WANG 2015) ne sont pas satisfaisants car ils n'ont pas de structure spécifique pour la classification des relations. La combinaison de deux ou plusieurs réseaux de neurones peut améliorer les résultats en tirant parti à la fois des RNN et des CNN. En outre, les expériences montrent que les modèles basés sur l'attention sont plus performants que les modèles basés sur les CNN et les RNN. Par exemple, Att-RCNN (X. GUO, Hui ZHANG, H. YANG et al. 2019a) atteint presque 4% de plus que CNN. Pour cette raison, pour la tâche de classification des relations selon le type nous utilisons le modèle Att-RCNN.

Att-RCNN utilise une combinaison de deux types de NN (RNN + CNN) pour capturer les caractéristiques en intégrant les informations de relation dans les textes. En outre,

Att-RCNN n'adopte aucune caractéristique fabriquée à la main et obtient de meilleures performances dans presque tous les ensembles de données afin d'éviter le bruit qui peut être produit par l'homme. En outre, ce modèle utilise une attention à plusieurs niveaux dans Att-RCNN.

4.6.3. Architecture du modèle « Att-RCNN » utilisé pour la classification des relations selon le type

L'architecture du modèle utilisé est présentée dans la figure 4.4 Le processus commence par une analyse de la relation en entrée afin de filtrer le bruit et de ne conserver que les éléments clés sur la base des informations SDP. Ensuite, un RNN bidirectionnel (CHO, VAN MERRIËNBOER, GULCEHRE et al. 2014) est utilisé avec des cellules GRU (Gated Recurrent Units) pour apprendre les caractéristiques contextuelles de chaque mot en utilisant un Word Embeddings. La sortie des cellules GRU (forward and backward) contient des informations qui seront, par la suite, concaténées avec le Word Embeddings original des mots. Par conséquent, l'ensemble de Word Embeddings est considéré comme une représentation d'un mot ou d'un texte d'une certaine manière.

Ensuite, un mécanisme d'attention au niveau du mot est appliqué et un CNN suivi d'une attention au niveau de la phrase est utilisé pour extraire les caractéristiques les plus importantes. Enfin, ces caractéristiques de haut niveau seront introduites dans une couche de calcul de score, qui inclut une matrice de classe et calcule les scores de texte de chaque classe de relation.

— *Phase 1 : suppression du bruit basée sur les informations SDP :* Dans cette phase, la partie plus importante sera extraite de la relation. Pour se faire, un algorithme est proposé permettant de se débarrasser du bruit qui est nuisible à la tâche de classification des relations. Pour une relation, nous analysons d'abord l'arbre de dépendance sémantique. En prenant l'exemple de la relation suivante : « *fifty essayes collected in this volume testify to most of the prominent themes from Professor Quispel's scholarly career.* », nous ne révélons qu'une partie de l'arbre de dépendance de cette relation. Le SDP entre « e_1 = *essayes* » et « e_2 = *volume* » peut être facilement obtenu, qui est « *essayes* $\rightarrow$ *collected* $\rightarrow$ *in* $\rightarrow$ *volume* ». Les informations SDP contiennent presque toutes les informations de la relation entre les entités. L'entrée finale pour les cellules GRU est « *essayes collected in this volume* ». De cette façon, nous obtenons un fragment réduit par rapport à la relation originale basé sur l'information SDP

— *Phase 2 : Représentation contextuelle de chaque mot :* Le modèle bi-GRU [21] est utilisé, dans cette phase, pour obtenir la représentation contextuelle de chaque mot. GRU possède moins de paramètres que LSTM, ce qui se traduit par une plus grande vitesse de convergence des calculs. De plus, d'après les expériences, GRU obtient des performances supérieures à celles de LSTM. Par la suite, la sortie des cellules GRU (forward and backward) contient des informations qui seront concaténées avec le Word Embeddings original des mots.

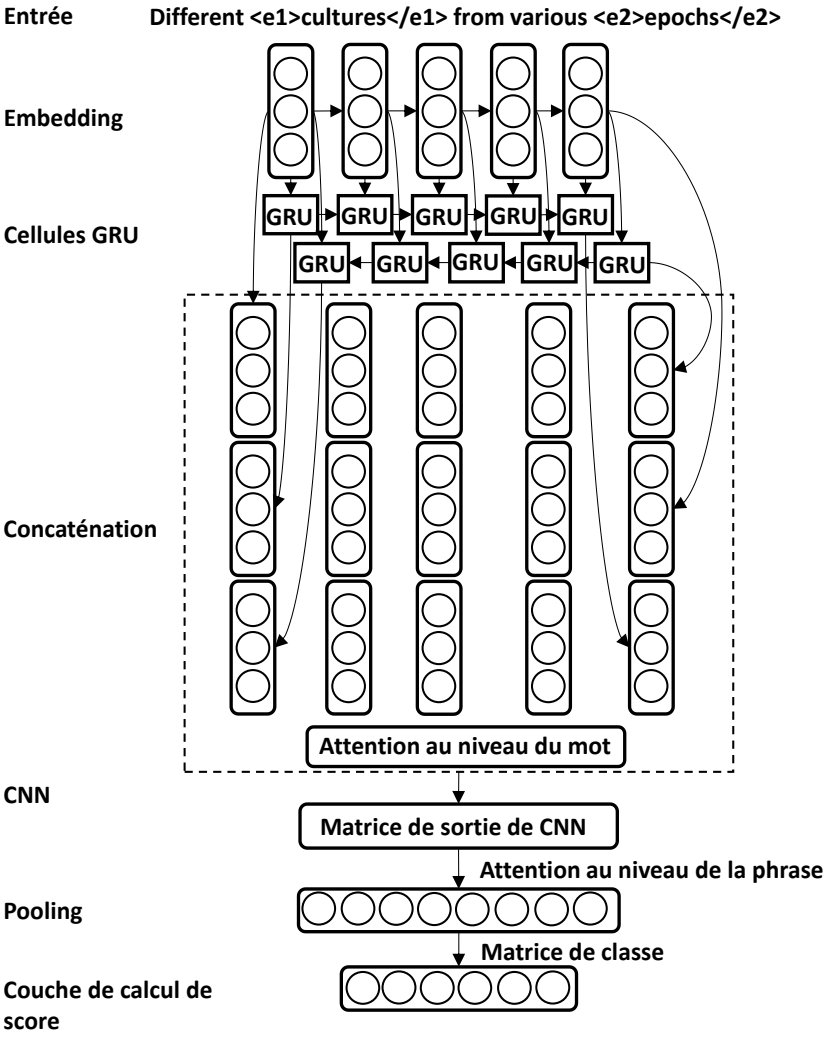


FIGURE 4.4. – Architecture du modèle Att-RCNN proposé par (X. GUO, Hui ZHANG, H. YANG et al. 2019b)

— *Phase 3 : Attention au niveau des mots :* Compte tenu de l'importance différente des mots dans le texte, une attention au niveau du mot est introduite pour modifier la représentation contextuelle originale du mot en multipliant différents poids. En détail, les mots en SDP (exactement dans le chemin) sont affectés d'une valeur de poids plus élevée et les autres mots sont affectés d'une valeur de poids plus faible. Ainsi, les vecteurs de mots modifiés peuvent être calculés par la formule suivante

$$w_l^{\text{modified}} = \begin{cases} \alpha_{\text{heigh}} \times w_l & \text{if } l \in S_{\text{SDP}} \\ \alpha_{\text{low}} \times w_l & \text{if } l \notin S_{\text{SDP}} \end{cases} \tag{4.2}$$

où SSDP représente l'ensemble des mots basés sur les informations SDP et $_h igh$,

$_{l}ow$ désignent respectivement les valeurs de poids supérieures et inférieures. Nous utilisons ces deux paramètres pour que notre modèle prenne conscience des différences entre les mots et attribue des poids plus élevés aux mots qui ont une grande importance pour les relations. Les valeurs de $_{h}igh$ et $_{l}ow$ sont attribuées selon une proportion de 2 :1. Après, la représentation complète du mot passe par un CNN pour extraire l'information contextuelle de la relation.

— *Phase 4 : Attention au niveau de la phrase et prédiction du type :* Avant d'appliquer le max pooling à la sortie du CNN, un mécanisme d'attention a été introduit au niveau de la phrase pour renforcer les caractéristiques. Finalement, une fonction softmax est appliquée pour prédire le type de la relation.

4.7. Expérimentation et évaluation

Cette section présente une évaluation de la proposition de classification de relations selon le type et le contexte après l'intégration de la méthode d'extraction générative de contexte proposée dans le chapitre 3. L'objectif de cette section est double : (1) montrer la validité de notre proposition sur deux collections de données en utilisant comme métrique d'évaluation la mesure F1-score, et (2) appliquer cette proposition sur une étude de cas réel relative à la visualisation automatique des données statistiques extraites d'un texte.

4.7.1. Environnement expérimental

Dans cette section, nous présentons l'environnement matériel que l'environnement logiciel qui a permis l'aboutissement de la mise en œuvre de la méthode de classification de relation selon le type et le contexte. Pour la mise en œuvre de cette méthode, nous avons travaillé sur la même machine présentée dans la section 3.7.1 avec la même configuration. Comme langage de programmation, nous avons aussi choisie Python pour implémenter notre solution. Parmi les bibliothèques utilisées, nous pouvons citer « NLTK », « Pandas », « Sklearn », « Keras », « TensorFlow-Hub » et « StanfordNLP ». La bibliothèque « StanfordNLP » est utilisée pour l'extraction des différentes relations à partir d'un document texte. Cette bibliothèque fournit plusieurs annotateurs à savoir « tokenize », « pos », « lemma », « openie », etc...

Dans notre cas nous avons utilisé l'annotateur « openie » qui permet de produire un nombre de fragments de phrase correspondant aux fragments impliqués de la phrase d'origine donnée. Ces fragments sont ensuite segmentés en triples OpenIE. « TensorFlow-Hub » est une bibliothèque de modèles d'apprentissage automatique entraînés, prêts à être affinés (fine-tuned) et pouvant être déployés. Grâce à cela nous sommes en mesure de réutiliser des modèles comme BERT avec seulement quelques lignes de code. Afin d'entraîner nos modèles de filtrage des relations et de classification de relation selon le type, nous avons aussi utilisé Google-Colab. De même, nous avons utilisé Jupyter notebook pour créer, manipuler et exécuter notre code et de faire des différentes visualisations facilement.

4.7.2. Protocole expérimental

Les contributions scientifiques de la proposition d'une méthode de classification des relations selon le type et le contexte seront démontrées à travers les expérimentations. Pour atteindre cet objectif, nous proposons de suivre un protocole expérimental présenté dans la figure 4.5

4.7.3. Pré-Evaluation

Dans le but de valider notre méthode de classification des relations selon le type et le contexte, il est nécessaire d'effectuer une série d'expérimentations sur deux corpus de données.

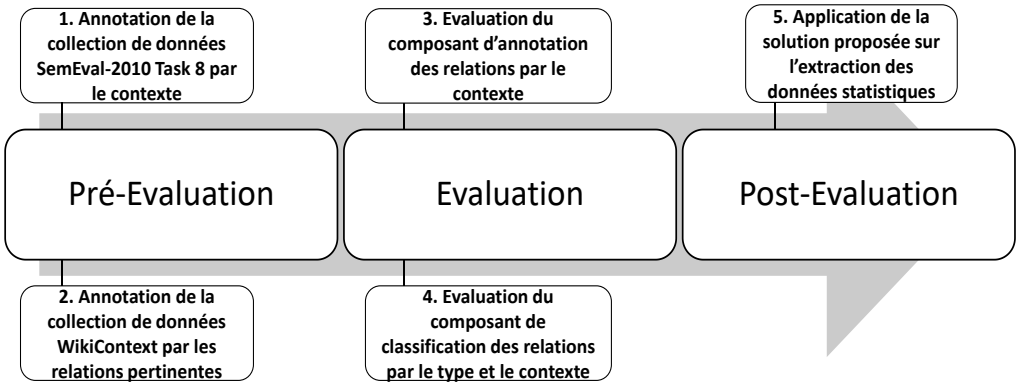


FIGURE 4.5. – Protocole expérimental retenu

1. **corpus de données SemEval 2010 Tak8 :** SemEval-2010 Task 8 [31] contient 10 717 instances annotées, dont 8 000 instances utilisées pour l'entraînement et 2 717 instances utilisées pour le test. Toutes les instances sont annotées avec 9 types de relations prédéfinies et une classe artificielle Other. Les neuf relations prédéfinies sont respectivement Cause-Effect, Instrument-Agency, Product-Producer, Content-Container, Entity-Origin, Entity-Destination, Component-Whole, Member-Collection, et Message-Topic. Nous prenons en compte la direction ainsi le nombre total de types de relations est de 19. Cette collection est utilisée pour évaluer le processus de classification des relations selon le type et le contexte. La figure 4.6 montre un exemple de 20 relations sélectionnées au hasard dans l'ensemble de tests. La première colonne contient l'identifiant de la phrase et la seconde contient la phrase annotée avec deux entités :

1 Avian <e1>influenza < /e1> is an infectious disease of birds caused by type A strains of the influenza <e2>virus</e2>
2 The <e1>ear</e1> of the African <e2>elephant</e2> is significantly larger– measuring 183 cm by 114 cm in the bush elephant
3 Skype, a free software, allows a <e1>hookup</e1> of multiple computer <e2>users</e2> to join in an online conference call without incurring any telephone costs.
4 This <e1>thesis</e1> defines the <e2>clinical characteristics</e2> of amyloid disease
5 An FTP server is an inexpensive and relatively simple to operate tool that works great for <e1>filesharing</e1> over the <e2>internet</e2>
6 The <e1>song</e1> was composed for a famous Brazilian <e2>musician</e2>
7 My <e1>cat</e1> has a problem with his <e2>paw</e2>
8 Essentially, the <e1>blisters</e1> that appear in the mouth are caused by the <e2>herpes simplex virus</e2> type 1, HSV-1 for short
9 The <e1>treaty</e1> establishes a double majority <e2>rule</e2> for Council decisions
10 Therefore, nowadays China is the complex mixture of different <e1>cultures</e1> from various <e2>epochs</e2>
11 The accident has spread <e1>oil</e1> into the <e2>ocean</e2>
12 By dividing the <e1>space</e1> in a kitchen <e2>drawer</e2> where you keep all your cooking utensils, you grouped items by size or purpose
13 <e1>News programs</e1> commented on the <e2>violence</e2> from the game and expressed worries on how it would affect the players' personalities
14 The <e1>subject</e1> of "imply" is the source of an <e2>implication</e2> while the subject of "infer" is the recipient of an implication
15 People now post their <e1>opinions</e1> to this <e2>blog</e2>
16 Traffic <e1>vibrations</e1> on the street outside had caused the <e2>movement</e2> of the light
17 A <e1>witch</e1> is able to change events by using <e2>magic</e2>
18 The <e1>family</e1> constructed some TCU Horned Frog supporting snow <e2>features</e2> as well as just a really nice picture of an old fashioned light post
19 In the <e1>article</e1>, the authors explore the <e2>use</e2> of technology in small pharmacy chains
20 The UBC graduate <e1>student</e1> spent his free time creating an <e2>application</e2> called ColorSplash that lets people play with the colour in their digital photos

FIGURE 4.6. – Exemple de 20 relations sélectionnées au hasard dans l'ensemble de test de SemEval-2010 task8

Afin d'évaluer la tâche de classification des relations selon le contexte, nous avons annoté manuellement chaque relation de la collection de données SemEval-2010 task8 par le contexte adéquat. Cette annotation a été effectuée par trois annotateurs. Nous adoptons le score F1 pour évaluer notre solution, qui représente la métrique d'évaluation officielle de SemEval-2010 task8.

2. **WikiContext :** Notre corpus de données WikiContext déjà défini et utilisé dans le chapitre précédent (section 3.7.3) va nous permettre d'effectuer une évaluation quantitative de deux composants associés aux deux étapes d'annotation des relations par le contexte et de classification des relations selon le type. Pour évaluer le premier composant, nous avons annoté notre collection par les relations pertinentes. La mesure F1-score est utilisée pour évaluer ces deux composants.

4.7.4. Evaluations : Résultats obtenus et interprétations

Nous présentons dans ce qui suit les paramètres d'expérimentation ainsi que l'évaluation du composant d'annotation des relations par le contexte du document.

1. **Résultats de l'évaluation de l'étape d'annotation des relations par le contexte du document :**

— **Paramètres des expériences :** Dans cette expérimentation, nous avons utilisé le modèle BERT-paire pour calculer la similarité entre deux phrases. Nous avons choisi le modèle « BERT-base ». Nous avons fixé la longueur maximale de la séquence d'entrée à 128, et la taille de lot à 8. Nous avons utilisé un taux d'apprentissage de 2e-5 pour mettre à jour les paramètres du modèle pendant l'entraînement. Ces valeurs ont été choisies en suivant les recommandations générales de l'état de l'art pour le calcul de similarité entre deux phrases à l'aide du modèle BERT-paire.
Pour le fine-tuning, nous avons choisi une longueur maximale de séquence de 128, une taille de lot de 8 et un taux d'apprentissage de 2e-5. Nous avons également utilisé un nombre de couches de 12, une dimension d'embedding cachée de 768, un nombre de têtes d'attention de 12 et une probabilité de dropout de 0,1. Ces valeurs ont été sélectionnées par expérimentation sur notre corpus « WikiContext ». Nous avons commencé par tester différentes combinaisons de paramètres, en ajustant chaque fois les valeurs pour voir comment elles affectent les performances du modèle.

— **Résultats de l'évaluation du composant d'annotation des relations par le contexte du document :** Afin d'évaluer ce processus, nous avons testé trois techniques différentes USE, BERT-paire et BERT-paire en utilisant un fine-tuning pour déterminer si une relation est pertinente pour le contexte d'utilisation. Le tableau 4.1 résume les résultats obtenus. Comme le montre

TABLEAU 4.1. – Performance du composant d'annotation des relations par le contexte du document sur la collection de données WikiContext

Modèle utilisé	**Précision(%)**	**Rappel(%)**	**F1–score**
Universal Sentence Encoder	59.15	68.25	63.37
BERT-paire	69.50	75.20	72.23
BERT-paire + transfer learning(notre approche)	**72.10**	**81.20**	**76.37**

le tableau 4.1, le modèle BERT-paire en utilisant le Fine-tuning sur notre collection de données « WikiContext » surpasse Universal Sentence Encoder et BERT-paire avec un score F1 de 76,37%.

2. **Résultats de l'évaluation de l'étape de classification des relations selon le type et le contexte :** Dans cette section, nous décrivons les étapes de préparation des corpus que nous avons utilisés, ainsi que les paramètres que nous avons choisis pour notre modèle de classification de relations en fonction de leur type. Nous évaluons également notre méthode de classification en prenant en compte le type et le contexte.
   — **Préparation des corpus :** Les corpus utilisés dans l'expérience pour évaluer la phase de classification des relations selon le type et le contexte sont Wiki-

Context et SemEval-2010 Task 8. Le corpus SemEval-2010 Task 8 a été divisé selon la proportion 75 :25, ce qui signifie que sur 10717 relations, 8000 sont utilisées pour l'entrainement et 2717 pour le test. La fréquence des relations est indiquée dans le tableau 4.2.

TABLEAU 4.2. – Fréquence des relations dans la corpus de données SemEval-2010 Task 8

Relation	Nombre d'instances d'entraînement	Nombre d'instances de test
Cause-Effect	1003	328
Instrument-Agency	941	312
Product-Producer	845	292
Content-Container	717	231
Entity-Origin	716	258
Entity-Destination	690	233
Component-Whole	634	261
Member-Collection	540	192
Message-Topic	504	156
Other	1410	454
Total	8000	2717

Concernant le corpus WikiContext, sur les 7592 documents collectés dans 30 contextes, 6067 documents (80% de chaque contexte) ont été utilisés pour l'entraînement et les 1525 restants (20% de chaque contexte) pour les tests.

— **Paramètres des expériences :** Tout d'abord, l'algorithme Adam est utilisé pour accélérer la procédure d'entraînement. Ensuite, un modèle de skip-gram est entraîné sur Wikipédia pour obtenir des embeddings de mots pré-entraînés. D'autres caractéristique ont été ajustées s telles que le coefficient de normalisation et le taux d'apprentissage qui a été fixé à 0.01 pour s'adapter à la taille de l'ensemble de données SemEval 2010 Task 8. Les hyperparamètres du modèle ont été ajustés en utilisant la méthode d'optimisation « dropout » pour la couche d'embedding. Le taux de dropout est de 0,5, ce qui signifie que chaque entrée de l'embedding a une probabilité de 0,5 d'être mise à zéro pendant l'entraînement.

— **Résultats de l'évaluation du composant de classification des relations selon le type et le contexte :** Au début de la phase d'apprentissage, nous créons le modèle du réseau de neurone Att-RCNN que nous devons entraîner avec la base d'apprentissage proposée par SemEval-2010 plusieurs fois et avec la totalité des relations de la base de données d'entraînement. Le but étant que le modèle classifie au mieux ces données. Lorsque le modèle a fini de mettre à jour ses poids, nous évaluons le modèle en lui présentant la base de données de validation (contenant de nouvelles relations). Un algorithme

TABLEAU 4.3. – Performance du composant de classification des relations selon le type et le contexte sur les collections de données SemEval-2010 Task8 et WikiContext

Collection de donnée	Type de classification	Précision	Rappel	F1-score
SemEval-2010 Task8	Type prédéfini Contexte	83.3% 73.5%	90.2% 78.2%	86.6% 75.7%
WikiContext	Type prédéfini Contexte	68.9% 75.2%	78.8% 81.3%	73.5% 78.1%

pourra alors réaliser des prédictions sur des données dont les types y ne sont pas connus.

Afin de vérifier que notre approche a la meilleure capacité d'apprendre automatiquement et d'obtenir de bonnes performances remarquables en matière de classification des relations selon le type et le contexte, nous présentons les résultats de performance en termes de F1- score sur les deux collections de données SemEval-2010 task8 et WikiContext dans le tableau 4.3 En effet, nous avons calculé des mesures d'évaluation pour 2717 et 6204 relations proposées respectivement par la base de test de SemEval-2010 task8 et WikiContext. Ces mesures sont obtenues après que le système ait retourné un classement pour les relations selon le type et le contexte. Ils sont obtenus en étudiant l'ensemble des relations restituées par le système et en distinguant les relations pertinentes des relations non pertinentes pour un contexte Ci en se basant sur les annotations offertes par nos collections de données.

a) **Quelques résultats de notre approche sur la collection SemEval-2010**
La figure 4.7 donne un aperçu de notre tâche de classification réalisée sur les 20 phrases présentées dans la figure 4.6

```
1   Cause-Effect(e2,e1) Influenza of birds
2   Component-Whole(e1,e2) ear of the African elephant
3   Member-Collection(e2,e1) Skype allows join online conference
4   Message-Topic(e1,e2) clinical characteristics of amployed disease
5   Other Inexpensive and simple FTP server
6   Product-Producer(e1,e2) Song of Brazilian musician
7   Component-Whole(e2,e1) Paw of cat
8   Product-Producer(e1,e2) Causeof the mouth blisters
9   Cause-Effect(e1,e2) Council decisions
10  Entity-Origin(e1,e2) Different cultures of China
11  Entity-Destination(e1,e2) Ocean accident
12  Other space in a kitchen
13  Message-Topic(e1,e2) violence in game affects players personalities
14  Cause-Effect(e1,e2) subject of infer imply and infer
15  Other opignons to blog
16  Cause-Effect(e1,e2) street vibrations
17  Instrument-Agency(e2,e1) witch change events using magic
18  Product-Producer(e2,e1) TCU Horned Frog
19  Message-Topic(e1,e2) use of technology in pharmacy chains
20  Product-Producer(e2,e1) student create ColorSplash application
```

FIGURE 4.7. – Classification des 20 relations présentées dans la figure 4.6 selon le type et le contexte

La première colonne représente l'identifiant de la phrase et la seconde donne le résultat de la tâche de classification : le type prédéfini annoté par le contexte. Ces résultats montrent que notre approche surpasse tous les autres systèmes en ajoutant la contextualisation des relations. L'ajout de capacités contextuelles précises aux relations pourrait être appliqué dans une variété d'applications décisionnelles. Ainsi la « *contextualisation des relations* », en particulier dans le cadre des moteurs de recherche, améliore l'efficacité de la recherche d'informations.

b) **Quelques résultats de notre approche sur la collection WikiContext**

Les expériences sur les documents ont demandé plus de travail. En effet, il était nécessaire de déterminer le contexte du document en même temps que la classification des relations. Nous illustrons cela par le traitement détaillé de 3 documents particuliers.

— **Document1 :** Considérons le document de notre jeu de données « WikiContext » présenté dans la Figure 3.3 Nous avons appliqué notre méthode d'extraction du contexte sur ce document, ce qui nous a permis d'obtenir le contexte « Animal attacks and bites are a public health problem that cause injuries and fatalities ».
Parallèlement, Parallèlement, nous avons utilisé l'outil Stanford OpenIE pour extraire les relations du document. Les relations extraites sont présentées dans la figure 4.8 Cependant, nous avons constaté

que plusieurs de ces relations étaient hors contexte ou redondantes. Pour remédier à cela, nous avons appliqué un processus de filtrage pour ne garder que les relations pertinentes pour le contexte du document. Ce processus de filtrage a permis d'identifier les relations qui ne correspondaient pas à l'aspect sémantique du document comme la relation « États-Unis en 1994 » qui n'est pas pertinente pour le contexte « Animal attacks and bites are a public health problem that cause injuries and fatalities ». Nous avons donc annoté cette relation avec le contexte « other ». Les relations restantes ont été annotées en fonction du contexte du document.
Finalement, avons classées les relations par type, ce qui nous a permis d'avoir à la fois une classification syntaxique et sémantique des relations comme présenté dans la figure 4.9. Cette classification nous a permis d'obtenir une vue plus précise et complète des relations entre l'aspect syntaxique et le sens.

1 [entity: 'Animal attacks', relation: 'are', object:'violent']
2 [entity: 'Bites', relation: 'are', object:'wounds caused as a result of an animal or human attack']
3 [entity: 'attacks', relation: 'are', object:'a cause of human injuries and fatalities worldwide']
4 [entity: 'United States citizens', relation: 'owned', object:'a pet']
5 [entity: 'United States', relation: 'in', object:'1994']
6 [entity: 'United States', relation: 'were bitten', object:'by dogs']
7 [entity: 'frequency of animal attacks', relation:'varies with', object:'geographical location']
8 [entity: 'Gonad glands', relation: 'found on', object:' the anterior side of the pituitary gland']
9 [entity: 'Animals', relation: 'with', object:'high levels of these hormones tend']
10 [entity: 'Animals', relation: 'tend to be', object:' more aggressive']
11 [entity: 'Animal attacks', relation: 'have been identified', object:' as a major public health problem']
12 [entity: 'Injuries', relation: 'caused by', object:'animal attacks']
13 [entity: 'Injuries', relation: ' result in', object: 'thousands of fatalities worldwide']
14 [entity:'causes of death', relation: 'are reported to', object:' Control and Prevention']
15 [entity: 'Medical injury code', relation: 'are used to ', object:'identify specific cases']

FIGURE 4.8. – Relations extraites du document présenté dans la figure 3.3 après application du système Stanford open-IE

1 Other Animal attacks and bites are a public health problem that cause injuries and fatalities
2 Content-Container(e1,e2) Animal attacks and bites are a public health problem that cause injuries and fatalities
3 Cause-Effect(e1,e2) AAnimal attacks and bites are a public health problem that cause injuries and fatalities
4 Member-Collection(e2,e1) Animal attacks and bites are a public health problem that cause injuries and fatalities
5 Other Other
6 Other Animal attacks and bites are a public health problem that cause injuries and fatalities
7 Instrument-Agency(e1,e2) Animal attacks and bites are a public health problem that cause injuries and fatalities
8 Content-Container(e1,e2) Animal attacks and bites are a public health problem that cause injuries and fatalities
9 Other Animal attacks and bites are a public health problem that cause injuries and fatalities
10 Other Animal attacks and bites are a public health problem that cause injuries and fatalities
11 Message-Topic(e1,e2) Animal attacks and bites are a public health problem that cause injuries and fatalities
12 Cause-Effect(e1,e2) Animal attacks and bites are a public health problem that cause injuries and fatalities
13 Cause-Effect(e1,e2) Animal attacks and bites are a public health problem that cause injuries and fatalities
14 Other Animal attacks and bites are a public health problem that cause injuries and fatalities
15 Instrument-Agency(e1,e2) Animal attacks and bites are a public health problem that cause injuries and fatalities

FIGURE 4.9. – Classification des relations selon le type et le contexte

Nous présentons les résultats du même processus pour deux autres documents de notre corpus de données WikiContext : le document 2 concerne le contexte de « traitement automatique des langues naturelles », et le document 3 concerne « la dépendance à la drogue ».

c) **Document2 :** Considérons un deuxième document de notre jeu de données « WikiContexte » qui traite du traitement automatique des langues naturelles. Ce document est présenté dans la figure 4.10 Nous appliquons notre processus d'extraction de contexte. Nous avons ainsi obtenu les mots clés suivants : Natural language, processing, computer science, artificial intelligence, history, challenges and subfield. Nous avons ensuite combiné, en utilisant notre LSTM, avec des mots de liaison pour former une phrase cohérente qui résume le contexte du document : « Natural language processing, a subfield of computer science, has a history and presents challenges in artificial intelligence. »
Après avoir identifié le contexte du document, nous avons poursuivi notre analyse en classifiant les relations en fonction de leur type et de leur contexte. Pour ce faire, nous avons d'abord identifié les relations à l'aide de « StandfordOpenIE », qui a produit les relations présentées dans la figure 4.11 Nous avons ensuite appliqué notre processus de filtrage pour

identifier les relations qui ne sont pas pertinentes pour le contexte. Dans le cas de ce document, les relations « Georgetown experiment in 1954 » et « ELIZA written by Joseph Weizenbaum » étaient hors contexte, nous les avons donc annotées avec le label « other ».
Notre processus se termine par la classification des relations selon leur type, ce qui nous a permis d'obtenir une double classification syntaxique et sémantique, comme illustré dans la figure 4.12. Cette classification nous a fourni une compréhension plus approfondie du document et nous a permis de mieux appréhender son contexte.

Natural language processing (NLP) is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.
Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural language generation.
The history of natural language processing (NLP) generally started in the 1950s, although work can be found from earlier periods. The Georgetown experiment in 1954 involved fully automatic translation of more than sixty Russian sentences into English. Some notably successful natural language processing systems developed in the 1960s were SHRDLU, a natural language system working in restricted "blocks worlds" with restricted vocabularies, and ELIZA, a simulation of a Rogerian psychotherapist, written by Joseph Weizenbaum between 1964 and 1966. During the 1970s, many programmers began to write "conceptual ontologies", which structured real-world information into computer-understandable data. Up to the 1980s, most natural language processing systems were based on complex sets of hand-written rules. Recent research has increasingly focused on unsupervised and semi-supervised learning algorithms. In the 2010s, representation learning and deep neural network-style machine learning methods became widespread in natural language processing, due in part to a flurry of results showing that such techniques[4][5] can achieve state-of-the-art results in many natural language tasks, for example in language modeling,[6] parsing,[7][8] and many others.

FIGURE 4.10. – Document de la corpus de données WikiContext qui concerne le contexte Le traitement automatique du Langage Naturel(NLP)

1 [entity: 'Natural language processing', relation: 'is a subfield of', object:'linguistics , computer science , information engineering , and artificial intelligence concerned with the interactions between computers and human']
2 [entity: 'program computers', relation: 'process and analyze', object:'large amounts of natural language data']
2 [entity: 'program computers', relation: 'process and analyze', object:'large amounts of natural language data']
3 [entity: 'Challenges', relation: 'in', object:'Natural language processing']
4 [entity: 'Challenges', relation: 'frequently involve', object:'speech recognition']
5 [entity: 'Georgetown experiment', relation: 'in', object:'1954']
6 [entity: 'Georgetown experiment', relation: 'involved' , object:'fully automatic translation of more than sixty Russian sentences into English']
7 [entity: 'successful natural language processing systems', relation: 'developed in', object:'1960s']
8 [entity: 'natural language system', relation: 'working in', object:'restricted blocks worlds with restricted vocabularies']
9 [entity: 'ELIZA', relation: 'simulation of', object:' Rogerian psychotherapist']
10 [entity: 'ELIZA', relation: 'written by, object:'Joseph Weizenbaum']
11 [entity: 'programmers', relation: 'began to write', object:'conceptual ontologies']
12 [entity: 'natural language processing systems', relation: 'were based on', object: 'complex sets of hand']
13 [entity:'research', relation: 'has increasingly focused on', object:'unsupervised and semi-supervised learning algorithms']
14 [entity: 'representation learning and deep neural network', relation: 'became', object:'widespread in natural language processing']
15 [entity: 'state - of - the - art results', relation: 'in', object:'many natural language tasks']
16 [entity: 'example', relation: 'in', object:'language modeling']

FIGURE 4.11. – Relations extraites du document présenté dans la figure 4.10 après application du système Stanford open-IE

1 Entity-Origin(e1,e2) Natural language processing a subfield of computer science has an history and presents artificial intelligence challenges
2 Other Natural language processing a subfield of computer science has an history and presents artificial intelligence challenges
3 Member-Collection(e2,e1) Natural language processing a subfield of computer science has an history and presents artificial intelligence challenges
4 Message-Topic(e1,e2) Natural language processing a subfield of computer science has an history and presents artificial intelligence challenges
5 Other Other
6 Message-Topic(e1,e2) Natural language processing a subfield of computer science has an history and presents artificial intelligence challenges
7 Other Natural language processing a subfield of computer science has an history and presents artificial intelligence challenges
8 Other Natural language processing a subfield of computer science has an history and presents artificial intelligence challenges
9 Other Natural language processing a subfield of computer science has an history and presents artificial intelligence challenges
10 Other Other
11 Other Natural language processing a subfield of computer science has an history and presents artificial intelligence challenge
12 Message-Topic(e1,e2) Natural language processing a subfield of computer science has an history and presents artificial intelligence challenge
13 Message-Topic(e1,e2) Natural language processing a subfield of computer science has an history and presents artificial intelligence challenge
14 Other Natural language processing a subfield of computer science has an history and presents artificial intelligence challenge
15 Component-Whole(e1,e2) Natural language processing a subfield of computer science has an history and presents artificial intelligence challenge
16 Component-Whole(e1,e2) Natural language processing a subfield of computer science has an history and presents artificial intelligence challenge

FIGURE 4.12. – Classification des relations selon le type et le contexte

— **Document3 :** En poursuivant notre expérimentation, nous nous sommes intéressés à un troisième document de notre jeu de données « WikiContext » (figure 4.13) portant sur la dépendance à la drogue. Après avoir appliqué notre processus d'extraction de contexte, nous avons identifié les mots clés suivants : « human brain, drug addiction, behavior, prescribed medication et use », que nous avons ensuite combinés pour obtenir le contexte « Drug addiction affects human brain and behavior, causing by abusive use of prescribed medication ». Nous avons ensuite poursuivi le processus en classifiant les relations en fonction de leur type et de leur contexte. StandfordOpenIE nous a permis d'obtenir une liste de relations pertinentes présenté dans la figure 4.14. Contrairement au document précédent, aucune relation n'a été filtrée car toutes étaient pertinentes. La classification des relations a ensuite été effectuée pour obtenir une double classification syntaxique et sémantique, comme présenté dans la figure 4.15.

Drug addiction, also called substance use disorder, is a disease that affects a person's brain and behavior and leads to an inability to control the use of a legal or illegal drug or medication. Substances such as alcohol, marijuana and nicotine also are considered drugs. When you're addicted, you may continue using the drug despite the harm it causes.
Drug addiction can start with experimental use of a recreational drug in social situations, and, for some people, the drug use becomes more frequent. For others, particularly with opioids, drug addiction begins with exposure to prescribed medications, or receiving medications from a friend or relative who has been prescribed the medication.
The risk of addiction and how fast you become addicted varies by drug. Some drugs, such as opioid painkillers, have a higher risk and cause addiction more quickly than others.
As time passes, you may need larger doses of the drug to get high. Soon you may need the drug just to feel good. As your drug use increases, you may find that it's increasingly difficult to go without the drug. Attempts to stop drug use may cause intense cravings and make you feel physically ill (withdrawal symptoms).

FIGURE 4.13. – Document de la corpus de données WikiContext qui concerne le contexte de la dépendance aux drogues

1 [entity: 'Drug addiction', relation: 'is', object:'a disease']
2 [entity: 'Drug addiction' relation:'affects', object:' person's brain and behavior']
3 [entity: 'Substances', relation: 'are considered', object:'drugs']
4 [entity: 'You', relation: 'are', object:'addicted']
5 [entity: 'You', relation: 'using', object:'drugs']
6 [entity: 'You', relation: 'using the drug despite', object:'harm']
7 [entity: 'Drug addiction', relation:'can start with', object:'experimental use of a recreational drug in social situations']
8 [entity: 'recreational drug', relation: 'in', object:'social situations']
9 [entity: 'drug', relation: 'becomes', object:'more frequent']
10 [entity: 'drug addiction', relation: 'begins with', object:'exposure to prescribed medications']
11 [entity: 'you become', relation: 'varies by', object:'drug']
12 [entity: 'drugs', relation: 'have', object:'higher risk']
13 [entity: 'drugs', relation: 'cause', object: 'addiction']
14 [entity:'You', relation: 'may need', object:'larger doses of the drug to get']
15 [entity: 'You', relation: 'may need', object:'drug']
16 [entity: 'You', relation: 'just to feel', object:'good']
17 [entity: 'it', relation: 'is', object:'increasingly difficult to go without the drug']
18 [entity: 'Attempts', relation: 'stop', object:'drug use']
19 [entity: 'you', relation: 'feel', object:'physical ill']

FIGURE 4.14. – Relations extraites du document présenté dans la figure 4.13 après application du système Stanford open-IE

1	Message-Topic(e1,e2) Drug addiction affects human brain and behavior causing by abusive use of prescribed medication
2	Other Drug addiction affects human brain and behavior causing by abusive use of prescribed medication
3	Message-Topic(e1,e2) Drug addiction affects human brain and behavior causing by abusive use of prescribed medication
4	Other Drug addiction affects human brain and behavior causing by abusive use of prescribed medication
5	Other other
6	Cause-Effect(e1,e2) Drug addiction affects human brain and behavior causing by abusive use of prescribed medication
7	Other Drug addiction affects human brain and behavior causing by abusive use of prescribed medication
8	Other Drug addiction affects human brain and behavior causing by abusive use of prescribed medication
9	Other Drug addiction affects human brain and behavior causing by abusive use of prescribed medication
10	Other Drug addiction affects human brain and behavior causing by abusive use of prescribed medication
11	Other Drug addiction affects human brain and behavior causing by abusive use of prescribed medication
12	Cause-Effect(e1,e2) Drug addiction affects human brain and behavior causing by abusive use of prescribed medication
13	Cause-Effect(e1,e2) Drug addiction affects human brain and behavior causing by abusive use of prescribed medication
14	Other Drug addiction affects human brain and behavior causing by abusive use of prescribed medication
15	Other Drug addiction affects human brain and behavior causing by abusive use of prescribed medication
16	Other Drug addiction affects human brain and behavior causing by abusive use of prescribed medication
17	Other Drug addiction affects human brain and behavior causing by abusive use of prescribed medication
18	Other Drug addiction affects human brain and behavior causing by abusive use of prescribed medication
19	Other Drug addiction affects human brain and behavior causing by abusive use of prescribed medication

FIGURE 4.15. – Classification des relations extraites du document 3 selon le type et le contexte

Dans l'état de l'art actuel, la classification des relations est un élément clé pour l'analyse de texte. Les méthodes de classification syntaxique se basent généralement sur la structure grammaticale des phrases et des mots, en utilisant des règles préétablies pour classifier les relations selon des types prédéfinis tels que content-container ou cause-effet. Bien que ces méthodes soient utiles pour identifier les relations de base entre les mots, elles ne permettent pas toujours de comprendre le sens profond des phrases.

C'est ici que l'ajout de la sémantique prend tout son sens. Notre approche novatrice permet non seulement la classification syntaxique, mais aussi la classification sémantique des relations en se basant sur une nouvelle notion : le contexte de la relation. Cette classification des relations en fonction de leur sémantique est particulièrement intéressante pour les humains, car elle leur permet de comprendre et de tirer des

conclusions à partir des données, ce qui est souvent difficile lorsqu'il s'agit de données non structurées. En effet, la sémantique se concentre sur le sens des mots et des phrases, et permet de mieux comprendre le sens des phrases et d'extraire des informations plus riches à partir du texte. Cela offre un avantage considérable pour l'analyse et l'interprétation des données, et ouvre de nouvelles perspectives pour la recherche en traitement automatique du langage naturel.

Notre méthode de classification sémantique est applicable dans plusieurs domaines, tels que la visualisation automatique des données statistiques extraites du texte. En effet, notre approche permet de résoudre efficacement ce problème en identifiant les relations sémantiques entre les différents concepts et en les organisant de manière cohérente. Dans la section suivante, nous présentons plus en détail les avantages et l'utilité de notre approche de classification des relations selon le contexte pour la visualisation automatique des données statistiques extraites du texte. Nous montrons comment elle permet de résoudre des problèmes complexes qui étaient jusqu'à présent difficiles à traiter avec les méthodes traditionnelles.

4.7.5. Post-Evaluation : visualisation automatique des données statistiques extraites d'un texte

Cette Post-Evaluation porte sur une application spécifique, consistant à utiliser l'approche de classification des relations selon le contexte pour la visualisation automatique des données statistiques extraites d'un texte.

1. **Motivation**

   La visualisation des données statistiques est un outil efficace pour communiquer les résultats d'enquêtes d'opinion, d'études épidémiologiques, de statistiques sur les habitudes de consommation, etc. Les représentations graphiques des données facilitent généralement le traitement de l'information par l'homme et permettent une lecture plus rapide de l'information (en agissant sur les formes, les directions, les couleurs...). Les données susceptibles d'être représentées graphiquement peuvent être structurées ou non structurées. Les données non structurées cachent souvent des informations importantes, voire vitales pour la société et les entreprises. Il faut donc beaucoup de travail pour extraire des informations précieuses des données non structurées. S'il est plus facile de comprendre un message par des données structurées, comme un tableau, que par un long texte narratif, il est encore plus facile de faire passer un message par un graphique que par un tableau. A notre avis, il est souvent très utile de synthétiser les données non structurées sous forme de représentations graphiques.

   Les données non structurées peuvent souvent cacher, entre autres, des données variables qui peuvent représenter des informations statistiques vitales. Prenons l'exemple d'un forum sur un réseau social où les gens expriment des opinions politiques lors d'une élection donnée. L'analyse statistique du forum peut donner des tendances sérieuses et fiables sur le résultat des élections. L'analyse statistique du forum est un exercice facile pour l'esprit humain tant que le volume de données

reste raisonnable pour être traité par un être humain. Les technologies de l'information peuvent facilement traiter d'énormes volumes de données. Cependant, l'extraction de nouvelles informations précieuses n'est facile pour un processus automatique que lorsque les données sont structurées.

L'expression des opinions n'est pas une donnée dénombrable, et il est donc nécessaire d'analyser les différentes opinions et de compter, pour chaque candidat ou parti politique, celles qui lui sont favorables. Deux problèmes fondamentaux se posent dans ce cas. Le premier consiste à reconnaître les variables qui constituent l'objet des calculs statistiques et le second consiste à construire une donnée structurée véhiculant une information statistique équivalente à celle cachée dans les données non structurées.

Le travail présenté dans cette section fait partie d'un travail de recherche qui est en phase d'expérimentation et d'évaluation dont l'objectif est la visualisation automatique des données statistiques extraites d'un document textuel non structuré. Pour pouvoir représenter graphiquement l'information statistique d'un texte non structuré, il faut d'abord identifier le contexte ainsi d'extraire et de classer les relations selon le type et le contexte. Par la suite d'extraire les variables statistiques, puis construire une donnée statistique structurée « équivalente » à celle encapsulée dans le texte. La représentation graphique ne pose pas de problème lorsque les données sont structurées dans un tableau par exemple. Cette étude de cas est basée sur **notre méthode de classification des relations selon le type et le contexte**.

2. **Méthode de visualisation automatique des données statistiques extraites d'un texte**

   Le traitement de données non structurées pour l'extraction de statistiques est un travail fastidieux. D'une part, il faut analyser le texte pour en identifier le contexte et classifier les relations selon le type et le contexte. D'autre part, il est nécessaire d'extraire les données statistiques, lorsqu'elles existent, de les traiter et d'effectuer des calculs afin de transformer les données non structurées en données structurées telles qu'un tableau. Le processus de l'approche de visualisation automatique des données statistiques est présenté dans la figure 4.16.

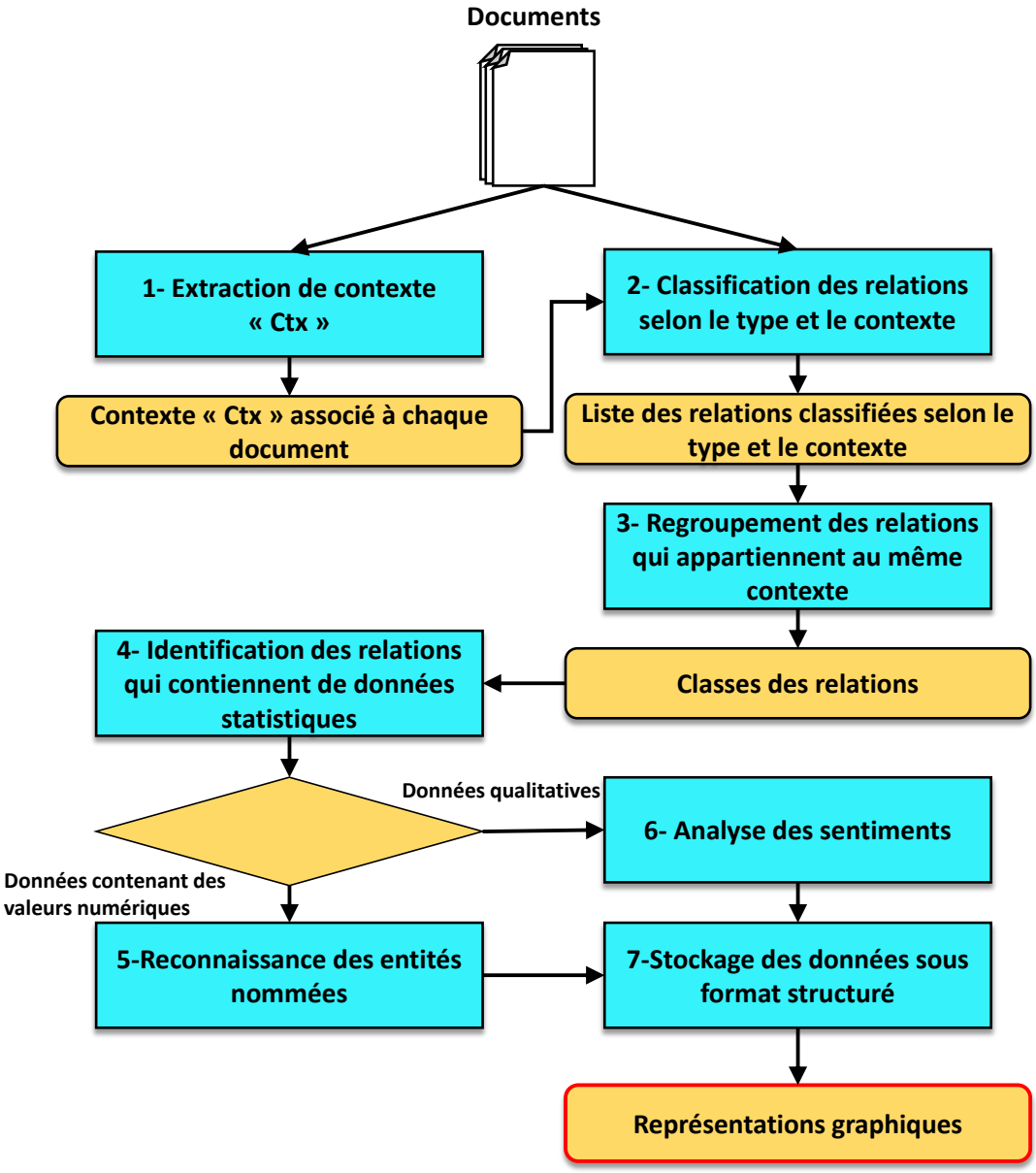


FIGURE 4.16. – Processus de l'approche de la visualisation automatique des données statistiques extraites d'un texte

Comme le montre la figure 4.16, la première étape du processus de l'approche de la visualisation automatique des données statistiques extraites du texte consiste à appliquer notre approche de classification des relations selon le type et le contexte. Cette approche permet non seulement de classer une relation syntaxiquement selon un type prédéfini mai aussi sémantiquement en annotant la relation par une phrase clé qui donne une brève idée du contenu du texte. Vu qu'on travaille sur un corpus qui contient plusieurs documents dans différents contextes, les relations qui sont annotées par le même contexte sont regroupées en une seule

entité appelée « classe de relations ». Une fois les différentes classes de relations extraites, le problème se pose pour l'identification des relations qui contiennent de données statistiques. Pour résoudre ce problème, deux cas de figure peuvent se présenter : Données contenant des valeurs numériques ou données qualitatives.

a) **Données contenant des valeurs numériques :** Le processus d'identification des données numériques se déroule en trois étapes principales : (1) *l'Extraction et regroupement des entités nommées*, (2) *le regroupement des entités nommées en classes de concepts*, et (3) *la construction d'une donnée structurée*.

Prenons l'exemple du document suivant (Figure 4.17) pour illustrer ce processus.

The 2017 French presidential election will occur in different French cities to allow the winner candidate to access the presidency. Until now, the candidate with the estimated highest scores of votings is Marine le Pen. She is credited with 35% of voting intentions in Paris. Polls give her 32% of voting intentions in Toulouse. However, she is ranked third in other cities such as Marseille with 18% of voting intentions. We asked the question "who would you voting for?" for two childhood friends Fabien Martin and Marcel Laurent. Although they lived in the same neighborhood, the two friends vote for two different candidates. Fabien says he will voting for François Fillon. Marcel is in favor of Marine Le Pen

FIGURE 4.17. – Document contenant des valeurs numériques

Comme le montre la figure 4.16, nous avons besoin d'une identification préalable du contexte du document et de déterminer ses mots-clés ainsi que la classification des relations selon le type et le contexte.

- **Extraction du contexte :** Voting intentions of the French presidential election
- **Identification des relations :** La figure 4.18 montre les relations extraites du document présenté dans la figure 4.17

1 [entity: '2017 French presidential election', relation: 'will occur in', object:'different French cities']
2 [entity: '2017 French presidential election', relation: 'allow', object:'the winner candidate to access the presidency']
3 [entity: 'candidate with the estimated highest scores of votings', relation: 'is', object:'Marine Le Pen']
4 [entity: 'Marine Le Pen', relation: 'credited with', object:'35% of voting intentions in Paris']
5 [entity: '35% of voting intentions', relation:'in', object:'Paris']
6 [entity: 'Polls', relation: 'give her', object:' 32% of voting intentions in Toulouse']
7 [entity: '32% of voting intentions', relation: 'in', object:'Toulouse']
8 [entity: 'she', relation: 'is ranked', object:' third in other cities such as Marseille with 18% of voting intentions.']
9 [entity: 'We', relation: 'asked', object:'the question "who would you voting for?" for two childhood friends Fabien Martin and Marcel Laurent']
10 [entity: 'they', relation: 'lived in', object:'the same neighborhood']
11 [entity: 'the two friends', relation: 'voting for', object: 'two different candidates']
12 [entity:'Fabien', relation: 'says', object:'he will voting for François Fillon']
13 [entity: 'he', relation: 'will vote for', object:'François Fillon']
14 [entity: 'Marcel', relation: 'is in', object:'favor of Marine Le Pen']

FIGURE 4.18. – Relations extraites du document présenté dans la figure 4.17 après application du système Stanford open-IE

- **Classification des relations selon le contexte et le type :** La figure 4.19 présente les relations classées par le type et le contexte

1 Content-Container(e1,e2) voting intentions of the French presidential election
2 Other voting intentions of the French presidential election
3 Other voting intentions of the French presidential election
4 Other voting intentions of the French presidential election
5 Content-Container(e1,e2) voting intentions of the French presidential election
6 Other voting intentions of the French presidential election
7 Content-Container(e1,e2) voting intentions of the French presidential election
8 Other voting intentions of the French presidential election
9 Other voting intentions of the French presidential election
10 Member-collection(e1,e2) Other
11 Other voting intentions of the French presidential election
12 Other voting intentions of the French presidential election
13 Other voting intentions of the French presidential election
14 Other voting intentions of the French presidential election

FIGURE 4.19. – Classification des relations selon le type et le contexte

Après avoir classé les relations selon le type et le contexte, nous procédons à l'identification des relations qui contiennent des variables numériques :

- **Extraction et regroupement des entités nommées :** Cette étape consiste à regrouper les entités nommées appartenant à une même « Classe de catégorie » telles que les noms de personnes, d'organisations, pourcentage, etc. Par exemple, dans l'exemple de la figure 4.17, trois Classes de catégorie ont été identifiées à savoir une classe pour les noms propres, une classe pour les lieux et une classe pour les pourcentages (Figure 4.20).

FIGURE 4.20. – Différentes classes de catégorie identifiées

- **Regroupement des entités nommées en classes de concepts :** Les entités nommées d'une même catégorie peuvent appartenir à des concepts différents. Un concept est une idée générale sur certaines entités ou classes d'entités distinctes. Considérons une classe de catégorie qui contient différents noms de candidats et d'électeurs, le processus doit être capable de différencier les deux concepts « candidat » et « électeur ». Par conséquent, la classe de catégories PERSON comprend les candidats à l'élection présidentielle et d'autres noms qui peuvent être considérés comme des électeurs comme le montre la figure 4.21. De même, la classe qui représente les pourcentages est nommée « voting intentions » et la classe des lieux est nommée « cities »

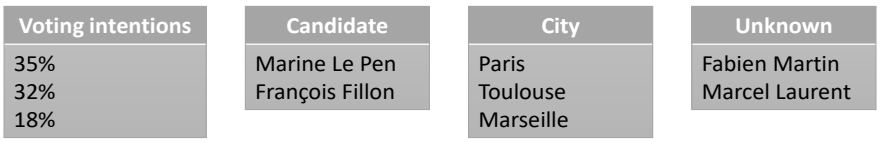


FIGURE 4.21. – Regroupement des entités nommées en classes de concepts

- **Construction d'une donnée structurée :** Dans cette méthode, les données structurées sont limitées aux « variables statistiques ». Ces variables statistiques sont exprimées sous forme de couples (valeur, variable), qui sont obtenues à partir de l'extraction des relations. Certaines associations sont multicritères, c'est-à-dire que les valeurs de la variable statistique varient elles-mêmes en fonction d'autres critères. Par exemple, le taux

de vote peut dépendre de la date, de la ville, etc. La valeur de la variable statistique est ainsi liée, par une seconde association à chacun de ces sous-critères : (valeur, critère). Cette association est également obtenue à partir de l'extraction des relations. Dans l'exemple présentée dans la figure 4.17, La paire (voting intentions, cities) est identifiée comme (valeur, critère). A partir de l'exemple présenté dans la figure 4.17 en appliquant le processus de construction d'une donnée structurée, on obtient le tableau 4.4.

TABLEAU 4.4. – Structuration des données statistiques

Candidate	Voting intentions	Cities
Marine Le Pen	36% 32% 18%	Paris Toulouse Marseille

b) **Données qualitatives :** Les données qualitatives apparaissent comme des déclarations descriptives qui peuvent être faites sur un sujet à partir d'observations, d'entretiens ou d'évaluations, par exemple un forum politique.

Le forum politique ne contient que des expressions d'intentions de vote et aucune donnée numérique. Chaque ligne de celui-ci est une expression d'opinion dans laquelle le « locuteur » soutient le candidat qu'il a choisi. Dans un texte non structuré où les gens expriment leurs opinions, il est important de déterminer l'auteur du texte d'une opinion ainsi que le contenu de cette opinion. L'opinion exprime généralement un sentiment positif ou négatif envers un contexte ou une situation particulière. La reconnaissance du « locuteur » permet d'éviter de compter plusieurs fois la même opinion. Plusieurs travaux ont été réalisés pour la reconnaissance du « locuteur » dans un texte. K. Glass et S. Bangay (GLASS et BANGAY 2007) proposent « une méthode naïve, basée sur la saillance, pour l'identification du locuteur dans les livres de fiction ». Ils ont montré qu'il est possible d'identifier différents locuteurs dans un texte sans utiliser de techniques complexes d'apprentissage automatique ou de logique.

L'analyse des sentiments (HASAN et ADJEROH 2011) (W. JIN, HO et SRIHARI 2009) , également connue sous le nom de « Opinion Mining », cherche à reconnaître et à caractériser les aspects de l'opinion ou de l'appréciation d'un auteur sur un sujet à partir d'informations contenues dans des textes écrits en langage naturel. Une opinion peut être exprimée par une personne ou une organisation sur un produit, un service, un sujet, un événement, etc. Nous utilisons le mot « entité » pour qualifier le sujet qui est évalué par l'expression d'un sentiment. Par exemple, dans un forum politique, l'entité est un candidat. L'analyse des sentiments se concentre aujourd'hui sur l'attribution d'une polarité aux expressions subjectives (mots et phrases qui expriment des opinions, des émotions, des sentiments, etc.) afin de décider de l'orientation d'une opinion ou de la valeur positive / négative / neutre de celle-ci dans un document (H. YU et

HATZIVASSILOGLOU 2003)(S.-M. KIM et HOVY 2004) . La plupart des techniques utilisées pour la classification du sentiment dans les textes en langues naturelles sont basées sur l'apprentissage supervisé, par exemple la classification bayésienne naïve, les machines à vecteurs de support (SVM), ou toute autre méthode d'apprentissage supervisé.

La représentation graphique d'un ensemble de données structurées représentées sous la forme d'un tableau ne présente aucune difficulté particulière. Tous les outils de traitement des données, les tableurs modernes et les applications statistiques le font très bien. Le seul petit problème à résoudre dans ce cas est de déterminer quelle variable sera représentée sur l'axe des ordonnées et celle qui occupera l'axe des abscisses. Ce problème est résolu à partir du contexte du document qui est déterminé de la même manière que les données sont structurées ou non.

4.8. Conclusion

Dans ce chapitre, nous avons proposé une nouvelle méthode de classification des relations selon le type de relations et le contexte du document, contexte extrait selon la méthode proposée dans le chapitre précédent.

Dans la première partie du chapitre, nous avons présenté cette méthode de classification de relations avec ses principales étapes : (1) Identification des relations, (2) Annotation de ces relations par le contexte du document et (3) Classification de ces relations par le type.

Dans une deuxième partie de ce chapitre, nous avons présenté nos résultats dans une étude expérimentale relative à la visualisation automatique des données statistiques extraites d'un texte. Pour se faire, nous avons opté pour les collections de test WikiContext et SemEval-2010 task8. Les résultats obtenus lors de notre expérimentation ont démontré la performance de notre système en termes de F1-score. En effet, concernant la classification selon le type, notre système a obtenu un taux de F1-score de 86.6% pour le jeu de données « SemEval-2010 task8 », et de 73.5% pour notre corpus « WikiContext ». Pour la tâche de classification selon le contexte, nous avons également obtenu des résultats encourageants, avec un F1-score de 75.7% pour « SemEval-2010 task8 », et de 78.1% pour « WikiContext ». Nous avons apporté une approche innovante, qui n'avait pas encore été explorée dans l'état de l'art, qui porte sur la classification contextuelle des relations. Cette nouvelle classification permet d'améliorer plusieurs domaines d'application, en offrant une meilleure compréhension des relations et leur contexte. En somme, notre étude a démontré la pertinence de la prise en compte du contexte dans le processus de classification de relations, ainsi que l'importance de l'apport de la nouvelle classification sémantique pour améliorer les performances des systèmes de classification.

Comme tout travail de recherche, le présent travail présente un certain nombre de limitations et est sujet à des extensions, des améliorations pouvant engendrer de nouvelles problématiques de recherche qui seront développées dans la conclusion

générale.

Conclusion générale

Dans cette conclusion, nous présentons tout d'abord une synthèse de nos travaux relatifs à la classification des relations selon le type et le contexte, et rappelons la nature de nos contributions. Ensuite nous précisons un certain nombre de limites de nos travaux de recherche. Enfin, nous introduisons quelques perspectives de recherche possibles.

Synthèse

Les supports d'information modernes véhiculent des données hétérogènes, en quantités tellement importantes que les moyens de traitement traditionnels deviennent inefficaces pour répondre aux besoins actuels. En plus de la quantité de données, la nature non structurée de ces données nécessite de nouvelles techniques de traitement intelligentes, efficaces et automatisées. Afin de produire des systèmes automatiques capables de gérer ces données et d'en extraire des connaissances pertinentes, un certain nombre de problèmes doivent être résolus, notamment l'extraction et la classification de relations à partir de données textuelles.

Les systèmes de classification de relations existants ne traitent que quelques types prédéfinis. Ces systèmes classent les relations en un certain nombre de types prédéfinis, essentiellement basés sur les aspects syntaxiques du texte. Ces méthodes négligent les aspects sémantiques et peuvent être trompeuses lors de la classification. Un des piliers de la sémantique est de prendre en compte le contexte dans la classification des relations. C'est précisément l'objectif principal que nous nous sommes fixés au début de ce travail de thèse. Notre objectif premier est de proposer une approche pour la classification des relations à partir de texte brut selon le type et le contexte.

Afin d'atteindre cet objectif, nous avons adopté une démarche qui se compose des cinq étapes suivantes :

1. Elaboration d'une revue systématique de la littérature sur les travaux d'extraction et de classification des relations.
2. Réalisation d'une revue systématique de la littérature pour identifier les travaux connexes qui ont abordé des aspects similaires à ceux liés à la notion de « contexte ». Notre étude s'est concentrée sur les méthodes proposées pour extraire le contenu à partir de documents textuels.
3. Proposition d'une contribution permettant de résoudre la problématique d'extraction du contexte à partir d'un document texte non structuré.

4. Proposition d'une contribution permettant de résoudre la problématique de classification des relations selon le type et le contexte.
5. Elaboration d'études expérimentales sur les contributions proposées.

   L'originalité de notre proposition réside dans notre vision selon laquelle l'extraction et classification se fait selon type et le contexte à la fois. En conséquence, les principales contributions scientifiques apportées par notre travail s'articulent autour des aspects suivants :

   a) La proposition d'***une nouvelle classification des méthodes en extraction et classification des relation0***s. Cette nouvelle classification devrait pouvoir aider chercheurs à appréhender l'état de l'art dans ce domaine afin de trouver le modèle adéquat à leurs besoins.
   b) La proposition d'***une méthode d'extraction de contexte***. Cette méthode permet d'extraire d'une manière précise et pertinente un contexte qui couvre le sujet d'un document. Elle se repose sur deux étapes principales : l'extraction des mots- clés et l'extraction du contexte du document. Nous avons identifié deux approches différentes de l'extraction du contexte : l'une *extractive* et l'autre *générative*, donnant lieu ainsi à deux variantes de la méthode.
   c) La proposition d'une nouvelle ***méthode de classification de relations binaires*** qui consiste alors à classer les relations en fonction de leurs types tout en tenant compte du contexte. Elle permet d'ajouter de capacités sémantiques et d'une contextualisation plus précise aux relations extraites de documents texte non structurés. Cette méthode utilise les contextes extraits par la méthode précédente pour réaliser un filtrage afin d'éliminer les relations qui ne sont pas significatives pour le contexte du document.

   Un cadre d'expérimentation basé sur ces propositions a été développé. Deux études expérimentales ont été élaborées sur la méthode de d'extraction de contexte et la méthode de classification des relations selon le type et le contexte. L'objectif principal des expérimentations menées a consisté à montrer d'une part la faisabilité des méthodes proposées, et d'autre part à évaluer chacune de ses différentes contributions et leurs limites.

Limitations

Notre travail de recherche présente un certain nombre de limitations que nous présentons.

La première limitation concerne le corpus WikiContext. Ce corpus constitue une collection de documents et contextes, que nous avons construit pour l'extraction du contexte d'un document textuel particulier pour la classification des relations selon le contexte. Ce corpus est de petite taille ce qui est insuffisant pour une validité solide de notre contribution. En effet, ce corpus ne contient que 600

documents répartis sur 30 contextes. De plus, l'application du processus d'augmentation de données n'a pas amélioré significativement l'entrainement de notre modèle de génération de l'étiquette de contexte.

La deuxième limitation concerne la validation expérimentale de notre méthode de classification des relations selon le type et le contexte est effectuée sur uniquement deux corpus de données SemEval 2010 Task 8 et le corpus WikiContext. Afin d'enrichir notre protocole de validation de la méthode de classification de relations selon le type et le contexte, d'autres expérimentations sur plusieurs autres corpus de données ayant différents contextes est nécessaire.

Enfin, une autre limitation de notre travail réside dans le fait que nous n'avons pas pu expérimenter et évaluer la solution de visualisation des données statistiques en appliquant notre méthode de classification des relations selon le type et le contexte et lancer sa mise en œuvre avant la fin de cette thèse. Une telle mise en œuvre aurait pu montrer l'intérêt et l'apport de notre méthode dans l'amélioration d'un tel processus. Toutefois comme nous l'avons déjà évoqué, nous considérons que cette limite est à relativiser car le développement de cette solution devrait être assuré dans les mois à venir.

Perspectives

Les limites de nos travaux de recherche évoqués précédemment ainsi que les résultats satisfaisants obtenus par les méthodes proposées lors des expérimentations ouvrent de nouvelles perspectives de recherche.

Perspectives à court terme

En réponses aux limitations précédemment évoquées :

— Il nous semble tout d'abord nécessaire d'augmenter le nombre de documents du corpus WikiContext que nous avons constitué en tant que ressource pour la méthode d'extraction de contexte d'un document textuel.

— De même pour une validation expérimentale plus solide de notre méthode de classification des relations selon le type et le contexte d'un document textuel, il nous apparaît nécessaire mener de nouvelles expérimentations sur d'autres corpus de données que les deux corpus de données « SemEval 2010 Task 8 » et « WikiContext » utilisés dans notre travail. Par exemple nous pourrions utiliser les corpus « TAC Relation Extraction Dataset » et Knowledge Base Population (KBP).

Perspectives à moyen terme

Les solutions proposées dans ce travail de thèse pourraient être étendues selon plusieurs directions :

— Comme première extension immédiate, pourrait être réalisée une étude plus systématique pour d'autres langues pour lesquelles il existe moins d'outils de traitement automatique des langues que la langue Anglaise, comme le Français, l'espagnol, etc...Nous pourrions aussi nous intéresser à la classification des relations dans des langues plus complexes comme l'Arabe.

— Une deuxième extension à moyen terme consisterait à essayer de classer selon le type et le contexte des relations n-aires et non plus binaires comme traitées dans notre recherche, c'est-à-dire des relations impliquant plus de deux entités, caractérisant un événement.

Perspectives à long terme

Les solutions proposées dans ce travail de thèse pourraient être étendues selon plusieurs directions :

— Il serait intéressant d'évaluer l'intérêt d'intégrer en amont notre méthode dans des tâches de traitement automatique des langues plus complexes afin de les améliorer, comme par exemple le résumé automatique de texte, l'extraction de topic, la classification de documents utiles pour la classification des brevets et des marques, ainsi que le développement de Systèmes de Gestion Electronique de Documents intelligents.

— Pourrait être aussi envisager l'usage de notre méthode pour d'autres tâches plus spécifiques comme la classification de documents utiles pour la classification des brevets et des marques, le développement de Systèmes de Gestion Electronique de Documents intelligents, ou encore l'extraction et formalisation automatique de relations de type Entité-Association à partir de spécifications de projets informatique.

Bibliographie

[Aba+21a] Ammar Kamal ABASI, Ahamad Tajudin KHADER, Mohammed Azmi AL-BETAR, Syibrah NAIM, Zaid Abdi Alkareem ALYASSERI et al. « An ensemble topic extraction approach based on optimization clusters using hybrid multi-verse optimizer for scientific publications ». In : *Journal of Ambient Intelligence and Humanized Computing* 12.2 (2021), p. 2765-2801 (cf. p. 208).

[Aba+21b] Ammar Kamal ABASI, Ahamad Tajudin KHADER, Mohammed Azmi AL-BETAR, Syibrah NAIM, Sharif Naser MAKHADMEH et al. « A novel ensemble statistical topic extraction method for scientific publications based on optimization clustering ». In : *Multimedia Tools and Applications* 80.1 (2021), p. 37-82 (cf. p. 208).

[Akb+12] Alan AKBIK, Larysa VISENGERIYEVA, Priska HERGER et al. « Unsupervised discovery of relations and discriminative extraction patterns ». In : *Proceedings of COLING 2012*. 2012, p. 17-32 (cf. p. 198).

[APN21a] Ghada ALFATTNI, Niels PEEK et Goran NENADIC. « Attention-based bidirectional long short-term memory networks for extracting temporal relationships from clinical discharge summaries ». In : *Journal of Biomedical Informatics* 123 (2021), p. 103915 (cf. p. 27).

[APN21b] Ghada ALFATTNI, Niels PEEK et Goran NENADIC. « Attention-based bidirectional long short-term memory networks for extracting temporal relationships from clinical discharge summaries ». In : *Journal of Biomedical Informatics* 123 (2021), p. 103915 (cf. p. 196).

[Alf+12] Enrique ALFONSECA, Katja FILIPPOVA, Jean-Yves DELORT et al. « Pattern learning for relation extraction with a hierarchical topic model ». In : *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*. 2012, p. 54-59 (cf. p. 18).

[Ana+18] Shakil Ashraful ANAM, AM Muntasir RAHMAN, Nasif Noor SALEHEEN et al. « Automatic text summarization using fuzzy c-means clustering ». In : *2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*. IEEE. 2018, p. 180-184 (cf. p. 205).

[APM15] Gabor ANGELI, Melvin Jose Johnson PREMKUMAR et Christopher D MANNING. « Leveraging linguistic structure for open domain information extraction ». In : *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*. 2015, p. 344-354 (cf. p. 126, 127).

[ACS13] MA ANGROSH, Stephen CRANEFIELD et Nigel STANGER. « Conditional random field based sentence context identification : enhancing citation services for the research community ». In : *Proceedings of the First Australasian Web Conference-Volume 144*. 2013, p. 59-68 (cf. p. 206).

[ATV13] Nguyen Kim ANH, Nguyen The TAM et Ngo VAN LINH. « Document clustering using dirichlet process mixture model of von mises-fisher distributions ». In : *Proceedings of the Fourth Symposium on Information and Communication Technology*. 2013, p. 131-138 (cf. p. 206).

[AD18] Chandrakala ARYA et Sanjay k DWIVEDI. « Keyphrase Extraction of News Web Pages ». In : *International Journal of Education and Management Engineering* 8.1 (2018), p. 48 (cf. p. 208).

[Baa01] R Harald BAAYEN. *Word frequency distributions*. T. 18. Springer Science & Business Media, 2001 (cf. p. 56).

[BAM19] P Samson Anosh BABU, Chandra Sekhar Rao ANNAVARAPU et Abhilash MOHAPATRA. « A novel method for next-generation sequence data analysis using PLSA topic modeling technique ». In : *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*. IEEE. 2019, p. 1-6 (cf. p. 68).

[BB07] Nguyen BACH et Sameer BADASKAR. « A review of relation extraction ». In : *Literature review for Language and Statistics II* 2 (2007), p. 1-15 (cf. p. 2).

[Bad+21] Ahmed BADAWY, Jesus A FISTEUS, Tarek M MAHMOUD et al. « Topic Extraction and Interactive Knowledge Graphs for Learning Resources ». In : *Sustainability* 14.1 (2021), p. 226 (cf. p. 208).

[BCB14] Dzmitry BAHDANAU, Kyunghyun CHO et Yoshua BENGIO. « Neural machine translation by jointly learning to align and translate ». In : *arXiv preprint arXiv :1409.0473* (2014) (cf. p. 26).

[BS14] Marion BARANES et Benoıt SAGOT. « Normalisation de textes par analogie : le cas des mots inconnus ». In : *TALN-Traitement Automatique du Langage Naturel*. 2014, p. 137-148 (cf. p. 61).

[Bee+] J BEEL, B GIPP, A SHAKER et al. « Extracting Titles from Scientific PDF Documents by Analyzing Style Information ». In : *International Conference on Theory and Practice of Digital Libraries*, p. 413-416 (cf. p. 207).

[Bee+13] Joeran BEEL, Stefan LANGER, Marcel GENZMEHR et al. « Docear's PDF inspector : title extraction from PDF files ». In : *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*. 2013, p. 443-444 (cf. p. 206).

[Bee+10] Jöran BEEL, Bela GIPP, Ammar SHAKER et al. « SciPlore Xtract : extracting titles from scientific PDF documents by analyzing style information (Font Size) ». In : *International Conference on Theory and Practice of Digital Libraries*. Springer. 2010, p. 413-416 (cf. p. 206).

[Bhi+07] Manish A BHIDE, Ajay GUPTA, Rahul GUPTA et al. « Liptus : Associating structured and unstructured information in a banking environment ». In : *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*. 2007, p. 915-924 (cf. p. 11).

[BB05] Patrick BLACKBURN et Johannes BOS. *Representation and inference for natural language : A first course in computational semantics*. Center for the Study of Language et Information Stanford, 2005 (cf. p. 78).

[Bla+04] C BLASCHKE et al. « Biocreative : Critical Assessment for Information Extraction in Biology ». In : *Granada, Spain (URL : http ://www. pdg. cnb. uam. es/Biolink/Workshop_BioCreative_04/handout/index. html)* (2004) (cf. p. 11).

[Bor05] Christian BORGELT. « An Implementation of the FP-growth Algorithm ». In : *Proceedings of the 1st international workshop on open source data mining : frequent pattern mining implementations*. 2005, p. 1-5 (cf. p. 89).

[Bri98] Sergey BRIN. « Extracting patterns and relations from the world wide web ». In : *International workshop on the world wide web and databases*. Springer. 1998, p. 172-183 (cf. p. 12).

[BP98] Sergey BRIN et Lawrence PAGE. « The anatomy of a large-scale hypertextual web search engine ». In : *Computer networks and ISDN systems* 30.1-7 (1998), p. 107-117 (cf. p. 57).

[Bun+05] Razvan BUNESCU, Ruifang GE, Rohit J KATE et al. « Comparative experiments on learning information extractors for proteins and their interactions ». In : *Artificial intelligence in medicine* 33.2 (2005), p. 139-155 (cf. p. 11).

[CBE11] Michael J CAFARELLA, Michele BANKO et Oren ETZIONI. *Open information extraction from the Web*. US Patent 7,877,343. Jan. 2011 (cf. p. 9).

[CZW16a] Rui CAI, Xiaodong ZHANG et Houfeng WANG. « Bidirectional recurrent convolutional neural network for relation classification ». In : *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. 2016, p. 756-765 (cf. p. 25, 47, 200).

[CZW16b] Rui CAI, Xiaodong ZHANG et Houfeng WANG. « Bidirectional recurrent convolutional neural network for relation classification ». In : *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. 2016, p. 756-765 (cf. p. 196).

[CZW16c] Rui CAI, Xiaodong ZHANG et Houfeng WANG. « Bidirectional recurrent convolutional neural network for relation classification ». In : *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. 2016, p. 756-765 (cf. p. 196).

[Cai+05] Yuhan CAI, Xin Luna DONG, Alon HALEVY et al. « Personal information management with SEMEX ». In : *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. 2005, p. 921-923 (cf. p. 11).

[Cha+06] Venkatesan T CHAKARAVARTHY, Himanshu GUPTA, Prasan ROY et al. « Efficiently linking text documents with relevant structured information ». In : *Proceedings of the 32nd international conference on Very large data bases*. 2006, p. 667-678 (cf. p. 11).

[CMN05] Soumen CHAKRABARTI, Jeetendra MIRCHANDANI et Arnab NANDI. « Spin : searching personal information networks ». In : *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. 2005, p. 674-674 (cf. p. 11).

[CL21] Chih-Yao CHEN et Cheng-Te LI. « ZS-BERT : Towards Zero-Shot Relation Extraction with Attribute Representation Learning ». In : *arXiv preprint arXiv :2104.04697* (2021) (cf. p. 201).

[Che+20] Tiansi CHEN, Shihao JI, Zhijiang GUO et al. « Graph-based Relation Extraction with Multimodal Transformer and Hierarchical Graph Pooling ». In : *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, p. 3019-3030 (cf. p. 29).

[Che+12] Yan CHEN, Yang YANG, Huisan ZHANG et al. « A topic detection method based on Semantic Dependency Distance and PLSA ». In : *Proceedings of the 2012 IEEE 16th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE. 2012, p. 703-708 (cf. p. 68).

[CM17] Fei CHENG et Yusuke MIYAO. « Classifying temporal relations by bidirectional LSTM over dependency paths ». In : *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*. 2017, p. 1-6 (cf. p. 202).

[Che15] Peter A CHEW. « 'Linguistics-Lite'Topic Extraction from Multilingual Social Media Data ». In : *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*. Springer. 2015, p. 276-282 (cf. p. 206).

[Chi98] Nancy A CHINCHOR. *Overview of muc-7/met-2*. Rapp. tech. SCIENCE APPLICATIONS INTERNATIONAL CORP SAN DIEGO CA, 1998 (cf. p. 10).

[CP18] Anupama CHINGACHAM et Denis PAPERNO. « Generalizing Representations of Lexical Semantic Relations. » In : *CLiC-it*. 2018 (cf. p. 196).

[Cho+14] Kyunghyun CHO, Bart VAN MERRIËNBOER, Caglar GULCEHRE et al. « Learning phrase representations using RNN encoder-decoder for statistical machine translation ». In : *arXiv preprint arXiv :1406.1078* (2014) (cf. p. 133).

[CMT05] William W COHEN, Einat MINKOV et Anthony TOMASIC. « Learning to Understand Web Site Update Requests. » In : *IJCAI*. 2005, p. 1028-1033 (cf. p. 11).

[CHL20] Trevor COHN, Yulan HE et Yang LIU. « Findings of the Association for Computational Linguistics : EMNLP 2020 ». In : *Findings of the Association for Computational Linguistics : EMNLP 2020*. 2020 (cf. p. 201).

[Col+11] Ronan COLLOBERT, Jason WESTON, Léon BOTTOU et al. « Natural language processing (almost) from scratch ». In : *Journal of machine learning research* 12.ARTICLE (2011), p. 2493-2537 (cf. p. 78).

[CB12] Francisco COSTA et António BRANCO. « Aspectual type and temporal relation classification ». In : *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 2012, p. 266-275 (cf. p. 201).

[Cro+14] Richard CROUCH, Martin Henk van den BERG, Franco SALVETTI et al. *Coreference resolution in an ambiguity-sensitive natural language processing system*. US Patent 8,712,758. Avr. 2014 (cf. p. 79).

[Dai+18a] Qin DAI, Naoya INOUE, Paul REISERT et al. « Improving scientific relation classification with task specific supersense ». In : *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*. 2018 (cf. p. 202).

[Dai+18b] Yuanfei DAI, Wenzhong GUO, Xing CHEN et al. « Relation classification via LSTMs based on sequence and tree structure ». In : *IEEE Access* 6 (2018), p. 64927-64937 (cf. p. 199).

[Dai+18c] Yuanfei DAI, Wenzhong GUO, Xing CHEN et al. « Relation classification via LSTMs based on sequence and tree structure ». In : *IEEE Access* 6 (2018), p. 64927-64937 (cf. p. 199).

[DR08] Dmitry DAVIDOV et Ari RAPPOPORT. « Classification of semantic relationships between nominals using pattern clusters ». In : *Proceedings of ACL-08 : HLT*. 2008, p. 227-235 (cf. p. 197).

[DL13] Oier Lopez DE LACALLE et Mirella LAPATA. « Unsupervised relation extraction with general domain knowledge ». In : *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 2013, p. 415-425 (cf. p. 198).

[DG13] Luciano DEL CORRO et Rainer GEMULLA. « Clausie : clause-based open information extraction ». In : *Proceedings of the 22nd international conference on World Wide Web*. 2013, p. 355-366 (cf. p. 126-128).

[DW20] Kun DENG et Shaochun WU. « Improving relation classification by incorporating dependency and semantic information ». In : *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2020, p. 1-6 (cf. p. 197).

[Dev+18a] Jacob DEVLIN, Ming-Wei CHANG, Kenton LEE et al. « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding ». In : *arXiv preprint arXiv :1810.04805* (2018) (cf. p. 28).

[Dev+18b] Jacob DEVLIN, Ming-Wei CHANG, Kenton LEE et al. « Bert : Pre-training of deep bidirectional transformers for language understanding ». In : *arXiv preprint arXiv :1810.04805* (2018) (cf. p. 130).

[DD16] Kartik DHIWAR et Abhishek Kumar DEWANGAN. « A Review of Relation Classification with Convolutional Neural Network ». In : *International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Print ISSN* (2016), p. 2395-1990 (cf. p. 2).

[DJ19] Juncheng DING et Wei JIN. « A Prior Setting that Improves LDA in both Document Representation and Topic Extraction ». In : *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2019, p. 1-8 (cf. p. 205).

[DR10] Quang DO et Dan ROTH. « Constraints based taxonomic relation classification ». In : *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. 2010, p. 1099-1109 (cf. p. 197).

[Dod+04] George R DODDINGTON, Alexis MITCHELL, Mark A PRZYBOCKI et al. « The automatic content extraction (ace) program-tasks, data, and evaluation. » In : *Lrec*. T. 2. 1. Lisbon. 2004, p. 837-840 (cf. p. 10).

[DKJ20] Sri Nath DWIVEDI, Harish KARNICK et Renu JAIN. « Relation Classification : How Well Do Neural Network Approaches Work ? » In : *Iberoamerican Knowledge Graphs and Semantic Web Conference*. Springer. 2020, p. 102-112 (cf. p. 200).

[EMS10] Andrea ESULI, Diego MARCHEGGIANI et Fabrizio SEBASTIANI. « ISTI@ SemEval-2 Task 8 : Boosting-Based Multiway Relation Classification ». In : *Proceedings of the 5th International Workshop on Semantic Evaluation*. 2010, p. 218-221 (cf. p. 200).

[Etz+08a] Oren ETZIONI, Michele BANKO, Stephen SODERLAND et al. « Open information extraction from the web ». In : *Communications of the ACM* 51.12 (2008), p. 68-74 (cf. p. 18).

[Etz+08b] Oren ETZIONI, Michele BANKO, Stephen SODERLAND et al. « Open information extraction from the web ». In : *Communications of the ACM* 51.12 (2008), p. 68-74 (cf. p. 126).

[Etz+11] Oren ETZIONI, Anthony FADER, Janara CHRISTENSEN et al. « Open information extraction : The second generation ». In : *Twenty-Second International Joint Conference on Artificial Intelligence*. 2011 (cf. p. 126).

[FSE11a] Anthony FADER, Stephen SODERLAND et Oren ETZIONI. « Identifying relations for open information extraction ». In : *Proceedings of the 2011 conference on empirical methods in natural language processing*. 2011, p. 1535-1545 (cf. p. 126, 127).

[FSE11b] Anthony FADER, Stephen SODERLAND et Oren ETZIONI. « Open relation extraction with bootstrapping ». In : *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2011, p. 280-290 (cf. p. 32, 33).

[FF16] Jun FENG et Yu FANG. « Research on hot topic discovery technology of micro-blog based on biterm topic model ». In : *International Conference on Geo-Informatics in Resource Management and Sustainable Ecosystem*. Springer. 2016, p. 234-244 (cf. p. 206).

[Fen+18] Jun FENG, Minlie HUANG, Li ZHAO et al. « Reinforcement learning for relation classification from noisy data ». In : *Proceedings of the aaai conference on artificial intelligence*. T. 32. 1. 2018 (cf. p. 32, 198).

[Fra+17] Valentina FRANZONI, Yuanxi LI, Paolo MENGONI et al. « Clustering facebook for biased context extraction ». In : *International Conference on Computational Science and Its Applications*. Springer. 2017, p. 717-729 (cf. p. 206).

[FM15] Valentina FRANZONI et Alfredo MILANI. « Semantic context extraction from collaborative networks ». In : *2015 IEEE 19th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE. 2015, p. 131-136 (cf. p. 205).

[Gáb+18a] Kata GÁBOR, Davide BUSCALDI, Anne-Kathrin SCHUMANN et al. « Semeval-2018 task 7 : Semantic relation extraction and classification in scientific papers ». In : *Proceedings of The 12th International Workshop on Semantic Evaluation*. 2018, p. 679-688 (cf. p. 197).

[Gáb+18b] Kata GÁBOR, Davide BUSCALDI, Anne-Kathrin SCHUMANN et al. « Semeval-2018 task 7 : Semantic relation extraction and classification in scientific papers ». In : *Proceedings of The 12th International Workshop on Semantic Evaluation*. 2018, p. 679-688 (cf. p. 202).

[GF16] Najlah GALI et Pasi FRÄNTI. « Content-based Title Extraction from Web Page. » In : *WEBIST (2)*. 2016, p. 204-210 (cf. p. 207).

[GMF16] Najlah GALI, Radu MARIESCU-ISTODOR et Pasi FRÄNTI. « Similarity measures for title matching ». In : *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE. 2016, p. 1548-1553 (cf. p. 206).

[GMF17] Najlah GALI, Radu MARIESCU-ISTODOR et Pasi FRÄNTI. « Using linguistic features to automatically extract web page title ». In : *Expert Systems with Applications* 79 (2017), p. 296-312 (cf. p. 205).

[Gan13] Aldo GANGEMI. « A comparison of knowledge extraction tools for the semantic web ». In : *Extended semantic web conference*. Springer. 2013, p. 351-366 (cf. p. 205).

[Gao+19a] Tianyu GAO, Xu HAN, Zhiyuan LIU et al. « Hybrid attention-based prototypical networks for noisy few-shot relation classification ». In : *Proceedings of the AAAI Conference on Artificial Intelligence*. T. 33. 01. 2019, p. 6407-6414 (cf. p. 198).

[Gao+20] Tianyu GAO, Xu HAN, Ruobing XIE et al. « Neural snowball for few-shot relation learning ». In : *Proceedings of the AAAI Conference on Artificial Intelligence*. T. 34. 05. 2020, p. 7772-7779 (cf. p. 201).

[Gao+19b] Tianyu GAO, Xu HAN, Hao ZHU et al. « FewRel 2.0 : Towards more challenging few-shot relation classification ». In : *arXiv preprint arXiv :1910.07124* (2019) (cf. p. 198).

[Gen+20] Xiaoqing GENG, Xiwen CHEN, Kenny Q ZHU et al. « MICK : A meta-learning framework for few-shot relation classification with small training data ». In : *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020, p. 415-424 (cf. p. 200).

[Gha+06] Rayid GHANI, Katharina PROBST, Yan LIU et al. « Text mining for product attribute extraction ». In : *ACM SIGKDD Explorations Newsletter* 8.1 (2006), p. 41-48 (cf. p. 11).

[Gir+19] Praveen Kumar Badimala GIRIDHARA, Chinmaya MISHRA, Reddy Kumar Modam VENKATARAMANA et al. « A Study of Various Text Augmentation Techniques for Relation Classification in Free Text. » In : *ICPRAM* 3 (2019), p. 5 (cf. p. 196).

[GB07] Kevin GLASS et Shaun BANGAY. « A naive salience-based method for speaker identification in fiction books ». In : *Proceedings of the 18th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA'07)*. 2007, p. 1-6 (cf. p. 155).

[Gou12] Ronald GOULD. *Graph theory*. Courier Corporation, 2012 (cf. p. 61).

[Goy+13] Rahul GOYAL, Ravee MALLA, Amitabha BAGCHI et al. « Esthete : a news browsing system to visualize the context and evolution of news stories ». In : *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2013, p. 2529-2532 (cf. p. 207).

[GQZ16] Jinghang GU, Longhua QIAN et Guodong ZHOU. « Chemical-induced disease relation extraction with various linguistic features ». In : *Database* 2016 (2016) (cf. p. 15).

[GM17] Ramzi GUETARI et Maha MALLEK. « Graphics on demand : the automatic data visualization on the WEB ». In : *Advances in Science, Technology and Engineering Systems Journal (ASTESJ)* 2.3 (2017), p. 951-957 (cf. p. 78).

[Guo+16a] Nan GUO, Yuan HE, ChunGang YAN et al. « Multi-level topical text categorization with Wikipedia ». In : *Proceedings of the 9th International Conference on Utility and Cloud Computing*. 2016, p. 343-352 (cf. p. 112, 114, 115).

[Guo+16b] Nan GUO, Yuan HE, ChunGang YAN et al. « Multi-level topical text categorization with Wikipedia ». In : *Proceedings of the 9th International Conference on Utility and Cloud Computing*. 2016, p. 343-352 (cf. p. 207).

[Guo+19a] Xiaoyu GUO, Hui ZHANG, Haijun YANG et al. « A single attention-based combination of CNN and RNN for relation classification ». In : *IEEE Access* 7 (2019), p. 12467-12475 (cf. p. 30, 48, 132, 199).

[Guo+19b] Xiaoyu GUO, Hui ZHANG, Haijun YANG et al. « A single attention-based combination of CNN and RNN for relation classification ». In : *IEEE Access* 7 (2019), p. 12467-12475 (cf. p. 134).

[GDS07] Rahul GUPTA, Ajit A DIWAN et Sunita SARAWAGI. « Efficient inference with cardinality-based clique potentials ». In : *Proceedings of the 24th international conference on Machine learning*. 2007, p. 329-336 (cf. p. 10).

[Ham50] Richard W HAMMING. « Error detecting and error correcting codes ». In : *The Bell system technical journal* 29.2 (1950), p. 147-160 (cf. p. 91).

[Han+18a] Xu HAN, Hao ZHU, Pengfei YU, Ziyang LIU et al. « FewRel : A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation ». In : *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018, p. 4803-4813 (cf. p. 31).

[Han+18b] Xu HAN, Hao ZHU, Pengfei YU, Ziyun WANG et al. « Fewrel : A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation ». In : *arXiv preprint arXiv :1810.10147* (2018) (cf. p. 197).

[HA11] Shamimul SM HASAN et Donald A ADJEROH. « Detecting human sentiment from text using a proximity-based approach ». In : *Journal of Digital Information Management* 9.5 (2011), p. 206-213 (cf. p. 155).

[HSG04] Takaaki HASEGAWA, Satoshi SEKINE et Ralph GRISHMAN. « Discovering relations among named entities from large corpora ». In : *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*. 2004, p. 415-422 (cf. p. 10).

[Heg07] Markus HEGLAND. « The apriori algorithm–a tutorial ». In : *Mathematics and computation in imaging science and information processing* (2007), p. 209-262 (cf. p. 89).

[Hen+19] Iris HENDRICKX, Su Nam KIM, Zornitsa KOZAREVA et al. « Semeval-2010 task 8 : Multi-way classification of semantic relations between pairs of nominals ». In : *arXiv preprint arXiv :1911.10422* (2019) (cf. p. 20).

[Het+18] Lena HETTINGER, Alexander DALLMANN, Albin ZEHE et al. « Claire at SemEval-2018 task 7 : classification of relations using embeddings ». In : *Proceedings of The 12th International Workshop on Semantic Evaluation*. 2018, p. 836-841 (cf. p. 202).

[Hil16] Martin HILBERT. « Big data for development : A review of promises and challenges ». In : *Development Policy Review* 34.1 (2016), p. 135-174 (cf. p. 2).

[Hol11] Andreas HOLZINGER. « Weakly structured data in health-informatics : The challenge for human-computer interaction ». In : *Proceedings of INTERACT 2011 Workshop : Promoting and supporting healthy living by design.* . 2011, p. 5-7 (cf. p. 1).

[Hol12] Andreas HOLZINGER. « On knowledge discovery and interactive intelligent visualization of biomedical data ». In : *Proceedings of the Int. Conf. on Data Technologies and Applications DATA*. 2012, p. 5-16 (cf. p. 1).

[Hol+13] Andreas HOLZINGER, Christof STOCKER, Bernhard OFNER et al. « Combining HCI, natural language processing, and knowledge discovery-potential of IBM content analytics as an assistive technology in the biomedical field ». In : *International Workshop on Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*. Springer. 2013, p. 13-24 (cf. p. 1).

[HZ12a] Bao HONG et Deng ZHEN. « An extended keyword extraction method ». In : *Physics Procedia* 24 (2012), p. 1120-1127 (cf. p. 79).

[HZ12b] Bao HONG et Deng ZHEN. « An extended keyword extraction method ». In : *Physics Procedia* 24 (2012), p. 1120-1127 (cf. p. 207).

[Hu+19] Linmei HU, Luhao ZHANG, Chuan SHI et al. « Improving distantly-supervised relation extraction with joint label embedding ». In : *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, p. 3821-3829 (cf. p. 28, 29).

[Hua+21] Yuan HUANG, Zhixing LI, Wei DENG et al. « D-BERT : Incorporating dependency-based attention into BERT for relation extraction ». In : *CAAI Transactions on Intelligence Technology* 6.4 (2021), p. 417-425 (cf. p. 200).

[Hui+20] Bei HUI, Liang LIU, Jia CHEN et al. « Few-shot relation classification by context attention-based prototypical networks with BERT ». In : *EURASIP Journal on Wireless Communications and Networking* 2020.1 (2020), p. 1-17 (cf. p. 198).

[Hul03] Anette HULTH. « Improved automatic keyword extraction given more linguistic knowledge ». In : *Proceedings of the 2003 conference on Empirical methods in natural language processing*. 2003, p. 216-223 (cf. p. 58).

[Ins23] Qatar Computing Research INSTITUTE. *Rayyan*. https://rayyan.qcri.org/. Accessed : 2 April 2023 (cf. p. 38).

[Jac01] Paul JACCARD. « Étude comparative de la distribution florale dans une portion des Alpes et des Jura ». In : *Bull Soc Vaudoise Sci Nat* 37 (1901), p. 547-579 (cf. p. 91).

[JR13] Y JAHNAVI et Y RADHIKA. « Hot topic extraction based on frequency, position, scattering and topical weight for time sliced news documents ». In : *2013 15th International Conference on Advanced Computing Technologies (ICACT)*. IEEE. 2013, p. 1-6 (cf. p. 205).

[Jam+09] M Shoaib JAMEEL, Nilesh SINGH, Nitin Kumar SINGH et al. « An Intelligent Automatic Text Summarizer ». In : *Proceedings of the First International Conference on Intelligent Human Computer Interaction*. Springer. 2009, p. 223-230 (cf. p. 205).

[Jam11] Emily JAMISON. « Using grammar rule clusters for semantic relation classification ». In : *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*. 2011, p. 46-53 (cf. p. 200).

[JA02] Martin JANSCHE et Steven ABNEY. « Information extraction from voicemail transcripts ». In : *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*. 2002, p. 320-327 (cf. p. 11).

[Jar19] Marwan JARRAH. « Factivity and subject extraction in Jordanian Arabic ». In : *Lingua* 219 (2019), p. 106-126 (cf. p. 208).

[Jia+18a] Shengbin JIA, Shijia E, Maozhen LI et al. « Chinese open relation extraction and knowledge base establishment ». In : *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 17.3 (2018), p. 1-22 (cf. p. 203).

[Jia+19] Haixin JIANG, Rui ZHOU, Limeng ZHANG et al. « Sentence level topic models for associated topics extraction ». In : *World Wide Web* 22.6 (2019), p. 2545-2560 (cf. p. 206).

[Jia+20] Ming JIANG, Jennifer D'SOUZA, Sören AUER et al. « Improving scholarly knowledge representation : evaluating BERT-based models for scientific relation classification ». In : *International Conference on Asian Digital Libraries*. Springer. 2020, p. 3-19 (cf. p. 202).

[Jia+18b] Yishun JIANG, Gongqing WU, Chenyang BU et al. « Chinese entity relation extraction based on syntactic features ». In : *2018 IEEE International Conference on Big Knowledge (ICBK)*. IEEE. 2018, p. 99-105 (cf. p. 201).

[Jin+18a] Di JIN, Franck DERNONCOURT, Elena SERGEEVA et al. « MIT-MEDG at SemEval-2018 task 7 : Semantic relation classification via convolution neural network ». In : *Proceedings of the 12th international workshop on semantic evaluation*. 2018, p. 798-804 (cf. p. 22, 202).

[Jin+18b] Di JIN, Franck DERNONCOURT, Elena SERGEEVA et al. « MIT-MEDG at SemEval-2018 task 7 : Semantic relation classification via convolution neural network ». In : *Proceedings of the 12th international workshop on semantic evaluation*. 2018, p. 798-804 (cf. p. 202).

[JHS09] Wei JIN, Hung Hay HO et Rohini K SRIHARI. « A novel lexicalized HMM-based learning framework for web opinion mining ». In : *Proceedings of the 26th annual international conference on machine learning*. T. 10. 1553374.1553435. Citeseer. 2009 (cf. p. 155).

[Jin+05] Chen JINXIU, Ji DONGHONG, Tan Chew LIM et al. « Automatic relation extraction with model order selection and discriminative label identification ». In : *International Conference on Natural Language Processing*. Springer. 2005, p. 390-401 (cf. p. 198).

[JL15] Taemin JO et Jee-Hyong LEE. « Latent keyphrase extraction using deep belief networks ». In : *International Journal of Fuzzy Logic and Intelligent Systems* 15.3 (2015), p. 153-158 (cf. p. 59).

[Kam+21] S KAMATH, KG KARIBASAPPA, Anvitha REDDY et al. « Improving the Relation Classification Using Convolutional Neural Network ». In : *IOP Conference Series : Materials Science and Engineering*. T. 1187. 1. IOP Publishing. 2021, p. 012004 (cf. p. 199).

[Kam04] Nanda KAMBHATLA. « Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction ». In : *Proceedings of the ACL interactive poster and demonstration sessions*. 2004, p. 178-181 (cf. p. 14).

[Kam+04] Nanda KAMBHATLA, Dayne FREITAG, Andrew MCCALLUM et al. « Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations ». In : *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*. 2004, p. 22-25 (cf. p. 15).

[Kan11] Mehmed KANTARDZIC. *Data mining : concepts, models, methods, and algorithms*. John Wiley & Sons, 2011 (cf. p. 64).

[KWR15] Andrew KARL, James WISNOWSKI et W Heath RUSHING. « A practical guide to text mining with topic extraction ». In : *Wiley Interdisciplinary Reviews : Computational Statistics* 7.5 (2015), p. 326-340 (cf. p. 207).

[KG12] Kamaldeep KAUR et Vishal GUPTA. « A survey of topic tracking techniques ». In : *Int J* 5 (2012) (cf. p. 207).

[KGM16] Preet Chandan KAUR, Tushar GHORPADE et Vanita MANE. « Topic extraction and sentiment classification by using latent dirichlet Markov allocation and sentiwordnet ». In : *Proceedings of the International Conference on Advances in Information Communication Technology & Computing*. 2016, p. 1-6 (cf. p. 207).

[Kho+11] ML KHODRA, DH WIDYANTORO, EA AZIZ et al. « Free Model of Sentence Classifier for Automatic Extraction of Topic Sentences ». In : *ITB Journal of Information and Communication Technology* 5 (2011), p. 17-34 (cf. p. 208).

[KH04] Soo-Min KIM et Eduard HOVY. « Determining the sentiment of opinions ». In : *COLING 2004 : Proceedings of the 20th International Conference on Computational Linguistics*. 2004, p. 1367-1373 (cf. p. 156).

[Kim+12] Sungchul KIM, Sungho JEON, Jinha KIM et al. « Finding core topics : Topic extraction with clustering on tweet ». In : *2012 Second International Conference on Cloud and Green Computing*. IEEE. 2012, p. 777-782 (cf. p. 205).

[KC07] Barbara KITCHENHAM et Stuart CHARTERS. « Guidelines for performing systematic literature reviews in software engineering ». In : (2007) (cf. p. 34, 37, 38).

[KÖ20] Abdullatif KÖKSAL et Arzucan ÖZGÜR. « The relx dataset and matching the multilingual blanks for cross-lingual relation classification ». In : *arXiv preprint arXiv :2010.09381* (2020) (cf. p. 197).

[KSC05] Rakesh KUMAR, PK SURI et RK CHAUHAN. « Search engines evaluation ». In : *DESIDOC Journal of Library & Information Technology* 25.2 (2005) (cf. p. 63).

[KPC95] Julian KUPIEC, Jan PEDERSEN et Francine CHEN. « A trainable document summarizer ». In : *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. 1995, p. 68-73 (cf. p. 64).

[Lah+13a] Fatima Zahra LAHLOU, Houda BENBRAHIMAND, Asmaa MOUNTASSIR et al. « Context extraction from reviews for Context Aware Recommendation using Text Classification techniques ». In : *2013 ACS International Conference on Computer Systems and Applications (AICCSA)*. IEEE. 2013, p. 1-4 (cf. p. 205).

[Lah+13b] Fatima Zahra LAHLOU, Asmaa MOUNTASSIR, Houda BENBRAHIM et al. « A text classification based method for context extraction from online reviews ». In : *2013 8th International Conference on Intelligent Systems : Theories and Applications (SITA)*. IEEE. 2013, p. 1-5 (cf. p. 205).

[Lee+13a] Sungwoo LEE, Jaedong LEE, Chang-Yong PARK et al. « Blog topic analysis using TF smoothing and LDA ». In : *Proceedings of the 7th International Conference on Ubiquitous Information Management and Communication*. 2013, p. 1-6 (cf. p. 68).

[Lee+13b] Sungwoo LEE, Jaedong LEE, Chang-Yong PARK et al. « Blog topic analysis using TF smoothing and LDA ». In : *Proceedings of the 7th International Conference on Ubiquitous Information Management and Communication*. 2013, p. 1-6 (cf. p. 206).

[Lev+66] Vladimir I LEVENSHTEIN et al. « Binary codes capable of correcting deletions, insertions, and reversals ». In : *Soviet physics doklady*. T. 10. 8. Soviet Union. 1966, p. 707-710 (cf. p. 91).

[Li+17] Bo LI, Xiang ZHAO, Shuai WANG et al. « Relation classification using revised convolutional neural networks ». In : *2017 4th International Conference on Systems and Informatics (ICSAI)*. IEEE. 2017, p. 1438-1443 (cf. p. 200).

[Li+16] Fei LI, Meishan ZHANG, Guohong FU et al. « A Bi-LSTM-RNN model for relation classification using low-cost sequence features ». In : *arXiv preprint arXiv :1608.07720* (2016) (cf. p. 30, 196).

[Li+19] Luoqin LI, Jiabing WANG, Jichang LI et al. « Relation classification via keyword-attentive sentence mechanism and synthetic stimulation loss ». In : *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.9 (2019), p. 1392-1404 (cf. p. 48, 199).

[LZC18] Ning LI, Hui ZHANG et Yong CHEN. « Convolutional neural network with sdp-based attention for relation classification ». In : *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE. 2018, p. 615-618 (cf. p. 27, 197).

[LQH19] Peng LI, Xipeng QIU et Xuanjing HUANG. « Relation Classification via Multi-Level Attention CNNs with Hierarchical Attention Transfer ». In : *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, p. 1388-1397 (cf. p. 29).

[Li+21a] Xiangju LI, Shi FENG, Yifei ZHANG et al. « Multi-level Emotion Cause Analysis by Multi-head Attention Based Multi-task Learning ». In : *China National Conference on Chinese Computational Linguistics*. Springer. 2021, p. 77-93 (cf. p. 199).

[Li+21b] Ximing LI, Yang WANG, Jihong OUYANG et al. « Topic extraction from extremely short texts with variational manifold regularization ». In : *Machine Learning* 110.5 (2021), p. 1029-1066 (cf. p. 207).

[Li+21c] Xin LI, Yang LIU, Peng ZHANG et al. « Dual-Attention Network for Distantly Supervised Relation Extraction ». In : *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 2021, p. 3747-3758 (cf. p. 16).

[LEF15] Rinaldo LIMA, Bernard ESPINASSE et Fred FREITAS. « Relation extraction from texts with symbolic rules induced by inductive logic programming ». In : *2015 IEEE 27th international conference on tools with artificial intelligence (ICTAI)*. IEEE. 2015, p. 194-201 (cf. p. 14).

[LEF18] Rinaldo LIMA, Bernard ESPINASSE et Fred FREITAS. « Ontoilper : an ontology-and inductive logic programming-based system to extract entities and relations from text ». In : *Knowledge and Information Systems* 56.1 (2018), p. 223-255 (cf. p. 14).

[LEF19] Rinaldo LIMA, Bernard ESPINASSE et Fred FREITAS. « A logic-based relational learning approach to relation extraction : The OntoILPER system ». In : *Engineering Applications of Artificial Intelligence* 78 (2019), p. 142-157 (cf. p. 2).

[Lin+18] Chen LIN, Timothy MILLER, Dmitriy DLIGACH et al. « Self-training improves recurrent neural networks performance for temporal relation extraction ». In : *Proceedings of the ninth international workshop on health text mining and information analysis*. 2018, p. 165-176 (cf. p. 201).

[Lin+20] Zhimin LIN, Dajiang LEI, Yuting HAN et al. « Siamese BERT Model with Adversarial Training for Relation Classification ». In : *2020 IEEE International Conference on Knowledge Graph (ICKG)*. IEEE. 2020, p. 291-296 (cf. p. 197).

[Liu+19] Baitao LIU, Jingxuan ZHU, Xiaoyan LI et al. « Multi-Task Learning for Few-Shot Relation Classification ». In : *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, p. 4783-4793 (cf. p. 31).

[LHC05] Bing LIU, Minqing HU et Junsheng CHENG. « Opinion observer : analyzing and comparing opinions on the web ». In : *Proceedings of the 14th international conference on World Wide Web*. 2005, p. 342-351 (cf. p. 12).

[Liu+21a] Fangchao LIU, Xinyan XIAO, Lingyong YAN et al. « From Learning-to-Match to Learning-to-Discriminate : Global Prototype Learning for Few-shot Relation Classification ». In : *China National Conference on Chinese Computational Linguistics*. Springer. 2021, p. 193-208 (cf. p. 198).

[Liu+21b] Fangchao LIU, Xinyan XIAO, Lingyong YAN et al. « From Learning-to-Match to Learning-to-Discriminate : Global Prototype Learning for Few-shot Relation Classification ». In : *China National Conference on Chinese Computational Linguistics*. Springer. 2021, p. 193-208 (cf. p. 199).

[Liu+14] Guolong LIU, Xiaofei XU, Ying ZHU et al. « An improved latent dirichlet allocation model for hot topic extraction ». In : *2014 IEEE Fourth International Conference on Big Data and Cloud Computing*. IEEE. 2014, p. 470-476 (cf. p. 205).

[Liu+21c] Jianyi LIU, Xi DUAN, Ru ZHANG et al. « Relation classification via BERT with piecewise convolution and focal loss ». In : *Plos one* 16.9 (2021), e0257092 (cf. p. 202).

[Liu15a] Qihua LIU. « A novel Chinese text topic extraction method based on LDA ». In : *2015 4th International Conference on Computer Science and Network Technology (ICCSNT)*. T. 1. IEEE. 2015, p. 53-57 (cf. p. 68).

[Liu15b] Qihua LIU. « A novel Chinese text topic extraction method based on LDA ». In : *2015 4th International Conference on Computer Science and Network Technology (ICCSNT)*. T. 1. IEEE. 2015, p. 53-57 (cf. p. 207).

[Liu+17] Qingtang LIU, Mingbo SHAO, Linjing WU et al. « Main content extraction from web pages based on node characteristics ». In : *Journal of Computing Science and Engineering* 11.2 (2017), p. 39-48 (cf. p. 208).

[LR12a] Song LIU et Fuji REN. « Relation extraction from wikipedia articles by entities clustering ». In : *2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems*. T. 3. IEEE. 2012, p. 1491-1495 (cf. p. 198).

[LR12b] Song LIU et Fuji REN. « Relation extraction from wikipedia articles by entities clustering ». In : *2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems*. T. 3. IEEE. 2012, p. 1491-1495 (cf. p. 198).

[Liu+16] Yang LIU, Sujian LI, Furu WEI et al. « Relation classification via modeling augmented dependency paths ». In : *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.9 (2016), p. 1589-1598 (cf. p. 199).

[Liu+15] Yang LIU, Furu WEI, Sujian LI et al. « A dependency-based neural network for relation classification ». In : *arXiv preprint arXiv :1507.04646* (2015) (cf. p. 25, 47, 197).

[Liu+10] Zhiyuan LIU, Wenyi HUANG, Yabin ZHENG et al. « Automatic keyphrase extraction via topic decomposition ». In : *Proceedings of the 2010 conference on empirical methods in natural language processing*. 2010, p. 366-376 (cf. p. 206).

[Lon+21] Jun LONG, Ye WANG, Xiangxiang WEI et al. « Entity-Centric Fully Connected GCN for Relation Classification ». In : *Applied Sciences* 11.4 (2021), p. 1377 (cf. p. 200).

[LC21] Shengfei LYU et Huanhuan CHEN. « Relation Classification with Entity Type Restriction ». In : *arXiv preprint arXiv :2105.08393* (2021) (cf. p. 199).

[Ma11] Hui-Fang MA. « Hot topic extraction using time window ». In : *2011 International Conference on Machine Learning and Cybernetics*. T. 1. IEEE. 2011, p. 56-60 (cf. p. 205).

[Mal+17] Maha MALLEK, Ramzi GUETARI, Nejmeddine ETTEYEB et al. « Graphical representation of statistics hidden in unstructured data : a software application ». In : *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE. 2017, p. 653-658 (cf. p. 78).

[Mar94] Mary Ann MARCINKIEWICZ. « Building a large annotated corpus of English : The Penn Treebank ». In : *Using Large Corpora* 273 (1994) (cf. p. 126).

[McC05] Andrew MCCALLUM. « Information extraction : Distilling structured data from unstructured text ». In : *Queue* 3.9 (2005), p. 48-57 (cf. p. 9).

[MT04] Rada MIHALCEA et Paul TARAU. « Textrank : Bringing order into text ». In : *Proceedings of the 2004 conference on empirical methods in natural language processing*. 2004, p. 404-411 (cf. p. 57).

[Mik+13] Tomas MIKOLOV, Kai CHEN, Greg CORRADO et al. « Efficient estimation of word representations in vector space ». In : *arXiv preprint arXiv :1301.3781* (2013) (cf. p. 91).

[Min06a] What Is Data MINING. « Data mining : Concepts and techniques ». In : *Morgan Kaufinann* 10 (2006), p. 559-569 (cf. p. 62).

[Min06b] What Is Data MINING. « Data mining : Concepts and techniques ». In : *Morgan Kaufinann* 10 (2006), p. 559-569 (cf. p. 64).

[MPK19] Asha Rani MISHRA, VK PANCHAL et Pawan KUMAR. « Extractive Text Summarization-An effective approach to extract information from Text ». In : *2019 International Conference on contemporary Computing and Informatics (IC3I)*. IEEE. 2019, p. 252-255 (cf. p. 67).

[Moh+12] Hadi MOHAMMADZADEH, Thomas GOTTRON, Franz SCHWEIGGERT et al. « TitleFinder : Extracting the headline of news web pages based on cosine similarity and overlap scoring similarity ». In : *Proceedings of the twelfth international workshop on Web information and data management*. 2012, p. 65-72 (cf. p. 207).

[MB05] Raymond MOONEY et Razvan BUNESCU. « Subsequence kernels for relation extraction ». In : *Advances in neural information processing systems* 18 (2005) (cf. p. 15).

[MDG21] Jose G MORENO, Antoine DOUCET et Brigitte GRAU. « Relation Classification via Relation Validation ». In : *Proceedings of the 6th Workshop on Semantic Deep Learning (SemDeep-6)*. 2021, p. 20-27 (cf. p. 200).

[Muh+19] Bello Aliyu MUHAMMAD, Rahat IQBAL, Anne JAMES et al. « Convolutional Neural Network for Core Sections Identification in Scientific Research Publications ». In : *International Conference on Intelligent Data Engineering and Automated Learning*. Springer. 2019, p. 265-273 (cf. p. 202).

[Mur17] Hendri MURFI. « Accuracy of separable nonnegative matrix factorization for topic extraction ». In : *Proceedings of the 3rd International Conference on Communication and Information Processing*. 2017, p. 226-230 (cf. p. 206).

[Nal+16] Ramesh NALLAPATI, Bowen ZHOU, Caglar GULCEHRE et al. « Abstractive text summarization using sequence-to-sequence rnns and beyond ». In : *arXiv preprint arXiv :1602.06023* (2016) (cf. p. 65, 66).

[NMI07] Dat PT NGUYEN, Yutaka MATSUO et Mitsuru ISHIZUKA. « Relation extraction from wikipedia using subtree mining ». In : *Proceedings of the National Conference on Artificial Intelligence*. T. 22. 2. Menlo Park, CA ; Cambridge, MA ; London ; AAAI Press ; MIT Press ; 1999. 2007, p. 1414 (cf. p. 15).

[NSY16] Khanh-Ly NGUYEN, Byung-Joo SHIN et Seong Joon YOO. « Hot topic detection and technology trend tracking for patents utilizing term frequency and proportional document frequency and semantic information ». In : *2016 international conference on big data and smart computing (BigComp)*. IEEE. 2016, p. 223-230 (cf. p. 208).

[Ni+22] Jian NI, Gaetano ROSSIELLO, Alfio GLIOZZO et al. « A Generative Model for Relation Extraction and Classification ». In : *arXiv preprint arXiv :2202.13229* (2022) (cf. p. 201).

[NKP17] Supattra NIBOONKIT, Worarat KRATHU et Praisan PADUNGWEANG. « Automatic discovering success factor relationship entities in articles using named entity recognition ». In : *2017 9th International Conference on Knowledge and Smart Technology (KST)*. IEEE. 2017, p. 238-241 (cf. p. 78).

[Nov92] Patricio NOVOA GREEN. « Corpus, Concordance, Collocation ». In : (1992) (cf. p. 56).

[Nun+14] Bernardo Pereira NUNES, Alexander MERA, Ricardo KAWASE et al. « A topic extraction process for online forums ». In : *2014 IEEE 14th International Conference on Advanced Learning Technologies*. IEEE. 2014, p. 541-543 (cf. p. 207).

[OV19] Abiola OBAMUYIDE et Andreas VLACHOS. « Model-agnostic meta-learning for relation classification with limited supervision ». In : *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, p. 5873-5879 (cf. p. 201).

[Pan+13] Shimei PAN, Michelle X ZHOU, Yangqiu SONG et al. « Optimizing temporal topic segmentation for intelligent text visualization ». In : *Proceedings of the 2013 international conference on Intelligent user interfaces*. 2013, p. 339-350 (cf. p. 207).

[Pan+18] Shuanshuan PANG, Wenjia NIU, Jiqiang LIU et al. « An approach to generate topic similar document by seed extraction-based SeqGAN training for bait document ». In : *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*. IEEE. 2018, p. 803-810 (cf. p. 207).

[Par09] Constituency PARSING. « Speech and language processing ». In : *Power Point Slides* (2009) (cf. p. 90).

[PT18] Ruchi PATEL et Sanjay TANWANI. « TEMPORAL RELATION IDENTIFICATION FROM CLINICAL TEXT USING LSTM BASED DEEP LEARNING MODEL ». In : (2018) (cf. p. 201).

[PH10] Mari-Sanna PAUKKERI et Timo HONKELA. « Likey : Unsupervised language-independent keyphrase extraction ». In : *Proceedings of the 5th international workshop on semantic evaluation*. 2010, p. 162-165 (cf. p. 90).

[PDG15] Fabio PETRONI, Luciano DEL CORRO et Rainer GEMULLA. « Core : Context-aware open relation extraction with factorization machines ». In : *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015, p. 1763-1773 (cf. p. 200).

[Pla+06] Conrad PLAKE, Torsten SCHIEMANN, Marcus PANKALLA et al. « AliBaba : PubMed as a graph ». In : *Bioinformatics* 22.19 (2006), p. 2444-2445 (cf. p. 11).

[PE07] Ana-Maria POPESCU et Orena ETZIONI. « Extracting product features and opinions from reviews ». In : *Natural language processing and text mining*. Springer, 2007, p. 9-28 (cf. p. 12).

[PS19] Rayehe Hosseini POUR et Mehrnoush SHAMSFARD. « EoANN : Lexical Semantic Relation Classification Using an Ensemble of Artificial Neural Networks ». In : *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. 2019, p. 481-486 (cf. p. 203).

[QXG16] Pengda QIN, Weiran XU et Jun GUO. « An empirical convolutional neural network approach for semantic relation classification ». In : *Neurocomputing* 190 (2016), p. 1-9 (cf. p. 2, 22, 199).

[QXG17] Pengda QIN, Weiran XU et Jun GUO. « Designing an adaptive attention mechanism for relation classification ». In : *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2017, p. 4356-4362 (cf. p. 48, 197).

[QXW18] Pengda QIN, Weiran XU et William Yang WANG. « Robust distant supervision relation extraction via deep reinforcement learning ». In : *arXiv preprint arXiv :1805.09927* (2018) (cf. p. 32, 201).

[QZ14] Likun QIU et Yue ZHANG. « ZORE : A syntax-based system for chinese open relation extraction ». In : *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, p. 1870-1880 (cf. p. 200).

[Rad+18] Alec RADFORD, Karthik NARASIMHAN, Tim SALIMANS et al. « Improving language understanding by generative pre-training ». In : *URL https ://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervis understanding paper. pdf* (2018) (cf. p. 29).

[RM13] Ahmed RAFEA et Nada A MOSTAFA. « Topic extraction in social media ». In : *2013 International Conference on Collaboration Technologies and Systems (CTS)*. IEEE. 2013, p. 94-98 (cf. p. 205).

[RA21] Engels RAJANGAM et Chitra ANNAMALAI. « Topic extraction using local graph centrality and semantic similarity ». In : *Concurrency and Computation : Practice and Experience* 33.7 (2021), p. 1-1 (cf. p. 208).

[RCV17] Chandrasekhar RANGU, Shuvojit CHATTERJEE et Srinivasa Rao VALLURU. « Text mining approach for product quality enhancement :(Improving product quality through machine learning) ». In : *2017 IEEE 7th International Advance Computing Conference (IACC)*. IEEE. 2017, p. 456-460 (cf. p. 78).

[] « Rayyan Systems Inc. Available online : https ://www.rayyan.ai/ (accessed on 1 August 2022) ». In : () (cf. p. 69, 70).

[RJ16] Amal REKIK et Salma JAMOUSSI. « Deep learning for hot topic extraction from social streams ». In : *International Conference on Hybrid Intelligent Systems*. Springer. 2016, p. 186-197 (cf. p. 207).

[Ren+17] Feiliang REN, Rongsheng ZHAO, Xiao HU et al. « Embedding Syntactic Tree Structures into CNN Architecture for Relation Classification ». In : *China Conference on Knowledge Graph and Semantic Computing*. Springer. 2017, p. 104-116 (cf. p. 197).

[RH10] Bryan RINK et Sanda HARABAGIU. « Utd : Classifying semantic relations by combining lexical and semantic resources ». In : *Proceedings of the 5th international workshop on semantic evaluation*. 2010, p. 256-259 (cf. p. 20).

[RV11] Marian-Andrei RIZOIU et Julien VELCIN. « Topic extraction for ontology learning ». In : *Ontology Learning and Knowledge Discovery Using the Web : Challenges and Recent Advances*. IGI Global, 2011, p. 38-60 (cf. p. 206).

[RSR15] Tim ROCKTÄSCHEL, Sameer SINGH et Sebastian RIEDEL. « Injecting logical background knowledge into embeddings for relation extraction ». In : *Proceedings of the 2015 conference of the north American Chapter of the Association for Computational Linguistics : Human Language Technologies*. 2015, p. 1119-1129 (cf. p. 19).

[Ros+10] Stuart ROSE, Dave ENGEL, Nick CRAMER et al. « Automatic keyword extraction from individual documents ». In : *Text mining : applications and theory* (2010), p. 1-20 (cf. p. 59).

[RF06a] Binjamin ROZENFELD et Ronen FELDMAN. « High-performance unsupervised relation extraction from large corpora ». In : *Sixth International Conference on Data Mining (ICDM'06)*. IEEE. 2006, p. 1032-1037 (cf. p. 10).

[RF06b] Binjamin ROZENFELD et Ronen FELDMAN. « High-performance unsupervised relation extraction from large corpora ». In : *Sixth International Conference on Data Mining (ICDM'06)*. IEEE. 2006, p. 1032-1037 (cf. p. 15).

[Sab+21] Ofer SABO, Yanai ELAZAR, Yoav GOLDBERG et al. « Revisiting Few-shot Relation Classification : Evaluation Data and Classification Schemes ». In : *Transactions of the Association for Computational Linguistics* 9 (2021), p. 691-706 (cf. p. 201).

[SJS17a] Anamta SAJID, Sadaqat JAN et Ibrar A SHAH. « Automatic topic modeling for single document short texts ». In : *2017 International Conference on Frontiers of Information Technology (FIT)*. IEEE. 2017, p. 70-75 (cf. p. 112-116).

[SJS17b] Anamta SAJID, Sadaqat JAN et Ibrar A SHAH. « Automatic topic modeling for single document short texts ». In : *2017 International Conference on Frontiers of Information Technology (FIT)*. IEEE. 2017, p. 70-75 (cf. p. 205).

[SB88] Gerard SALTON et Christopher BUCKLEY. « Term-weighting approaches in automatic text retrieval ». In : *Information processing & management* 24.5 (1988), p. 513-523 (cf. p. 56, 90).

[SD03] Erik F SANG et Fien DE MEULDER. « Introduction to the CoNLL-2003 shared task : Language-independent named entity recognition ». In : *arXiv preprint cs/0306050* (2003) (cf. p. 10).

[SXZ15] Cicero Nogueira dos SANTOS, Bing XIANG et Bowen ZHOU. « Classifying relations by ranking with convolutional neural networks ». In : *arXiv preprint arXiv :1504.06580* (2015) (cf. p. 22, 47, 196).

[SNG12] Kamal SARKAR, Mita NASIPURI et Suranjan GHOSE. « Machine learning based keyphrase extraction : comparing decision trees, naıve Bayes, and artificial neural networks ». In : *Journal of Information Processing Systems* 8.4 (2012), p. 693-712 (cf. p. 59).

[SB14] Jordan SCHMIDEK et Denilson BARBOSA. « Improving open relation extraction via sentence re-structuring ». In : *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. 2014, p. 3720-3723 (cf. p. 201).

[Sch+12] Michael SCHMITZ, Stephen SODERLAND, Robert BART et al. « Open language learning for information extraction ». In : *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. 2012, p. 523-534 (cf. p. 126, 127).

[SLM17] Abigail SEE, Peter J LIU et Christopher D MANNING. « Get to the point : Summarization with pointer-generator networks ». In : *arXiv preprint arXiv :1704.04368* (2017) (cf. p. 65, 66).

[Sek+20] Yohei SEKI, Kangkang ZHAO, Masaki OGUNI et al. « A framework for classifying temporal relations with question encoder ». In : *International Conference on Asian Digital Libraries*. Springer. 2020, p. 20-32 (cf. p. 199, 202).

[Sek+21] Yohei SEKI, Kangkang ZHAO, Masaki OGUNI et al. « CNN-based framework for classifying temporal relations with question encoder ». In : *International Journal on Digital Libraries* (2021), p. 1-11 (cf. p. 202).

[She+18] Shirong SHEN, Yang WEN, Lijuan ZHOU et al. « Customized Attention Mechanism for Relation Classification ». In : *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*. 2018 (cf. p. 30, 48, 196).

[She+11] Wei SHEN, Jianyong WANG, Ping LUO et al. « REACTOR : a framework for semantic relation extraction and tagging over enterprise data ». In : *Proceedings of the 20th international conference companion on World wide web*. 2011, p. 121-122 (cf. p. 198).

[SRP14] Yongwook SHIN, Chuhyeop RYO et Jonghun PARK. « Automatic extraction of persistent topics from social text streams ». In : *World Wide Web* 17.6 (2014), p. 1395-1420 (cf. p. 206).

[SD16] Vered SHWARTZ et Ido DAGAN. « The roles of pathbased and distributional information in recognizing lexical semantic relations ». In : *CoRR, abs/1608.05014* (2016) (cf. p. 203).

[Sin+13] Sameer SINGH, Sebastian RIEDEL, Brian MARTIN et al. « Joint inference of entities, relations, and coreference ». In : *Proceedings of the 2013 workshop on Automated knowledge base construction*. 2013, p. 1-6 (cf. p. 17).

[ST19] Padipat SITKRONGWONG et Atsuhiro TAKASU. « Unsupervised context extraction via region embedding for context-aware recommendations ». In : *Proceedings of the 23rd International Database Applications & Engineering Symposium*. 2019, p. 1-10 (cf. p. 207).

[Soa+19a] Livio Baldini SOARES, Nicholas FITZGERALD, Jeffrey LING et al. « Matching the blanks : Distributional similarity for relation learning ». In : *arXiv preprint arXiv :1906.03158* (2019) (cf. p. 48, 203).

[Soa+19b] Livio Baldini SOARES, Nicholas FITZGERALD, Jeffrey LING et al. « Matching the blanks : Distributional similarity for relation learning ». In : *arXiv preprint arXiv :1906.03158* (2019) (cf. p. 198).

[Soc+12] Richard SOCHER, Brody HUVAL, Christopher D MANNING et al. « Semantic compositionality through recursive matrix-vector spaces ». In : *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. 2012, p. 1201-1211 (cf. p. 23, 47, 199).

[SRS18] Yuan SONG, Ruo Nan RAO et Jun SHI. « Relation classification in knowledge graph based on natural language text ». In : *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*. IEEE. 2018, p. 1104-1107 (cf. p. 197).

[Spi74] Baruch SPINOZA. *Éthique, démontrée selon l'ordre géométrique*. Trad. par Charles APPUHN. Paris : Éditions de Minuit, 1974. ISBN : 2707300506 (cf. p. 53).

[SSA18] Víctor SUÁREZ-PANIAGUA, Isabel SEGURA-BEDMAR et Akiko AIZAWA. « UC3M-NII Team at SemEval-2018 Task 7 : Semantic Relation Classification in Scientific Papers via Convolutional Neural Network ». In : *Proceedings of The 12th International Workshop on Semantic Evaluation*. 2018, p. 793-797 (cf. p. 202).

[Sun+18] Yiping SUN, Yu CUI, Jinglu HU et al. « Relation classification using coarse and fine-grained networks with SDP supervised key words selection ». In : *International Conference on Knowledge Science, Engineering and Management*. Springer. 2018, p. 514-522 (cf. p. 197).

[SB11] Zhan Feng SUN et Kong Jun BAO. « Research of Text Topic Automatic Extraction Method Based on Rough Set Theory ». In : *Advanced Materials Research*. T. 268. Trans Tech Publ. 2011, p. 1127-1131 (cf. p. 208).

[Sut16] S SUTHAHARAN. *Decision tree learning, in Machine Learning Models and Algorithms for Big Data Classification*. 2016 (cf. p. 63).

[TC14] Leonardo Sameshima TABA et Helena CASELI. « Automatic semantic relation extraction from portuguese texts ». In : *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. 2014, p. 2739-2746 (cf. p. 19).

[TOI15] Sho TAKASE, Naoaki OKAZAKI et Kentaro INUI. « Fast and large-scale unsupervised relation extraction ». In : *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*. 2015, p. 96-105 (cf. p. 198).

[Tan+21] Yubao TANG, Zhezhou LI, Cong CAO et al. « Knowledge-Based Diverse Feature Transformation for Few-Shot Relation Classification ». In : *International Conference on Knowledge Science, Engineering and Management*. Springer. 2021, p. 101-114 (cf. p. 199).

[Tat12] Kayo TATSUKAWA. « Topic extraction based on prior knowledge obtained from target documents ». In : *Proceedings of ACL 2012 Student Research Workshop*. 2012, p. 31-36 (cf. p. 207).

[Tia+21] Fei TIAN, Lei ZHANG, Wei HUANG et al. « Improving Few-Shot Relation Classification with Multiple Hypergraphs ». In : *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021, p. 2413-2423 (cf. p. 31).

[TR10] Juan-Manuel TORRES-MORENO et J RAMIREZ. « REG : un algorithme glouton appliqué au résumé automatique de texte ». In : *Proceedings of the 10th Int. Conference on the Statistical Analysis of Textual. Roma, Italia*. T. 84. 2010 (cf. p. 60).

[Vie+18] Felipe VIEGAS, Washington LUIZ, Christian GOMES et al. « Semantically-enhanced topic modeling ». In : *Proceedings of the 27th ACM international conference on information and knowledge management*. 2018, p. 893-902 (cf. p. 207).

[WHZ19] Chengyu WANG, Xiaofeng HE et Aoying ZHOU. « Spherere : Distinguishing lexical relations with hyperspherical relation embeddings ». In : *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, p. 1727-1737 (cf. p. 203).

[Wan+20a] Chengyu WANG, Minghui QIU, Jun HUANG et al. « KEML : A knowledge-enriched meta-learning framework for lexical relation classification ». In : *arXiv preprint arXiv :2002.10903* (2020) (cf. p. 203).

[WTA12] Di WANG, Marcus THINT et Ahmad AL-RUBAIE. « Semi-supervised latent Dirichlet allocation and its application for document classification ». In : *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. T. 3. IEEE. 2012, p. 306-310 (cf. p. 207).

[Wan+21a] Hongru WANG, Zhijing JIN, Jiarun CAO et al. « Inconsistent Few-Shot Relation Classification via Cross-Attentional Prototype Networks with Contrastive Learning ». In : *arXiv preprint arXiv :2110.08254* (2021) (cf. p. 200).

[Wan+16] Linlin WANG, Zhu CAO, Gerard DE MELO et al. « Relation classification via multi-level attention cnns ». In : *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. 2016, p. 1298-1307 (cf. p. 27, 200).

[WXH17] Pengfei WANG, Zhipeng XIE et Junfeng HU. « Relation Classification via CNN, Segmented Max-pooling, and SDP-BLSTM ». In : *International Conference on Neural Information Processing*. Springer. 2017, p. 144-154 (cf. p. 30, 48, 197).

[Wan+21b] Qian WANG, Peng ZHANG, Yang LIU et al. « DSGCN : A Dynamic Semantic Graph Convolutional Network for Distantly Supervised Relation Extraction ». In : *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021, p. 1752-1762 (cf. p. 16).

[Wan+13] Xikui WANG, Yang LIU, Donghui WANG et al. « Cross-media topic mining on wikipedia ». In : *Proceedings of the 21st ACM international conference on Multimedia*. 2013, p. 689-692 (cf. p. 206).

[WYL14] Yanjun WANG, Xinyan YAO et Ting LIU. « Unsupervised feature selection for clustering and its application to relation extraction ». In : *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. 2014, p. 977-986 (cf. p. 33).

[Wan+20b] Yingyao WANG, Junwei BAO, Guangyi LIU et al. « Learning to decouple relations : Few-shot relation classification with entity-guided attention and confusion-aware training ». In : *arXiv preprint arXiv :2010.10894* (2020) (cf. p. 198).

[Wan+20c] Yingyao WANG, Junwei BAO, Guangyi LIU et al. « Learning to decouple relations : Few-shot relation classification with entity-guided attention and confusion-aware training ». In : *arXiv preprint arXiv :2010.10894* (2020) (cf. p. 199).

[Wan+18] Zihuan WANG, Kyusup HAHN, Youngsam KIM et al. « A news-topic recommender system based on keywords extraction ». In : *Multimedia Tools and Applications* 77.4 (2018), p. 4339-4353 (cf. p. 205).

[Won19] Papis WONGCHAISUWAT. « Automatic keyword extraction using textrank ». In : *2019 IEEE 6th International Conference on Industrial Engineering and Applications (ICIEA)*. IEEE. 2019, p. 377-381 (cf. p. 61).

[WXZ+11] Chanle WU, Ming XIE, Yunlu ZHANG et al. « A new intelligent topic extraction model on web ». In : (2011) (cf. p. 207).

[WWW16] Chunzi WU, Bin WU et Bai WANG. « Event evolution model based on random walk model with hot topic extraction ». In : *International Conference on Advanced Data Mining and Applications*. Springer. 2016, p. 591-603 (cf. p. 206).

[WW10] Fei WU et Daniel S WELD. « Open information extraction using wikipedia ». In : *Proceedings of the 48th annual meeting of the association for computational linguistics*. 2010, p. 118-127 (cf. p. 18).

[Wu+20a] Linfang WU, Hua-Ping ZHANG, Yaofei YANG et al. « Dynamic prototype selection by fusing attention mechanism for few-shot relation classification ». In : *Asian Conference on Intelligent Information and Database Systems*. Springer. 2020, p. 431-441 (cf. p. 198).

[Wu+17] Menglong WU, Lin LIU, Wenxi YAO et al. « Semantic relation classification by bi-directional lstm architecture ». In : *Advanced Sciences and Technology Letters* (2017) (cf. p. 199).

[WH19a] Shanchan WU et Yifan HE. « Enriching pre-trained language model with entity information for relation classification ». In : *Proceedings of the 28th ACM international conference on information and knowledge management*. 2019, p. 2361-2364 (cf. p. 48, 200).

[WH19b] Shanchan WU et Yifan HE. « Enriching pre-trained language model with entity information for relation classification ». In : *Proceedings of the 28th ACM international conference on information and knowledge management*. 2019, p. 2361-2364 (cf. p. 197).

[Wu+20b] Siyuan WU, Yuwei WU, Sheng WANG et al. « Self-supervised Relation Extraction with Multi-Head Attention and Label Refinery ». In : *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 2020, p. 6732-6741 (cf. p. 17, 18).

[Xia+12] Huan XIA, Juanzi LI, Jie TANG et al. « Plink-LDA : using link as prior information in topic modeling ». In : *International Conference on Database Systems for Advanced Applications*. Springer. 2012, p. 213-227 (cf. p. 206).

[XL16] Minguang XIAO et Cong LIU. « Semantic relation classification via hierarchical recurrent neural network with attention ». In : *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*. 2016, p. 1254-1263 (cf. p. 196).

[XJH21] Yan XIAO, Yaochu JIN et Kuangrong HAO. « Adaptive prototypical networks with label words and joint representation learning for few-shot relation classification ». In : *IEEE Transactions on Neural Networks and Learning Systems* (2021) (cf. p. 200).

[Xu+20] Guixian XU, Ziheng YU, Changzhi WANG et al. « Research on topic discovery technology for Web news ». In : *Neural Computing and Applications* 32.1 (2020), p. 73-83 (cf. p. 206).

[Xu+16a] Jing XU, Liang GAN, Zhou YAN et al. « Open relation extraction from chinese microblog text ». In : *2016 IEEE First International Conference on Data Science in Cyberspac0e (DSC)*. IEEE. 2016, p. 673-677 (cf. p. 202).

[Xu+15a] Kun XU, Yansong FENG, Songfang HUANG et al. « Semantic relation classification via convolutional neural networks with simple negative sampling ». In : *arXiv preprint arXiv :1506.07650* (2015) (cf. p. 26, 47, 196).

[Xu+18] TaoLing XU, YaJun DU, ChunLong FU et al. « Incorporating forward and backward instances in a bi-lstm-cnn model for relation classification ». In : *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*. IEEE. 2018, p. 2133-2137 (cf. p. 30, 48).

[Xu+16b] Yan XU, Ran JIA, Lili MOU et al. « Improved relation classification by deep recurrent neural networks with data augmentation ». In : *arXiv preprint arXiv :1601.03651* (2016) (cf. p. 23, 203).

[Xu+15b] Yan XU, Lili MOU, Ge LI et al. « Classifying relations via long short term memory networks along shortest dependency paths ». In : *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2015, p. 1785-1794 (cf. p. 2, 24, 47, 196).

[Yan+19a] Kaijia YANG, Liang HE, Xinyu DAI et al. « Exploiting noisy data in distant supervision relation classification ». In : *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, p. 3216-3225 (cf. p. 32, 198).

[Yan+18] Shanliang YANG, Qi SUN, Huyong ZHOU et al. « A topic detection method based on KeyGraph and community partition ». In : *Proceedings of the 2018 International Conference on Computing and Artificial Intelligence*. 2018, p. 30-34 (cf. p. 206).

[Yan+19b] Yinfei YANG, Daniel CER, Amin AHMAD et al. « Multilingual universal sentence encoder for semantic retrieval ». In : *arXiv preprint arXiv :1907.04307* (2019) (cf. p. 130).

[Yan+16] Yunlun YANG, Yunhai TONG, Shulei MA et al. « A position encoding convolutional neural network based on dependency tree for relation classification ». In : *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2016, p. 65-74 (cf. p. 196).

[Yao+19] Yuan YAO, Demin YE, Peng LI et al. « DocRED : A Large-Scale Document-Level Relation Extraction Dataset ». In : *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, p. 764-773 (cf. p. 28).

[Ye+19] Qinyuan YE, Liyuan LIU, Maosen ZHANG et al. « Looking beyond label noise : Shifted label distribution matters in distantly supervised relation extraction ». In : *arXiv preprint arXiv :1904.09331* (2019) (cf. p. 198).

[YL19a] Zhi-Xiu YE et Zhen-Hua LING. « Multi-level matching and aggregation network for few-shot relation classification ». In : *arXiv preprint arXiv :1906.06678* (2019) (cf. p. 198).

[YL19b] Zhi-Xiu YE et Zhen-Hua LING. « Multi-level matching and aggregation network for few-shot relation classification ». In : *arXiv preprint arXiv :1906.06678* (2019) (cf. p. 199).

[Yin+18] Zhongbo YIN, Zhunchen LUO, Wei LUO et al. « IRCMS at SemEval-2018 task 7 : evaluating a basic CNN method and traditional pipeline method for relation classification ». In : *Proceedings of The 12th International Workshop on Semantic Evaluation*. 2018, p. 811-815 (cf. p. 202).

[YY10] Takeru YOKOI et Hidekazu YANAGIMOTO. « Topic extraction for a large document set with the topic integration ». In : *2010 Third International Conference on Knowledge Discovery and Data Mining*. IEEE. 2010, p. 46-49 (cf. p. 205).

[Yon+18] Wang YONGLI, Yuan HUANHUAN, Gong XIAOZE et al. « QH-K Algorithm for News Text Topic Extraction ». In : *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*. IEEE. 2018, p. 610-614 (cf. p. 205).

[YFB13] Wei YOU, Dominique FONTAINE et Jean-Paul BARTHÈS. « An automatic keyphrase extraction system for scientific documents ». In : *Knowledge and information systems* 34.3 (2013), p. 691-724 (cf. p. 205).

[YH03] Hong YU et Vasileios HATZIVASSILOGLOU. « Towards answering opinion questions : Separating facts from opinions and identifying the polarity of opinion sentences ». In : *Proceedings of the 2003 conference on Empirical methods in natural language processing*. 2003, p. 129-136 (cf. p. 155).

[YL10] Xiaofeng YU et Wai LAM. « Jointly identifying entities and extracting relations in encyclopedia text via a graphical model approach ». In : *Coling 2010 : Posters*. 2010, p. 1399-1407 (cf. p. 17).

[Yun+11] Jiali YUN, Liping JING, Jian YU et al. « Document topic extraction based on wikipedia category ». In : *2011 Fourth International Joint Conference on Computational Sciences and Optimization*. IEEE. 2011, p. 852-856 (cf. p. 205).

[Zak02] Mohammed J ZAKI. « Efficiently mining frequent trees in a forest ». In : *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2002, p. 71-80 (cf. p. 89).

[Zar+16] Haıfa ZARGAYOUNA, Isabelle TELLIER, Davide BUSCALDI et al. « Unsupervised relation extraction in specialized corpora using sequence mining ». In : *International Symposium on Intelligent Data Analysis*. Springer. 2016, p. 237-248 (cf. p. 200).

[ZAR03a] Dmitry ZELENKO, Chinatsu AONE et Anthony RICHARDELLA. « Kernel methods for relation extraction ». In : *Journal of machine learning research* 3.Feb (2003), p. 1083-1106 (cf. p. 2).

[ZAR03b] Dmitry ZELENKO, Chinatsu AONE et Anthony RICHARDELLA. « Kernel methods for relation extraction ». In : *Journal of machine learning research* 3.Feb (2003), p. 1083-1106 (cf. p. 13).

[Zen+14] Daojian ZENG, Kang LIU, Siwei LAI et al. « Relation classification via convolutional deep neural network ». In : *Proceedings of COLING 2014, the 25th international conference on computational linguistics : technical papers*. 2014, p. 2335-2344 (cf. p. 22, 47, 132, 196).

[Zen+19] Xingdi ZENG, Yankai LIU, Zhiyuan CHEN et al. « Unsupervised relation extraction with multi-head selection and self-supervised learning ». In : *Proceedings of the 57th Conference of the Association for Computational Linguistics*. 2019, p. 3211-3222 (cf. p. 33).

[Zha+21a] Chenwei ZHANG, Chen WANG, Yang YU et al. « Self-Supervised Relation Extraction via Multi-Task Learning with Pre-Trained Language Models ». In : *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2021, p. 7426-7438 (cf. p. 18).

[ZW15] Dongxu ZHANG et Dong WANG. « Relation classification via recurrent neural network ». In : *arXiv preprint arXiv :1508.01006* (2015) (cf. p. 23, 47, 132, 196).

[ZXW19] Haoyu ZHANG, Jianjun XU et Ji WANG. « Pretraining-based natural language generation for text summarization ». In : *arXiv preprint arXiv :1902.09243* (2019) (cf. p. 66).

[Zha+21b] Jiawen ZHANG, Jiaqi ZHU, Yi YANG et al. « Knowledge-Enhanced Domain Adaptation in Few-Shot Relation Classification ». In : *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021, p. 2183-2191 (cf. p. 199).

[Zha+21c] Jiawen ZHANG, Jiaqi ZHU, Yi YANG et al. « Knowledge-Enhanced Domain Adaptation in Few-Shot Relation Classification ». In : *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021, p. 2183-2191 (cf. p. 201).

[ZX18] Lei ZHANG et Fusheng XIANG. « Relation classification via bilstm-cnn ». In : *International conference on data mining and big data*. Springer. 2018, p. 373-382 (cf. p. 197).

[Zha+17] Qin ZHANG, Jianhua LIU, Ying WANG et al. « A convolutional neural network method for relation classification ». In : *2017 International Conference on Progress in Informatics and Computing (PIC)*. IEEE. 2017, p. 440-444 (cf. p. 197).

[Zha+18a] Runyan ZHANG, Fanrong MENG, Yong ZHOU et al. « Relation classification via recurrent neural network with attention and tensor layers ». In : *Big Data Mining and Analytics* 1.3 (2018), p. 234-244 (cf. p. 2, 48, 199).

[Zha+15] Shu ZHANG, Dequan ZHENG, Xinchen HU et al. « Bidirectional long short-term memory networks for relation classification ». In : *Proceedings of the 29th Pacific Asia conference on language, information and computation*. 2015, p. 73-78 (cf. p. 24, 47, 196).

[ZCH18] Xiaobin ZHANG, Fucai CHEN et Ruiyang HUANG. « A combination of RNN and CNN for attention-based relation classification ». In : *Procedia computer science* 131 (2018), p. 911-917 (cf. p. 201).

[Zha+20a] Xiaoya ZHANG, Xu HAN, Jianxing MA et al. « Entity and Relation Extraction using Multi-Task Learning and BERT-based Models ». In : *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online : Association for Computational Linguistics, juill. 2020, p. 1780-1791. URL : https://www.aclweb.org/anthology/2020.acl-main.162 (cf. p. 29).

[ZH18] Xuefei ZHANG et Ruifang HE. « Topic extraction of events on social media using reinforced knowledge ». In : *International Conference on Knowledge Science, Engineering and Management*. Springer. 2018, p. 465-476 (cf. p. 206).

[ZQ21] Ya ZHANG et Shuai QIN. « Improving Relation Classification with Multi-graph GCN ». In : *2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML)*. IEEE. 2021, p. 190-194 (cf. p. 199).

[Zha+16] Ye ZHANG, Luheng YANG, Zhiyuan LIN et al. « A novel hybrid neural network approach for relation extraction via leveraging dependency information and word embeddings ». In : *Proceedings of the 26th International Conference on Computational Linguistics*. Association for Computational Linguistics. 2016, p. 2565-2574 (cf. p. 30).

[Zha+18b] Yi ZHANG, Jie LU, Feng LIU et al. « Does deep learning help topic extraction ? A kernel k-means clustering method with word embedding ». In : *Journal of Informetrics* 12.4 (2018), p. 1099-1117 (cf. p. 207).

[ZLZ18] Yijie ZHANG, Peifeng LI et Guodong ZHOU. « Classifying temporal relations between events by deep BiLSTM ». In : *2018 International Conference on Asian Language Processing (IALP)*. IEEE. 2018, p. 267-272 (cf. p. 201).

[ZZ12] Yue ZHANG et Hongli ZHANG. « Social Topic Detection for Web Forum ». In : *2012 International Conference on Computer Science and Service System*. IEEE. 2012, p. 955-959 (cf. p. 205).

[Zha+18c] Guifen ZHAO, Yanjun LIU, Wei ZHANG et al. « TFIDF based feature words extraction and topic modeling for short text ». In : *Proceedings of the 2018 2Nd International Conference on Management Engineering, Software Engineering and Service Sciences*. 2018, p. 188-191 (cf. p. 207).

[ZLL16] Jiuru ZHAO, Xinguang LI et Xia LI. « The text mining model building of open questionnaire based on LSA ». In : *2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*. IEEE. 2016, p. 435-438 (cf. p. 67).

[Zha+20b] Lulu ZHAO, Weiran XU, Sheng GAO et al. « Cross-sentence N-ary relation classification using LSTMs on graph and sequence structures ». In : *Knowledge-Based Systems* 207 (2020), p. 106266 (cf. p. 203).

[ZG05a] Shubin ZHAO et Ralph GRISHMAN. « Extracting relations with integrated information using kernel methods ». In : *Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05)*. 2005, p. 419-426 (cf. p. 16).

[ZG05b] Shubin ZHAO et Ralph GRISHMAN. « Extracting relations with integrated information using kernel methods ». In : *Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05)*. 2005, p. 419-426 (cf. p. 202).

[ZFH11] Ying ZHAO, Wanyu FU et Shaobin HUANG. « Topic discovery and topic-driven clustering for audit method datasets ». In : *International Conference on Advanced Data Mining and Applications*. Springer. 2011, p. 346-358 (cf. p. 206).

[ZLW10] Ning ZHONG, Yuefeng LI et Sheng-Tang WU. « Effective pattern discovery for text mining ». In : *IEEE transactions on knowledge and data engineering* 24.1 (2010), p. 30-44 (cf. p. 78).

[ZZL11] Erzhong ZHOU, Ning ZHONG et Yuefeng LI. « Hot topic detection in professional blogs ». In : *International Conference on Active Media Technology*. Springer. 2011, p. 141-152 (cf. p. 206).

[ZZL14] Erzhong ZHOU, Ning ZHONG et Yuefeng LI. « Extracting news blog hot topics based on the W2T Methodology ». In : *World Wide Web* 17.3 (2014), p. 377-404 (cf. p. 206).

[Zho+05] GuoDong ZHOU, Jian SU, Jie ZHANG et al. « Exploring various knowledge in relation extraction ». In : *Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05)*. 2005, p. 427-434 (cf. p. 15).

[Zho+16a] Peng ZHOU, Wei SHI, Jun TIAN et al. « Attention-based bidirectional long short-term memory networks for relation classification ». In : *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2 : Short papers)*. 2016, p. 207-212 (cf. p. 197).

[Zho+16b] Peng ZHOU, Wenhao SHI, Jinjun TIAN et al. « Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification ». In : *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. Berlin, Germany : Association for Computational Linguistics, août 2016, p. 2077-2087. DOI : 10.18653/v1/P16-1234. URL : https://www.aclweb.org/anthology/P16-1234 (cf. p. 27).

[Zho+20a] Xinyu ZHOU, Peifeng LI, Qiaoming ZHU et al. « Incorporating temporal cues and ac-gcn to improve temporal relation classification ». In : *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer. 2020, p. 580-592 (cf. p. 201).

[Zho+20b] Xinyu ZHOU, Peifeng LI, Qiaoming ZHU et al. « Incorporating temporal cues and ac-gcn to improve temporal relation classification ». In : *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer. 2020, p. 580-592 (cf. p. 201).

[ZBK07] Guangyu ZHU, Timothy J BETHEA et Vikas KRISHNA. « Extracting relevant named entities for automated expense reimbursement ». In : *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2007, p. 1004-1012 (cf. p. 11).

[Zhu+20] Wei ZHU, Xipeng QIU, Yuan NI et al. « AutoRC : Improving BERT based relation classification models via architecture search ». In : *arXiv preprint arXiv :2009.10680* (2020) (cf. p. 201).

[Zou+21] Bo-Wei ZOU, Rong-Tao HUANG, Zeng-Zhuang XU et al. « Language adaptation for entity relation classification via adversarial neural networks ». In : *Journal of Computer Science and Technology* 36.1 (2021), p. 207-220 (cf. p. 200).

ANNEXE A

A. Liste des papiers liés à la classiication des relations

Papier	Référence	Titre
A1	GIRIDHARA, C. MISHRA, VENKATARAMANA et al. 2019	Claire at SemEval-2018 task 7 : classification of relations using embeddings
A2	S. ZHANG, D. ZHENG, Xinchen HU et al. 2015	Bidirectional long short-term memory networks for relation classification
A3	Y. XU, MOU, G. LI et al. 2015	Classifying relations via long short term memory networks along shortest dependency paths
A4	S. SHEN, WEN, L. ZHOU et al. 2018	Customized Attention Mechanism for Relation Classification
A5	CHINGACHAM et PAPERNO 2018	Generalizing Representations of Lexical Semantic Relations
A6	R. CAI, Xiaodong ZHANG et Houfeng WANG 2016b	Classifying relations via long short term memory networks along shortest dependency paths
A7	R. CAI, Xiaodong ZHANG et Houfeng WANG 2016c	Bidirectional recurrent convolutional neural network for relation classification
A8	ALFATTNI, PEEK et NENADIC 2021b	Attention-based bidirectional long short-term memory networks for extracting relationships from clinical discharge summaries
A9	D. ZENG, K. LIU, LAI et al. 2014	Relation classification via convolutional deep neural network
A10	D. ZHANG et Dong WANG 2015	Relation classification via recurrent neural network
A11	M. XIAO et C. LIU 2016	Semantic relation classification via hierarchical recurrent neural network with attention
A12	F. LI, Meishan ZHANG, G. FU et al. 2016	A Bi-LSTM-RNN model for relation classification using low-cost sequence features
A13	K. XU, Y. FENG, Songfang HUANG et al. 2015	Semantic relation classification via convolutional neural networks with simple negative sampling
A14	SANTOS, B. XIANG et B. ZHOU 2015	Classifying relations by ranking with convolutional neural networks
A15	Yunlun YANG, TONG, S. MA et al. 2016	A position encoding convolutional neural network based on dependency tree for relation classification

A16	Yang LIU, F. WEI, S. LI et al. 2015	A dependency-based neural network for relation classification
A17	Q. ZHANG, Jianhua LIU, Ying WANG et al. 2017	A convolutional neural network method for relation classification
A18	DAVIDOV et RAPPOPORT 2008	Classification of semantic relationships between nominals using pattern clusters
A19	N. LI, Hui ZHANG et Yong CHEN 2018	Convolutional neural network with sdp-based attention for relation classification
A20	P. QIN, W. XU et J. GUO 2017	Designing an adaptive attention mechanism for relation classification
A21	Yuan SONG, RAO et J. SHI 2018	Relation classification in knowledge graph based on natural language text
A22	Shanchan WU et Yifan HE 2019b	Enriching pre-trained language model with entity information for relation classification
A23	DO et ROTH 2010	Constraints based taxonomic relation classification
A24	GÁBOR, BUSCALDI, SCHUMANN et al. 2018a	Semeval-2018 task 7 : Semantic relation extraction and classification in scientific papers
A25	P. WANG, Z. XIE et Junfeng HU 2017	Relation Classification via CNN, Segmented Max-pooling, and SDP-BLSTM
A26	Lei ZHANG et F. XIANG 2018	Relation classification via bilstm-cnn
A27	Feiliang REN, R. ZHAO, Xiao HU et al. 2017	Embedding Syntactic Tree Structures into CNN Architecture for Relation Classification
A28	Y. SUN, CUI, Jinglu HU et al. 2018	Relation classification using coarse and fine-grained networks with SDP supervised key words selection
A29	KÖKSAL et ÖZGÜR 2020	The relx dataset and matching the multilingual blanks for cross-lingual relation classification
A30	K. DENG et Shaochun WU 2020	Improving relation classification by incorporating dependency and semantic information
A31	P. ZHOU, Wei SHI, Jun TIAN et al. 2016	Attention-based bidirectional long short-term memory networks for relation classification
A32	Zhimin LIN, LEI, Y. HAN et al. 2020	Siamese BERT Model with Adversarial Training for Relation Classification
A33	X. HAN, H. ZHU, P. YU, Ziyun WANG et al. 2018	Fewrel : A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation

A34	T. GAO, X. HAN, Zhiyuan LIU et al. 2019	Hybrid attention-based prototypical networks for noisy few-shot relation classification
A35	Z.-X. YE et Z.-H. LING 2019a	Multi-level matching and aggregation network for few-shot relation classification
A36	Q. YE, Liyuan LIU, Maosen ZHANG et al. 2019	Looking beyond label noise : Shifted label distribution matters in distantly supervised relation extraction
A37	SOARES, FITZGERALD, J. LING et al. 2019b	Matching the blanks : Distributional similarity for relation learning
A38	Linfang WU, H.-P. ZHANG, Yaofei YANG et al. 2020	Dynamic prototype selection by fusing attention mechanism for few-shot relation classification
A39	J. FENG, M. HUANG, Li ZHAO et al. 2018	Reinforcement learning for relation classification from noisy data
A40	K. YANG, L. HE, X. DAI et al. 2019	Exploiting noisy data in distant supervision relation classification
A41	HUI, Liang LIU, J. CHEN et al. 2020	Few-shot relation classification by context attention-based prototypical networks with BERT
A42	T. GAO, X. HAN, H. ZHU et al. 2019	FewRel 2.0 : Towards more challenging few-shot relation classification
A43	Yingyao WANG, J. BAO, Guangyi LIU et al. 2020a	Learning to decouple relations : Few-shot relation classification with entity-guided attention and confusion-aware training
A44	Fangchao LIU, X. XIAO, L. YAN et al. 2021a	From Learning-to-Match to Learning-to-Discriminate : Global Prototype Learning for Few-shot Relation Classification
A45	S. LIU et Fuji REN 2012a	Relation extraction from wikipedia articles by entities clustering
A46	S. LIU et Fuji REN 2012b	Relation extraction from wikipedia articles by entities clustering
A47	W. SHEN, Jianyong WANG, P. LUO et al. 2011	REACTOR : a framework for semantic relation extraction and tagging over enterprise data
A48	JINXIU, DONGHONG, LIM et al. 2005	Automatic relation extraction with model order selection and discriminative label identification
A49	TAKASE, OKAZAKI et INUI 2015	Fast and large-scale unsupervised relation extraction
A50	AKBIK, VISENGERIYEVA, HERGER et al. 2012	Unsupervised discovery of relations and discriminative extraction patterns
A51	DE LACALLE et LAPATA 2013	Unsupervised relation extraction with general domain knowledge

A52	Yingyao WANG, J. BAO, Guangyi LIU et al. 2020b	Learning to decouple relations : Few-shot relation classification with entity-guided attention and confusion-aware training
A53	Y. DAI, W. GUO, Xing CHEN et al. 2018a	Relation classification via LSTMs based on sequence and tree structure
A54	KAMATH, KARIBASAPPA, REDDY et al. 2021	Improving the Relation Classification Using Convolutional Neural Network
A55	Xiangju LI, S. FENG, Yifei ZHANG et al. 2021	Multi-level Emotion Cause Analysis by Multi-head Attention Based Multi-task Learning
A56	LYU et H. CHEN 2021	Relation Classification with Entity Type Restriction
A57	Ya ZHANG et S. QIN 2021	Improving Relation Classification with Multi-graph GCN
A58	Z.-X. YE et Z.-H. LING 2019b	Multi-level matching and aggregation network for few-shot relation classification
A59	Jiawen ZHANG, Jiaqi ZHU, Yi YANG et al. 2021a	Knowledge-Enhanced Domain Adaptation in Few-Shot Relation Classification
A60	Y. TANG, Zhezhou LI, C. CAO et al. 2021	Knowledge-Based Diverse Feature Transformation for Few-Shot Relation Classification
A61	Fangchao LIU, X. XIAO, L. YAN et al. 2021b	From Learning-to-Match to Learning-to-Discriminate : Global Prototype Learning for Few-shot Relation Classification
A62	SEKI, K. ZHAO, OGUNI et al. 2020	A framework for classifying temporal relations with question encoder
A63	P. QIN, W. XU et J. GUO 2016	An empirical convolutional neural network approach for semantic relation classification
A64	SOCHER, HUVAL, MANNING et al. 2012	Semantic compositionality through recursive matrix-vector spaces
A65	M. WU, Lin LIU, W. YAO et al. 2017	Semantic relation classification by bi-directional lstm architecture
A66	X. GUO, Hui ZHANG, H. YANG et al. 2019a	A single attention-based combination of CNN and RNN for relation classification
A67	L. LI, Jiabing WANG, Jichang LI et al. 2019	Relation classification via keyword-attentive sentence mechanism and synthetic stimulation loss
A68	Y. DAI, W. GUO, Xing CHEN et al. 2018b	Relation classification via LSTMs based on sequence and tree structure
A69	Yang LIU, S. LI, F. WEI et al. 2016	Relation classification via modeling augmented dependency paths
A70	Runyan ZHANG, MENG, Y. ZHOU et al. 2018	Relation classification via LSTMs based on sequence and tree structure

A71	B. LI, X. ZHAO, Shuai WANG et al. 2017	Relation classification using revised convolutional neural networks
A72	S. N. DWIVEDI, KARNICK et JAIN 2020	Relation Classification : How Well Do Neural Network Approaches Work?
A73	GENG, Xiwen CHEN, K. Q. ZHU et al. 2020	MICK : A meta-learning framework for few-shot relation classification with small training data
A74	PETRONI, DEL CORRO et GEMULLA 2015	Core : Context-aware open relation extraction with factorization machines
A75	ZARGAYOUNA, TELLIER, BUSCALDI et al. 2016	Unsupervised relation extraction in specialized corpora using sequence mining
A76	L. QIU et Yue ZHANG 2014	Chinese open relation extraction and knowledge base establishment
A77	LONG, Ye WANG, X. WEI et al. 2021	Entity-Centric Fully Connected GCN for Relation Classification
A78	Hongru WANG, Z. JIN, J. CAO et al. 2021	Inconsistent Few-Shot Relation Classification via Cross-Attentional Prototype Networks with Contrastive Learning
A79	Y. XIAO, Y. JIN et HAO 2021	Adaptive prototypical networks with label words and joint representation learning for few-shot relation classification
A80	ZOU, R.-T. HUANG, Z.-Z. XU et al. 2021	Language adaptation for entity relation classification via adversarial neural networks
A81	Y. HUANG, Zhixing LI, W. DENG et al. 2021	D-BERT : Incorporating dependency-based attention into BERT for relation extraction
A82	JAMISON 2011	Using grammar rule clusters for semantic relation classification
A83	ESULI, MARCHEGGIANI et SEBASTIANI 2010	ISTI@ SemEval-2 Task 8 : Boosting-Based Multiway Relation Classification
A84	L. QIU et Yue ZHANG 2014	ZORE : A syntax-based system for chinese open relation extraction
A85	MORENO, DOUCET et GRAU 2021	Relation Classification via Relation Validation
A86	R. CAI, Xiaodong ZHANG et Houfeng WANG 2016a	Bidirectional recurrent convolutional neural network for relation classification
A87	L. WANG, Z. CAO, DE MELO et al. 2016	Relation classification via multi-level attention cnns
A88	Shanchan WU et Yifan HE 2019a	Enriching pre-trained language model with entity information for relation classification

A89	OBAMUYIDE et VLACHOS 2019	Model-agnostic meta-learning for relation classification with limited
A90	W. ZHU, X. QIU, Y. NI et al. 2020	AutoRC : Improving BERT based relation classification models via
A91	T. GAO, X. HAN, R. XIE et al. 2020	Neural snowball for few-shot relation learning
A92	P. QIN, W. XU et W. Y. WANG 2018	Robust distant supervision relation extraction via deep reinforcement learning
A93	Jiawen ZHANG, Jiaqi ZHU, Yi YANG et al. 2021b	Knowledge-Enhanced Domain Adaptation in Few-Shot Relation Classification
A94	C.-Y. CHEN et C.-T. LI 2021	ZS-BERT : Towards Zero-Shot Relation Extraction with Attribute Representation Learning
A95	SCHMIDEK et BARBOSA 2014	Improving open relation extraction via sentence re-structuring
A96	SABO, ELAZAR, GOLDBERG et al. 2021	Revisiting Few-shot Relation Classification : Evaluation Data and Classification Schemes
A97	W. ZHU, X. QIU, Y. NI et al. 2020	AutoRC : Improving BERT based relation classification models via
A98	Y. JIANG, G. WU, BU et al. 2018	Chinese entity relation extraction based on syntactic features
A99	C. LIN, MILLER, DLIGACH et al. 2018	Self-training improves recurrent neural networks performance for temporal relation extraction
A100	Yijie ZHANG, Peifeng LI et Guodong ZHOU 2018	Classifying temporal relations between events by deep BiLSTM
A101	PATEL et TANWANI 2018	TEMPORAL RELATION IDENTIFICATION FROM CLINICAL TEXT USING LSTM BASED DEEP LEARNING MODEL
A102	COSTA et BRANCO 2012	Aspectual type and temporal relation classification
A103	Xiaobin ZHANG, Fucai CHEN et R. HUANG 2018	A combination of RNN and CNN for attention-based relation classification
A104	J. NI, ROSSIELLO, GLIOZZO et al. 2022	A Generative Model for Relation Extraction and Classification
A105	X. ZHOU, Peifeng LI, Q. ZHU et al. 2020a	Incorporating temporal cues and ac-gcn to improve temporal relation classification
A106	COHN, Yulan HE et Yang LIU 2020	Findings of the Association for Computational Linguistics : EMNLP 2020
A107	X. ZHOU, Peifeng LI, Q. ZHU et al. 2020b	Incorporating temporal cues and ac-gcn to improve temporal relation classification

A108	Jing XU, GAN, Z. YAN et al. 2016	Open relation extraction from chinese microblog text
A109	SEKI, K. ZHAO, OGUNI et al. 2020	A framework for classifying temporal relations with question encoder
A110	SEKI, K. ZHAO, OGUNI et al. 2021	CNN-based framework for classifying temporal relations with question encoder
A111	F. CHENG et MIYAO 2017	Classifying temporal relations by bidirectional LSTM over dependency paths
A112	Q. DAI, INOUE, REISERT et al. 2018	Improving scientific relation classification with task specific supersense
A113	D. JIN, DERNONCOURT, SERGEEVA et al. 2018a	MIT-MEDG at SemEval-2018 task 7 : Semantic relation classification via convolution neural network
A114	MUHAMMAD, IQBAL, JAMES et al. 2019	Convolutional Neural Network for Core Sections Identification in Scientific Research Publications
A115	M. JIANG, D'SOUZA, AUER et al. 2020	Improving scholarly knowledge representation : evaluating BERT-based models for scientific relation classification
A116	Jianyi LIU, DUAN, Ru ZHANG et al. 2021	Relation classification via BERT with piecewise convolution and focal loss
A117	HETTINGER, DALLMANN, ZEHE et al. 2018	Claire at SemEval-2018 task 7 : classification of relations using embeddings
A118	YIN, Z. LUO, W. LUO et al. 2018	IRCMS at SemEval-2018 task 7 : evaluating a basic CNN method and traditional pipeline method for relation classification
A119	D. JIN, DERNONCOURT, SERGEEVA et al. 2018b	MIT-MEDG at SemEval-2018 task 7 : Semantic relation classification via convolution neural network
A120	SUÁREZ-PANIAGUA, SEGURA-BEDMAR et AIZAWA 2018	UC3M-NII Team at SemEval-2018 Task 7 : Semantic Relation Classification in Scientific Papers via Convolutional Neural Network
A121	S. ZHAO et GRISHMAN 2005b	Improving relation classification by entity pair graph
A122	GÁBOR, BUSCALDI, SCHUMANN et al. 2018b	Semeval-2018 task 7 : Semantic relation extraction and classification in scientific papers

A123	SHWARTZ et DAGAN 2016	The roles of pathbased and distributional information in recognizing lexical semantic relations
A124	Y. XU, R. JIA, MOU et al. 2016	Improved relation classification by deep recurrent neural networks with data augmentation
A125	POUR et SHAMSFARD 2019	EoANN : Lexical Semantic Relation Classification Using an Ensemble of Artificial Neural Networks
A126	Chengyu WANG, M. QIU, J. HUANG et al. 2020	KEML : A knowledge-enriched meta-learning framework for lexical relation classification
A127	Chengyu WANG, X. HE et A. ZHOU 2019	Spherere : Distinguishing lexical relations with hyperspherical relation embeddings
A128	S. JIA, E, M. LI et al. 2018	Chinese open relation extraction and knowledge base establishment
A129	SOARES, FITZGERALD, J. LING et al. 2019a	Matching the blanks : Distributional similarity for relation learning
A130	Lulu ZHAO, W. XU, S. GAO et al. 2020	Cross-sentence N-ary relation classification using LSTMs on graph and sequence structures

ANNEXE B

B. Liste des papiers liés à l'extraction de contexte

Papier	Référence	Titre
B1	LAHLOU, MOUNTASSIR, BENBRAHIM et al. 2013	A Text Classification Based Method for Context extraction
B2	Guolong LIU, X. XU, Y. ZHU et al. 2014	An Improved Latent Dirichlet Allocation Model for hot topic extraction
B3	ANAM, RAHMAN, SALEHEEN et al. 2018	Automatic Text Summarization using Fuzzy C-Means
B4	SAJID, JAN et SHAH 2017b	Automatic Topic Modeling for Single Document Short Texts
B5	LAHLOU, BENBRAHIMAND, MOUNTASSIR et al. 2013	Context Extraction from Reviews for Context Aware
B6	YUN, JING, J. YU et al. 2011	Document Topic Extraction Based on Wikipedia Category
B7	S. KIM, JEON, J. KIM et al. 2012	Finding Core Topics Topic Extraction with Clustering on Tweet
B8	JAHNAVI et RADHIKA 2013	Hot Topic Extraction based on Frequency
B9	H.-F. MA 2011	HOT TOPIC EXTRACTION USING TIME WINDOW
B10	DING et W. JIN 2019	Improves LDA in both document representation and topic extraction
B11	YONGLI, HUANHUAN, XIAOZE et al. 2018	QH-K Algorithm for News Text Topic Extraction
B12	FRANZONI et MILANI 2015	Semantic Context Extraction From Collaborative Networks
B13	Yue ZHANG et Hongli ZHANG 2012	Social Topic Detection for Web Forum
B14	YOKOI et YANAGIMOTO 2010	Topic Extraction for a Large Document Set with the Topic Integration
B15	RAFEA et MOSTAFA 2013	Topic extraction in social media
B16	GANGEMI 2013	A Comparison of Knowledge Extraction Tools for semantic web
B17	Zihuan WANG, HAHN, Y. KIM et al. 2018	A news-topic recommender system based on keywords extraction
B18	YOU, FONTAINE et BARTHÈS 2013	An automatic keyphrase extraction system for scientific
B19	JAMEEL, N. SINGH, N. K. SINGH et al. 2009	An Intelligent Automatic Text Summarizer
B20	GALI, MARIESCU-ISTODOR et FRÄNTI 2017	Using linguistic features to automatically extract web page title

B21	Y. SHIN, RYO et J. PARK 2014	Automatic extraction of persistent topics
B22	FRANZONI, Yuanxi LI, MENGONI et al. 2017	Clustering Facebook for Biased Context Extraction
B23	RIZOIU et VELCIN 2011	Topic Extraction for Ontology Learning
B24	GALI, MARIESCU-ISTODOR et FRÄNTI 2016	Similarity Measures for Title Matching
B25	Chunzi WU, B. WU et B. WANG 2016	Event Evolution Model with Hot Topic Extraction
B26	E. ZHOU, ZHONG et Yuefeng LI 2014	Extracting news blog hot topics based
B27	Jöran BEEL, Bela GIPP, Ammar SHAKER et al. 2010	Extracting Titles from Scientific PDF
B28	E. ZHOU, ZHONG et Yuefeng LI 2011	Hot Topic Detection in Professional Blogs
B29	XIA, Juanzi LI, J. TANG et al. 2012	Plink-LDA Using Link as Prior Information in Topic Modeling
B30	J. FENG et FANG 2016	Research on Hot Topic Discovery Technology of Micro-blog Based
B31	G. XU, Z. YU, Changzhi WANG et al. 2020	Research on topic discovery technology for Web news
B32	H. JIANG, R. ZHOU, Limeng ZHANG et al. 2019	Sentence level topic models for associated topics extraction
B33	Y. ZHAO, W. FU et Shaobin HUANG 2011	Topic Discovery and Topic-Driven Clustering
B34	CHEW 2015	Topic Extraction from Multilingual Social Media Data
B35	Xuefei ZHANG et R. HE 2018	Topic Extraction of Events on Social Media Using Reinforced Knowledge
B36	S. YANG, Q. SUN, H. ZHOU et al. 2018	A Topic Detection Method Based on KeyGraph and Community Partition
B37	MURFI 2017	Accuracy of separable nonnegative matrix factorization for topic extraction
B38	Zhiyuan LIU, Wenyi HUANG, Y. ZHENG et al. 2010	Automatic keyphrase extraction via topic decomposition
B39	S. LEE, J. LEE, C.-Y. PARK et al. 2013b	Blog topic analysis using TF smoothing and LDA
B40	ANGROSH, CRANEFIELD et STANGER 2013	Conditional random field based sentence context identification
B41	X. WANG, Yang LIU, Donghui WANG et al. 2013	Cross-media topic mining on wikipedia
B42	Joeran BEEL, LANGER, GENZMEHR et al. 2013	Docear's PDF inspector title extraction from PDF files
B43	ANH, TAM et VAN LINH 2013	Document clustering using dirichlet process mixture model

B44	GOYAL, MALLA, BAGCHI et al. 2013	ESTHETE a news browsing system to visualize the context
B45	N. GUO, Yuan HE, C. YAN et al. 2016b	Multi-level topical text categorization with wikipedia
B46	PAN, M. X. ZHOU, Yangqiu SONG et al. 2013	Optimizing temporal topic segmentation for intelligent text visualization
B47	VIEGAS, LUIZ, GOMES et al. 2018	Semantically-Enhanced Topic Modeling
B48	Di WANG, THINT et AL-RUBAIE 2012	Semi-Supervised Latent Dirichlet Allocation and Its Application for dc
B49	G. ZHAO, Yanjun LIU, W. ZHANG et al. 2018	TFIDF based Feature Words Extraction and Topic Modeling for Short Text
B50	MOHAMMADZADEH, GOTTRON, SCHWEIGGERT et al. 2012	Title Finder extracting the headline of news web pages based on cosine similarity
B51	P. C. KAUR, GHORPADE et MANE 2016	Topic Extraction and Sentiment Classification by using Latent Dirichlet
B52	TATSUKAWA 2012	Topic extraction based on prior knowledge obtained from target documents
B53	SITKRONGWONG et TAKASU 2019	Unsupervised context extraction via region embedding for context
B54	Ximing LI, Yang WANG, OUYANG et al. 2021	Topic extraction from extremely short texts with variational manifold regularization
B55	Yi ZHANG, LU, Feng LIU et al. 2018	A kernel k-means clustering method with word embedding
B56	Chanle WU, M. XIE, Yunlu ZHANG et al. 2011	A New Intelligent Topic Extraction Model on Web
B57	Qihua LIU 2015b	A Novel Chinese Text Topic Extraction Method based on LDA
B58	KARL, WISNOWSKI et RUSHING 2015	A practical guide to text mining with topic extraction
B59	K. KAUR et V. GUPTA 2012	A Survey of Topic Tracking Techniques
B60	NUNES, MERA, KAWASE et al. 2014	A Topic Extraction Process for Online Forums
B61	PANG, NIU, Jiqiang LIU et al. 2018	An Approach to Generate Topic Similar Document by Seed Extraction-based SeqGAN Training for Bait Document
B62	HONG et ZHEN 2012b	An Extended Keyword Extraction Method
B63	GALI et FRÄNTI 2016	Content-based Title Extraction from Web Page
B64	REKIK et JAMOUSSI 2016	Deep Learning for Hot Topic Extraction from Social Streams
B65	BEEL, GIPP, SHAKER et al. s. d.	Extracting Titles from Scientific PDF Documents by Analyzing Style Information

B66	JARRAH 2019	Factivity and subject extraction in Jordanian Arabic
B67	KHODRA, WIDYANTORO, AZIZ et al. 2011	Free Model of Sentence Classifier for Automatic Extraction of Topic Sentences
B68	BADAWY, FISTEUS, MAHMOUD et al. 2021	Topic Extraction and Interactive Knowledge Graphs for Learning Resources
B69	K.-L. NGUYEN, B.-J. SHIN et YOO 2016	Hot Topic Detection and Technology Trend Tracking
B70	ARYA et S. k. DWIVEDI 2018	Keyphrase Extraction of News Web Pages
B71	Qingtang LIU, SHAO, Linjing WU et al. 2017	Main Content Extraction from Web Pages Based on Node Characteristic
B72	Z. F. SUN et K. J. BAO 2011	Research of Text Topic Automatic Extraction Method Based on Rough set theory
B73	ABASI, KHADER, AL-BETAR, NAIM, MAKHADMEH et al. 2021	A novel ensemble statistical topic extraction method for scientific publications based on optimization clustering
B74	ABASI, KHADER, AL-BETAR, NAIM, ALYASSERI et al. 2021	An ensemble topic extraction approach based on optimization clusters using hybrid multi-verse optimizer for scientific publications
B75	RAJANGAM et ANNAMALAI 2021	Topic extraction using local graph centrality and semantic similarity