
*ACADÉMIE D'AIX-MARSEILLE*
*UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE*

---

# THÈSE

présentée à l'Université d'Avignon et des Pays de Vaucluse
pour obtenir le diplôme de DOCTORAT

**SPÉCIALITÉ : Informatique**

École Doctorale 380 « Sciences et Agronomie »
Laboratoire d'Informatique (EA 931)

# *Résumé automatique de parole pour un accès efficace aux bases de données audio*

par
**Benoît FAVRE**

**Soutenue publiquement le 19 mars 2007 devant un jury composé de :**

| | | | |
|---|---|---|---|
| M$^{me}$ | Catherine BERRUT | Professeur, LIG, Grenoble | Présidente du jury |
| M. | Guy LAPALME | Professeur, RALI, Montréal | Rapporteur |
| M. | François YVON | Maître de Conférences, ENST, Paris | Rapporteur |
| M. | Patrick GALLINARI | Professeur, LIP6, Paris | Examinateur |
| M. | François CAPMAN | Ingénieur, Thales, Colombes | Examinateur |
| M. | Jean-François BONASTRE | Maître de Conférences, LIA, Avignon | Directeur de thèse |
| M. | Patrice BELLOT | Maître de Conférences, LIA, Avignon | Co-directeur de thèse |

---


Laboratoire d'Informatique d'Avignon

# Remerciements

Je tiens tout d'abord à remercier Jean-François Bonastre, Patrice Bellot et François Capman pour leur encadrement, leurs nombreux conseils et leur soutien constant tout au long de ma thèse. Je remercie Catherine Berrut pour avoir accepté d'être présidente de mon jury. J'ai également été très honoré par Guy Lapalme et François Yvon qui ont accepté d'être rapporteurs de ma thèse. Ils m'ont apporté de précieux conseils sur l'ensemble de mon travail. J'exprime ma profonde gratitude à Patrick Gallinari pour sa participation à mon jury.

Je dédie ce document à mes proches Laure, Floriane, Martine et Roger et à la mémoire de mon frère Julien. Leur soutien tout au long de ce travail a été inestimable.

Je tiens aussi à remercier ceux qui m'ont tant apporté durant mes journées et mes soirées par leur joie, leur gentillesse et leur amitié. Par ordre alphabétique, ça donne quelque chose comme : Alex, Anakin, Andrea, Annypog, Antho, Audrey, Ben, Bertrand, Cathy, Christophe, Cissou, Corinne, Denz, Dju, Domi, Driss, Eric, Florian, Fred B., Fred D., Fred W., Gayp, Georges, Gilles, J.-P., Joce, Jocelyne, Lapo, Laurent, Laurianne, Lolo, Louisa, M.-J., Maman Ours, Marc P., Max, Med, Mimi, Nanou, Nath, Nenex, Neug, Nick, Nico, Nicolas F., Nicole, Nimaan, Olivier, Ourselin, Papa Ours, Pascal, Phanou, Phillou, Pierrot, Quang, Ralph, Rico, Riton, Romane, Sarah, Simone, Stan, Steph, Tania, Ted, Thierry S., Thierry V., Tom, Virginie, et Will.

# Résumé

L'avènement du numérique permet de stocker de grandes quantités de parole à moindre coût. Malgré les récentes avancées en recherche documentaire audio, il reste difficile d'exploiter les documents à cause du temps nécessaire pour les écouter. Nous tentons d'atténuer cet inconvénient en produisant un résumé automatique parlé à partir des informations les plus importantes. Pour y parvenir, une méthode de résumé par extraction est appliquée au contenu parlé, transcrit et structuré automatiquement. La transcription enrichie est réalisée grâce aux outils Speeral et Alize développés au LIA. Nous complétons cette chaîne de structuration par une segmentation en phrases et une détection des entités nommées, deux caractéristiques importantes pour le résumé par extraction. La méthode de résumé proposée prend en compte les contraintes imposées par des données audio et par des interactions avec l'utilisateur. De plus, cette méthode intègre une projection dans un espace pseudo-sémantique des phrases. Les différents modules mis en place aboutissent à un démonstrateur complet facilitant l'étude des interactions avec l'utilisateur. En l'absence de données d'évaluation sur la parole, la méthode de résumé est évaluée sur le texte lors de la campagne DUC 2006. Nous simulons l'impact d'un contenu parlé en dégradant artificiellement les données de cette même campagne. Enfin, l'ensemble de la chaîne de traitement est mise en œuvre au sein d'un démonstrateur facilitant l'accès aux émissions radiophoniques de la campagne ESTER. Nous proposons, dans le cadre de ce démonstrateur, une frise chronologique interactive complémentaire au résumé parlé.

## Mots-clés

Résumé Automatique de Parole, Recherche d'Information Parlée, Reconnaissance Automatique de la Parole, Transcription Enrichie, Frontière de Phrase, Entité Nommée, Maximal Marginal Relevance, MMR, Conditional Random Fields, CRF, Latent Semantic Analysis, LSA, Document Understanding Conference, DUC.

# Abstract

The digital era has revealed new ways to store great quantities of speech at a low cost. Whereas recent advances in spoken document retrieval, exploiting audio documents is still difficult because of the time necessary to listen to them. We try to attenuate this disadvantage by producing an automatic spoken abstract from the most important information. For that purpose, an extractive summarization algorithm is applied to the spoken content thanks to automatic speech structuring. The rich transcription is carried out thanks to Speeral and Alize toolkits developed at LIA. We complement this structuring chain by sentence segmentation and named entities detection, two important features for extractive summarization. The proposed summarization approach includes constraints imposed by audio data and interactions with the user. Moreover, the method integrates a projection of sentences in pseudo-semantic-space. We integrated the various modules in a coherent prototype that ease the study of user interactions. Due to the lack of evaluation data for the speech summarization task, we evaluate our approach on the textual documents from the DUC 2006 campaign. We simulate the impact of spoken content structuring by artificially degrading the textual content provided for DUC. Finally, the whole processing sequence is implemented within a demonstrator facilitating the access radio broadcasts from the ESTER evaluation campain. Within the framework of this prototype, we present an interactive timeline that aims at recontextualizing the spoken summary.

## Keywords

Automatic speech summarization, Spoken Document Retrieval, Automatic Speech Recognition, Rich Transcription, Sentence Boundaries, Named Entities, Maximal Marginal Relevance, MMR, Conditional Random Fields, CRF, Latent Semantic Analysis, LSA, Document Understanding Conference, DUC.

# Résumé long

L'avènement du numérique permet de stocker de grandes quantités de parole à moindre coût. Afin de les exploiter, la recherche documentaire audio tire parti de la transcription automatique du discours parlé. Malgré la compensation des erreurs de transcription, les moteurs de recherche sur la parole nécessitent une écoute des documents car il est plus difficile d'obtenir un aperçu d'un contenu audio que d'un contenu écrit ou visuel. Cette caractéristique provoque une réduction de la quantité d'information perçue par l'utilisateur à cause de longues et fastidieuses écoutes limitant l'efficacité des moteurs de recherche audio. Pour répondre à cette difficulté, nous proposons de générer un résumé parlé des informations les plus importantes retrouvées par le moteur de recherche. Pour y parvenir, une méthode de résumé par extraction est appliquée au contenu parlé, transcrit et structuré automatiquement.

La transcription enrichie est réalisée grâce aux outils Speeral et Alize développés au LIA. Nous complétons cette chaîne de structuration par une segmentation en phrases et une détection des entités nommées, deux caractéristiques importantes pour le résumé par extraction. Les frontières de phrases sont retrouvées par étiquetage de séquence grâce à une modélisation Conditional Random Fields (CRF) fondée sur des caractéristiques prosodiques et linguistiques. L'approche est validée sur le corpus radiophonique ESTER (précision de 0.77). Par ailleurs, les entités nommées sont détectées directement dans le graphe d'hypothèses de transcription pour essayer d'atténuer l'influence des erreurs commises lors de la phase de transcription. Le cadre des transducteurs pondérés (Weighted Finite State Transducers, WFST) permettent l'application de grammaires locales au treillis d'hypothèses, puis sa pondération par un modèle d'étiquetage HMM. Testée lors de la tâche expérimentale de la campagne ESTER, la méthode obtient les meilleurs résultats ($F_1$-mesure de 0.63).

Le modèle proposé pour le résumé de parole prend en compte la nature du contenu parlé et les contraintes imposées par les interactions avec l'utilisateur. Ce modèle est implémenté dans *Maximal Marginal Relevance* (MMR), sous la forme d'une séparation des caractéristiques d'une phrase en une partie indépendante du besoin de l'utilisateur (pouvant être calculée en temps différé) et une partie dépendante de ce besoin (soumise à des contraintes de complexité). La longueur des phrases ou les scores de confiance de la structuration peuvent être considérés comme indépendants du besoin. La similarité d'une phrase avec le besoin de l'utilisateur est calculée après projection des phrases dans un espace pseudo-sémantique construit par *Latent Semantic Analysis* (LSA).

Évaluée sur le résumé de texte, au travers d'une soumission conjointe LIA-Thales, lors de la campagne *Document Understanding Conference* (DUC 2006), la méthode obtient des résultats au niveau de l'état de l'art (avec un rang de 6/34, selon l'évaluation automatique Rouge). Aucune donnée d'évaluation n'existant pour la parole, nous simulons l'impact d'un contenu parlé sur les données textuelles de DUC. Les erreurs de structuration sont simulées à l'aide d'insertions, de suppressions et de substitutions de mots dans les documents. Ces erreurs sont introduites uniformément pour limiter la tendance du système de résumé à sélectionner des phrases contenant moins d'erreurs. Une limitation du vocabulaire aux mots les plus fréquents n'a pas d'impact significatif sur le critère d'évaluation Rouge. Par contre, la dégradation systématique des entités nommées provoque une chute de ce critère. Une étude de l'évolution de Rouge par rapport au taux d'erreurs de mots dans les documents d'origine montre que dans une condition où le résumé est « écouté », le système est robuste jusqu'à environ 40% d'erreurs. Au contraire, lorsque le résumé est « lu », les performances chutent dès 10% d'erreurs. Ces résultats sont à relativiser car le comportement du critère Rouge lorsque les données sont dégradées reste encre peu étudié.

La chaîne de structuration et le système de résumé sont mis en œuvre sur les données radiophoniques de la campagne ESTER. Le prototype ainsi conçu démontre la faisabilité des méthodes proposées et permettra une évaluation directe auprès d'utilisateurs. Une première étude à été réalisée pour valider l'ergonomie du démonstrateur et tester le concept de frise chronologique interactive, une vue complémentaire au résumé parlé. Cette vue permet d'améliorer la perception, par l'utilisateur, de l'organisation temporelle des informations retrouvées et ouvre la voie vers de nouveaux outils de navigation fondés sur le résumé.

# Résumé généré automatiquement

Les recherches sur les interfaces d'accès à une base de données audio ont convergé vers l'utilisation de la même métaphore que celle permettant l'accès à des documents textuels, ou des documents indexés par des métadonnées textuelles. Le besoin utilisateur et l'information parlée sont projetés dans un espace sémantique, puis des méthodes de recherche documentaire et de résumé automatique permettent la génération d'un résumé parlé. Les premières approches de la recherche d'information dans un contenu parlé ont d'abord utilisé des techniques similaires à celle développées pour les documents textuels, appliquées à la transcription automatique du flux de parole. Un résumé automatique de parole est constitué à partir d'un flux audio parlé (entrées) et généré sous forme écrite ou parlée (sorties). L'objectif de ces travaux est de faciliter l'accès à l'information audio à l'aide du résumé de parole et les éléments de structuration présentés au chapitre précédent ne sont pas suffisants pour obtenir un résumé de qualité. Ces derniers travaux représentent une ouverture vers la recherche de méthodes pour l'accès à l'information parlée pour aller plus loin qu'une simple amélioration des méthodes de structuration et de sélection de l'information. Nous avons proposé dans cette étude d'améliorer l'efficacité de l'accès à des bases de données parlées à l'aide d'une approche reposant sur le résumé automatique de parole. Toutefois, le manque de données dédiées au résumé automatique de parole dans le cadre d'ESTER (cette campagne ne proposait pas de tâche directement liée au résumé automatique), n'a pas permis une évaluation exhaustive de la méthode proposée.

*Ce résumé a été produit par la méthode exposée dans ce document. Des détails sur le procédé sont disponibles dans l'annexe B.*

# Table des matières

# Chapitre 1

# Introduction

**Sommaire**



## 1.1 Recherche d'information

La recherche d'information a pour objectif de délester l'utilisateur d'une exploration exhaustive d'un ensemble de données en lui fournissant directement l'information qui l'intéresse, dans une représentation utile. L'intérêt pour une information varie dans le temps et il est parfois impossible d'explorer l'ensemble des données avant qu'elles ne deviennent obsolètes. L'outil *recherche d'information* est principalement considéré comme un gain de temps vis à vis d'une recherche exhaustive. Cette définition fait apparaître différentes notions : l'*information*, les *données*, l'*utilisateur*, l'intérêt de l'utilisateur, plus généralement appelé *besoin*, et enfin la manière de représenter la réponse à ce besoin, nommée *résultat*.

L'*information* est souvent définie comme une forme évoluée de *donnée*. Alors que les *données* dénotent une composante de l'univers (état, absence d'état ou changement d'état), l'*information* correspond à une notion d'ordre par opposition au chaos : elle apporte une valeur ajoutée par rapport à une simple *donnée*. De façon plus générale, l'*information* représente une « connaissance » extraite de la *donnée*, qui lui sert de support, de représentation dans le monde réel. Cette définition peut susciter de nombreuses interrogations (Floridi, 2005) comme : l'absence de données est-elle une information ? L'information a-t-elle nécessairement un support physique ? L'information existe-elle indépendamment de la pensée humaine ? Quel est le lien entre vérité et information ? Bien que la réponse à ces questions soit loin d'être triviale, nous simplifions en établissant une hypothèse nécessaire à la recherche d'information : l'*information* est

une *donnée* qui a une importance à l'échelle de l'homme. Elle n'existe que si un *utilisateur* s'y intéresse et cet intérêt est appelé *besoin*. Le *besoin* est une expression de ce que recherche l'*utilisateur*, ce qui lui est important dans un contexte donné. Il indique à l'outil de recherche d'information comment trouver un îlot d'informations *pertinentes* au milieu d'un océan d'informations *non pertinentes*. Cette notion de pertinence est très importante car elle permet d'évaluer la qualité de l'outil de recherche d'information. Le *résultat* d'une recherche d'information est la représentation dans le monde réel de l'*information* intéressante répondant au *besoin* de l'*utilisateur*. Ce résultat prend généralement la forme d'un sous ensemble d'objets servant à stocker les informations (une liste de documents, par exemple), mais peut aussi prendre une forme indépendante de ces objets, créée spécifiquement à l'intention de l'*utilisateur* (une liste de faits, de dates...).

Bien que la recherche d'information dans sa généralité soit un domaine de recherche très intéressant, nous nous focalisons sur l'information audio et, plus précisément le sous domaine de la parole. Depuis l'apparition de l'espèce humaine, les phénomènes sociaux ont été les plus grands moteurs de son évolution cognitive (Donald, 1991). La parole y a joué un grand rôle car elle est le moyen de communication favori entre les personnes et surtout une concrétisation de la pensée grâce au langage. La parole est apparue bien avant l'écriture et reste dans certaines cultures la principale forme d'expression et de mémoire collective. Par exemple, la langue somalienne est parlée par environ 15 millions de personnes et écrite par moins de la moitié de cette population, dans une forme translitérée qui n'existe que depuis 1972 (il faut d'ailleurs se demander si l'écriture aurait été inventée, si, au moment où le besoin de conserver la parole est né, un objet avait permis d'enregistrer un message et de le rediffuser). En considérant qu'une personne produit en moyenne 1 heure de parole par jour, soit environ 10000 mots, 6 milliards d'humains prononcent $6 \times 10^{16}$ mots par jour, une quantité bien plus grande que le volume de documents écrits produit dans le même temps (qui pourrait être estimé à 100 mots en moyenne par jour et par personne, sachant que 16% de la population mondiale est analphabète, selon une étude des Nations Unies en 1998). Le côté éphémère de la parole agit-il plus comme frein à l'évolution de la pensée, par la quantité d'information utile perdue, ou comme un catalyseur de cette évolution, par la sélection « darwinienne » de ce qui est important ? Dans le cadre de la première hypothèse, comment une société réagirait-elle si toutes les conversations étaient enregistrées et analysées pour un usage ultérieur ? En ignorant le problème de la réutilisation d'un discours sorti de son contexte, doit-on imaginer que le droit à l'erreur (parlée) devienne un droit fondamental de la personne ?

Le langage parlé est bien différent du langage écrit (Biber, 1991) et véhicule beaucoup plus d'informations, qui peuvent être classées en fonction du degré d'intention du locuteur lié à leur existence. Les informations non intentionnelles sont le contexte, l'état du locuteur et les contraintes de la langue. Le contexte inclut par exemple un environnement acoustique qui force le locuteur à parler plus ou moins distinctement ; l'état du locuteur regroupe les émotions (dont les principales sont selon Laukka (2004) : la joie, la tristesse, la peur, la colère, le dégoût et la surprise) et les attitudes véhiculées par la voix, les pathologies de la voix et du langage (bégaiement, dysphasie...) ; enfin, la langue

contraint les idées à être exprimées sous formes de mots, alors qu'elles le seraient peut être mieux par des images ou des gestes. D'un autre côté, les informations intentionnelles incluent l'état volontaire du locuteur et le message en lui même. Par exemple, le locuteur peut adapter son élocution à un auditeur non natif en parlant lentement, ou utiliser des émotions et des attitudes jouées afin d'étayer son message. Contrairement à la parole, le langage écrit ne contient en plus du message que les informations issues des contraintes du langage et de la mise en forme (agencement, couleurs, police, taille), ce qui appauvrit la communication tout en limitant sa variabilité. Par exemple, pour pallier le manque d'émotions dans l'expression écrite, des frimousses (*smileys* ;-) sont employées dans une conversation textuelle. L'écrit agit comme une clarification permettant de se concentrer sur le message dont l'interprétation est facilitée par l'ensemble des règles de la langue, normalisées et largement enseignées. Il n'existe aucun système symbolique répandu pour représenter l'état du locuteur ou l'environnement acoustique. De plus, les hésitations, les reprises, les répétitions, les coupures, ou les éléments phatiques (remplacement d'une pause par des mots pour gagner du temps sans créer de discontinuité dans le flux vocal, comme par exemple « eh bien, tu vois », ou encore « truc », « machin » pour remplacer un échec d'accès lexical) sont des défauts d'élocution très courants et considérés comme une pollution du message à l'écrit, donc supprimés lors de la transcription. Il n'existe pas non plus de système de représentation de ces phénomènes très répandus. Le passage du parlé à l'écrit correspond à une réduction d'information, mais il ne faut pas oublier que le message est aussi appauvri en ne traitant que l'information audio, car la gestuelle, la posture et le contexte de la locution sont ignorés.

Les informations contenues dans un message parlé se traduisent en paramètres et en phénomènes mesurables. La voix est un signal acoustique résultant du passage de l'air par le conduit vocal (cordes vocales du larynx, pharynx, cavité buccale et cavité nasale), de sa modification par un environnement (écho, bruit, canal...), et de sa perception par un capteur (oreille, micro). Ce signal est tout d'abord caractérisé par des paramètres perceptifs comme par exemple la hauteur, le timbre, le rythme, la vitesse, la clarté ou le tremblement. Au niveau physique, ces paramètres peuvent être observés par étude du spectre, une analyse fréquentielle du signal par rapport au temps. La recherche d'information audio consiste donc en une analyse des propriétés acoustiques du signal audio pour déterminer ce qui pourrait intéresser l'utilisateur. La quantité d'informations à traiter peut être calculée facilement : généralement, on considère que les fréquences utiles de la voix sont inférieures à 8000 Hz, ce qui nécessite un échantillonnage à 16000 Hz selon le théorème de Shannon ; une heure d'audio correspond alors à 5 millions d'informations à réduire en 10000 mots pour approcher la sémantique du message et comparer sa substance au besoin de l'utilisateur.

## 1.2 Problématique

Le cycle de vie de l'information audio se compose de son *acquisition*, de son *stockage*, de sa *transmission*, de sa *recherche* et de sa *restitution*. L'*acquisition* de l'information

audio est effectuée par le microphone, inventé en 1876 par Edison, dont le rôle est de transformer le signal audio en signal électrique. Ce dernier peut être traité par un système électronique tel que l'ordinateur grâce à une numérisation du signal électrique par un convertisseur analogique-numérique. Dans le cadre de l'information parlée, la qualité de l'*acquisition* du message informatif est améliorée en diminuant l'influence de l'environnement ou en supprimant l'écho. Le *stockage* des données audio devient alors aussi simple que le stockage de données numériques génériques. Mais lorsque ces données doivent être stockées en grandes quantités, les supports numériques, de taille limitée, imposent une compression du signal audio. La solution la plus connue pour la compression audio est le mp3 (Brandenburg, 1999), qui applique un modèle psychoacoustique pour ne dégrader que les fréquences les moins audibles, afin de limiter la détérioration de la qualité générale. Les codages spécifiques à la voix permettent de réduire encore plus la quantité de données tout en conservant un maximum d'information, au détriment d'un signal restauré peu similaire au signal d'origine. La numérisation a aussi facilité la *transmission* de la voix en autorisant l'exploitation des réseaux de données (voix sur IP) tout en limitant le délai de transmission et la variation du délai (gigue), principales sources de dégradation du message (Goode, 2002). La *restitution* de l'information audio est simplement le renversement du processus d'acquisition, par une conversion numérique vers analogique, puis la transformation du signal électrique en onde acoustique par un haut-parleur. Ici encore, de nombreuses solutions permettent d'améliorer la qualité de la restitution du message dans des conditions d'environnement bruité, ou sous contraintes. La *recherche* d'information audio n'a pas connu autant d'avancées que les autres composantes du cycle de vie de l'information : elle se limite généralement à l'utilisation de méta-informations (étiquettes) décrivant le contenu audio, générées manuellement au moment de l'*acquisition*.

Les recherches sur les interfaces d'accès à une base de données audio ont convergé vers l'utilisation de la même métaphore que celle permettant l'accès à des documents textuels, ou des documents indexés par des métadonnées textuelles. En effet, le processus habituel est de transcrire le contenu parlé et d'exploiter cette transcription comme un texte classique. Ce phénomène s'explique en partie par l'orientation des campagnes de recherche documentaire textuelle (comme par exemple *Text REtrieval Conference*, TREC, organisées par National Institute of Standards and Technologies, NIST) qui se sont intéressées à l'application des méthodes textuelles sur des données dégradées comme celles issues de systèmes de transcription de la parole. Des prototypes comme SpeechBot (Thong et al., 2000) ont vu le jour suite à ces évaluations et ont montré que, bien que les performances de recherche documentaire soient acceptables compte tenu du taux d'erreur lors de la transcription, le manque de structuration des résultats et leur quantité était loin de satisfaire l'utilisateur. En effet, le modèle de la recherche documentaire prend pour cible des experts et des documentalistes recherchant l'information de façon exhaustive. Pour ce type d'utilisateur, le découpage des résultats est binaire, entre documents pertinents et documents non-pertinents. Afin d'éviter d'omettre un document pertinent, les systèmes de recherche documentaire génèrent une liste de documents classés par pertinence estimée. L'utilisateur parcourt les documents dans l'ordre et fait appel à ses capacités à déterminer si le document est pertinent sans l'assimiler dans son intégralité.

Cette capacité à discerner rapidement le contenu d'un document est essentielle aux systèmes traitant de grandes quantités de données audio car contrairement à l'information textuelle qui peut être parcourue rapidement, l'information audio impose un temps d'écoute égal à sa durée. Ce phénomène est dû à la nature éphémère du signal sonore : il ne persiste pas à travers le temps. En effet, s'il suffit de tourner la tête pour voir plus de choses, il faut attendre pour en entendre d'avantage. De plus, la vision permet une analyse des images à divers niveaux de détails, facilitant une approximation rapide du contenu qui n'est pas possible avec l'audio. Par exemple, il est très facile de lire un document en diagonale pour savoir s'il parle d'un sujet en particulier, mais beaucoup plus dur d'*écouter un document en diagonale*, même avec un dispositif adapté. Cette impossibilité d'appréhender rapidement un document audio empêche l'utilisation des techniques de recherche d'information textuelle directement sur la transcription en langage écrit du contenu vocal. Le problème ne se limite pas à la discrimination rapide des informations non pertinentes, mais s'étend aussi à l'appréciation globale du contenu des résultats. Cette problématique, commune à tous les systèmes utilisant une liste de documents ordonnés comme résultats d'une recherche, est basée sur la constatation que l'utilisateur va néanmoins assimiler l'ensemble des informations jugées pertinentes, ce qui peut être très long dans le cas d'un média lié au temps comme la parole. Le système doit alors être capable de donner un aperçu de l'ensemble des résultats en plus d'aider à la discrimination de chaque résultat.

### Le résumé de parole : un accès efficace aux données audio

Nous proposons de répondre à ces problématiques de la recherche d'informations parlées par la notion de résumé automatique de parole. Cette notion apporte l'idée de reproduire le comportement développé par les humains pour faire face à une grande quantité d'informations parlées : rapporter à un auditoire l'essentiel du discours d'un tiers. Cette définition est proche de celle du résumé de texte dans le sens où elle implique une interprétation de l'information, suivie d'une synthèse formulée en langue naturelle. Par contre, dans le cadre de la recherche d'information, nous considérons qu'il faut y ajouter l'étude du besoin de l'utilisateur et de la manière de répondre à ce besoin. Il s'agit finalement de l'étude du triplet *information parlée*, *besoin de l'utilisateur* et *système automatique* (voir la figure 1.1).

Le résumé automatique de parole n'est pas une nouveauté en soi, mais ce domaine émergeant représente un point où coïncident de nombreuses autres disciplines liées à l'étude de la parole et de la langue. Le travail présenté dans ce document a d'abord comme objectif de faire un tour d'horizon de ces disciplines, puis de construire une première solution complète fonctionnant dans des conditions réelles à partir de techniques existantes. Ce prototype facilitera l'étude globale du système sur une application, en plus des sous-parties du traitement, habituellement évaluées de manière indépendante. Le but est de poser les briques qui permettront d'explorer la question générale : le résumé automatique de parole est-il la solution à la croissance rapide de l'information parlée ? Cette question en induit quelques autres :

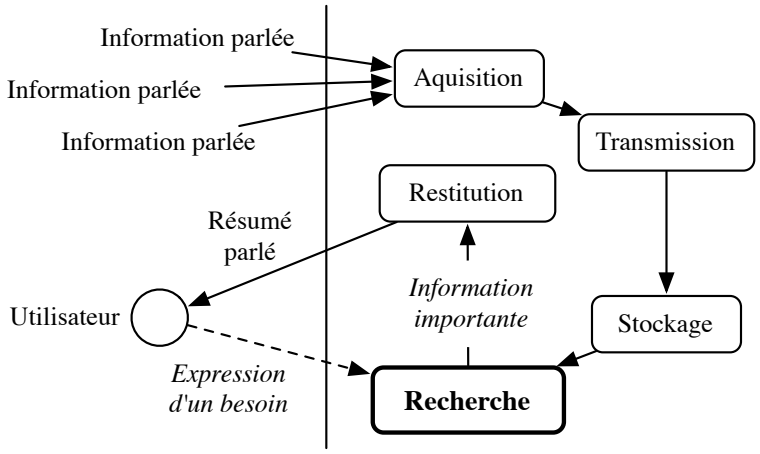



**FIG. *1.1:*** *Dans le cycle de vie de l'information, nous nous intéressons plus particulièrement à la recherche d'information à partir d'un besoin utilisateur. Nous prenons pour hypothèse que ce besoin peut être satisfait par la génération d'un résumé automatique parlé.*

- la parole numérisée est une séquence de nombres ; comment extraire une signification à partir de ces données ?
- quel est l'impact d'erreurs lors de cette extraction ?
- l'information rapportée dans un résumé doit être la plus importante aux yeux de l'utilisateur ; comment inférer cette importance à partir d'un discours parlé ?
- quel est l'impact d'une mauvaise décision dans l'inférence de cette importance ?
- comment s'exprime le besoin de l'utilisateur et comment le satisfaire ?
- le besoin est-il uniquement explicite, comment faire apparaître la part implicite du besoin (inexprimée, ou inexprimable) ?
- quelles sont les formes possibles de résumé de parole, dans quels contexte sont-elles optimales ?

Le travail présenté dans ce document ne se targue pas de répondre définitivement à ces questions, bien qu'il esquisse une réponse à certaines d'entre elles, mais plutôt de donner les moyens d'étudier les problématiques sur un système complet, en plus des parties séparées. En effet, le résumé automatique de parole fait appel à de nombreux sous-domaines fortement étudiés pour lesquels il est tentant de se restreindre à des conditions expérimentales éloignées du problème visé. Une approche de bout-en-bout sur une application réelle peut valider la faisabilité des concepts et méthodes développés.

## 1.3 Schéma général

Les travaux présentés dans ce document sont délimités par un cadre précis. L'objectif général est de réduire le temps d'écoute lors d'une recherche d'informations parlées. La méthode proposée pour atteindre cet objectif est de résumer les résultats d'un moteur de recherche sur des données audio. Cette approche, similaire à ce qui est fait pour

les documents textuels (Jenhani, 2006), trouve tout son intérêt dans le cadre d'un média difficile à explorer comme l'audio. Notre travail se concentre sur le résumé automatique selon un besoin utilisateur comme une extension de la recherche d'information classique, adaptée aux problématiques de la parole dans un cadre interactif. Nous choisissons de traiter des journaux radio-diffusés car ce type de données est relativement éprouvé dans les domaines de la parole et de la recherche d'information. Adapter les techniques qui vont être présentées à d'autres types de données audio n'est généralement qu'une question de ressources et de paramétrage. Comme la parole est relativement variable, nous suivons les approches classiques en nous concentrant sur le contenu linguistique, source de signification la plus exploitée dans des données parlées. Cette restriction est due à la quantité de données potentiellement traitées et facilite le traitement de l'expression en langue naturelle du besoin de l'utilisateur. Elle représente une première étape vers du résumé intégrant tous les paramètres structurels, discursifs et acoustiques.

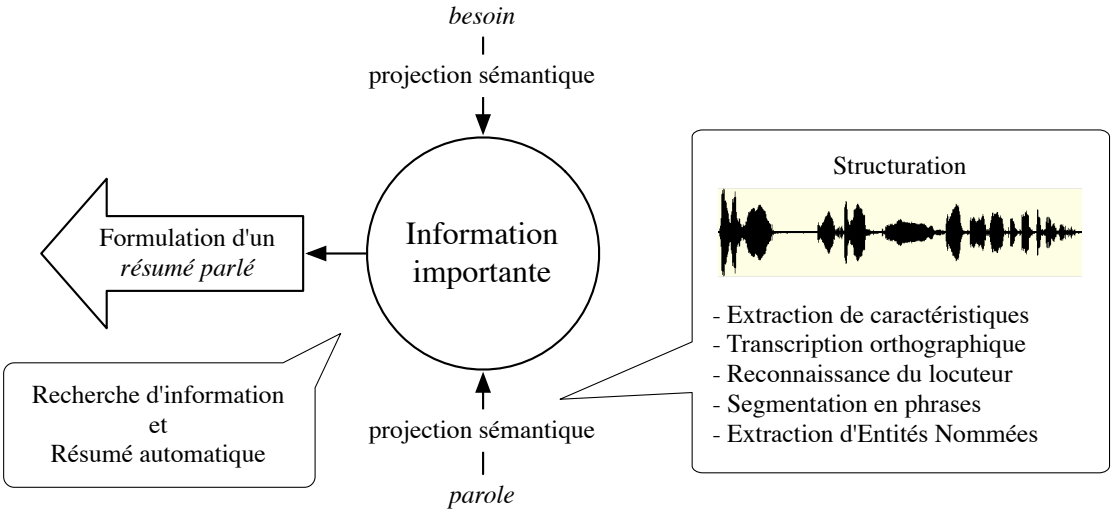



**FIG. 1.2:** *Schéma général pour une réduction de l'information parlée à l'aide du résumé automatique. Le besoin utilisateur et l'information parlée sont projetés dans un espace sémantique, puis des méthodes de recherche documentaire et de résumé automatique permettent la génération d'un résumé parlé. La projection de la parole dans un espace sémantique nécessite une structuration préalable.*

La figure 1.2 illustre le schéma général dirigeant la résolution du problème. Le besoin de l'utilisateur est comparé à des informations extraites de données parlées dans un espace sémantique. Les informations les plus importantes sont conservées et formulées en un résumé parlé. La projection sémantique du contenu parlé est rendue possible par une structuration du signal acoustique. La méthode de résumé suit les approches textuelles par extraction en procédant par la sélection puis la lecture des extraits porteurs de l'information importante.

## 1.4 Organisation du document

Ce travail est organisé en deux grandes parties, consacrées à la recherche d'information parlée et au résumé automatique de parole. Dans la première partie, le chapitre 2 décrit un état de l'art des méthodes de recherche d'information et de résumé automatique. L'étude inclut les contraintes spécifiques à l'utilisation d'un média parlé dans ces domaines. Le chapitre 3 présente la chaîne de structuration existante. Cette chaîne permet d'extraire des descripteurs structurels et sémantiques : la segmentation en macro-classes acoustiques, l'indexation en locuteurs et la transcription orthographique. Par la suite, la chaîne de structuration est complétée par deux éléments indispensables au résumé automatique de parole. La segmentation en phrases et l'extraction d'entités nommées sont traitées dans le chapitre 4. La seconde partie est focalisée sur le résumé automatique de parole. Le chapitre 5 étudie l'ajout de contraintes liées à la parole et à l'interactivité dans un modèle classique de résumé par extraction (*Maximal Marginal Relevance*). Dans le chapitre 6, le modèle proposé est évalué de façon indirecte lors de *Document Understanding Conference* 2006, une campagne d'évaluation du résumé textuel reconnue. Le chapitre traite aussi de l'impact des erreurs commises par la structuration par simulation sur les données textuelles. Ce chapitre est suivi de quelques pistes pour pallier les problèmes majeurs de structure et de cohérence inhérents au résumé par extraction (chapitre 7). Ces pistes incluent une frise chronologique interactive dont le rôle est de montrer à l'utilisateur la structure temporelle des résultats. Le document se termine par une conclusion reprenant les différentes contributions et décrivant les perspectives de ce travail.

# Première partie

# Recherche d'information et Structuration de la parole

# Chapitre 2

# Recherche d'information parlée

**Sommaire**



Ce chapitre traite la recherche d'information à travers les thèmes complémentaires de la recherche documentaire (section 2.1) et du résumé automatique (section 2.2). Les approches présentées sont étudiées dans le cadre d'un média textuel, puis selon un média audio, en mettant l'accent sur les adaptations induites par la parole.

## 2.1 Recherche documentaire

La notion de recherche d'information (*information retrieval*), introduite pour la première fois par Mooers (1950), a tout d'abord été l'apanage de documentalistes ayant besoin d'un classement efficace de leurs ouvrages. Leur but était d'étendre la notion d'index, présente dans les livres, à une bibliothèque entière. Le concept d'index a été inventé, dès 1230, lorsque Hugo de St. Cher employa 500 moines pour créer une *concordance* de la Bible (Wheatley, 1879). La recherche d'information est donc née de l'exploitation du contenu d'un document pour le retrouver ; cette tâche est connue sous

le nom de recherche documentaire. Par opposition, la classification décimale de Dewey (1876) permet de retrouver des documents grâce à des méta-informations externes à l'ouvrage, selon une annotation réalisée par le documentaliste. La spécificité de la recherche documentaire est de ne réaliser qu'une partie du travail en ne présentant comme résultat non pas l'information en elle-même, mais son interprétation au sein d'un document. Le besoin de l'utilisateur est exprimé sous la forme « J'aimerais tous les documents qui parlent de ... ; je les lirai tous afin de me forger une idée exhaustive de ce sujet ». L'utilisateur est assimilé à un documentaliste recherchant non pas une information précise, mais demandant à acquérir des connaissances sur un thème donné. Cette vision du problème de la recherche d'information a l'avantage de ne nécessiter qu'une formalisation précaire des thèmes abordés dans un ouvrage : un index à base de mots devrait suffire. Pour cela, les premiers modèles de recherche d'information ont suivi un schéma simple : transformer le besoin de l'utilisateur en une série de mots-clés, puis générer la liste des documents dont l'index contient ces mots-clés. Cette approche est fonctionnelle lorsque le nombre de documents retrouvés est limité et lorsque les mots-clés choisis ne mènent pas à un trop grand nombre de documents hors-sujet (pour des raisons de polysémie). En effet, ces deux cas augmentent le temps que l'utilisateur passe à explorer les documents sans forcément obtenir une réponse à son besoin.

Pour remédier à perte de temps, il faut abandonner la problématique documentaliste et faire une étude plus approfondie du contenu des documents. Tout d'abord, les documents hors-sujet peuvent être écartés en générant non pas un ensemble de documents, mais une liste ordonnée par pertinence estimée en fonction du besoin de l'utilisateur (en comptant par exemple le nombre d'occurrences dans un document des mots-clés utilisés pour retrouver les documents). Les documents au début de cette liste sont censés contenir des informations plus intéressantes pour l'utilisateur et devraient être explorés en premier lieu. Dans un second temps, le contenu de chaque document peut être résumé en fonction du besoin utilisateur pour lui permettre de juger rapidement du potentiel informatif de ce document (en présentant par exemple le contexte d'utilisation des mots-clés déduits du besoin utilisateur). La dernière idée est de s'affranchir du document et de répondre directement au besoin de l'utilisateur, en donnant une réponse exacte à la question qu'il se pose (problématique Questions-Réponses décrite et évaluée par Voorhees, 2003). Cette notion se rapproche beaucoup plus du sens premier de la recherche d'information, mais ce domaine très intéressant demande une analyse approfondie des questions et de leurs réponses potentielles. Toutefois, elle n'est traitée relativement efficacement que pour des questions fermées ou factuelles dont la réponse est une ou plusieurs entités ou quantités (Qui ont été les présidents des États-Unis ? Combien d'habitants la France compte-elle ? ...). Les questions non factuelles du type *pourquoi* et *comment* demandent des développements construits, approchés actuellement par le résumé de documents multiples guidé par un besoin utilisateur (voir section 2.2 sur ce sujet). Il faut tout de même noter que toutes les approches pour la recherche d'information sont construites autour d'une base de connaissances (corpus, bibliothèque, base de données) constituant la Vérité et contraignant toute réponse. Bien que le raisonnement par inférence (Raina et al., 2005) puisse donner des réponses à des questions non traitées dans le socle de connaissances exploité, certaines questions

métaphysiques n'auront certainement jamais de réponse fondée de la part d'un système informatique (Il paraîtrait qu'un ordinateur ne peut répondre que 42 à la question « Quel est le sens de la vie ? », Adams, 1979).

Les problématiques de la recherche d'information sont avant tout de représenter les informations et de déduire celles qui correspondent au besoin de l'utilisateur. Mais il ne faut pas oublier que l'information est conservée sur un support dont elle doit être extraite. De plus, l'expression du besoin de l'utilisateur se fait généralement en langue naturelle. Cependant, ce besoin peut prendre d'autres formes et se retrouver étroitement lié aux résultats de la recherche d'information. Dans ce cas, l'évolution du besoin doit être analysée au travers de son reflet dans les interactions entre l'utilisateur et le système. Ce type d'analyse est primordial pour mieux estimer le besoin de l'utilisateur. Un autre problème lié à la recherche d'information réside dans la quantité de données traitées, car cette dernière impose des contraintes sur l'ensemble des problématiques précédentes (Callan, 2000).

### 2.1.1 Définition de la tâche

La tâche la plus répandue en recherche d'information est la recherche documentaire (*Document Retrieval*). Dans ce cadre, les informations sont matérialisées sous forme de *documents* dans une ou plusieurs modalités. Un ensemble de *documents* est appelé *corpus* et la tâche consiste à extraire d'un corpus l'ensemble des documents correspondant au besoin de l'utilisateur, exprimé sous forme d'une *requête*. La tâche est définie de façon à rendre possible une répétition des résultats car un système doit se comporter de façon déterministe dans des conditions fixées à l'avance. Historiquement, les documents et les requêtes ont été d'abord textuels, puis différents médias ont été pris en compte (son, image, vidéo). Afin de trouver les documents répondant au besoin de l'utilisateur, la plupart des approches font une analyse du contenu des documents et de la requête. L'étude de ce contenu met en jeu l'extraction d'*unités informatives* (ou *descripteurs*), le support observable de l'information. Les *unités informatives* les plus évidentes sont les mots pour un contenu textuel, les histogrammes de couleurs pour une image et les fréquences pour un signal sonore. Cette notion d'*unité informative* est dérivée du processus de généralisation, ou conceptualisation, propre au système cognitif humain. Elle implique une hypothèse d'existence de motifs représentant une même idée, une même classe d'objets, un même concept sémantique. Smoliar et al. (1996) nomment *expressives* les approches fondées sur des *unités informatives* proches des données observées et *sémantiques* les approches réalisant une analyse poussée du contenu. Nous nous intéressons dans cette partie uniquement aux unités informatives issues d'une analyse du contenu linguistique, dans l'optique d'analyser la parole extraite de documents audio.

La recherche d'information textuelle repose sur la capacité à représenter le fond (niveau sémantique) de façon indépendante de la forme (niveau syntaxique), puis d'effectuer des opérations de comparaison dans l'espace de représentation ainsi formé. Cette opération est nécessaire car la langue offre de nombreuses façons d'exprimer une idée et montre une forte variabilité de forme. Il n'existe pas de bijection entre les mots et les sens associés, un mot pouvant avoir plusieurs sens (polysémie) et plusieurs mots

pouvant avoir le même sens (synonymie). En fait, de nombreuses relations lient les concepts dénotés par les mots, comme la relation de généralisation (hyperonymie), ou de spécialisation (hyponymie). De plus, des mots peuvent agir comme représentants d'autres mots, afin d'alléger le discours. Les pronoms sont un bon exemple d'utilisation d'une forme plus courte pour faire référence à un objet que seul le contexte peut définir. Cette utilisation de plusieurs formes pour représenter un même objet ou une même idée s'appelle une anaphore ou cataphore grammaticale (à ne pas confondre avec l'anaphore rhétorique) et le phénomène est connu sous le nom de coréférence. À un plus haut niveau, les nombreuses figures de style, comme la métaphore ou l'euphémisme, altèrent le sens en offrant plusieurs niveaux d'interprétation dépendant du contexte et de la culture. Les nombreux modèles de recherche d'information essaient tous de traiter ces phénomènes de façon plus ou moins implicite, en prenant pour hypothèse qu'un champ lexical donné caractérise suffisamment bien le contenu sémantique associé. Toutefois, de plus en plus d'approches associent à ces modèles des pré-traitements linguistiques pour détecter ces phénomènes de variabilité de la forme et retrouver le fond sous-jacent.

Cette section commence par une description de la tâche de recherche documentaire et de son évaluation. Puis, les pré-traitements linguistiques les plus courants sont abordés. Ensuite, les principaux modèles pour estimer la pertinence d'un document à une requête sont présentés. L'expansion de requête vient compléter ces modèles. Enfin, l'impact de l'ensemble des méthodes précédentes sur un média parlé et un aperçu des interactions avec l'utilisateur dans ce cadre sont étudiés.

### 2.1.2 Évaluation

Plusieurs campagnes d'évaluation sont organisées chaque année afin de suivre les avancées dans le domaine de la recherche d'information. Ces campagnes fournissent un protocole et des données d'évaluation aux participants et réalisent un jugement de leurs performances objectif et indépendant. Les plus importantes sont *Text REtrieval Conference* (TREC[1], Voorhees et Harman, 1999), *Cross-Language Evaluation Forum* (CLEF[2], Braschler et Peters, 2004), *NII Test Collection for IR Systems* (NTCIR[3], Kando, 2005). Ces campagnes évaluent la qualité des systèmes de recherche d'information sous diverses conditions (tâche, média, langue, quantité...), selon une souche commune. Pour un besoin utilisateur donné (requête), un système doit générer une liste de réponses (documents) ordonnées par pertinence estimée. Les documents ayant le meilleur score sont considérés comme les plus susceptibles de répondre au besoin utilisateur. Dans le cadre de la recherche documentaire, les références sont constituées d'une annotation binaire[4] (pertinent / non-pertinent) de chaque document du corpus pour chaque requête évaluée. Les mesures d'évaluation utilisent la répartition entre documents pertinents et non-pertinents à un rang donné de la liste de résultats (figure 2.1).

[1] http://trec.nist.gov/, visité en octobre 2006.
[2] http://clef.iei.pi.cnr.it/, visité en octobre 2006.
[3] http://research.nii.ac.jp/ntcir/index-en.html, visité en octobre 2006.
[4] Certaines évaluations ajoutent une troisième classe pour les documents partiellement pertinents.

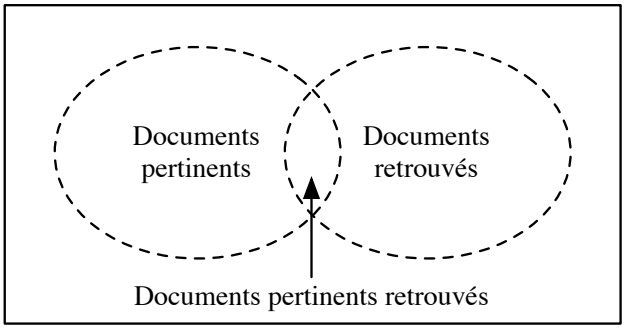



**FIG. 2.1:** *Illustration des différents ensembles de documents utilisés pour l'évaluation de la recherche d'information. Les mesures d'évaluations comme le rappel et la précision sont basées sur ces ensembles.*

$$R = \frac{\text{nb}(\{retrouvés\} \cap \{pertinents\})}{\text{nb}(\{pertinents\})} \tag{2.1}$$

$$P = \frac{\text{nb}(\{retrouvés\} \cap \{pertinents\})}{\text{nb}(\{retrouvés\})} \tag{2.2}$$

$$F_\beta = \frac{(1+\beta^2)PR}{\beta^2 P + R} \tag{2.3}$$

La précision ($P$, équation 2.2) est le nombre de documents pertinents retrouvés par rapport au nombre de documents retrouvés, alors que le rappel ($R$, équation 2.1) est le nombre de documents pertinents retrouvés par rapport au nombre de documents pertinents du corpus. La précision est tracée en fonction du rappel afin de comparer ces deux mesures au sein d'une courbe (courbe de précision-rappel dont un exemple est donné par la figure 2.2). Un score général d'un système sera donné par la précision moyenne sur onze points de rappel $(0.0, 0.1..., 1.0)$, sur l'ensemble des requêtes (*Mean Average Precision*, MAP). Une autre façon de combiner précision et rappel est d'utiliser la $F_1$-mesure, moyenne harmonique de $P$ et $R$ (équation 2.3, avec $\beta = 1$). Ces mesures d'évaluation sont focalisées sur une tâche répondant à des documentalistes car elles prennent en compte un parcours exhaustif des documents. Afin de rendre compte d'une utilisation plus actuelle des moteurs de recherche, il faut employer une mesure intégrant le fait que les utilisateurs ne parcourent que les deux ou trois premières pages des résultats avant d'estimer que l'information recherchée ne sera pas retrouvée. La précision à $n$ résultats (avec $n < 20$) est une bonne estimation de la qualité de la réponse au besoin de l'utilisateur, en ignorant toutefois la difficulté de la requête (Carmel et al., 2005).

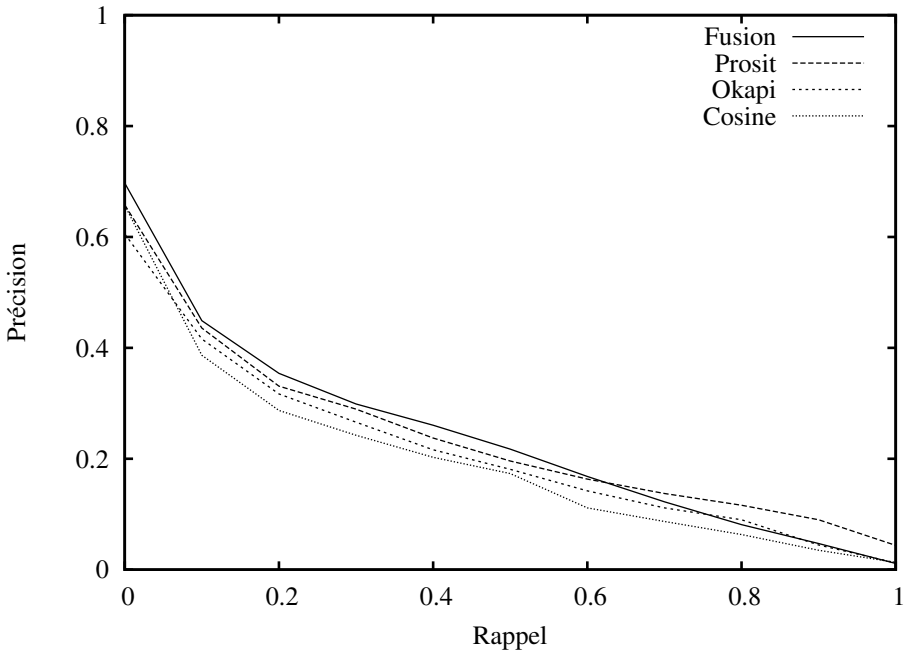



**FIG. 2.2:** *Exemple de comparaison de systèmes en recherche documentaire à l'aide de courbes de précision-rappel sur la campagne TREC-8. Les systèmes présentés sont le modèle vectoriel classique (Cosine), un modèle fonction de la quantité d'information (Prosit), un modèle probabiliste (Okapi), et la moyenne pour chaque document des scores normalisés des trois modèles précédents (Fusion). Ces différents systèmes suivent les formulations proposées par Savoy et Berger (2005).*

### 2.1.3 Pré-traitements linguistiques

Des pré-traitements linguistiques sont appliqués avant la transposition des documents dans un espace sémantique dans le but de minimiser l'impact des ambiguïtés. Lorsque le média traité est le texte, une première étape consiste à normaliser la forme de surface (correction orthographique, réaccentuation, normalisation de la casse, expansion des abréviations, normalisation des valeurs numériques, suppression du contenu non informatif...). Ce premier traitement correspond à un nettoyage des données, généralement implémenté à l'aide de règles de réécriture et de lexiques. Après ce traitement, les mots outils (déterminants, conjonctions, certains adverbes...) sont identifiés et supprimés car ils ne sont pas discriminants pour représenter le thème d'un document.

Les différents mots-formes d'un texte peuvent être regroupés sur la base de leur lemme, afin d'éliminer une partie de la variabilité morphologique et de créer des représentants plus « robustes » des concepts. Une première technique, appelée lemmatisation, remplace les mots par leur forme canonique (indépendante du genre, du nombre, de la personne et du mode). En général, elle est appliquée après étiquetage morpho-

syntaxique à l'aide d'un dictionnaire de triplets (*mot, étiquette, lemme*). Une autre technique cherche à réduire les mots à leur racine de façon algorithmique : la *racinisation* (Porter, 1980) est une approche heuristique dépendante de la langue et contenant un risque lié à de nombreux cas particuliers[5] créant des erreurs de confusion.

Toutefois, l'approche précédente ne remplace pas une désambiguïsation des mots. Par désambiguïsation, il est suggéré que le sens d'un mot peut être découvert parmi ses différents sens possibles, en utilisant le contexte autour de ce mot. Les différentes techniques pour la désambiguïsation[6] sémantique automatique n'ont pas de très bonnes performances dans le cas général (Agirre et Edmonds, 2006). Connaître le sens dans lequel les mots ont été employés permet de les projeter dans une ressource sémantique décrivant leurs relations (synonymie, antonymie, hyperonymie, hyponymie...), comme Wordnet (Miller, 1995).

Afin d'aller plus loin, des entités informatives liées au domaine peuvent être annotées (détection des entités nommées, sujet traité dans la section 4.2), les expressions anaphoriques pronominales peuvent être résolues (Lappin et Leass, 1994) et les figures de style peuvent être reconnues (Markert et Nissim, 2002). Ces domaines florissants tentent de résoudre des problèmes difficiles mais il est évident qu'ils auront de plus en plus d'applications en recherche d'information.

Les approches pour annoter ces phénomènes reposent souvent sur une analyse statistique des propriétés des objets en jeu, et des algorithmes d'apprentissage capables de reconnaître des classes d'objets en fonction de leurs caractéristiques. Les gains que peuvent apporter ces approches sur une tâche donnée sont très dépendants du domaine traité et des corpus utilisés. Ces gains ne sont pas toujours cumulatifs en raison de l'introduction d'erreurs ou d'ambiguïtés dans les différents pré-traitements. Dans l'optique de traiter des documents audio, il serait souhaitable d'appliquer ces techniques afin d'améliorer la perception du message parlé. Les traitements de plus haut niveau, ayant déjà de faibles performances sur le texte, risquent de se dégrader à cause de la variabilité de la parole.

### 2.1.4 Modèles

La plupart des modèles de recherche d'information sont soit définis pour une tâche de recherche documentaire (comparaison d'une requête à un document), soit pour une tâche de catégorisation (comparaison d'un document avec un autre document, ou avec un pseudo-document). Bien que ces deux types de tâches impliquent des hypothèses différentes lors de la modélisation, elles sont souvent considérées comme équivalentes dans la pratique, en considérant un document comme étant la requête, généralement au détriment de légers changements dans le comportement des modèles. De plus, les modèles se réfèrent aux *unités informatives* observées sous différents noms (mots, termes,

[5] Par exemple, le *stemmer* Snowball (`http://snowball.tartarus.org`, visité en décembre 2006) pour le français fait correspondre « guérir » et « guerre » à travers le radical « guer ».

[6] Plus d'informations peuvent être obtenues sur `http://www.senseval.org`, campagne d'évaluation des systèmes de désambiguïsation sémantique, visité en octobre 2006.

concepts...) selon les pré-traitements appliqués. Nous présentons dans cette section les modèles les plus répandus, dont une taxonomie est donnée par la table 2.1. Des ouvrages comme (Baeza-Yates et Ribeiro-Neto, 1999) ou (Ihadjadene, 2004) détaillent les formulations des modèles présentés dans cette section.

| **Type** / **Cadre** | **Sans interdépendance** | **Interdépendance** | |
|---|---|---|---|
| | | *Intrinsèque* | *Extrinsèque* |
| *Ensembliste* | Booléen<br>Booléen étendu | | Ensembles flous |
| *Algébrique* | VSM | GVSM<br>Réseaux de neurones<br>LSA<br>*Random Indexing* | TVSM<br>Infomap |
| *Probabiliste* | BIR<br>Réseaux bayesiens | Modèles de langage<br>Modèles de pertinence<br>pLSA, LDA | RbI |

**TAB. 2.1:** *Taxonomie des principaux modèles en recherche d'information inspirée par Kuropka (2004), en fonction de leurs propriétés et de leur origine mathématique. Ces modèles reposent sur des hypothèses fortes liées au degré de dépendance entre les unités de surfaces observées (mots, termes ou concepts), et la façon de modéliser cette dépendance (intrinsèque ou extrinsèque). Les acronymes sont détaillés tout au long de la section.*

### Modèles ensemblistes

Le modèle booléen s'appuie sur la logique de Boole en permettant à l'utilisateur d'exprimer son besoin par une formule logique sur les mots. Les documents satisfaisant cette formule sont considérés comme pertinents et renvoyés à l'utilisateur. Lorsque cette satisfaction est binaire, il n'y a pas de notion de degré de pertinence, donc le modèle n'est utile que dans les cas où le nombre de résultats retrouvés correspond aux attentes de l'utilisateur. Afin d'y remédier, le modèle booléen a été étendu de nombreuses manières. Par exemple, Salton et al. (1983) ajoutent un degré de satisfaction de la requête logique fonction des propriétés statistiques des mots et une pondération des opérateurs logiques « ET » et « OU ». Ces approches s'apparentent autant au modèle booléen qu'au modèle vectoriel. Parallèlement aux modèles fondés sur la théorie des ensembles « stricts », il existe d'autres modèles basés sur la théorie des ensembles flous qui prennent en compte la corrélation entre les mots pour définir l'appartenance de chaque document aux ensembles des mots qui les composent. Les performances de cette classe de modèles ont rarement été comparées à celles des modèles les plus répandus.

### Modèles algébriques

Les modèles algébriques utilisent une projection des documents et des requêtes dans un espace vectoriel dans lequel ces vecteurs peuvent être comparés. La pertinence

estimée d'un document est directement proportionnelle à une mesure de sa similarité à la requête. Le modèle vectoriel (*Vector Space Model*, VSM, Salton et al., 1975) est le modèle le plus populaire en recherche d'information car il permet d'obtenir des performances intéressantes en nécessitant peu de ressources. L'idée de cette modélisation est d'exprimer chaque documents ($\vec{d}$) et la requête ($\vec{q}$) comme des vecteurs dans l'espace formé par le vocabulaire (équations 2.4 et 2.5). Chaque dimension de cet espace représente un mot du vocabulaire ($u_i$) ; la composante du vecteur document (ou requête) pour ce mot, $w_{i,d}$ est donnée dans l'équation 2.6, en fonction de la fréquence du mot dans le document, $f_i(d)$, par rapport à sa fréquence dans le corpus ($N$ est le nombre de documents, et $n_i$ le nombre de documents où apparaît le mot $u_i$). Cette normalisation par rapport à la fréquence dans le corpus est appelée *Inverse Document Frequency* (IDF) et permet de spécifier que les événements peu fréquents sont plus susceptibles d'intéresser l'utilisateur. Cette notion est présente dans la recherche d'information depuis ses débuts (Bar-Hillel, 1958), mais a fait l'objet de nombreuses formulations. A partir de la représentation vectorielle d'un document, sa pertinence à la requête est estimée en calculant sa similarité avec le vecteur requête. La similarité la plus répandue est la similarité *cosine*, cosinus de l'angle entre les deux vecteurs $\vec{q}$ et $\vec{d}$, comme le détaille l'équation 2.7. Savoy et Berger (2005) comparent les formulations de pondération pour $tf$ et $idf$ et les différentes similarités sur la campagne CLEF 2005. Certaines par exemple prennent en compte le biais de la longueur des documents et de la fréquence de mots intra-document pour permettre des améliorations significatives au sens de l'évaluation.

$$\vec{d} = (w_{0,d}..., w_{|W|,d})^T \tag{2.4}$$

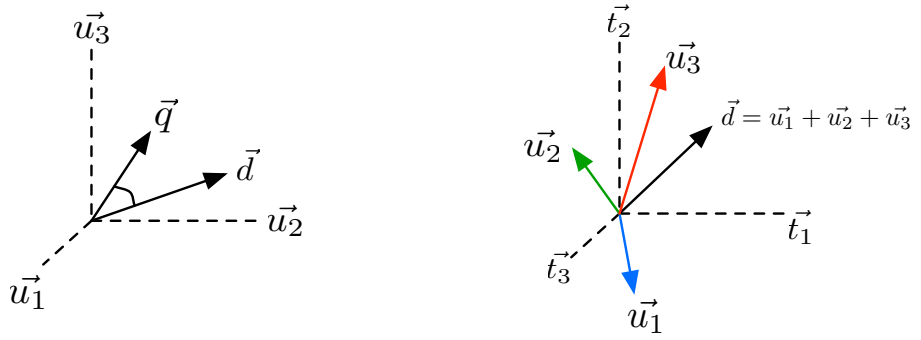
$$\vec{q} = (w_{0,q}..., w_{|W|,q})^T \tag{2.5}$$

$$w_{i,d} = tf_{i,d} \times idf_i = log(1 + f_i(d)) \times \log \frac{N}{n_i} \tag{2.6}$$

$$cosine(\vec{q}, \vec{d}) = \frac{\vec{d} \cdot \vec{q}}{|\vec{d}| \times |\vec{q}|} = \frac{\sum_{i=0}^{|W|} w_{i,d} w_{i,q}}{\sqrt{\sum_{i=0}^{|W|} w_{i,d}^2} \sqrt{\sum_{i=0}^{|W|} w_{i,q}^2}} \tag{2.7}$$

Malgré la faible complexité du calcul de la similarité entre deux documents qui rend celui-ci adapté aux grands corpus, il est souvent reproché à ce modèle de ne pas prendre en compte l'ordre des mots (modèles à sac-de-mots), ni même la relation entre les mots (« maison » et « blanche » sont des mots très répandus, donc faiblement pondérés alors que « maison blanche » devrait avoir un impact beaucoup plus fort sur la similarité).

Wong et al. (1985) ont proposé le modèle vectoriel généralisé (*Generalized Vector Space Model*, GVSM) pour prendre en compte les corrélations inter-mots. Le modèle vectoriel impose une base orthonormale de l'espace des documents. Cette base implique que chaque vecteur représentant un mot est orthogonal à tous les autres vecteurs représentant des mots. Dans le modèle vectoriel généralisé, le vecteur représentant un mot est défini selon sa corrélation avec les autres mots du lexique. Ainsi, comme l'illustre la figure 2.3, un vecteur document (somme des vecteurs représentant les mots qu'il contient) prendra en compte les affinités des mots à apparaître ensemble. La complexité de ce modèle est beaucoup plus grande que celle du modèle classique, pour un gain de




modèle vectoriel classique (VSM) modèle vectoriel généralisé (GVSM)

**FIG. 2.3:** *Illustration de la différence entre le modèle vectoriel classique et le modèle vectoriel généralisé et autres modèles impliquant des vecteurs mots non orthogonaux. $\vec{q}$, $\vec{d}$, $\vec{u_i}$ et $\vec{t_j}$ représentent respectivement la requête, un document, un mot (unité informative) et une dimension sous-jacente aux mots (thème).*

performance pas toujours convaincant (augmentation du rappel, diminution de la précision) .

*Latent Semantic Analysis* (LSA), également référencé sous le nom de *Latent Semantic Indexing* (LSI), peut être vu comme une extension de GVSM offrant une réduction de la taille de l'espace de comparaison et donc de la complexité (Deerwester et al., 1990). La méthode repose sur une réduction de la matrice documents-mots $X$ à ses dimensions principales en utilisant une décomposition en valeurs singulières (SVD). La matrice est décomposée en une multiplication de la matrice de vecteurs singuliers gauche $U$, de la matrice diagonale de valeurs singulières $\Sigma$ et de la matrice de vecteurs singuliers droite $V^T$ (équation 2.8). Lorsque les valeurs singulières sont ordonnées de la plus grande à la plus petite, réduire le rang $k$ de la matrice $\Sigma$ correspond à approximer la matrice $X$ en minimisant l'erreur au sens de la norme $L^2$ entre les mots (équation 2.9, dans laquelle $\hat{X}$ est l'approximation de $X$ et $\Sigma_k$ la matrice de valeurs propres réduite au rang $k$). De plus, les dimensions de l'espace réduit font apparaître des « thématiques » selon lesquelles sont exprimés les documents. Une projection de la requête dans cet espace permet de calculer une similarité prenant en compte les caractéristiques thématiques de la requête. Ce modèle est plus largement décrit dans la section 5.3 où il est appliqué au résumé automatique orienté par une requête.

$$X = U\Sigma V^T \tag{2.8}$$

$$\hat{X} = U\Sigma_k V^T \tag{2.9}$$

LSA offre une réduction de dimension de bonne qualité, mais nécessite d'évaluer la matrice d'occurrences entre mots et documents et de décomposer cette matrice. Afin d'éviter cette réduction coûteuse en ressources, Kanerva et al. (2000) proposent une technique nommée *Random Indexing*. Cette approche consiste à associer aux mots des vecteurs aléatoires quasi-orthogonaux (contenant un grand nombre de 0 et un petit nombre de $-1$ et $+1$), de dimension fixe et de construire un équivalent de la matrice

de coocurrences grâce à des accumulateurs. La réduction est d'aussi bonne qualité que pour LSA, avec l'avantage de supporter le passage à l'échelle et de pouvoir être mise à jour de façon incrémentale. De nombreux autres modèles algébriques sont décrits dans la littérature tels que la recherche documentaire à base de réseaux de neurones (Wilkinson et Hingston, 1991), utilisant les mots de la requête en entrée et générant les documents en sortie ; l'emploi de LSA pour créer les vecteurs de mots du modèle vectoriel généralisé (Infomap, voir section 5.3) ; et les modèles vectoriels thématiques (TVSM, Becker et Kuropka, 2003) associés à une classification en thèmes des mots et/ou des documents.

### Modèles probabilistes

Le cadre probabiliste est très attirant pour la recherche documentaire et a été appliqué de nombreuses façons. Trois événements entrent en jeu dans ce cadre : la requête $Q$, un document $D$ et la pertinence binaire $R$. Les modèles proposés diffèrent par leur estimation de la probabilité qu'un document soit pertinent pour une requête donnée $P(R = 1|D, Q)$. La figure 2.4 illustre les principales approches, à savoir : le modèle classique, les modèles de langage et les modèles de pertinence. Le modèle classique de Robertson et Spärck-Jones (1988) est fondé sur le ratio de vraisemblance entre $P(R = 1|D, Q)$ et $P(R = 0|D, Q)$, les modèles de langage de Ponte et Croft (1998) estiment la probabilité que la requête soit issue de la même distribution qu'un document $P(Q|D)$ et les modèles de pertinence (Lavrenko, 2002) utilisent la divergence entre les distributions $P(Q|R)$ et $P(D|R)$. Les probabilités incluant la pertinence sont difficiles à estimer dans le cas où cette variable n'est pas observée (schéma de recherche d'information classique). Par contre, lorsqu'un *a priori* sur la pertinence est fourni (par exemple grâce à des interactions utilisateur), ces modèles ont un avantage certain.

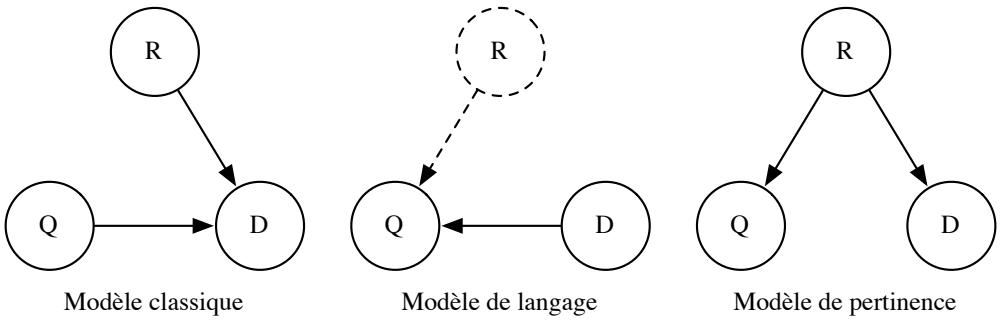



**FIG. 2.4:** *Différentes modélisations probabilistes pour la recherche documentaire selon le formalisme des modèles graphiques. Ces modèles estiment la pertinence d'un document pour la requête en fonction des événements Q (la requête), D (le document) et R (la pertinence). Dans cette figure, une absence de flèche correspond à une hypothèse d'indépendance entre deux événements. Les éléments en pointillés sont implicites au modèle et n'apparaissent pas dans sa formulation.*

Le modèle classique, nommé *Binary Independence Retrieval* (BIR), fait l'hypothèse que les mots sont indépendants deux à deux. Ses dernières extensions sont détaillées dans

(Spärck-Jones et al., 2000). La probabilité qu'un document soit pertinent est exprimée comme le produit des probabilités de pertinence de chaque mot qui le compose. Cette probabilité est modélisée par une distribution 2-poisson prenant en compte l'« élitisme » d'un mot pour un document (sur-pondération des mots très représentatifs du thème) en fonction de son nombre d'occurrences (Robertson et Walker, 1994).

Les modèles de langage (Ponte et Croft, 1998) tentent d'introduire une meilleure estimation de la distribution des mots dans les documents la modélisant par une loi multinomiale. Toutefois, Church et Gale (1995) prouvent que ce type de distribution ne reflète pas complètement la réalité. En général, le cadre du maximum de vraisemblance est utilisé pour estimer la probabilité pour un modèle de document de générer la requête. Les idées du modèle algébrique LSI/LSA ont été reprises par pLSI/pLSA, une approche probabiliste de l'analyse en sémantique latente (Hofmann, 2000). Cette approche définit un modèle génératif de la matrice de comptes selon lequel les mots proviennent de $k$ thèmes mis en relation par des variables latentes de mélange. pLSI a le désavantage de nécessiter l'estimation de beaucoup de paramètres et d'aboutir à du sur-apprentissage. Pour lever ces problèmes, Blei et al. (2003) proposent *Latent Dirichlet Allocation* (LDA), un modèle où les mots des documents sont toujours générés par une distribution multinomiale. Néanmoins, les thèmes latents sont considérés comme des variables régissant le comportement des paramètres de la multinomiale à travers une distribution de Dirichlet. Ce modèle implique une estimation coûteuse de ses paramètres (généralement approximée) sans toutefois apporter les gains en performance escomptés sur une tâche de recherche documentaire (Wei et Croft, 2006).

Il existe d'autres modèles probabilistes fondés sur une topologie de réseaux bayesiens, comme par exemple (Callan et al., 1992), à la base du système Inquery. Ce modèle cherche à estimer l'inférence $d \rightarrow u \rightarrow q$. Cette inférence est aussi estimée dans le cadre plus général des réseaux de croyances (Ribeiro-Neto et Muntz, 1996). Une autre méthode, *Retrieval by Logical Imaging* (RbI, Crestani et van Rijsbergen, 1995) apporte une vision intéressante en utilisant une logique modale pour étendre à un cadre probabiliste les idées du modèle booléen (*Retrieval by Logical Imaging*, RbI).

La prochaine section est dédiée à l'expansion de requête, un processus indispensable pour les modèles probabilistes.

### 2.1.5 Expansion de requête

Une requête est censée représenter le besoin de l'utilisateur. Or, la langue naturelle offre de nombreuses possibilités de varier la forme de surface tout en conservant un contenu sémantique proche. L'expansion de requête consiste à améliorer la représentation du besoin utilisateur grâce aux interactions issues d'une première recherche d'information. Par exemple, l'utilisateur peut explicitement désigner parmi les résultats des documents pertinents et des documents non-pertinents, ce qui peut être utilisé pour favoriser (resp. défavoriser) les mots apparaissant dans ces documents. Pour prendre un autre exemple, un moteur de recherche offrant un aperçu des documents retrouvés pourra tirer avantage de l'ordre dans lequel l'utilisateur visualise les documents. Les

modèles de recherche documentaire doivent intégrer ces informations pour améliorer les prochaines recherches, au travers d'un processus d'expansion de requête.

Dans le cadre des modèles vectoriels, les mots des documents désignés comme pertinents sont ajoutés avec un facteur positif pour la requête, alors que les mots des documents désignés non-pertinents sont ajoutés avec un facteur négatif (Buckley et al., 1994). Ce processus permet d'améliorer efficacement les performances de la recherche. Les modèles probabilistes qui intègrent la notion de pertinence ont alors une observation directe des probabilités $P(R|D)$ et $P(\overline{R}|D)$, qui évite d'utiliser des approximations de mauvaise qualité pour ces probabilités.

Cette utilisation est limitée aux scénarios incluant des utilisateurs. Il est heureusement possible de réaliser une expansion en aveugle en estimant que les documents en haut du classement ont une plus grande probabilité d'être pertinents que ceux en bas du classement. Dans la pratique, les $n$ premiers documents sont considérés comme pertinents. Ce processus peut être itéré, mais il n'a jamais été prouvé qu'il convergeait vers les documents pertinents. Dans le meilleur des cas, un petit nombre d'itérations permet de fortement améliorer le rappel, au détriment de la précision (ce qui correspond à relâcher les contraintes désignant les documents recherchés).

D'autres types d'expansion ont été proposés, comme l'utilisation d'ontologies pour augmenter les mots de la requête avec leurs synonymes (Gonzalo et al., 1998), ou encore étendre le champ lexical d'un document à l'aide de ses voisins dans un corpus « propre » de grande taille (Singhal et Pereira, 1999) .

L'expansion de requête et de document fonctionne bien pour augmenter le rappel, mais la précision peut être améliorée en effectuant un partitionnement non supervisé des meilleurs documents en thèmes, et en supprimant les partitions de petite taille correspondant à des résultats aberrants (de Loupy et al., 1998). Une autre possibilité pour augmenter la précision consiste en la fusion de plusieurs listes de résultats générées selon différentes combinaisons de pré-traitements, de modèles et de paramètres (Bellot, 2000). La moyenne, le minimum ou le maximum des classements de chaque document sont des approches simples dans le cas où aucun *a priori* n'est connu sur la qualité des classements (Bellot et El-Bèze, 2000). Dans le cas contraire, il est possible d'apprendre les paramètres du reclassement en fonction de la source, de l'espace des scores ou à l'aide de propriétés des documents et des requêtes (Croft, 2000).

### 2.1.6 Extension à la parole

Les premières approches de la recherche d'information dans un contenu parlé ont d'abord utilisé des techniques similaires à celle développées pour les documents textuels, appliquées à la transcription automatique du flux de parole. La recherche documentaire audio (*Spoken Document Retrieval*, SDR) est la première formalisation de la tâche au travers de la campagne TREC 7. Cette tâche est associée à la recherche d'information dans des documents papier numérisés par *Optical Caracter Recognition* (OCR) car, dans les deux cas, les erreurs introduites peuvent être assimilées à un bruitage du contenu linguistique originel. La tâche SDR de TREC (Garofolo et al., 1999) consiste à

indexer 500 heures d'émissions radio en anglais, en utilisant les transcriptions automatiques (à différents taux d'erreur de mots, *Word Error Rate*, WER) des documents issus de la campagne Hub 4 (Przybocki et al., 1998) organisée par NIST. L'information recherchée dans les documents audio est exprimée sous la forme d'une requête textuelle semblable à celles exploitées dans TREC *ad-hoc*. Pour plus d'informations, le lecteur pourra se référer à l'atelier spécial sur l'adaptation des techniques de recherche documentaire aux applications impliquant de la parole (Coden et al., 2002), organisé lors de la conférence SIGIR 2001. Il faut remarquer que la plupart des systèmes de recherche documentaire fonctionnent soit sur du texte soit sur l'audio, mais ne mélangent pas les deux modalités. Sanderson et Shou (2002); Favre (2003) soulignent qu'en général ce mélange défavorise l'audio et qu'aucune technique permettant de réduire cet écart n'a été proposée à ce jour.

Les évaluations TREC montrent que le taux d'erreur de mots est linéairement corrélé aux performances en recherche documentaire et qu'un taux d'erreur inférieur à 40% permet d'obtenir des résultats acceptables par l'utilisateur (Garofolo et al., 1999). Cette bonne réussite s'explique d'abord par la longueur des requêtes TREC et la quantité d'informations qu'elles contiennent (environ 10 mots porteurs de sens, à comparer à des requêtes WEB de moins de 2 mots en moyenne). L'impact du taux d'erreur peut être limité à 10% des performances sur la transcription manuelle en utilisant des techniques d'expansion de requête et de document comme celles présentées dans la section 2.1.5. Johnson et al. (2000) notent que le gain des différentes techniques n'est pas cumulatif et que l'utilisation de corpus externes propres (et thématiquement proches des données traitées) pour l'expansion est bénéfique. Toutefois, Hansen et al. (2004) font face à des conditions moins favorables sur les données de la *National Gallery of the Spoken Word* (NGSW) avec des taux d'erreur de mots de 40% et observent qu'un bon choix des paramètres utilisés lors de l'expansion permet d'obtenir un cumul des gains (20% relatif au total).

Des données audio sont utilisées dans le cadre d'une autre tâche intéressante lors de la campagne *Topic Detection and Tracking* (TDT), pour laquelle il faut faire du suivi de thème et détecter les nouveautés dans le flux d'informations (Allan, 2002). Cette tâche a impliqué la mise au point de nouvelles méthodes de détection de coupures thématiques en utilisant aussi bien le contenu liguistique que le contenu audio. Les techniques de recherche d'information audio tentent maintenant d'aller plus loin que la parole des flux radio, en se focalisant sur la parole spontanée et sur les applications temps réel. Brown et al. (2001), par exemple, s'attachent à annoter des flux télévisuels et des conférences avec des informations susceptibles d'intéresser le spectateur. Pour ce qui est de la parole spontanée, Byrne et al. (2004) ont annoté un corpus de 10000 heures d'interview puisées dans les enregistrements de la *Shoah Visual History Foundation*. Ce corpus est à ce jour le plus grand corpus de parole spontanée réunissant de nombreux locuteurs sur le même thème ; ce corpus permettra certainement de mieux tester les approches de recherche d'information et de segmentation que les corpus téléphoniques de Switchboard (Godfrey et al., 1992).

Le taux d'erreur de mots n'est pas le seul problème lié à la transcription automatique du contenu parlé, les systèmes de transcription ont en effet un vocabulaire limité aux

mots les plus fréquents (dans le but de minimiser le taux d'erreur de mots, tout en limitant les ressources nécessaires). Les mots les moins fréquents sont considérés comme des mots hors vocabulaire (*Out of Vocabulary*, OOV) et ignorés lors du décodage du signal de parole. Ils ne pourront être retrouvés et paradoxalement, ce sont justement les événements peu fréquents et inattendus qui sont le plus susceptibles de sélectionner les documents pertinents. En effet, le moteur de recherche SpeechBot (Thong et al., 2000) a offert pendant plusieurs années l'accès à du contenu parlé transcrit automatiquement sur le web et il a été observé que plus de 12% des mots utilisés dans les requêtes étaient hors vocabulaire. Le problème est aussi lié aux modèles de langages nécessairement mal estimés pour les langues à ressources minoritaires comme les langues africaines (Abdillahi et al., 2006). Des techniques basées sur l'utilisation de sous-parties des mots comme les phonèmes ou les radicaux sont apparues pour essayer de remédier au problème des mots hors vocabulaire (Wechsler et al., 1998). Ces approches demandent une phonétisation de la requête, puis la comparaison de cette séquence de phonèmes avec les hypothèses de transcription phonétique du système de transcription. Une mesure de confiance basée sur l'adéquation entre la modélisation phonétique et le contenu acoustique est utilisée afin de ne rapporter que des séquences proches de la meilleure hypothèse (probabilité *a posteriori* du sous-graphe d'hypothèses passant par le chemin étudié). L'utilisation du treillis[7] de phonèmes apporte un gain intéressant en rappel au détriment de la précision car de nombreux passages ont une transcription phonétique similaire à la requête sans pour autant impliquer la présence des mêmes mots. Yu et Seide (2004) intègrent la recherche dans le treillis de phonèmes avec une recherche dans le treillis de mots afin de profiter de l'augmentation à la fois du rappel et de la précision. Face à un taux d'erreur de mots de l'ordre de 43% à 60% selon les conditions, ils observent un gain de 10% en performance sur la détection de mots (*word spotting*) par rapport à l'utilisation d'une des deux méthodes isolément. Les mots hors vocabulaire ont des effets de bord sur la qualité de la transcription, car ils sont remplacés par une séquence de mots acoustiquement proches, mais qui diverge du contenu réel et provoque des erreurs autour du mot inconnu. Bazzi et Glass (2000) proposent par exemple d'introduire un mot « INCONNU » dans le vocabulaire et d'utiliser un modèle phonétique complet pour représenter son acoustique. Cette approche par cas particulier n'entre pas dans les cadres mathématiques utilisés en transcription et demande un contrôle fin de son activation.

Nous retiendrons que la recherche d'information audio s'est surtout concentrée sur l'interaction transcription/recherche. Pour preuve, le standard MPEG 7 (Manjunath et al., 2002) adopte, entre autre, la représentation par treillis d'hypothèses phonétiques pour le stockage des transcriptions automatiques de flux structurés afin d'autoriser la remise en cause du lexique de reconnaissance.

### 2.1.7 Interaction avec l'utilisateur

Gilbert et Zhong (2003) observent que des interfaces efficaces pour la recherche d'information audio restent peu développées, car l'effort de recherche s'est concentré sur

[7] le mot « treillis » est utilisé dans le sens de graphe d'hypothèses, de l'anglais *lattice*.

des modèles correspondant à une application générique.

Une première interface a été proposée par Arons (1993) avec pour objectif de faciliter la navigation dans un signal sonore. Des fonctions d'avance rapide et retour rapide sont étudiées en réduisant d'abord les silences inter-mots, puis en supprimant les mots à faible intonation. Elles permettent de naviguer 5 fois plus rapidement dans le flux de parole, mais certainement pas d'aborder des bases de grande taille. Le principe de focalisation, grâce auquel l'homme est capable de suivre un locuteur au milieu de plusieurs conversations, a été exploité par Kobayashi et Schmandlt (1997) qui spatialisent le son pour diffuser plusieurs signaux en mouvement. Ainsi, l'axe temporel est projeté dans l'espace et l'utilisateur fait alors appel à sa mémoire spatiale (généralement plus développée que la mémoire temporelle). L'utilisateur peut à tout moment réécouter un segment intéressant tout en continuant la lecture du flux principal. Cette approche n'a connu que peu de succès compte tenu de l'infrastructure nécessaire pour une spatialisation efficace du son. Avec l'arrivée des méthodes de recherche d'information appliquées à la parole, des interfaces intégrant un moteur de recherche se sont développées (Hirschberg et al., 1999; Thong et al., 2000) afin de mieux cibler le besoin utilisateur. Malheureusement, il est difficile sur les grandes bases de données audio d'aller à l'essentiel et une requête simple peut engendrer une grande quantité de segments longs à écouter. Pourtant, l'utilisation de la transcription reste le moyen le plus efficace d'extraire une sémantique du signal audio. Cependant, Hirschberg et al. (2001) rapportent que les utilisateurs de SCANMail, un système de présentation de messages vocaux, ont tendance à trop faire confiance à la transcription automatique et à ne pas écouter le message réel. De plus, tous leurs efforts pour créer des présentations alternatives à la transcription (mise en valeur des segments importants sur une frise chronologique...) sont généralement inutilisés car beaucoup moins intuitifs. Dans le même domaine que la recherche documentaire textuelle, la problématique Questions-Réponses (Q/A) se transpose naturellement à l'audio sous la forme d'un dialogue homme-machine (Galibert et Rosset, 2005; Hori et al., 2003c; Varges et al., 2006; Stenchikova et al., 2006). L'avantage principal du dialogue est de pouvoir demander des précisions sur la question ou une reformulation en cas de transcription peu confiante. Par contre, il faut être capable de formuler la réponse la plus courte possible et d'utiliser les mêmes artifices que l'être humain pour faire patienter son auditeur. Le dialogue permet d'optimiser la granularité de la réponse en étant d'abord généraliste et en demandant à l'utilisateur s'il souhaite plus de détails. La problématique temps réel est impérative pour une bonne interaction et le système doit gagner du temps pour trouver sa réponse. Il peut le faire en tirant parti de la fonction phatique (ajout d'éléments lexicaux pour ne pas rompre le discours), ou par des formulations longues (choix de synonymes plus longs à prononcer). Les applications issues de Q/A sont pour l'instant surtout limitées à des questions dont la réponse directement accessible dans les documents sources, mais les approches devraient, comme ça a été le cas pour Q/A sur le texte, rejoindre le résumé automatique pour être étendues aux questions non factuelles.

## 2.2 Résumé automatique

Cette section présente les approches majeures[8] en résumé automatique de texte et leur extension à la parole. Le lecteur trouvera de nombreuses informations sur le sujet dans (Mani, 2001).

Historiquement, le résumé automatique a d'abord été appliqué au texte. En effet, la première approche a été proposée lorsque les premiers ordinateurs ont été capables de numériser des documents. Luhn (1958) extrait des mots-clés représentatifs du contenu d'un document à l'aide de statistiques et se sert de ces mots-clés pour sélectionner les phrases les plus importantes du document. L'idée du résumé par extraction est restée une des approches les plus répandues pendant les 50 années qui suivirent. Edmundson (1969) introduit l'idée d'utiliser les caractéristiques des phrases afin d'évaluer leur importance. Cette idée se traduit par une analyse des documents traités suggérant que les phrases importantes sont caractérisées par la présence d'indicateurs lexicaux (« Ce document décrit ... », « Pour conclure... »), par la présence de mots-clés, par le nombre de mots en commun avec le titre, et par la position de la phrase dans le document. Une combinaison de ces différents paramètres dans le choix des phrases extraites permet d'améliorer le nombre de mots en commun avec un résumé écrit à la main. Elle est aussi une des premières approches de l'évaluation de la qualité d'un résumé. Après quelques échecs de méthodes linguistiques observés par Paice (1990), Kupiec et al. (1995) formulent le problème de sélection de phrases candidates au résumé comme un problème de classification. Des paires (*document d'origine – résumé*) sont nécessaires pour apprendre les paramètres optimaux du classifieur. Parallèlement à ces méthodes statistiques, de nombreuses approches utilisent des connaissances du domaine traité pour extraire les informations importantes (McKeown et Radev, 1995). Par exemple, un résumé sur la vie d'une personne nécessite de détecter sa date de naissance, des informations sur sa famille, ses études, ses activités professionnelles...Une fois ces informations extraites, un système de génération les utilise pour synthétiser la séquence linguistique correspondant au résumé. La qualité des résumés produits est fortement dépendante de la qualité des patrons utilisés et reste très dépendante du domaine.

Le résumé automatique n'a pas été appliqué uniquement au texte. Dans le domaine de la vidéo, une bande annonce de film peut s'apparenter à un résumé dont le but est de donner envie au spectateur de voir le film. Dans ce domaine, Nam et Tewfik (1999) génèrent des résumés de séquences vidéo par extraction des zones de forte activité des modalités observées dans le temps, sans prendre en compte l'intérêt du spectateur. L'objectif d'un résumé ne doit pas être négligé. Il est impensable, par exemple, que le résumé au dos d'un livre dévoile l'intrigue. En fait, les résumés peuvent être classés selon différents critères :

1. **Le but**
   Le résumé *indicatif* permet de savoir s'il faut approfondir un sujet ; le résumé

[8] Le site http://summarization.com, visité en octobre 2006, est une bonne source d'informations sur le résumé automatique de texte. Maintenu par D. Radev, il propose par ailleurs une liste de 700 références bibliographiques dans le domaine.

*informatif* donne une information généraliste et objective ; le résumé *critique* donne le point de vue de son auteur.

2. **La forme**
   Un résumé peut prendre la forme d'une succession d'*extraits* représentatifs, ou consister en une formulation complète après *abstraction*. Un parallèle peut être fait avec la classification de Smoliar et al. (1996) entre les méthodes *expressives* et *sémantiques* pour la recherche d'information.
3. **La dimension**
   Le résumé *mono-document* sera plutôt *indicatif* alors que résumé *multi-document* essaie d'éviter un examen de tous les documents.
4. **Le contexte**
   Le résumé peut être *générique* ou prendre en compte le besoin de l'utilisateur au travers d'une *requête*.

Un résumé automatique peut prendre en compte chacun des paramètres précédents au travers du besoin de l'utilisateur, mais dans la plupart des approches, ils sont fixés en fonction de l'application. Cependant, le cas du résumé de parole offre plus de possibilités car le média de destination peut être textuel ou parlé, impliquant une synthèse vocale et/ou la participation des locuteurs d'origine.

### 2.2.1 Évaluation

Nous allons présenter l'évaluation du résumé automatique sous le regard de la campagne *Document Understanding Conference* (DUC) organisée par NIST[9]. Les tâches liées à cette évaluation sont décrites plus en détail dans le chapitre 6. La tâche principale est de résumer une vingtaine de documents en fonction d'un besoin utilisateur avec une longueur maximale de 250 mots. Comme pour la plupart des tâches simulant une part de compréhension, l'évaluation d'un résumé est difficile. En effet, deux personne ne produisent pas le même résumé sur des mêmes contraintes de données, de besoin, de temps : il est difficile de définir le résumé parfait que doit générer un système. En effet, un résumé de même qualité peut être produit en utilisant des mots différents mais de même sens ; ou au contraire il est possible d'utiliser les mêmes mots tout en changeant leur ordre, ou en introduisant des négations, afin de détourner le sens. Les figures 2.3, 2.4 et 2.5 présentent un *topic* DUC, deux résumés et un document associés à ce *topic*. La table 2.2 montre les propriétés de quelques méthodes d'évaluation du résumé automatique.

Il existe des évaluations de sous-objectifs, très dépendantes de l'approche implémentée. Ces méthodes sont généralement peu représentatives du résultat final, mais facilitent l'interprétation des performances du système. Le résumé par extraction, par exemple, nécessite un module de sélection des phrases importantes dans les documents. Cette sous-tâche peut être évaluée en comparant la sélection d'un système auto-

[9]Site web : `http://doc.nist.gov`, visité en décembre 2006. DUC est organisée chaque année depuis 2001 et au moins jusqu'en 2007. Une autre évaluation, SUMMAC (Mani et al., 2002), est moins récente.

| Tâche | Obj. | Rep. | Op. | Auto. | Type |
|---|---|---|---|---|---|
| Sélection de phrase | sous | - - | - - | + | intrinsèque |
| Compression de phrase | sous | - - | - - | + | |
| Tâches d'analyse linguistique | sous | - - | - - | + | |
| Questions linguistiques | semi | + | - | - | intrinsèque |
| Rouge | quasi | - | + + | + | |
| *Pyramids* | quasi | + | + | - | |
| Besoin utilisateur (*responsiveness*) | pseudo | + + | + + | - | extrinsèque |
| Application | réel | + + + | + + + | - | |

**TAB. 2.2:** *Différentes évaluations du résumé automatique et leurs caractéristiques principales telles que le réalisme de l'objectif (Obj.), la représentativité de l'évaluation (Rep.), l'opacité par rapport aux performances des composantes du système sous-jacent (Op.), la possibilité d'automatiser l'évaluation (Auto.), et le type d'évaluation (Type). Cette classification est inspirée d'un travail de groupe lors de l'atelier DUC 2005.*

matique avec celle d'un opérateur humain. Le problème provient de l'observation que deux phrases mises ensemble n'apportent pas la même information que lorsqu'elles sont séparées. La compression de phrase correspond à éliminer l'information inutile afin de gagner de la place dans un résumé de taille fixe et y introduire plus d'informations (Hori et al., 2003a). Cette sous-tâche reste difficile à évaluer comme pour le résumé automatique car même si aucune reformulation n'est autorisée, différents évaluateurs ont tendance à produire différentes formes compressées de référence. Plus généralement, l'ensemble des tâches d'analyse linguistique — comme la résolution des références ou la détection de l'argumentation — peuvent être évaluées comme un sous-objectif. De bonnes performances sur l'ensemble de ces tâches n'induisent pas forcément la qualité *in fine* du résumé.

**Topic :** D0604D
**Title :** anticipation of and reaction to the premier of Star Wars Episode I – The Phantom Menace
**Narr :** How did fans, media, the marketplace, and critics prepare for and react to the movie ? Include preparations and reactions outside the United States.

**TAB. 2.3:** *Exemple de topic (sujet) DUC 2006 sur la sortie du film « Star Wars : la Menace Fantôme » (D0604). Le champ Topic indique l'identifiant du topic ; le champ Title représente le titre et le champ Narr explicite les sous-thèmes que le résumé devra aborder.*

La campagne DUC évalue les résumés sur leur forme et leur fond. Pour la forme, le logiciel *Summarization Evaluation Environment*[10] (SEE) permet à des juges de noter manuellement la qualité des résumés selon les critères linguistiques suivants :

Q1 : la *grammaticalité* (problèmes de formatage, de capitalisation, d'omissions, qui rendent difficile la lecture du texte) ;
Q2 : la *redondance* (répétitions non nécessaires d'expressions ou de noms d'entités lorsqu'un pronom aurait suffi) ;

[10] http ://www.isi.edu/ cyl/SEE

Q3 : la *clarté* des références (difficulté à déterminer le référent d'un pronom, connecteurs logiques non satisfaits...) ;
Q4 : la *focalisation* (le résumé ne doit pas contenir d'informations hors de sa thématique) ;
Q5 : la *structure* et la *cohérence* (le résumé doit être organisé et non un amas d'informations).

Prior to its premiere on May 21, 1999, "Star Wars : Episode 1 : The Phantom Menace," the film's producer, Lucasfilm Ltd., and its distributor, 20th Century Fox, engaged in a massive campaign of pre-release hype. A 2-minute 10-second trailer for the film was publicized on the Lucasfilm Web site and screened in late 1998 in 26 states on 75 screens. It was shown on "Entertainment Tonight" and "Access Hollywood." Fans watched the trailer several times. Fans worldwide exchanged information about the film on the Internet. The press reported on buzz about the film. The Atlanta Journal-Constitution covered the trailer's debut and featured articles on the film, as did USA Today.

Lucasfilm Ltd. orchestrated pre-release efforts of fast food companies, toy manufactures, and video game manufactures As the film's release date approached, merchants geared up for what was expected to be one of the biggest rollouts of merchandise in history. Fans began frantic purchases of merchandise from stores and on-line. Media attention also intensified. George Lucas, the film's director appeared on two segments of "60 Minutes." Vanity Fair devoted cover stories to the film and its stars.

Fans lined up at theaters, often in "Star Wars" costumes, to see the film. This scene was repeated in other countries, as Japan, when the film was released. Initial box office sales indicated that receipts for the film could set a record. The reviews of the film were generally negative. Critics praised the special effects but felt that the film's characters were less than compelling.

In preparation for the premier of "Star Wars : Episode I - The Phantom Menace" in May 1999, a 2-minute 10-second trailer (or preview) was released in November 1998. "Star Wars" fans flocked to theaters throughout the United States, not to see the featured films, but to see the "Star Wars" trailer repeatedly. A German college student, not content to wait until the show's opening overseas in June, proposed a trip to the U.S. to see the movie in May and received eager responses from dozens of fans throughout Europe. Licensed merchandise from dolls to shampoo, featuring "Star Wars" characters and themes was marketed with anticipated sales of $1 billion. As May approached some fans stood in line as long as three weeks in advance to get tickets. Estimates of anticipated ticket sales were as high as $1.78 billion. Enthusiasm was shared by fans in Japan including 30 who flew to Los Angeles for the show's opening there.

The opening days were impressive. "The Phantom Menace" reached the $100 million benchmark in its first five days-a record. Crowds diminished after the first week and some reviews were less than completely laudatory. It had grossed $205 million by the 13th day, but by the 23rd day its take of $271.3 million placed it only 11th among cinema moneymakers. In England 13 million "Star Wars" books were printed to meet anticipated demand, but only 3 million were sold. The hype was over and the headline read "'Star Wars' Bombs in Britain."

**TAB. 2.4:** *Comparatif de deux résumés de référence DUC (provenant respectivement des experts J et E) pour le topic D0604 sur la sortie du film « Star Wars : la Menace Fantôme ». Alors que les thèmes abordés dans ces résumés sont similaires (la bande annonce, le merchandizing...), la structure générale est différente et les informations présentées ne se recouvrent pas complètement.*

Une note de 1 à 5 est donnée pour chacune de ces questions linguistiques, 1 correspondant à une très mauvaise qualité et 5 à l'équivalent d'une production humaine. Cette évaluation par un expérimentateur humain est très importante dans le cas d'un résumé textuel car elle reste à notre connaissance la seule façon de juger la forme d'un

résumé. Par contre, les critères sont dépendants du média dans lequel est formulé le résumé : l'audio ou la vidéo ne sont pas soumis aux même contraintes. Par exemple, nous proposons quelques critères plus précis qu'imposerait un résumé parlé :

Q6 : Le contenu vocal doit être agréable à écouter (rythme, clarté, environnement). Une voix de synthèse, par exemple, peut ne pas paraître naturelle ; un fond sonore trop présent dégrade la clarté du message ; une voix trop aiguë peut agacer l'auditeur.

Q7 : La prosodie doit être régulière et en accord avec le message. Il est plus difficile de comprendre une question ayant la prosodie d'une phrase déclarative. Une phrase coupée dont la prosodie n'a pas la forme attendue est désagréable à écouter.

Q8 : Si le résumé contient de la parole rapportée, celle-ci doit être différenciée de la narration. Sa présence doit être justifiée en montrant par exemple une prise de position.

Q9 : La locution du narrateur doit être claire (pas d'hésitations, de reprises, de réparations ...). Ces phénomènes sont souvent dus à un état émotionnel ou une pathologie du locuteur. Le narrateur doit montrer un état neutre. Cette neutralité permet de le différencier des intervenants en condition spontanée.

Q10 : L'identité des intervenants doit être explicitée.

Q11 : Les événements sonores hors parole doivent être décrits (ou attendus) et avoir une pertinence vis-à-vis du message. Par exemple, l'auditeur s'attend à entendre des explosions lors de l'intervention d'un reporter de guerre, mais ne comprend pas une accroche musicale inattendue.

De tels critères ont une incidence certaine sur les méthodes utilisées pour la génération du résumé, comme par exemple la synthèse vocale ou l'identification du locuteur, tâches qui peuvent elles-mêmes être évaluées séparément. Il n'existe pas à notre connaissance d'évaluation du résumé parlé, mais les critères Q6 à Q11 consitituent une base intéressante pour en construire une.

La réponse au besoin est une évaluation de fond du résumé. DUC a choisi de l'évaluer selon des critères manuels et automatiques (afin de réduire les ressources nécessaires à l'évaluation de l'amélioration d'un système). Des juges évaluent le fond en se posant la question : « le résumé répond-t-il au besoin de l'utilisateur ? ». Ce besoin est exprimé dans DUC sous la forme d'un thème et de requêtes en langue naturelle, mais il peut être exprimé dans une forme totalement différente et impliquer des contraintes beaucoup plus complexes à évaluer. DUC propose aussi une évaluation manuelle qui implique le fond et la forme. Les juges doivent répondre à la question « combien d'argent auriez-vous donné pour ce résumé ? ». Ces deux questions donnent lieu à une note entre 1 et 5 et s'appellent dans DUC, *content responsiveness* et *overall content quality*.

DUC propose ensuite une évaluation du fond moins opaque que la note générale de *responsiveness*, sous le nom de *Pyramids* (Nenkova et Passonneau, 2004). Cette évaluation manuelle demande le découpage des résumés de références en informations élémentaires (*Summary Content Units*, SCU ; « Bush a été battu par les démocrates » et « Les démocrates ont gagné la chambre des représentants et le sénat » seront considé-

<DOC>
<DOCNO> APW20000124.0232 </DOCNO>
<DOCTYPE> NEWS STORY </DOCTYPE>
<DATE_TIME> 2000-01-24 22:27 </DATE_TIME>
<BODY>
<HEADLINE> 'Star Wars' Bombs in Britain </HEADLINE>
<TEXT>
<P>
LONDON (AP) – A failed bet on a "Star Wars" craze in Britain has left a publisher with 10 million unsold books and a vacancy for the post of chief executive.
</P>
<P>
British children's books publisher Dorling Kindersley announced Monday that its chief executive has resigned because of "a seriously misjudged" overinvestment in books tied to the launch of the latest movie in the saga, "Star Wars : Episode I, The Phantom Menace."
</P>
<P>
Expecting a Christmas rush, James Middlehurst arranged for the group to print 13 million of the books in the 18 months to Dec. 31 last year. Sales totaled 3 million.
</P>
<P>
Company chairman Peter Kindersley blamed the "Star Wars" debacle for more than half of a pretax loss of $41 million which it expects to announce in the spring.
</P>
</TEXT>
</BODY>
<TRAILER> AP-NY-01-24-00 2227 </TRAILER>
</DOC>

**TAB. 2.5:** *Exemple de document DUC pour le topic D0604 sur la sortie du film « Star Wars : la Menace Fantôme ». Le document illustre un aspect traité dans le résumé de l'expert E.*

rées comme équivalentes), dont une pondération est donnée par le nombre de résumés qui les contiennent. Puis, ces informations élémentaires sont mises en relation manuellement avec celles contenues dans un résumé automatique, en autorisant des imprécisions, des généralisations et des spécialisations. Le score du résumé est donné par le nombre d'unités en commun, sachant que les unités de poids supérieur (les plus importantes) doivent apparaître avant d'obtenir la validation des unités de poids inférieur.

Lin (2004) propose une évaluation automatique fortement corrélée avec l'évaluation humaine à travers *Recall-Oriented Understudy for Gistering Evaluation* (Rouge). Cette mesure fait intervenir la différence entre la distribution des mots d'un résumé candidat et celle d'un ensemble de résumés de référence. Copeck et Szpakowicz (2004) estiment qu'il faut 30 résumés pour construire un résumé moyen de référence représentatif ; DUC n'en produit que 4 pour l'évaluation (pour des raisons de coût), n'ayant pas observé de différence fondamentale dans les mesures Rouge lorsque plus de résumés sont disponibles. Fortement utilisé durant les campagnes DUC pour lesquelles elle représente l'évaluation automatique du fond, ce genre de mesure est en constante amélioration et tend à intégrer des éléments conceptuels (notion de *Basic Elements* : entités-

relations). Il s'agit de se rapprocher d'un espace de représentation de l'information et de s'éloigner de l'espace d'instanciation linguistique. Une mesure automatique est indispensable à l'amélioration des systèmes car elle prend beaucoup moins de temps et ne demande pas d'opérateur humain[11]. Rouge-2 et Rouge-SU-4 se sont imposées dans DUC pour l'évaluation automatique. Rouge-2 correspond au nombre de bigrammes en commun entre un résumé automatique et l'ensemble des résumés écrits à la main ($R_n$ avec $n = 2$ dans l'équation 2.10). Rouge-SU-4 correspond au rappel en « bigrammes à trous » (*skip units*) de taille maximum 4 ($RSU_n$ avec $n = 4$ dans l'équation 2.11). La table 2.6 regroupe quelques exemples d'éléments utilisés dans Rouge.

$$R_n(hyp, ref) = \frac{|\{w_i, \cdots, w_{i+n-1}\}_{ref} \cap \{w_i, \cdots, w_{i+n-1}\}_{hyp}|}{|\{w_i, \cdots, w_{i+n-1}\}_{ref}|} \tag{2.10}$$

$$RSU_n(hyp, ref) = \frac{|\{w_i, w_{j<i+n+1}\}_{ref} \cap \{w_i, w_{j<i+n+1}\}_{hyp}|}{|\{w_i, w_{j<i+n+1}\}_{ref}|} \tag{2.11}$$

| **Phrase** | **le chat boit du lait** |
|---|---|
| 2-gram | le-chat, chat-boit, boit-du, du-lait |
| 2-gram (lemmes) | chat-boire, boire-lait |
| SU-2 | *2-grams* + le-boit, le-du, chat-du, chat-lait, boit-lait |
| BE | chat-le-dét., boire-chat-sujet, lait-du-prép., boire-lait-comp. |

**TAB. 2.6:** *Illustration de différents découpages pour le calcul de Rouge. Les éléments de base sont : les n-grammes (lemmatisés ou non), les bigrammes à trous (Skip Units, SU) et les triplets d'entités-relations (Basic Elements, BE).*

Des pré-traitements peuvent être appliqués, comme la suppression des mots-outils (qui représentent en général 50% des occurrences des unigrammes), ou l'utilisation de dictionnaires de synonymes pour améliorer la corrélation avec les résumés écrits à la main. Un des inconvénients de Rouge est qu'un bon choix des sous-séquences de mots autorise la génération de résumés ayant un meilleur score que les résumés écrits manuellement au détriment de la forme. Notamment, les méthodes à base d'apprentissage maximisant le critère Rouge doivent absolument contrebalancer cet effet par des contraintes linguistiques.

Les méthodes proposées pour évaluer la qualité d'un résumé requièrent soit un jugement sur le moment (évaluation manuelle), soit la création de références *a priori* (évaluation automatique). Dans les deux cas, il est observé que l'accord entre les juges est généralement faible (Minel, 2004) et qu'il faut multiplier les jugements pour obtenir des résultats significatifs. La mesure Kappa (Fleiss, 1971) représente le degré d'accord entre les juges $P(A)$ au delà du hasard $P(C)$ (équation 2.12).

$$\kappa = \frac{P(A) - P(C)}{1 - P(C)} \tag{2.12}$$

[11] Les organisateurs de DUC estiment à 3000 heures le temps passé pour réunir les documents, créer les références et évaluer les soumissions (30 participants et 50 *topics*).

La même problématique se retrouve dans les mesures automatiques comme Rouge, pour lesquelles une méthode de *Jacknife* améliore la robustesse. Lorsqu'un résumé est évalué par rapport à $N$ résumés de référence, l'évaluation est répétée $N$ fois en retirant à chaque fois un résumé de référence. Le score final est défini par la moyenne des scores de chaque itération. Cette approche permet de calculer des intervalles de significativité par analyse de la variance ANOVA (Lindman et al., 1976).

Après ce tour d'horizon de l'évaluation du résumé, nous allons voir les approches les plus populaires pour le résumé textuel et les contraintes spécifique à l'utilisation d'un média parlé dans ce domaine.

### 2.2.2 Résumé par extraction

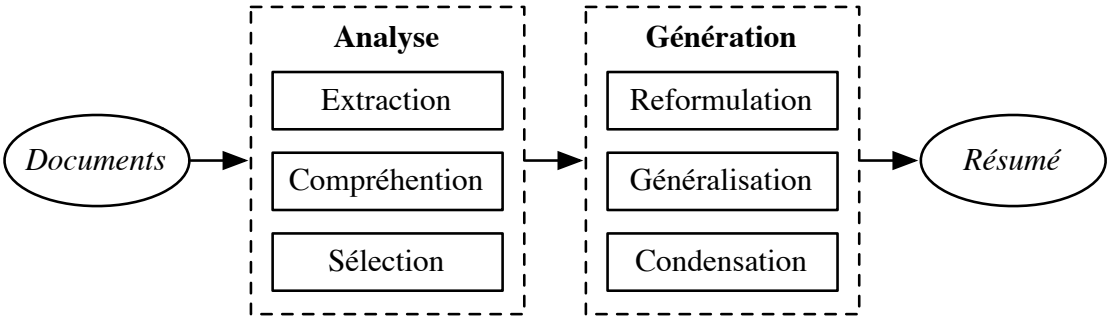



**FIG. 2.5:** *Processus de création d'un résumé, en deux étapes : une étape d'analyse de l'information contenue dans les documents d'origine, suivie d'une étape de génération dans le média cible.*

Le texte véhicule des idées, des concepts et des informations. Le résumé automatique de texte nécessite avant tout de dégager les informations importantes et de les resynthétiser avec un ratio de réduction fixé. Il a été observé qu'environ 70% des phrases utilisées dans des résumés réalisés par des opérateurs humains étaient empruntées au texte d'origine sans modification (Lin et Hovy, 2003; Jing, 2002). Il est dans ce cas tout à fait envisageable de rechercher avant tout ces phrases transposables sans nécessiter de «compréhension/synthèse» complète. Nous avons déjà vu qu'il existe deux grandes approches pour le résumé automatique. L'approche par extraction vise à trouver parmi les phrases d'origine, les plus susceptibles d'être réutilisées, en partie ou en totalité, dans le résumé généré. Ce genre d'approche implique une étape de sélection des phrases suivie d'une étape de placement des phrases dans le résumé et d'amélioration de la forme. L'alternative est d'identifier les informations importantes du domaine et de les mettre en correspondance avec des patrons de résumé permettant de générer des phrases et les assembler en un discours cohérent. La première méthode est la plus appropriée pour des résumés indépendants du domaine car elle utilise généralement des statistiques alors que la seconde demande des connaissances du domaine.

### Analyse

Les méthodes d'analyse de l'information comprennent souvent une phase d'extraction d'unités sémantiques suivie d'une phase de réduction du contenu à l'essentiel. Cependant, il existe des approches focalisées sur le contenu sémantique et ignorant la forme de surface des phrases. Ou, au contraire, indépendantes du contenu mais utilisant des caractéristiques de surface indicatrices de l'importance d'une phrase. Toutefois, l'utilisation conjointe de ces critères est la plus répandue.

Les traitements linguistiques permettant l'extraction d'unités sémantiques à partir des mots sont généralement les mêmes que ceux appliqués en recherche documentaire avec pour spécificité que l'unité est la phrase au lieu d'être le document (voir section 2.1). Il n'est plus possible de compter sur l'effet quantitatif et redondant du document ; le résumé automatique est plus fortement demandeur de précision dans l'extraction des unités sémantiques et dans les divers pré-traitements, car les erreurs sont directement répercutées sur le résultat. Par exemple, Blair-Goldensohni et al. (2004) appliquent un étiquetage morpho-syntaxique des mots pour dégager des patrons de phrases et marquer les mots porteurs de sens en fonction de leur catégorie syntaxique. Les entités nommées sont des marqueurs de l'information du domaine utiles pour inférer la tendance d'une phrase à apparaître dans le résumé (Bergler et al., 2004). Dans un souci de précision de l'information présentée au sein d'une phrase, Vanderwende et al. (2004) et Witte et Bergler (2003) appliquent une résolution des anaphores pronominales et des groupes nominaux. Ces anaphores et ces entités nommées sont également utiles lors de la phase de génération du résumé pour améliorer la clarté des références. Le contenu de surface (les mots) peut être projeté dans un réseau sémantique tel Wordnet, comme le font Doran et al. (2004) pour tirer parti des relations entre les éléments (synonymie, généralisation...). Finalement, une analyse grammaticale complète aide à l'extraction de paires d'entités-relations représentatives du message (Hovy et al., 2005).

Les descripteurs sémantiques représentent chaque phrase sous la forme d'un point dans un « espace sémantique », permettant de sélectionner celles qui répondent le mieux au besoin de l'utilisateur. Cette sélection est réalisée de façon similaire à la tâche de recherche documentaire avec pour contrainte supplémentaire qu'il ne faut pas proposer plusieurs fois la même information. Pour cela, la sélection de phrases est formulée comme un problème d'optimisation consistant à maximiser la couverture de l'information présentée tout en minimisant sa redondance. Ce problème est en partie résolu par des algorithmes de classification non supervisée afin de partitionner l'espace en $N$ classes et de sélectionner la phrase la plus représentative de chaque classe. Seki et al. (2004) implémentent un *clustering* hiérarchique ascendant et comparent diverses méthodes d'agrégation : le critère de Ward semble donner les résultats les plus équilibrés. Dans la même optique, *Maximal Marginal Relevance* (MMR) est un algorithme glouton qui sélectionne les phrases en fonction de leur similarité avec le besoin utilisateur et leur dissimilarité avec ce qui a déjà été sélectionné (Goldstein et al., 2000; Erkan et Radev, 2004; Mori et Sasaki, 2002). La formule de MMR est explicitée par l'équation 2.14, dans laquelle $mmr_k$ correspond à la sélection de phrases à l'étape $k$, $s_i$ au représentant dans l'espace sémantique de la phrase $i$, $c$ au représentant du besoin utilisateur, $sim(a, b)$ à

un opérateur de similarité dans l'espace sémantique, et $\lambda$ à un hyperparamètre fixé de manière empirique.

$$mmr_k = mmr_{k-1} \cup \{\hat{s}_k\} \tag{2.13}$$

$$\hat{s}_k = \underset{s_i \notin mmr_{k-1}}{\operatorname{argmax}} \left( \lambda sim(s_i, c) - (1-\lambda) \max_{s_j \in mmr_{k-1}} sim(s_i, s_j) \right) \tag{2.14}$$

Les implémentations classiques de MMR reposent sur le modèle vectoriel (VSM, voir section 2.1.4), en utilisant des vecteurs de mots et la similarité *cosine*. Au-delà de cette méthode rapide, le problème d'optimisation peut être résolu grâce à diverses techniques d'optimisation, comme le montrent (Alfonseca et al., 2004) en appliquant un algorithme génétique pour produire et sélectionner des résumés maximisant la couverture et minimisant la redondance. Gong et Liu (2001) ont une approche similaire à LSA, en remarquant que la décomposition en valeurs singulières effectue un partitionnement flou des phrases (voir section 2.1.4). La SVD effectue une décomposition de la matrice des occurrences de mots dans les phrases en trois matrices $U\Sigma V^T$. La matrice $V$ donne une projection des phrases selon la base quasi orthogonale trouvée par la SVD. En comparant les axes de cette base à des thèmes latents, les phrases les plus représentatives de chaque thème (de projection maximale sur les axes principaux de la base) sont retenues pour construire le résumé.

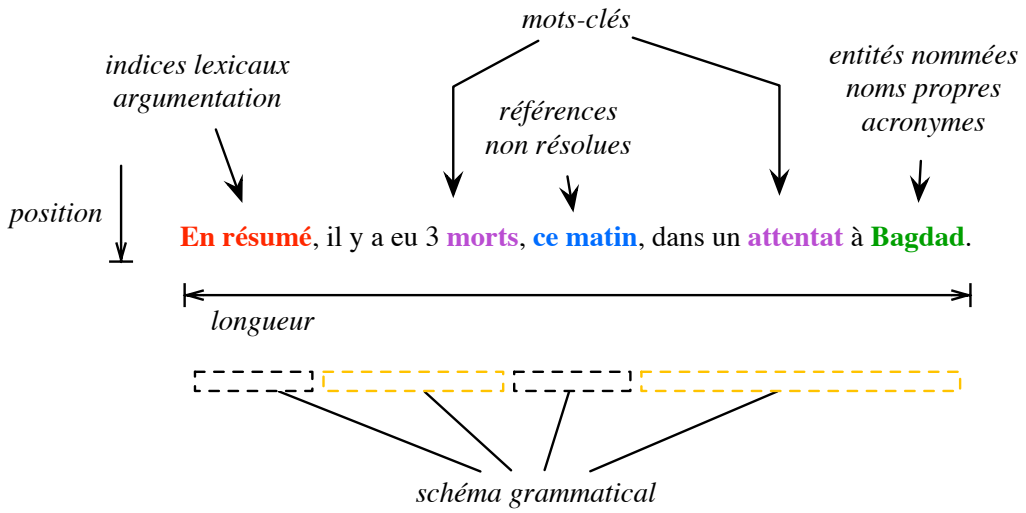



**FIG. 2.6:** *Quelques caractéristiques de la phrase ayant une influence sur sa sélection dans le résumé (position dans le document, longueur, indices lexicaux, références anaphoriques, mots-clés porteurs du contenu, entités spécifiques, schéma grammatical).*

L'information portée par une phrase n'est pas la seule caractéristique utile pour intégrer une phrase dans un résumé (Nobata et Sekine, 2004). La figure 2.6 illustre les caractéristiques pertinentes pour le résumé par extraction. Par exemple, il est observé que les phrases en début et en fin de document décrivent bien son contenu général (*position*). Les phrases courtes sont trop générales et les phrases longues sont trop spé-

cifiques (*longueur*). Il faut ajouter que les phrases qui contiennent des pronoms non résolus risquent de détériorer la cohérence finale du résumé (*références non résolues*). Des expressions permettent aussi de situer la phrase dans l'argumentation et de détecter les descriptions générales du contenu (*indices lexicaux*). Des mots-clés du domaine, ou tout simplement des mots-clés fréquents, ainsi que les mots morphologiquement indicateurs de contenu, comme les noms propres ou les acronymes, sont caractéristiques de phrases denses en informations (*mots-clés*, *entités nommées*, *acronymes*). Enfin, l'étude de la forme grammaticale de la phrase facilite la détection des propositions déclaratives les plus propices au résumé (*schéma grammatical*). Les caractéristiques utiles sont celles qui indiquent un fort contenu et qui permettent de maximiser la qualité linguistique du résumé produit. Les approches supervisées utilisent un corpus d'apprentissage (couples documents-résumés) pour résoudre le problème de classification consistant à sélectionner ou non une phrase pour le résumé. Par exemple, Kupiec et al. (1995) effectuent une classification bayesienne naïve et Daumé III et Marcu (2004) utilisent des machines à vecteur support (*Support Vector Machines*, SVM). Le problème est résolu de façon non supervisée par Torres-Moreno et al. (2002) grâce à une règle de décision prenant en compte une version normalisée de chaque statistique.

Des méthodes plus originales ont été proposées comme la création d'un graphe d'entités-actions (noms-verbes), sur lequel est appliqué un algorithme de popularité comme *PageRank* (Vanderwende et al., 2004). Il faut aussi noter l'approche par débruitage du texte après estimation du bruit généré par le canal qui le sépare du résumé (Daumé III et Marcu, 2001). Dans le cadre de l'évaluation DUC, Lacatusu et al. (2006) profitent des sous questions contenues dans l'expression du besoin utilisateur — à l'aide d'un système question-réponse — et appliquent une technique d'inférence textuelle (*textual entailment*) pour déterminer la redondance d'une phrase par rapport à une autre.

#### Génération

Le résumé par remplissage de patrons demande des modules de génération complexes, souvent à base de règles. Par exemple, Radev et McKeown (1998) effectuent une combinaison de patrons pour générer des transitions entre les phrases et utilisent des opérateurs pour construire l'argumentation. D'un autre côté, la génération la plus simple pour le résumé par extraction correspond à la juxtaposition des phrases sélectionnées jusqu'à l'obtention de la quantité d'informations voulue (ratio de réduction, nombre de phrases, nombres de mots). Les méthodes un peu plus avancées utilisent tout de même quelques post-traitements de surface pour améliorer la forme du résumé (Zajic et al., 2004) :

- compression des phrases par un modèle supprimant des sous-arbres syntaxiques ;
- remplacement des anaphores par les entités auxquelles elles se réfèrent (ou minimisation des références non résolues) ;
- normalisation des noms propres (leur première occurrence doit être le nom complet, les suivantes utiliseront uniquement le nom de famille...) ;
- suppression des annonces de discours rapporté (« *..., a dit le chef de la police* ») ;

– suppression des marques de l'argumentation générant des contre-sens lorsque les phrases sont utilisées hors contexte ;
– suppression du contenu entre parenthèses.

En plus de ces exemples de post-traitements, des efforts sont faits pour améliorer la structure et la cohérence des résumés. Daumé III et Marcu (2004) remarquent que l'ordre des phrases est important et qu'il peut être judicieux de sélectionner une phrase si elle est encadrée par deux phrases qui apparaîtront dans le résumé.

### 2.2.3 Spécificités de la parole

Un résumé automatique de parole est constitué à partir d'un flux audio parlé (entrées) et généré sous forme écrite ou parlée (sorties). La méthode la plus naturelle consiste à profiter des approches développées pour le texte, grâce à une étape de la transcription automatique du contenu parlé. Toutefois, dans ce cas, les traitements ne peuvent plus compter sur l'absolue précision de la forme écrite et doivent être capables d'intégrer les erreurs du processus automatique de transcription. Plusieurs questions doivent être explorées : à quel point les méthodes issues du texte sont-elles affectées par un contenu audio ? Est-il possible de les adapter à un contenu bruité ? Existe-il des différences fondamentales entre le texte et la parole qui permettent d'extraire des paramètres utiles supplémentaires ?

Si les recherches sur le résumé textuel se sont concentrées sur des contenus techniques et journalistiques[12], le résumé de parole est, pour l'instant, appliqué à trois sources de données : les émissions radiodiffusées, les dialogues téléphoniques et les réunions. Ces sources amènent néanmoins de nombreux challenges (McKeown et al., 2005). Christensen et al. (2003) montrent que les techniques classiques de résumé automatique sont portables aux journaux radiodiffusés et que les méthodes de sélection de phrases ont tendance à sélectionner des phrases dont le taux d'erreur de mots (provoqué par la transcription automatique) est plus faible que sur la globalité des données, un phénomène qui a aussi été observé par Murray et al. (2005). Cette différence en taux d'erreur s'explique par la relation entre l'importance d'une information et sa redondance dans les documents, selon la plupart des modèles de résumé automatique de textes.

Pour ce qui est de l'adaptation des méthodes de résumé et d'extraction d'informations, Zechner (2001) effectue une correction des malformations de l'élocution (*disfluencies*) comme les hésitations, les reprises, les coupures, les faux départs ou les pauses remplies. Il utilise des méthodes d'annotation de séquence sur le contenu linguistique et écarte les phrases ayant trop d'erreurs grâce à des mesures de confiance issues de la transcription. Il obtient ainsi une transcription propre, appropriée aux méthodes conçues pour le texte. Il reste malheureusement quelques erreurs qui réduisent les performances des techniques d'extraction de descripteurs sémantiques (entités nommées,

[12] Les évaluations DUC sont aujourd'hui concentrées sur les journaux, mais lors de l'atelier de la conférence en 2006 à New York, il a tout de même été discuté de migrer vers un contenu moins formel comme les *blogs*, et plus tard un contenu audio

relations, analyse grammaticale) et de nombreux travaux sont en cours dans ces domaines pour mieux coupler la transcription avec ces tâches intéressantes pour le résumé (Kubala et al., 1998; Van Noord et al., 2000).

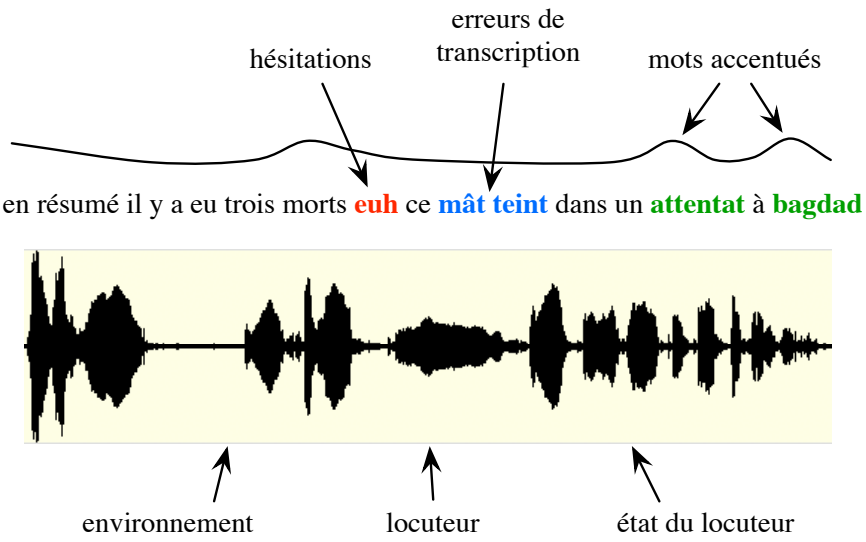



**FIG. 2.7:** *Paramètres supplémentaires de la parole par rapport au texte pour caractériser des phrases en vue de leur extraction pour créer un résumé (environnement acoustique, identité et type de locuteur, état du locuteur, hésitations, fiabilité de la transcription, et prosodie).*

La figure 2.7 donne quelques exemples des spécificités de la parole qui peuvent être ajoutées aux caractéristiques textuelles pour construire un résumé. Par exemple, Maskey et Hirschberg (2006) recherchent les paramètres permettant de se passer de la transcription pour faire un résumé. Plus précisément, ils tirent parti de la prosodie, et de la position des phrases dans le document à l'aide de Modèles de Markov Cachés (HMM). Bien que le modèle permette une amélioration significative par rapport au hasard, il ne s'applique que pour le résumé mono-document et les auteurs ne présentent pas de comparatif avec une méthode basée sur la transcription. Murray et al. (2005) utilisent un classifieur et des paramètres à la fois linguistiques, prosodiques et structurels pour générer des résumés de réunions et arrivent à des résultats intéressants en limitant la complexité des paramètres employés. Zhu et Penn (2005) comparent différentes approches pour cette tâche (MMR, une similarité sémantique, SVM et une régression logistique). Leurs conclusions sont que ces approches ont des performances très variables en fonction du degré de réduction et du taux d'erreur de mots de la transcription. La segmentation en phrases est un autre aspect important et non trivial du résumé par extraction. En effet, Mrozinski et al. (2006) comparent la segmentation de référence avec une segmentation automatique utilisant des caractéristiques linguistiques et une segmentation aléatoire. Ils observent que les performances en résumé baissent fortement lorsque la segmentation en phrases est de mauvaise qualité. Furui et al. (2004) de leur côté testent plusieurs unités de base pour l'extraction (mots, syntagmes «entre deux pauses remplies», phrases) dans le cadre d'un résumé parlé et observent que seules

les unités longues, comme la phrase, sont acceptables pour l'utilisateur. D'autres pistes sont explorées, comme l'intégration du rôle des locuteurs dans le résumé d'émissions radiophoniques (présentateur, reporter, invité, Barzilay et al., 2000) ou la cohérence des dialogues pour attacher les questions à leurs réponses (Zechner, 2002).

La plupart des méthodes appliquées à l'audio concentrent leurs efforts sur la partie analyse du processus de résumé. Seuls Hori et al. (2003b) essaient d'améliorer la partie génération en tirant parti des hypothèses de transcription pour réaliser une compression de phrase. Des modèles probabilistes d'informativité *a priori* des mots et de suppression de séquences de mots sont appliqués au graphe d'hypothèses de transcription dans le but de prendre en compte ces paramètres lors de la recherche de la meilleure hypothèse. Cette approche donne de bons résultats dans le cadre du sous-titrage d'emissions télévisées japonaises.

## 2.3 Conclusion

L'information parlée introduit de nombreux challenges comparé à la recherche d'information textuelle. La variabilité de la parole est le premier obstacle à l'extraction des descripteurs sémantiques d'un document. En outre, les nombreuses techniques d'analyse des phénomènes linguistiques sur le texte peuvent être appliquées à la transcription automatique du message parlé, au détriment d'erreurs proportionnelles au taux d'erreur de mots. Si l'impact de ces erreurs est relativement faible sur une tâche de recherche documentaire grâce à la redondance de l'information contenue dans les documents, il s'alourdit lors de l'extraction de descripteurs sémantiques de plus haut niveau et lorsque des structures plus petites que le document, comme la phrase, doivent être employées. Un premier challenge consiste à mieux coupler les tâches pré-transcription et post-transcription, en tirant parti des spécificités de la parole. Intégrer ces spécificités à la recherche d'information parlée représente un second verrou à lever. Toutefois, avant de s'intéresser à ces problèmes, il est indispensable de décrire les principes régissant la structuration de l'information parlée ; ceci est l'objectif du chapitre 3.

# Chapitre 3

# Structuration de l'information parlée

**Sommaire**



Le chapitre 2 était dédié à la recherche d'information au travers d'une description de la recherche documentaire (section 2.1) et du résumé automatique (section 2.2). Les méthodes présentées sont pour la plupart issues du traitement de la langue naturelle écrite avant d'être adaptées aux problématiques de la parole. Cette adaptation implique l'extraction de descripteurs structurels et sémantiques dans le flux de parole. Nous allons maintenant nous intéresser aux différentes étapes nécessaires pour structurer ce type de flux.

La figure 3.1 illustre les différentes étapes de la structuration. Une chaîne de structuration a été développée au LIA : elle regroupe un module de segmentation en macro-classes acoustiques, un module d'indexation en locuteur et un module de transcription orthographique. Cette chaîne, antérieure aux travaux décrits dans ce document, est présentée rapidement par la section 3.1. Les différentes composantes sont évaluées en section 3.2, au travers de la campagne ESTER. Nous avons développé d'autres éléments de structuration, la segmentation en phrases et l'extraction d'entités nommées, pour compléter cette chaîne de structuration. Ces éléments feront l'objet du chapitre 4.

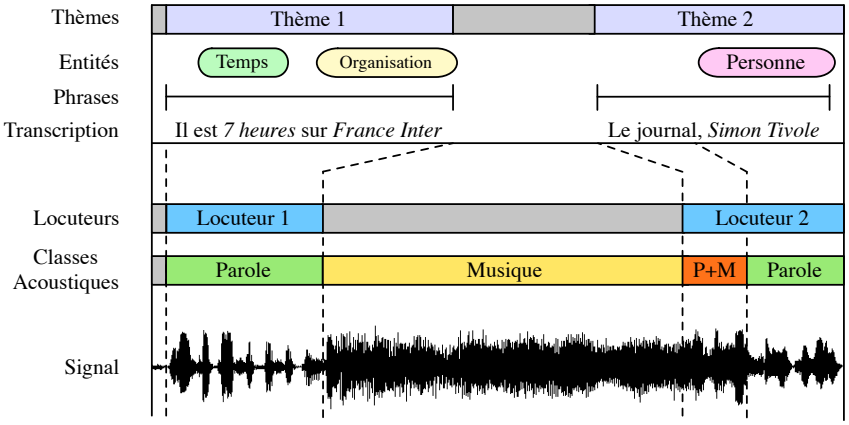



**FIG. 3.1:** *Principe de la structuration du flux de parole par tâches de plus en plus proches de la sémantique. Le signal sonore est découpé en classes acoustiques pour détecter les portions contenant de la parole ; puis les tours de parole et éventuellement l'identité des locuteurs sont retrouvés. Ces informations permettent d'améliorer la transcription orthographique du discours parlé. Cette dernière facilite l'extraction de descripteurs sémantiques de plus haut niveau (entités nommées, thèmes...).*

## 3.1 La chaîne de structuration Speeral

Le système de structuration de parole du LIA effectue la transcription du contenu parlé d'un document audio et génère une segmentation en locuteurs tout en étiquetant les zones de silence et de musique. La plupart des outils ont été développés au LIA et utilisent des techniques classiques d'apprentissage artificiel. La figure 3.2 détaille le fonctionnement de la chaîne étape par étape.

### 3.1.1 Paramétrisation acoustique

La parole est stockée sous la forme d'un signal numérique généralement quantifié sur 16 bits à une fréquence de 16000 échantillons par seconde. La plupart des tâches de structuration reposent sur une reconnaissance de forme dans cet espace. Pour rendre cette reconnaissance possible, les paramètres d'un modèle de production et/ou de perception sont représentés sous forme de vecteurs dans un « espace acoustique ». Les paramètres les plus répandus sont : *Linear Predictive Cepstral Coefficient* (LPCC, Rahim et Lee, 1996), *Mel Frequency Ceptstrum Coefficents* (MFCC, Davis et Mermelstein, 1980), ou encore *Perceptual Linear Predictive* (PLP) analysis (Hermansky, 1990).

Dans la chaîne présentée, le signal de parole est découpé en vecteurs acoustiques (d'une portée de 25 ms, avec un décalage de 10 ms) représentant les fréquences caractéristiques de la parole, ainsi que leur dynamique. Les paramètres extraits prennent en compte à la fois la production et la perception de la parole (12 coefficients PLP et l'énergie, leurs dérivées et dérivées secondes, soit 39 dimensions pour la transcription (Lévy et al., 2004) ; les autres tâches reposent sur des jeux de paramètres similaires). La

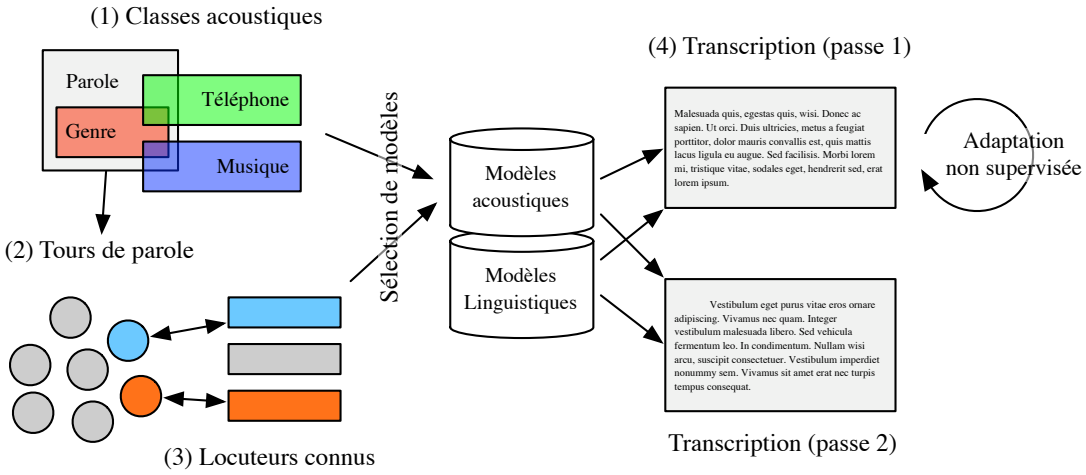



**FIG. 3.2:** *La chaîne de structuration audio Speeral procède en 4 étapes : segmentation en classes acoustiques (1), segmentation en tours de parole (2), identification de locuteurs connus (3), transcription (4) à l'aide de modèles adaptés aux caractéristiques de la parole détectées dans les étapes précédentes, seconde passe de transcription après adaptation non supervisée des modèles acoustiques.*

figure 3.3 illustre la paramétrisation acoustique.

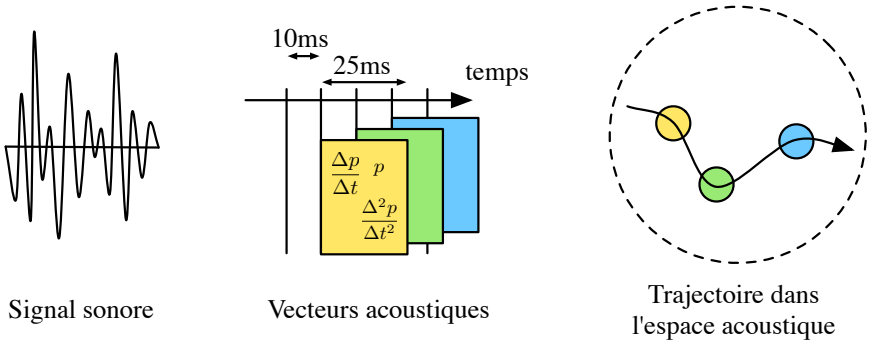



**FIG. 3.3:** *Passage d'une représentation numérique du signal sonore en un espace acoustique dédié à l'application de méthodes mathématiques et statistiques classiques pour les tâches de segmentation et d'identification.*

### 3.1.2 Segmentation en classes acoustiques

Un segmenteur en classes acoustiques permet de séparer les différents types de signaux (parole, silence, musique, bruit) et de classifier le type de parole (genre du locuteur, téléphone ou studio, parole sur musique) afin d'utiliser des modèles adaptés à l'environnement acoustique durant les phases suivantes. Cette approche est implémentée dans la chaîne de structuration sous forme d'un *Hidden Markov Model* (HMM) ergodique dont les états sont modélisés par des *Gaussian Mixture Models* (GMM, Fredouille

et al., 2004). Ces densités de probabilités sont estimées par l'algorithme Estimation-Maximisation (EM, Dempster et al., 1977) ; les probabilités de transition entre les états sont estimées par l'algorithme Baum-Welch (Baum et al., 1970). Cette approche est illustrée dans la figure 3.4. La séquence la plus probable est trouvée par programmation dynamique à l'aide de l'algorithme Viterbi (Viterbi, 1967).

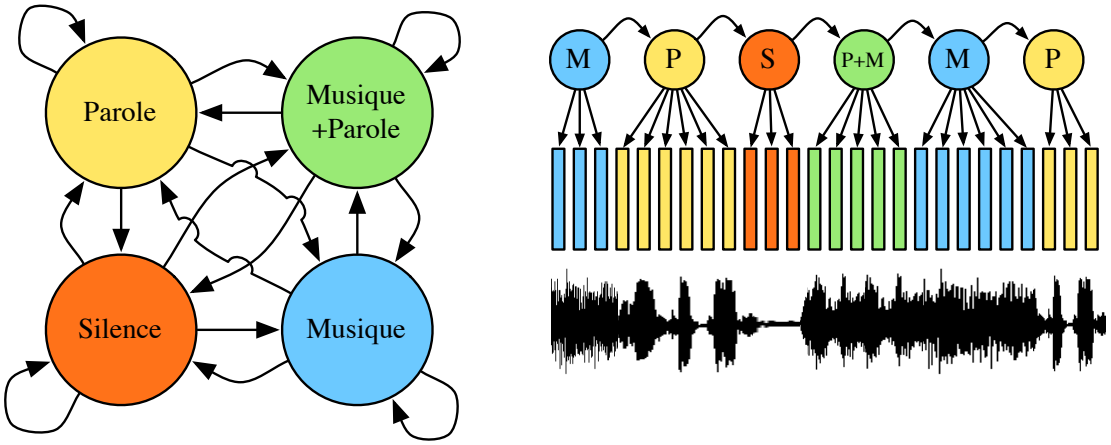



**FIG. 3.4:** *Illustration de la segmentation en classes acoustiques d'un signal sonore à l'aide d'un HMM ergodique dont les données générées par les états sont modélisés par des GMM.*

Ce type de modélisation a été étendu à d'autres tâches : Atrey et al. (2006) appliquent les techniques de modélisation GMM à la détection d'événements sonores (bruits de pas, course, pleurs, bruits de chute) pour la surveillance multimédia ; Dufaux et al. (2000) utilisent des techniques similaires, pour détecter des événements sonores dans un environnement bruité.

### 3.1.3 Indexation en locuteurs

L'indexation en locuteurs consiste en une étape de segmentation en tours de parole, suivie du regroupement des tours de parole en locuteur et d'une identification des locuteurs connus (suivi de locuteur). Connaître l'identité des locuteurs permet l'emploi de modèles adaptés pour les locuteurs fréquents lors de la phase de transcription (comme, par exemple, les présentateurs d'émissions radiophoniques).

Dans notre cas, la segmentation et le suivi de locuteur sont réalisés grâce aux outils LIA_SpkSeg et LIA_SpkDet, fondés sur Alize[1] (Istrate et al., 2005). La segmentation est générée par un HMM dynamique auquel est ajouté un état à chaque fois qu'un nouveau locuteur prend la parole. Les locuteurs connus sont ensuite recherchés parmi les regroupements de tours de parole. La décision de classification provient du rapport de vraisemblance entre un modèle de locuteur et un modèle générique (UBM). Comme

[1] disponible sous une licence GPL sur `http://lia.univ-avignon.fr/heberge/ALIZE/`, visité en septembre 2006

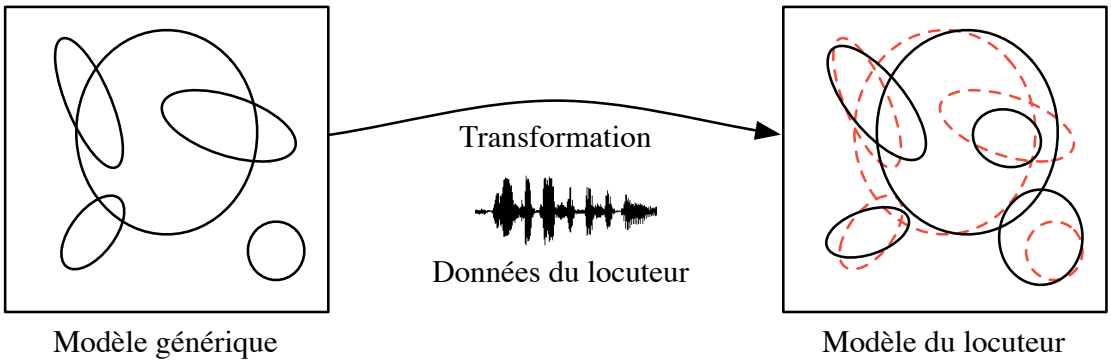



**FIG. 3.5:** *Modélisation de l'occupation de l'espace acoustique par un locuteur en estimant une transformation d'un modèle générique à partir des données observées de ce locuteur.*

peu de données sont disponibles sur l'occupation de l'espace acoustique par un locuteur, le modèle de ce locuteur est estimé à l'aide d'une adaptation du modèle générique aux données observées (voir figure 3.5).

### 3.1.4 Transcription automatique

La transcription orthographique consiste en une reconnaissance de la séquence de mots prononcée dans un flux de parole. Les systèmes de transcription actuels sont indépendants du locuteurs, traitent de la parole continue et reconnaissent un vocabulaire étendu (*Large Vocabulary Continuous Speech Recognition*, LVCSR).

Dans la chaîne de structuration présentée, la transcription automatique est effectuée en 2 passes, dont la première sert à générer rapidement une séquence de mots approximative. Celle-ci permet une adaptation en aveugle des modèles acoustiques. Les modèles ainsi adaptés sont utilisés en deuxième passe. Le système de transcription, Speeral (Nocéra et al., 2004), est un moteur de reconnaissance de la parole grand vocabulaire, multi-locuteurs utilisant une reconnaissance HMM des phonèmes, un lexique de phonétisation et des modèles linguistiques $n$-grammes. Le meilleur chemin dans le graphe d'hypothèses est déterminé grâce à une version modifiée d'A$^\star$ basée sur une estimation à moindre coût de la fin de la transcription (acoustique et prédiction linguistique) et une méthode d'élagage de l'arbre d'hypothèses.

Des ouvrages comme (Huang et al., 2001), (De Mori, 1998), ou (Haton et al., 2006) expliquent plus en détail le fonctionnement d'un système de transcription de la parole.

### 3.1.5 Traitements de plus haut niveau

La segmentation en macro-classes acoustiques, l'indexation en locuteurs et la transcription orthographique représentent les éléments de structuration fournis par la chaîne au début des travaux présentés dans ce document. Nous y avons ajouté des traitements

spécifiques au résumé automatique : une segmentation en phrases et une détection d'entités nommées. Ces modules sont décrits en détail dans les sections 4.1 et 4.2. Nous n'abordons pas la segmentation et le suivi de thèmes qui pourraient aussi être importants selon les traitements de haut niveau envisagés. Walls et al. (1999) proposent une détection de thème dans des émissions radiodiffusées. Shriberg et al. (2000) étudient des paramètres prosodiques pour une détection de frontières thématiques.

## 3.2 Évaluation lors de la campagne ESTER

La campagne d'Évaluation des Systèmes de Transcription d'Émissions Radiophoniques (ESTER) a été organisée par l'Association Francophone de la Communication Parlée (AFCP), la Délégation Générale pour l'Armement (DGA) et *Evaluation and Language resources Distribution Agency* (ELDA), dans le cadre du projet EVALDA (évaluation des technologies de la langue en français), un volet de l'action Technolangue, financée par le Ministère de la Recherche et de l'Industrie. Entre 2003 et 2005, différentes phases de la campagne ont dynamisé les sous-domaines du traitement de la parole, en fournissant les moyens d'évaluer les systèmes issus de la recherche sur des tâches bien définies et reconnues au niveau international.

### 3.2.1 Présentation des données et des tâches

Un premier corpus de 90 heures de radio en français, transcrit et annoté, a été fourni aux participants. Ce corpus est divisé en 2 parties aux fonctions différentes : la première, « ensemble d'entraînement » (*train*), sert à l'apprentissage des paramètres et des modèles utilisés dans les systèmes automatiques ; la seconde, « ensemble de développement » (*dev*), donne une estimation des performances des systèmes automatiques lors de leur développement. À ce corpus est ajouté un corpus de test, distribué sans annotation ni transcription quelques temps avant la date de soumission des résultats de l'évaluation. Ce corpus permet de comparer les performances des participants de façon raisonnablement équitable. Des ressources annexes facilitent le travail des participants : une grande quantité de corpus de texte pour l'apprentissage de modèles de langage (10 ans du journal Le Monde, soit 400 millions de mots) et un corpus audio non transcrit pour les approches non supervisées (1700 heures de radio).

Les données, dont la répartition par source et par corpus est détaillée dans la table 3.1, sont en majorité des journaux radio diffusés et des émissions radiophoniques impliquant des invités et des interventions d'auditeurs. De nombreuses difficultés sont présentes dans les données et peuvent détériorer la qualité de l'annotation automatique :

- des locuteurs parlant une variante nord-africaine du français avec notamment des prononciations de noms propres en arabe ;
- des tours de parole dans une langue étrangère avec éventuellement un doublage ;
- de nombreuses difficultés d'élocution comme des hésitations, des coupures, des reprises, et des lapsus ;

| Source | Entr. | Dév. | Test | Non-Trans |
|---|---|---|---|---|
| France Inter | 33h03 | 2h00 | 2h00 | 346h24 |
| France Info | 8h01 | 2h00 | 2h00 | 660h10 |
| RFI | 23h00 | 2h00 | 1h59 | 457h14 |
| RTM | 18h28 | 1h58 | 2h04 | - |
| France Culture | - | - | 1h01 | 260h29 |
| Radio Classique | - | - | 1h00 | - |
| Total | 82h | 8h | 10h | 1724h |

**TAB. 3.1:** *Répartition des données de la campagne ESTER entre les corpus d'entraînement (Entr.), de développement (Dév.), de test (Test) et non transcrit (Non-Trans).*

– de la parole sur un fond musical, systématiquement au moment des titres du journal ;
– des bruits de fond, comme des éternuements et des jingles courts structurant les émissions ;
– des segments où plusieurs locuteurs parlent en même temps ;
– des coupures de fréquences dues à l'enregistrement.

Certaines de ces difficultés sont des cas particuliers trop peu nombreux pour que les systèmes soient capables de les traiter (par exemple, les tours de parole de locuteurs dans une langue étrangère). Ces difficultés sont annotées dans le corpus pour être ignorées lors de l'évaluation. Les spécificités du corpus de tests sont détaillées dans la table 3.2 ; la table 3.3 donne la répartition des conditions acoustiques et du genre des locuteurs sur les segments évalués du corpus de test. Un exemple de l'annotation de référence ESTER est décrit dans La figure 3.6.

| | |
|---|---|
| Durée | 10h07 |
| Nombre de mots | 103203 |
| Nombre de locuteurs | 343 |
| Transcriptions ignorées en éval. | 5.99% |
| Segmentations ignorées en éval. | 2.47% |
| Parole simultanée | 0.43% |
| Musique et bruits | 4.95% |

**TAB. 3.2:** *Spécificités du corpus de test ESTER.*

Les tâches de l'évaluation ESTER, décrites dans (Galliano et al., 2005), sont de 3 formes : transcription, segmentation et extraction d'information.

Parmi les tâches de transcription, la première est la transcription orthographique du contenu parlé sans limites de ressources, alors que la seconde est limitée en temps de calcul (le système doit transcrire le corpus dans un temps équivalent à sa durée). La mesure d'évaluation de la tâche de transcription est le taux d'erreur de mots (*Word Error Rate*, WER).

Une première tâche de segmentation est focalisée sur le suivi d'événements sonores sous la forme d'une détection des classes acoustiques suivantes : parole, musique et

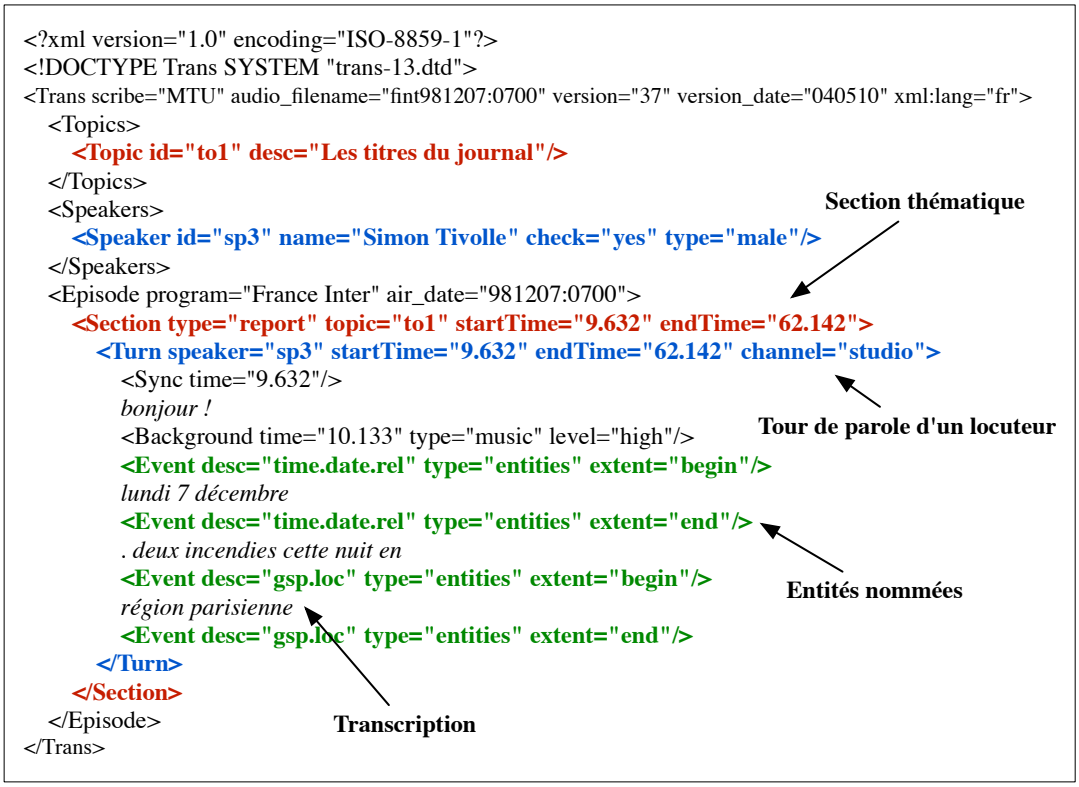



**FIG. 3.6:** *Exemple de structuration de référence des données ESTER : l'annotation est réalisée par un expert humain à l'aide du logiciel Transcriber et sauvegardée dans un format XML. Ce format inclut des définitions de Thèmes, de Locuteurs, puis des sections thématiques, des tours de parole, la transcription accompagnée d'événements sonores, de synchronisation et des débuts et fins d'entités nommées. L'annotation est synchronisée sur le flux audio en utilisant des attributs et des balises dédiés.*

parole sur musique. Puis la seconde tâche a pour but de segmenter le flux audio en tours de parole sur les changements de locuteurs, tout en identifiant les contributions consécutives d'un même locuteur. Enfin, une tâche d'identification de locuteurs connus parmi les locuteurs précédents complète la tâche de segmentation en locuteurs.

Une dernière tâche d'extraction d'information consiste en l'annotation des entités nommées (noms de personnes, lieux, organisations...). Cette tâche fait l'objet de la section 4.2.

### 3.2.2 Mesures d'évaluation

La segmentation en classes acoustiques est évaluée selon le nombre de frontières bien classées avec une tolérance de 0.25 seconde. La précision ($P$) est calculée comme le rapport entre le nombre de frontières correctes et le nombre de frontières de l'hypothèse ; et le rappel ($R$) comme le rapport entre le nombre de frontières correctes et le

| Conditions | Répartition (%) | WER |
|---|---|---|
| Studio | 61.1 | 23.5 |
| Téléphone | 4.2 | 24.8 |
| Fond musical | 7.6 | 23.2 |
| Acoustique dégradée | 3.6 | 38.3 |
| Locuteurs non natifs | 14.1 | 34.9 |
| Autre | 9.0 | 36.7 |
| Féminin | 27.9 | 22.9 |
| Masculin | 72.0 | 28.3 |
| Inconnu | 0.1 | 73.7 |
| Total | 100.0 | 26.7 |

**TAB. 3.3:** *Répartition des conditions acoustiques et genre des locuteurs dans le corpus de test ESTER. Les taux d'erreur de mots (WER) du système lors de la campagne sont donnés pour illustrer la dégradation des performances selon les conditions (des modèles adaptés sont utilisés en fonction des classes acoustiques détectées).*

nombre de frontières de la référence. La $F_\beta$-mesure est une moyenne harmonique du rappel et de la précision. La valeur de $\beta$ est fixée[2] en fonction de l'application pour donner plus de poids au rappel ou à la précision (équation 3.1).

$$\begin{aligned} P &= \frac{\text{nb}(correct)}{\text{nb}(hyp)} \\ R &= \frac{\text{nb}(correct)}{\text{nb}(ref)} \\ F_\beta &= \frac{(1+\beta^2)PR}{\beta^2 P + R} \end{aligned} \tag{3.1}$$

Le suivi de locuteur est évalué par une $F$-mesure similaire à celle utilisée pour la segmentation en classes acoustiques. Des mesures de performances alternatives comme les courbes *Detection Error Tradeoff* (DET) sont présentées dans les résultats officiels de la campagne (Galliano et al., 2005).

Le taux d'erreur de segmentation en locuteurs ($E_{seg}$) est détaillé par l'équation 3.2 dans laquelle : $dur(s)$ représente la durée du segment $s$ ; $nb_{ref}(s)$ est le nombre de locuteurs de la référence parlant dans $s$ ; $nb_{hyp}(s)$ est le nombre de locuteurs de l'hypothèse dans $s$ ; $nb_{correct}(s)$ est le nombre de nombre de locuteur de l'hypothèse parlant réellement dans $s$.

$$E_{seg} = \frac{\sum_s dur(s)\left(max(\text{nb}_{ref}(s), \text{nb}_{hyp}(s)) - \text{nb}_{correct}(s)\right)}{\sum_s dur(s)\,\text{nb}_{ref}(s)} \tag{3.2}$$

La qualité de la transcription est évaluée selon le taux d'erreur de mots (*Word Error Rate*, WER) explicité dans la formule 3.3.

[2]Dans ESTER, $\beta$=1.

$$WER = \frac{I + S + D}{R} \tag{3.3}$$

où $I$ est le nombre d'insertions, $S$ est le nombre de substitutions, $D$ est le nombre de suppressions et $R$ est le nombre de mots de la référence. Les types d'erreurs sont déterminés par alignement dynamique de la référence et de l'hypothèse.

### 3.2.3 Résultats du système LIA

Les performances de la chaîne de traitement du LIA sont détaillées dans la table 3.4. Les résultats sur chaque tâche sont comparés à la meilleure soumission lors de l'évaluation et aux performances correspondant aux améliorations des systèmes depuis l'évaluation. Il est à noter que des problèmes dans la détection du genre et la stratégie d'utilisation des classes acoustiques ont dégradé les performances en détection du locuteur (Istrate et al., 2005). De plus, le système de transcription a surtout été amélioré au niveau de la vitesse d'exécution (sa vitesse a été multipliée par 20 pour des performances identiques). Les différences entre le système LIA et le meilleur système sont principalement dues à la différence entre la quantité de données d'apprentissage utilisées, la taille de l'espace de recherche (65000 mots contre 200000 mots) et l'expérience dans le domaine de la transcription d'émissions radio.

| **Tâche** | **Perf.** | **Post.** | **Meill.** | **Unité** |
|---|---|---|---|---|
| Détection de parole | 99.2 | - | 99.2 | $f_1$-m |
| Détection de parole sur musique | 92.7 | - | 94.2 | $f_1$-m |
| Détection de musique | 54.8 | - | 54.8 | $f_1$-m |
| Segmentation en locuteurs | 19.2 | - | 11.5 | %err |
| Suivi de locuteurs | 66.0 | 75.5 | 84.3 | $f_1$-m |
| Transcription | 26.7 | 22.7 | 11.9 | WER |
| Transcription (temps limité) | 36.3 | - | 16.8 | WER |

**TAB. 3.4:** *Résultats du LIA sur les différentes tâches d'ESTER phase II (Perf.), post-évaluation (Post.) et du meilleur participant lors de l'évaluation (Meill.) selon l'unité correspondante : le taux d'erreur de mots (WER) pour la transcription, la $f_1$-mesure et le taux d'erreur de segmentation (%err) pour les tâches de segmentation.*

Une analyse des erreurs de transcription fait ressortir qu'une grande quantité d'entre elles provient de noms propres mal reconnus et d'homophonies dont la résolution est hors de portée des modèles tri-grammes. Des exemples extraits de la soumission illustrent cette observation :

*Réf. :* les grands titres de l' actualité **** **Maude Bailleux** bonjour
*Hyp. :* les grands titres de l' actualité **émaux de Bayeux** bonjour

*Réf. :* Nicolas **** **Pierron signait** la troisième édition de **Classique** Matin
*Hyp. :* Nicolas **Pierre ont signé** la troisième édition de **Classiques** Matin

L'amélioration des systèmes de structuration audio passe d'abord par une augmentation de la quantité de données d'apprentissage afin de couvrir un maximum d'événements mais aussi pour mieux estimer les paramètres des algorithmes statistiques employés. L'intégration des objectifs de la tâche finale et une analyse fine des erreurs donneront les voies de recherche à privilégier.

## 3.3 Conclusion

L'extraction de descripteurs sémantiques est indispensable pour la construction de résumés parlés. Nous avons présenté dans ce chapitre les différentes tâches impliquées dans une chaîne de structuration des données acoustiques. Les méthodes les plus répandues pour la résolution de ces tâches sont fondées sur l'apprentissage artificiel, formulées comme des problèmes de segmentation et d'identification. Leurs performances sont directement liées à la quantité de données d'apprentissage et à l'adéquation de ces dernières aux conditions d'utilisation. La chaîne de structuration présentée, développée au LIA, est une concrétisation de ces différentes tâches. Ses performances, validées sur la campagne ESTER sont suffisantes pour envisager l'emploi des descripteurs sémantiques et structuraux ainsi extraits dans une méthode de résumé automatique de parole. Toutefois, des éléments complémentaires à cette chaîne et indispensables pour le résumé sont présentés dans le chapitre suivant.

# Chapitre 4

# Compléments à l'extraction de descripteurs structurels et sémantiques

**Sommaire**



Le chapitre 2 a présenté les nombreuses méthodes de recherche d'information parlée et il s'avère que ce type d'information nécessite une extraction spécifique de descripteurs sémantiques à partir de l'acoustique. Par la suite, le chapitre 3 a introduit les différentes tâches de structuration et leur mise en œuvre dans la chaîne de structuration Speeral. L'objectif de ces travaux est de faciliter l'accès à l'information audio à l'aide du résumé de parole et les éléments de structuration présentés au chapitre précédent ne sont pas suffisants pour obtenir un résumé de qualité. Nous nous concentrons maintenant sur la présentation de deux compléments à la structuration pour le résumé automatique de parole. Tout d'abord, une segmentation en phrases de qualité est nécessaire pour résumer la parole avec une approche par extraction. En effet, du point de vue de l'utilisateur, cet aspect de la forme d'un résumé parlé est déterminant car une coupure inopportune au milieu d'une phrase peut fortement dégrader la compréhension. La méthode proposée pour la segmentation en phrases s'appuie sur un étiquetage

de séquence dans le cadre des *Conditional Random Fields* (section 4.1). Le second point de contribution réside dans l'extraction d'entités nommées dans le flux de parole. Ces entités liées au domaine (personnes, organisations, lieux...) dirigent la projection dans l'espace sémantique lors de la génération du résumé. L'approche développée pour cette tâche consiste en une recherche des entités nommées dans l'ensemble des hypothèses de transcription au lieu d'être restreinte à la meilleure hypothèse (section 4.2).

## 4.1 Segmentation en phrases par étiquetage de séquence

Il a été remarqué dans la section 2.2 que la segmentation en phrases demandait une attention particulière dans le cadre du résumé de parole par extraction (Rappelons que Mrozinski et al. (2006) ont observé une forte réduction de la qualité des résumés de parole fondés sur une segmentation automatique par rapport à une segmentation manuelle).

Dans la littérature, le problème de segmentation en phrases est généralement reformulé en un problème d'identification de frontières de phrases (étiquetage de séquence). La transcription automatique est employée pour générer une suite de mots et des frontières (événement binaire $B$) sont recherchées entre les mots. La décision est généralement issue d'une combinaison de paramètres prosodiques (événement $S$) et linguistiques (événement $L$). Trouver des frontières de phrases est loin d'être facile, en attestent par exemple Stevenson et Gaizauskas (2000), qui évaluent les performances d'annotateurs humains sur la reponctuation d'un texte, et qui observent qu'il est beaucoup plus facile de reponctuer un flux de mots contenant les majuscules d'origine ($F_1$-mesure de 0.95) qu'en l'absence de ces marqueurs ($F_1$-mesure de 0.80), comme dans le cas d'une transcription automatique.

La majorité des approches sont fondées sur des modèles probabilistes tentant de prédire la séquence $B$ en fonction de $S$ et $L$. Gotoh et Renals (2000) constituent un modèle pour chacune des modalités ($S$ et $L$) sur des ensembles de données séparés. La probabilité linguistique $P(B, L)$ qu'une frontière de phrase précède un mot est modélisée à partir de données textuelles disponibles en masse ; l'implication de la prosodie $P(B, S)$ est modélisée à partir des durées de pauses sur un corpus acoustique de plus petite taille. Les deux modèles sont fusionnés grâce à une heuristique[1]. Shriberg et al. (2000) étudient les différentes caractéristiques prosodiques en profondeur : les pauses, le rythme phonétique ou syllabique, la pente de fréquence fondamentale ($f_0$) et sa continuité, les sauts de $f_0$, l'écart à la moyenne de la $f_0$, et la qualité de voix. Les valeurs sont fonction du locuteur ou d'un locuteur moyen lorsque les données sont insuffisantes. En plus de ces paramètres, la décision repose sur la durée des phrases et les changements de locuteurs (segmentation manuelle en locuteurs). Un arbre de décision donne une sélection des paramètres les plus pertinents et ces derniers servent à construire un modèle de séquence génératif. Les paramètres les plus efficaces sur des données radio-diffusées semblent être les pauses et les changements de locuteurs. Liu et al. (2005) continuent ces

[1] $P(B, L, S) \simeq P(B, S)^{\alpha} P(B, L)$, $\alpha > 10$ donnant les meilleurs résultats.

travaux en comparant des approches HMM, maximum d'entropie et CRF pour l'étiquetage de la séquence : ce dernier modèle s'avérant être le plus efficace (une fusion des trois apporte un gain complémentaire). Il est intéressant de noter que la décision prosodique sur la frontière est prise avant l'inclusion dans le modèle de séquence. Des travaux similaires de (Kim et al., 2004) intègrent des arbres de décision avec un système de détection de difficultés de prononciation.

La tâche de détection de frontières de phrase (en anglais, *Sentence Unit Boundary Detection*, SUBD) a été évaluée lors des éditions 2002 à 2004 des campagnes *Rich Transcription* « automne » (RT-fall), organisées par NIST. Les données de référence reposent sur un guide d'annotation (Strassel, 2003)[2] précisant que la notion de phrase à l'oral (nommée « unité syntagmatique ») est différente de l'écrit. Les différences sont avant tout grammaticales ; les unités sont classées selon leur type (déclarations, questions, éléments phatiques et unités incomplètes). La mesure de performance NIST est le taux d'erreur sur les frontières (nombre de frontières oubliées, ajoutées ou de mauvais type, divisé par le nombre de frontières dans la référence : équation 4.1 dans laquelle $\mathrm{nb}(\cdot)$ est le cardinal d'un ensemble de frontières).

$$SB_{err} = \frac{\mathrm{nb}(\text{oubli}) + \mathrm{nb}(\text{ajout}) + \mathrm{nb}(\text{mauvais type})}{\mathrm{nb}(\text{référence})} \tag{4.1}$$

Sur des données radio-diffusées, Liu et al. (2005) aboutissent à un taux d'erreur de 0.54 (sans prendre en compte les erreurs de type). Cette valeur correspond à une $F_1$-mesure d'environ 0.70, proche des performances annoncées par les autres auteurs.

La détection de frontières de phrases que nous avons mise en place pour le résumé de parole est similaire à l'approche de Liu et al. (2005). En restant dans le cadre de l'étiquetage bi-classe de la séquence de mots, nous appliquons un modèle CRF sur des caractéristiques prosodiques et linguistiques. Ces dernières sont issues de la chaîne de structuration Speeral. Les frontières de phrases sont recherchées dans les émissions de radio en français de la campagne ESTER.

### 4.1.1 Conditional Random Fields

#### Définition

*Conditional Random Fields* (CRF, Lafferty et al., 2001) est un cadre probabiliste discriminant pour l'étiquetage de séquences. Au lieu de modéliser la probabilité jointe d'apparition des séquences d'observation et des séquences d'étiquettes comme le fait une approche générative telle que HMM, CRF repose sur la probabilité conditionnelle de l'étiquetage sachant l'ensemble de la séquence. Les méthodes à maximum d'entropie de Markov (MEMM) recherchent aussi à maximiser cette probabilité conditionnelle, mais de façon locale. Ceci pose des problèmes au niveau des hypothèses partielles débouchant sur un petit nombre de successeurs car ils sont systématiquement préférés

[2] disponible en ligne sur http://projects.ldc.upenn.edu/MDE/Guidelines/SimpleMDE_V5.0.pdf, visité en novembre 2006

aux chemins de plus grande entropie. Cet effet est décrit sous le nom d'effet du biais des étiquettes par Lafferty et al. (2001).

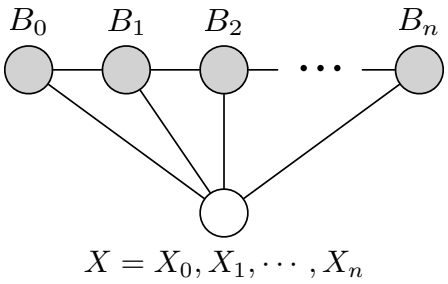



**FIG. 4.1:** *Détection de frontières de phrases par modélisation CRF. La séquence d'événements représentant la présence ou l'absence de frontières ($B = B_0, \cdots, B_n$) est globalement conditionnée par la séquence d'observations phonétiques et linguistiques ($X = X_0, \cdots, X_n$).*

Appliquons CRF à une tâche de segmentation en phrases : $B$ est une séquence d'étiquettes ($B = 1$ pour une frontière de phrase, $B = 0$ pour une absence de frontière) ; $X$ est une séquence d'observations prosodiques et linguistiques. Le modèle conditionne la séquence $B$ sur l'ensemble de la séquence $X$ (figure 4.1). La meilleure hypothèse d'étiquetage est celle qui maximise la probabilité $P(B|X)$. Cette probabilité est estimée par une distribution de forme exponentielle satisfaisant des caractéristiques sur des données d'apprentissage (équation 4.2).

$$\begin{aligned} \hat{B} &\operatorname*{argmax}_B P(B|X) \\ P(B|X) &\simeq \frac{1}{Z(X)} e^{\sum_k \lambda_k f_k(B,X)} \\ Z(X) &= \sum_B e^{\sum_k \lambda_k f_k(B,X)} \\ f_k(B,X) &\geq 0 \end{aligned} \tag{4.2}$$

Dans cette équation, les $\lambda_k$ sont les paramètres du modèle ; $Z(X)$ sert à la normalisation de la distribution ; $f_k(B,X)$ sont les fonctions caractéristiques sur les arcs et les sommets du modèle graphique associé au problème. Ces fonctions sont des relations entres les $B_i$ et $X$ et entre des $B_i$ voisins.

L'inférence des paramètres $\lambda_i$ se fait par maximisation de la vraisemblance conditionnelle sur un ensemble de données étiquetées. Le maximum de cette fonction log-concave est découvert par des méthodes de maximisation classiques, comme *Generalized Iterative Scaling* (GIS, Darroch et Ratcliff, 1972), *Improved Iterative Scaling* (IIS, Della Pietra et al., 1997), ou *Limited-memory Broyden-Fletcher-Goldfarb-Shanno* (LBFGS, Liu et Nocedal, 1989), qui s'avère être la plus rapide. Ces méthodes sont comparées dans (Malouf, 2002). La dépendance des étiquettes sur l'ensemble de la séquence d'observation rend l'apprentissage beaucoup plus coûteux que pour un maximum d'entropie local classique. L'étiquetage d'une séquence nouvelle se fait par programmation dynamique.

#### La boîte à outils CRF++

L'ensemble de nos expériences sur la détection de frontières de phrases repose sur CRF++[3], une boîte à outils pour l'étiquetage de séquences fondée sur CRF. CRF++ implémente un apprentissage dont l'optimisation repose sur une méthode de quasi-Newton (LBFGS) et un décodage grâce à l'algorithme Viterbi. Cette boîte à outils a été utilisée avec succès pour de nombreuses tâches de traitement automatique du langage naturel comme la désambiguïsation sémantique, la décomposition en groupes grammaticaux, l'étiquetage morpho-syntaxique ou encore la détection d'entités nommées (Kudo et al., 2004).

### 4.1.2 Traits acoustiques et linguistiques

Nous suivons les approches classiques pour la segmentation en phrases en recherchant des frontières potentielles uniquement entre les mots et en fixant l'événement $B = 1$ si une frontière a précédé un mot et $B = 0$ dans le cas contraire. La prédiction de la présence d'une frontière de phrase avant un mot dépend de caractéristiques linguistiques et acoustiques que nous allons décrire (voir table 4.1). Au niveau linguistique, les mots et leurs catégories morpho-syntaxiques modélisent les phénomènes grammaticaux de la séquence. La catégorie morpho-syntaxique des mots est trouvée grâce à lia_tagg[4]. Cet étiqueteur repose sur un dictionnaire d'étiquettes possibles par mots et effectue l'étiquetage dans un cadre HMM. Alors que certains couples syntaxiques, comme « le déterminant et le nom », qui ne doivent pas être séparés par une frontière de phrase, sont plutôt bien capturés par cette modélisation, d'autres groupes comme « le verbe et son complément » sont moins faciles à détecter sans une modélisation plus approfondie de la grammaire. Si les éléments linguistiques sont utiles pour reponctuer un texte, ils peuvent être faussés par les erreurs de transcription, d'étiquetage morpho-syntaxique et le manque relatif de grammaire de la langue parlée. Pour y remédier, il faut associer des caractères acoustiques aux indices linguistiques, comme les changements de locuteur et quelques éléments de prosodie. Les changements de locuteurs sont issus, comme la séquence de mots, de la chaîne de transcription et employés tels quels sans prendre en compte les identités retrouvées. En terme de prosodie, les pauses sont explorées à deux niveaux : avant le mot et à l'intérieur du mot pour essayer d'éviter de prendre les hésitations pour des fins de phrase. De plus, comme il est difficile de profiter des informations apportées par la courbe de fréquence fondamentale ($f0$), nous utilisons seulement sa pente globale, sur trois horizons temporels différents (le mot, une fenêtre allant de 4 secondes avant le début du mot jusqu'à sa fin et une fenêtre allant de 8 secondes avant le début du mot jusqu'à sa fin). Bien que cette approche ne soit pas optimale, elle permet tout de même de modéliser les grands phénomènes macro-prosodiques de la phrase. Toutefois, certaines caractéristiques sont perdues, comme les effets du rythme prosodique ou syllabique connus pour ralentir en fin de phrase. Les

[3]Disponible sur http://chasen.org/~taku/software/CRF++, visité en août 2006.

[4]Étiqueteur morpho-syntaxiaque du LIA, disponible sous licence GPL, sur http://www.univ-avignon.fr/chercheurs/bechet/download_fred.html, visité en octobre 2006.

frontières des mots de la référence sont extraites grâce à un alignement forcé sur le signal en utilisant un outil dérivé du système de transcription (gvalign).

| Mot | Étiquette | P1 | P2 | Loc. | F1 | F2 | F3 | Ponct. |
|---|---|---|---|---|---|---|---|---|
| avait | V3S | C0 | C0 | SPK | C0 | C0 | C0 | point |
| le | DETMS | C0 | C0 | n | C0 | C0 | C0 | x |
| salut | NMS | C0 | C4 | n | C4 | C0 | C0 | x |
| à | XSOC | C0 | C0 | n | C0 | C8 | C3 | x |
| tous | AINDMP | C0 | C0 | n | C4 | C0 | C0 | x |
| ceux | PDEMMP | C0 | C0 | n | C5 | C4 | C0 | x |
| en | PREP | C0 | C0 | n | C5 | C5 | C0 | x |
| bonne | AFS | C0 | C0 | n | C2 | C4 | C0 | point |
| journée | NFS | C0 | C0 | n | C5 | C3 | C0 | x |
| euh | ADV | C0 | C0 | n | C4 | C8 | C1 | x |

**TAB. 4.1:** *Exemple des paramètres extraits pour la segmentation en phrases. Au niveau linguistique : le mot de la transcription et son étiquette morpho-syntaxique. Au niveau prosodique : la durée de pause avant le mot (P1) et à l'intérieur du mot (P2), un éventuel changement de locuteur avant le mot (Loc.), la pente de F0 à divers horizons temporels (F1=le mot, F2=-4s, F3=-8s). La ponctuation qui précède le mot est prédite grâce à ces paramètres (Ponct.). Les valeurs numériques sont quantifiées uniformément selon les classes C0 à C9 (sur une fenêtre glissante de 300 valeurs, avec un jeu de classes par paramètre).*

CRF++ facilite la génération des fonctions caractéristiques en utilisant des patrons de conjonction d'événements de $X$ et $B$. Dans notre implémentation, une frontière de phrase potentielle est conditionnée par des séquences $n$-grammes de chaque type de caractères linguistiques et acoustiques autour du mot à étiqueter et par la conjonction des séquences précédentes (illustrées par la figure 4.2). La boîte à outils est cependant limitée dans sa version actuelle à des caractéristiques symboliques. Cette limitation implique la quantification des valeurs continues comme la durée des pauses ou la pente de fréquence fondamentale. La quantification se fait sur une fenêtre glissante en utilisant une répartition uniforme en $n$ classes[5]. Cette approche permet de normaliser les valeurs lors de changements de locuteurs et d'environnement.

### 4.1.3 Performances

Les performances en segmentation en phrases sont calculées sur la base du nombre de frontières bien placées par rapport au nombre de frontières erronées, en rappel, précision, et $f$-mesure (un exemple est donné par la table 4.2). Les expériences sont réalisées sur le corpus ESTER qui n'a malheureusement pas fait l'objet de directives d'annotation pour les frontières de phrases. Le guide d'annotation précise que « la ponctuation est facultative, mais peut être utilisée pour faciliter la tâche de transcription ». Cette dernière varie donc beaucoup d'un annotateur à l'autre ; les phrases peuvent être

[5] $n = 10$ dans les expériences qui suivent. La fenêtre glissante fait 300 valeurs. Ces valeurs sont fixées empiriquement, mais ne semblent pas avoir un impact important sur les performances.

| Mot | Étiquette | P1 | P2 | Loc. | F1 | F2 | F3 | Ponct. |
|---|---|---|---|---|---|---|---|---|
| avait | V3S | C0 | C0 | SPK | C0 | C0 | C0 | point |
| le | DETMS | C0 | C0 | n | C0 | C0 | C0 | x |
| salut | NMS | C0 | C4 | n | C4 | C0 | C0 | x |
| à | XSOC | C0 | C0 | n | C0 | C8 | C3 | x |
| tous | AINDMP | C0 | C0 | n | C4 | C0 | C0 | x |
| ceux | PDEMMP | C0 | C0 | n | C5 | C4 | C0 | x |
| en | PREP | C0 | C0 | n | C5 | C5 | C0 | x |
| bonne | AFS | C0 | C0 | n | C2 | C4 | C0 | point |
| journée | NFS | C0 | C0 | n | C5 | C3 | C0 | x |
| euh | ADV | C0 | C0 | n | C4 | C8 | C1 | x |

**FIG. 4.2:** *Illustration des groupes de paramètres utilisés pour la prédiction de la présence ou absence d'une frontière de phrase. En plus de ces événements, le modèle prend en compte les unigrammes dans une fenêtre de deux mots autour du mot courant et la conjonction de chacun des événements précédents sur l'ensemble de la séquence. Les données sont celles de la figure 4.1*

très longues, jusqu'à faire un tour de parole complet, contenant un grand nombre de virgules, alors que dans d'autres cas, chaque pause du locuteur a été annotée par une fin de phrase. Ce problème de fiabilité du corpus implique une nécessaire prudence dans l'interprétation des résultats d'évaluation.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Référence | * | * | * | p | * | p | * | * | * | * | * | p |
| Hypothèse | * | p | * | * | * | p | * | * | * | p | * | p |

**TAB. 4.2:** *Performances de la segmentation en phrases pour un exemple fictif. « p » représente une frontière de phrase et « * » une absence de frontière. Il y a 3 frontières à trouver dans la référence, 4 frontières ont été trouvées dans l'hypothèse, dont 2 bien placées. La précision est de $P = 2/4 = 0.5$, le rappel est de $R = 2/3 = 0.66$ et la $F_1$-mesure est de $F_1 = 2 * PR/(P + R) = 0.57$. Le taux d'erreur NIST est égal au nombre d'erreurs ($hyp_i \neq ref_i$) par rapport au nombre de frontères à trouver : $SB_{err} = 3/3 = 100\%$.*

Il est intéressant de noter que les journalistes des radios francophones du corpus ont tendance à utiliser une architecture prosodique très spéciale qui détériore la cohérence des événements caractérisant une frontière de phrase. En effet, pour captiver l'attention de l'auditeur, les journalistes reprennent leur souffle en milieu de phrase, pour provoquer un effet « d'attente ». Cet effet diminue la cohérence de l'annotation par l'insertion de pauses. Ces pauses ont les caractéristiques acoustiques d'une fin de phrase et les caractéristiques linguistiques d'un milieu de phrase.

#### Comparatif en structuration automatique et manuelle

Les données d'entraînement utilisées dans ces expériences correspondent aux 80 heures d'entraînement (environ 874000 mots) du corpus ESTER, alors que les performances sont rapportées pour les 10 heures de la partie développement du corpus (en-

| Données | Rappel | Précision | $F_1$-mesure |
|---|---|---|---|
| *Étiquetage : points* | | | |
| M+M | 0.42 | 0.80 | 0.55 |
| M+A | 0.34 | 0.84 | 0.49 |
| A+M | 0.62 | 0.74 | 0.67 |
| A+A | 0.61 | 0.77 | 0.68 |
| *Étiquetage : points et virgules* | | | |
| M+M | 0.41 | 0.64 | 0.50 |
| M+A | 0.30 | 0.72 | 0.42 |
| A+M | 0.50 | 0.65 | 0.56 |
| A+A | 0.49 | 0.74 | 0.59 |
| *Étiquetage : points et virgules fusionnés* | | | |
| M+M | 0.55 | 0.78 | 0.64 |
| M+A | 0.37 | 0.81 | 0.51 |
| A+M | 0.59 | 0.70 | 0.64 |
| A+A | 0.55 | 0.81 | 0.66 |

**TAB. 4.3:** *Performances de la segmentation en phrases selon le type d'étiquetage recherché et les données utilisées en apprentissage et en test. Par exemple, « M+A » représente un apprentissage sur les données extraites à la main (M) et un test sur les données transcrites et segmentées automatiquement (A).*

viron 88000 mots). La table 4.3 présente des comparatifs entre l'utilisation de données structurées automatiquement ou manuellement en apprentissage et en test, pour les tâches d'étiquetage sur les points («.») comme frontières de phrases, les points et les virgules sous forme d'un problème 3-classes («.», «,» et ∅) et la fusion des points et des virgules («,»=«.»).

Globalement, les tests sur les données structurées manuellement montrent que l'approche admet un faible rappel et une forte précision sur les frontières retrouvées. La différence est moins prononcée lors de l'utilisation de données structurées automatiquement lors du test. De plus, la méthode la plus performante consiste en l'utilisation de données structurées automatiquement en apprentissage et en test. En revanche, les données de référence mènent à de moins bonnes performances générales. Il semblerait que ceci soit dû à une différence dans la notion de pause entre l'algorithme d'alignement automatique et le système de transcription. Pour ce qui est des différents étiquetages possibles, étant donné qu'aucun guide d'annotation en frontières de phrases n'a été fourni lors de la création des données de référence, nous avons essayé de réduire les incohérences virgule-point en fusionnant ces 2 types de frontières et en les annotant séparément. Les performances ne sont néanmoins jamais au niveau de celles obtenues par l'annotation des « points ».

Finalement, nous déduisons de ces résultats que l'approche permet d'établir des performances de l'ordre de ce qui est donné dans la littérature (une $f_1$-mesure d'environ 0.70). De plus, il est bon de noter que la méthode a une bonne précision et une tendance à sous-générer les frontières de phrases. Ce type de comportement est béné-

fique pour le résumé automatique car le type d'erreur le plus pénalisant dans ce cadre reste l'insertion de frontières de phrases là où elles n'ont pas lieu d'être.

### 4.1.4 Améliorations envisagées

Nous avons proposé une détection des frontières de phrases par étiquetage d'une séquence d'« inter-mots » à l'aide de CRF. L'approche peut être améliorée en utilisant des caractéristiques continues (et non symboliques) — en prenant en compte les scores de confiance du système de transcription — et en calculant une courbe de $f0$ plus fine, normalisée pour chaque locuteur. Une des limitations de CRF est que cette approche ne peut tenir compte de paramètres au niveau global de la phrase, comme sa longueur ou sa cohérence syntaxique et sémantique. Une solution à ce problème peut être semi-CRF (Sarawagi et Cohen, 2005) qui tente de remettre en cause de l'hypothèse Markovienne[6] du processus sur un segment temporel de taille raisonnable de l'ordre de la phrase. D'autres pistes doivent être envisagées, comme un test sur la fiabilité du corpus afin de détecter et d'écarter les phrases mal annotées, ou une intégration complète de la segmentation en phrases dans la transcription du contenu parlé pour retarder la prise de décision sur les frontières de mots.

## 4.2 Extraction d'entités nommées dans le flux de parole

Les entités nommées sont des entités du monde « réel », dont la forme linguistique est une représentation directe dénuée d'ambiguïté. Notamment, lorsqu'une de ces entités se retrouve dans le discours de plusieurs personnes, il est considéré que ces différentes références ont le même antécédent. Bien que cette affirmation soit loin d'être vraie dans le cas général, les types d'entités recherchés doivent s'en approcher le plus possible. Par exemple, « une table » est un concept qui se réfère à un objet dans un contexte donné. Dans un autre contexte, le locuteur se référera généralement à une autre entité. En revanche, dans un domaine journalistique, les noms propres se réfèrent à des objets considérés comme uniques, dont la forme linguistique peut être séparée de son contexte sans rendre la référence ambiguë. Ce type de comportement est très intéressant dans le cadre de l'analyse sémantique indispensable pour le résumé car la projection depuis la linguistique devient transparente.

Dans le cadre de l'extraction de descripteurs sémantique de journaux radio diffusés, les entité nommées sont étendues à certaines quantités fortement porteuses d'information dans ce domaine. Les entités recherchées sont de deux types : entités uniques basées sur des noms propres (personnes, lieux, organisations...) et entités basées sur des séquences de noms communs (dates, quantités monétaires, distances...). Les majuscules des noms propres sont de bons indicateurs de la présence d'entités du premier type et les valeurs numériques sont de bons indicateurs de la présence du second type d'entité.

[6] L'hypothèse Markovienne est vérifiée pour un processus si et seulement si la distribution conditionnelle de probabilité des états futurs, étant donné l'instant présent, ne dépend que de ce même état présent et pas des états passés.

### 4.2.1 Introduction

Les deux approches majeures pour l'extraction d'entités nommées sont la création de grammaires spécifiques au domaine, à base de règles et de listes de mots, et l'étiquetage de séquence par apprentissage. Ce dernier est le plus efficace lorsque la séquence à étiqueter est bruitée et que des données d'apprentissage sont disponibles. Dans ce cadre, la séquence d'étiquettes est mise en correspondance avec la séquence de mots (les observations) en utilisant le formalisme *Begin Inside Outside* (BIO). Lorsqu'une étiquette s'étale sur plusieurs mots, elle est subdivisée en une méta-étiquette par mot, selon la position du mot dans l'étiquette. La méta-étiquette *Begin* (B) correspond à un mot en début d'entité, la méta-étiquette *Inside* (I) correspond à un mot en milieu ou fin d'entité et la méta-étiquette *Outside* (O) correspond à un mot à l'extérieur d'une entité. La table 4.4 donne un exemple de correspondance BIO. Celle-ci permet de s'affranchir du problème de segmentation et de se concentrer sur le problème d'étiquetage.

| ← **org** → | | | ← **pers** → | | | | ← **time** → | |
|---|---|---|---|---|---|---|---|---|
| B-org | I-org | O | B-pers | I-pers | O | O | B-time | I-time |
| *France* | *Inter* | bonjour | *Nicolas* | *Stoufflet* | il | est | *sept* | *heures* |

**TAB. 4.4:** *Illustration de la correspondance Begin Inside Outside (BIO) pour transformer le problème d'étiquetage en entités nommées en un problème de classification de séquence.Begin (B) correspond à un mot en début d'entité, Inside (I) à un mot au milieu ou en fin d'entité et Outside (O) à un mot à l'extérieur d'une entité.*

Dans un cadre probabiliste, le problème est formalisé de la façon suivante : pour une séquence de mots $W = w_0, \cdots, w_n$ donnée, la séquence d'étiquettes $L = l_0, \cdots, l_n$ la plus probable est recherchée (elle maximise la probabilité $P(L|W)$, équation 4.3).

$$\hat{L} = \underset{L}{\operatorname{argmax}}\, P(L|W) \tag{4.3}$$

La probabilité d'une séquence peut être exprimée en factorisant la probabilité à posteriori et sera approximée en ne prenant en compte qu'un contexte fixe pour déterminer l'étiquette associée à un mot (à travers $p(l_n|ctx_n)$, équation 4.4).

$$\begin{aligned} P(L|W) &= \prod_n P(l_n|l_0, \cdots, l_{n-1}, w_0, \cdots, w_n) \\ P(L|W) &\simeq \prod_n p(l_n|ctx_n) \end{aligned} \tag{4.4}$$

Bender et al. (2003) et Miller et al. (2000) appliquent une méthode à base de maximum d'entropie pour estimer $P(L|W)$. Dans ce cadre proposé par Berger et al. (1996)[7] pour le traitement de la langue naturelle, la probabilité d'une étiquette en contexte est exprimée sous forme d'une distribution satisfaisant des contraintes sur les données

[7]Une boîte à outils est disponible sur `http://maxent.sourceforge.net/`, visité en novembre 2006.

d'apprentissage $f(\cdot)$ (équation 4.5).

$$p(l_n|ctx_n) = \frac{1}{Z(ctx_n)} \prod_f e^{\lambda_f f(ctx_n, l_n)} \tag{4.5}$$
$$Z(ctx_n) = \sum_l \prod_f e^{\lambda_f f(ctx_n, l)}$$

Dans cette formulation, les $f(\cdot)$ sont des fonctions binaires de la présence de caractéristiques. $\lambda_f$ est le paramètre associé à chaque caractéristique. $Z(\cdot)$ est un facteur de normalisation dépendant uniquement du contexte. Le contexte est généralement constitué de mots autour du mot courant et des $m$ étiquettes attribuées aux mots précédents. Les caractéristiques sont fondées sur des propriétés des mots pour assurer la généralisation du procédé, comme la casse (capitalisé, majuscule, minuscule), la présence de caractères spécifiques (chiffres, ponctuations, tirets, virgule, dollar), la catégorie morpho-syntaxique (adjectif, nom, verbe) et l'appartenance à des listes de mots (prénom, ville, société...).

Les paramètres $\lambda_f$ sont appris sur un corpus d'entraînement étiqueté manuellement par maximisation de la vraisemblance conditionnelle des modèles produits en utilisant l'algorithme *Global Iterative Scaling* (GIS, Darroch et Ratcliff, 1972) ou des algorithmes d'optimisation convexe. L'algorithme Viterbi est utilisé pour déterminer la séquence la plus probable sachant que la fonction de normalisation $Z(\cdot)$ peut être ignorée (car elle est indépendante de l'étiquette). Pour éviter le sur-apprentissage, les paramètres des modèles sont contraints par une distribution *a priori* de type gaussienne (Chen, 1999). Chieu et Ng (2002) proposent d'ajouter des statistiques globales à cet étiquetage local afin de prendre en compte les affinités entre les entités et leur contexte d'utilisation. Collins et Singer (1999) appliquent une technique de co-apprentissage pour trouver les attributs efficaces d'extraction d'entités nommées sur un corpus partiellement étiqueté. Le principe est d'utiliser deux méthodes d'étiquetage faiblement couplées ; les étiquettes trouvées par la première méthode associées à une bonne confiance, sont utilisées comme référence pour la seconde méthode, puis le processus est itéré alternativement sur chaque méthode jusqu'à l'étiquetage complet du corpus. D'autres algorithmes obtiennent d'excellentes performances sur cette tâche, comme CRF (McCallum et Li, 2003), ou SVM (Kazama et al., 2002), en utilisant une formulation similaire du problème.

Les performances de ces approches sont directement liées à la quantité de données d'apprentissage disponibles. Haghighi et Klein (2006) n'utilisent pas de données d'apprentissage, mais un corpus non étiqueté et un petit nombre d'exemples d'étiquetages. La distribution globale des voisinages des exemples connus permet d'inférer l'étiquetage du reste du corpus. Cette méthode obtient environ 80% des performances d'une méthode avec des données complètement étiquetées (sur une tâche d'étiquetage morpho-syntaxique).

Contrairement à la recherche d'information sur un contenu parlé qui profite de la redondance des documents, l'annotation en entités nommées est beaucoup plus sensible aux erreurs de transcription. Par exemple, Kubala et al. (1998) étudient l'application

d'Identifinder (Bikel et al., 1997) sur un corpus de parole journalistique et notent que le taux d'erreur de mots de la transcription a un effet direct sur celui de l'annotation, les noms propres étant les plus touchés. Une première idée afin de prendre en compte les erreurs de transcriptions est d'adapter l'algorithme de détection pour qu'il autorise des insertions et des délétions (les substitutions sont des délétions suivies d'insertions, Grishman, 1998; Gotoh et Renals, 1999). Une autre proposition est de travailler directement sur les graphes d'hypothèses du système de transcription de la parole (Horlock et King, 2003; Béchet et al., 2004) et ainsi retrouver des entités bien formées parmi les hypothèses les plus probables. Enfin, Acero et al. (2004) adaptent le modèle de langage du système de transcription à la tâche finale en utilisant une grammaire probabiliste. Cette approche a le désavantage de nécessiter beaucoup de données d'apprentissage et ne peut s'appliquer qu'à des cas très particuliers (Wang et Acero, 2003). Pour éviter que les noms propres soient considérés comme des mots inconnus, Allauzen (2003) met à jour le lexique du système de transcription avec des mots de corpus textuels similaires au corpus de parole annoté.

L'évaluation de l'étiquetage en entités nommées est réalisé en utilisant principalement deux mesures : la $F_\beta$-mesure et le Slot Error Rate. La $F_\beta$-mesure est composée du rappel $R$ (nombre d'entités justes par rapport au nombre d'entités à trouver) et de la précision $P$ (nombre d'entités justes par rapport au nombre d'entités trouvées). La $F$-mesure est présentée en 3.2.2. Le *Slot Error Rate* (SER) se veut plus précis car il caractérise mieux les erreurs et les pénalise de façon plus fine. La formulation du SER est donnée par l'équation 4.6, dans laquelle $I$ est le nombre d'insertions, $D$ le nombre de délétions, $S$ le nombre de substitutions, $R$ le nombre d'entités de la référence et $\alpha_i$ le poids associé chaque type d'erreur.

$$SER = \sum_{e \in \{I,D,S\}} \frac{\alpha_e e}{Nb_{Ref}} \tag{4.6}$$

SER autorise une pondération des substitutions en fonction de leur origine : erreur sur le type de l'entité, son contenu, ou sa portée. La réalisation manuelle de références (et donc de corpus d'apprentissage) pose de nombreux problèmes car les classes peuvent être redondantes et l'annotation ambiguë (à cause d'une interprétation différente des règles d'annotation par les annotateurs). Les phénomènes d'anaphore et de métonymie entrent en jeu et bien définir les règles d'annotation prend une grande importance afin d'en conserver la cohérence. Ces problèmes sont abordés dans la convention d'annotation des entités nommées de la campagne ESTER (Le Meur et al., 2004).

### 4.2.2 Coopération avec le processus de transcription

La reconnaissance d'entités nommées dans un flux de parole en utilisant un système de transcription de parole est intrinsèquement limitée par le vocabulaire que ce dernier peut reconnaître et par les erreurs qu'il peut commettre en le générant. Plus précisément, les entités nommées sont généralement constituées de noms propres intéressants pour leur fréquence élevée dans un contexte local par rapport à leur rareté globale. Ces

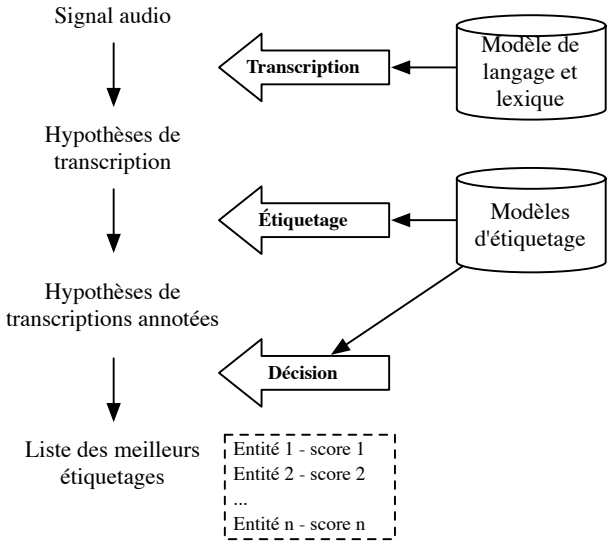



**FIG. 4.3:** *Illustration de l'extraction d'entités nommées dans un flux de parole. L'approche est fortement couplée au moteur de transcription, travaillant sur les treillis d'hypothèses de phrases. Un premier module détermine les annotations possibles de toutes les hypothèses de phrases ; le module de décision extrait les meilleurs entités en fonction des besoins de l'application visée.*

noms propres ont moins de chances d'apparaître dans le lexique du système de transcription et leurs probabilités d'apparition sont faibles dans les modèles grammaticaux du dit système. Les entités sont relativement mal reconnues ou tout simplement non reconnues. L'approche proposée consiste à diminuer l'influence des entités mal reconnues en recherchant des entités nommées directement dans le graphe des hypothèses générées par le reconnaisseur de parole. Le second problème, lié au taux de mots inconnus du lexique doit être résolu avant ou pendant la phase de transcription, car des approches *a posteriori* impliqueraient la localisation des frontières du mot inconnu avec précision.

### Modèle

L'objectif est de trouver le meilleur étiquetage $\hat{L}$ connaissant l'acoustique $A$. Il est intéressant de faire intervenir la séquence de mots $W$, portée par l'acoustique (équation 4.7) :

$$\begin{aligned} \hat{L} &= argmax_L P(L|A) \\ \hat{L} &\simeq argmax_{L,W} P(L,W|A) \\ \hat{L} &\simeq argmax_{L,W} P(A|W)P(W)P(L|W) \end{aligned} \tag{4.7}$$

Cette formulation laisse apparaître la probabilité de la séquence de mots issue de la transcription à travers $P(A|W)P(W)$ et la probabilité d'un étiquetage sur la séquence de

mots à travers $P(L|W)$. Ceci nous permet de modéliser les entités nommées séparément en utilisant un modèle appris sur des données textuelles sans nécessiter explicitement la connaissance du comportement des entités dans l'espace acoustique. Généralement, en reconnaissance automatique de la parole, la probabilité de la séquence entière est approximée en considérant que les dépendances entre éléments de la séquence sont limitées à un contexte proche (Markov). Cette approximation permet de résoudre le problème de maximisation en utilisant la programmation dynamique et, surtout, de représenter l'espace de recherche sous forme d'un graphe. Si la même approximation est faite pour la modélisation de l'étiquetage, il est possible de générer les hypothèses d'étiquetage par composition du modèle représenté sous la forme d'un transducteur pondéré ($\Omega_{P(L|W)}$, équation 4.8), avec le treillis de mots (accepteur pondéré) issu du reconnaisseur de parole ($\Omega_{P(A|W)P(W)}$).

$$\Omega_{P(L|A)} = \Omega_{P(A|W)P(W)} \oplus \Omega_{P(L|W)} \tag{4.8}$$

L'étiquetage d'entités nommées dans un flux de parole doit tenir compte des erreurs faites lors de la transcription. En effet, si l'étiquetage textuel utilisé fait entièrement confiance à la séquence de mots, une simple substitution peut faire apparaître ou disparaître une entité. Un bon compromis entre généralisation et précision doit être trouvé pour que l'algorithme soit robuste aux erreurs. Le modèle que nous proposons est constitué de deux composantes aux rôles bien distincts : une grammaire non contextuelle $\Omega_G$ dont le rôle est de contrôler la généralisation et un modèle $n$-gramme $\Omega_N$ qui permet de désambiguïser les cas incertains et de prendre en compte le contexte (équation 4.9).

$$\Omega_{P(L|W)} = \Omega_G \oplus \Omega_N \tag{4.9}$$

**Grammaire**

Dans le cadre d'une transcription à vocabulaire limité, aucun mot n'est inconnu lors de la phase d'étiquetage. Il devient possible d'attribuer à chacun des mots du lexique une classe ou un ensemble de classes spécifique aux entités dans lesquelles les mots peuvent apparaître. Par exemple, les mots ayant pour classe Prénom participent majoritairement à l'étiquetage des entités personne, mais peuvent faire partie de n'importe quelle entité impliquant un nom propre, comme une entité adresse.

Les entités nommées auxquelles nous nous intéressons sont généralement composées de noms propres mais peuvent également être formés par des séquences spécifiques comme des numéros de téléphone ou des dates. Ces dernières entités sont efficacement décrites par une grammaire car elles impliquent généralement des quantités numériques et des unités de mesure.

La grammaire proposée dans ces travaux est une grammaire non-contextuelle, régulière à droite, composée d'un ensemble de règles de réécriture impliquant des symboles terminaux $t$ (les mots du lexique) et des symboles non terminaux $NT$ (incluant les classes de mots et les entités nommées). Elle peut être exprimée sous forme du graphe

de l'ensemble des séquences reconnues. L'équation[8] 4.10 définit le transducteur $\Omega_G$ comme un ensemble de règles de réécriture.

$$\Omega_G = \{NT \leftarrow (NT|t)+\} \tag{4.10}$$

Pour être fonctionnelle, cette grammaire se voit imposer des contraintes précises : en tant que transducteur, elle doit prendre en entrée n'importe quelle séquence de mots et générer en sortie tous les étiquetages possibles, dont la séquence non étiquetée. Pour cela, l'axiome $S$ correspond à une alternative répétable de n'importe quelle entité $NE$ et d'un non-terminal spécial, appelé mange-mots $B_g$, qui accepte l'ensemble du vocabulaire (équation 4.11).

$$\begin{aligned} S &\leftarrow (B_g|NE)* \\ B_g &\leftarrow t+ \end{aligned} \tag{4.11}$$

Les étiquettes sont insérées lors de la transduction en ajoutant une transition, vide ($\varepsilon$) en entrée, génératrice d'un début d'étiquette $\langle tag \rangle$ avant le non-terminal correspondant à l'entité ; les fins d'étiquettes $\langle /tag \rangle$ sont insérées de la même façon après l'entité (équation 4.12). Des exemples de règles sont donnés dans la figure 4.4.

$$NE \leftarrow (\varepsilon \Rightarrow \langle tag \rangle)t + (\varepsilon \Rightarrow \langle /tag \rangle) \tag{4.12}$$

### Modèle N-grammes

La grammaire précédente permet de déterminer tous les étiquetages valides d'un treillis de mot, mais elle n'est pas capable de faire un choix entre ces étiquetages. Notamment, il faut choisir entre les différents étiquetages possibles d'une séquence (dont l'absence d'étiquetage).

Un modèle N-gramme (équation 4.13, $l_i$ est l'étiquette associée au mot $w_i$) approximant $P(L|W)$ a l'avantage de pouvoir être instancié sous forme d'un transducteur capable de probabiliser l'espace d'hypothèses. Pour que ce transducteur ne soit pas de taille exponentielle par rapport à la taille du vocabulaire, il est approximé grâce aux travaux de Mohri et al. (2002).

$$\begin{aligned} P(L|W) &= \frac{P(L,W)}{P(W)} \\ P(L,W) &= \prod_{l_i,w_i} P(l_i, w_i | l_{i-1}, w_{i-1}, \cdots, l_{i-n}, w_{i-n}) \end{aligned} \tag{4.13}$$

[8] Notations : « $\cdot \leftarrow \cdot$ » représente une règle de réécriture d'une partie droite par une partie gauche ; « $(\cdot)$ » est un regroupement ; « $\cdot+$ » est un opérateur de répétition (au moins une fois) ; « $\cdot*$ » est un opérateur de répétition (zéro ou plusieurs fois) ; « $\cdot|\cdot$ » représente l'alternative entre deux éléments ; $(a \Rightarrow b)$ sont les symboles d'entrée ($a$) et de sortie ($b$) portés par une transition d'un transducteur ; lorsque l'entrée et la sortie sont identiques, $(a \Rightarrow a)$ est remplacé par $a$ pour clarifier la notation ; $\varepsilon$ est une transition vide ne consommant pas de symbole d'entrée et/ou de sortie.

```
PERS_HUM:
($LEFT_CONTEXT_TITLE)? <pers.hum> $PERSON_NAME </pers.hum>
<pers.hum> $PERSON_NAME </pers.hum> ($RIGHT_CONTEXT_TITLE)?
$RELATIVE_TYPE de <pers.hum> $PERSON_NAME </pers.hum>

PERSON_NAME:
($FIRST_NAME)? ($FIRST_NAME)? (de|de la|du|le|des|les)?
     $FAMILY_NAME
$FIRST_NAME ($FIRST_NAME)?

LEFT_CONTEXT_TITLE:
monsieur|madame|mademoiselle|chef|président|présidente
     |responsable|chancelier|roi|reine|premier ministre
     |ministre|docteur

RIGHT_CONTEXT_TITLE:
(le|la)? (chef|président|présidente|responsable|chancelier
     |roi|reine|premier ministre|ministre|docteur)

RELATIVE_TYPE:
le (petit|beau)? fils|le (grand|beau)? père|la (petite|belle)?
     fille|la (grand|belle)? mère|le cousin|la cousine
     |l'oncle|la tante|la (demi|belle)? soeur|le
     (demi|beau)? frère
```

**FIG. 4.4:** *Exemple de règles contextuelles pour l'étiquetage des entités de type Personne, sous-type Humaine (pers.hum). La définition d'un non terminal débute par son nom en majuscules suivi de deux points ; une règle est définie par ligne ; dans la partie droite des règles, un non terminal est précédé du signe dollar ; le signe « | » représente une alternative ; un point d'interrogation désigne une partie facultative ; et les parenthèses facilitent le regroupement ; les débuts et fins d'étiquette sont exprimés par des balises ouvrantes et fermantes.*

Un modèle discriminant comme CRF pourrait certainement conduire à de meilleures performances, mais il est difficile à l'heure actuelle d'instancier ce modèle de façon efficace sous forme de transducteur. L'étiquetage en entités nommées est un problème de segmentation en plus d'être un problème d'étiquetage. Les étiquettes doivent avoir la même granularité que les mots, ce qui nécessite de mettre en place un étiquetage *Begin Inside Outside* (BIO), qui produit des sous-classes différentes pour les mots en début (B), milieu (I) ou hors entité (O).

#### Mélange avec les hypothèses de transcription

La méthode proposée pour l'annotation d'entités nommées dans le treillis d'hypothèses de transcription a l'avantage de permettre la mise en place de diverses techniques de fusion des hypothèses de transcription et d'étiquetage. Les modèles peuvent

être fusionnés en imposant un facteur de normalisation $\alpha$ entre les composantes issues de la transcription et de l'étiquetage (équation 4.14).

$$P(L|A) \simeq \prod_i p(a_i|w_i)p(w_i)p(l_i|w_i)^{\alpha} \tag{4.14}$$

L'hyperparamètre $\alpha$ est alors déterminé empiriquement en utilisant un corpus d'apprentissage. Ce type de fusion est appliqué entre les espaces acoustiques et linguistiques dans un système de transcription et permet de minimiser le taux d'erreur de mots moyen. Il n'est pas intuitif dans le cas de l'étiquetage car il s'agit de mélanger des événements de nature différente, dont les distributions de probabilités sont inférées dans des conditions et sur des données généralement différentes.

Un autre type de fusion utilise la probabilité *a posteriori* de transcription du *support* de chaque hypothèse d'étiquetage (équation 4.15). Le *support* d'une hypothèse d'étiquetage est formé du sous-graphe de mots de l'espace d'hypothèses qui produit une séquence d'étiquettes donnée. Cette probabilité est souvent utilisée comme mesure de confiance dans les tâches fondées sur la reconnaissance de la parole (Falavigna et al., 2002).

$$P(L|A) \simeq \frac{\sum_{W_L} P(A|W_L)P(W_L)}{\sum P(A|W)P(W)} P(L|W) \tag{4.15}$$

Enfin, lorsque la tâche permet de construire un modèle du type d'entité attendu, il est possible de filtrer le graphe d'hypothèses afin de construire la liste des $n$-meilleures séquences de mots correspondant à chaque type d'entité. La transcription n'est plus dans ce cas guidée par la maximisation de la probabilité d'une hypothèse, mais par la tâche post-transcription elle-même.

### 4.2.3 Performances

Cette section est dédiée à l'évaluation de la méthode proposée sur les données de la campagne ESTER.

#### Système de comparaison Lingpipe

Lingpipe[9], un système d'étiquetage en entités nommées optimisé pour l'annotation du texte, a été choisi pour établir des comparaisons avec l'approche utilisant le graphe d'hypothèses de transcription. Ce système offre l'avantage d'utiliser une modélisation HMM, proche de celle présentée dans ces travaux (équation 4.16, dans laquelle $L$ est

[9]Disponible sur http://www.alias-i.com/lingpipe/ visité en août 2006

une séquence d'étiquettes $l_i$, $W$ est une séquence de mots $w_i$).

$$\begin{aligned} \hat{L} &= \underset{L}{\operatorname{argmax}}\, P(L, W) \\ P(L, W) &= \prod_{n=0}^{N} P(w_n, l_n | w_{n-2}, w_{n-1}, l_{n-1}) \\ P(w_n, l_n | w_{n-2}, w_{n-1}, l_{n-1}) &= P(l_n | w_{n-2}, w_{n-1}, l_{n-1}) P(w_n | w_{n-2}, w_{n-1}, l_{n-1}, l_n) \end{aligned} \tag{4.16}$$

Lingpipe utilise une sous-classification BIO pour différencier les début, milieu et fin d'entité. Cette notation permet l'utilisation de modèles communs pour chaque entité pour remplacer les étiquettes conditionnant les probabilités. Finalement, un repli de Witten-Bell (Witten et Bell, 1991) et un lissage par une loi uniforme affinent la qualité du modèle.

Étant donné que les corpus disponibles pour l'apprentissage des paramètres de l'extracteur d'entités nommées ne sont pas très fournis, un processus de généralisation sur les mots inconnus est nécessaire. Lingpipe effectue cette généralisation sur les mots dont la fréquence est faible dans le corpus d'apprentissage. Ils sont remplacés par des classes qui prennent en compte leurs caractéristiques morphologiques (table 4.5).

| Classe | Description |
|---|---|
| 1-DIG | composé d'un seul chiffre |
| 2-DIG | composé d'exactement 2 chiffres |
| 3-DIG | composé d'exactement 3 chiffres |
| 4-DIG | composé d'exactement 4 chiffres |
| 5+-DIG | composé de 5 chiffres ou plus |
| DIG-LET | composé de chiffres et lettres |
| DIG– | composé de chiffres et tirets |
| DIG-/ | composé de chiffres et barres oblique |
| DIG-, | composé de chiffres et virgules |
| DIG-. | composé de chiffres et points |
| 1-LET-UP | composé d'une seul lettre majuscule |
| 1-LET-LOW | composé d'une seule lettre minuscule |
| LET-UP | composé uniquement de lettres majuscules |
| LET-LOW | composé uniquement de lettres minuscules |
| LET-CAP | commence par une majuscule suivie par des minuscules |
| LET-MIX | contient des majuscules et des minuscules |
| PUNC- | ponctuations |
| OTHER | tout le reste |

**TAB. 4.5:** *Les classes morphologiques utilisées par Lingpipe pour généraliser les mots à faible fréquence. La classification est appliquée dans l'ordre des règles.*

Lingpipe est utilisé dans de nombreux travaux car il est facile à mettre en œuvre et couvre de nombreuses tâches. Il est utilisé en question/réponse (Chen et al., 2004; Neumann et Sacaleanu, 2004), en résumé automatique (Schilder et al., 2006), et en résolution

d'anaphores (Vlachos et al., 2006). Ses performances ont été évaluées sur CoNLL 2002, où une $f_1$-mesure de 0.77 lui permet d'atteindre la troisième[10] place sur 14 participants.

### Évaluation ESTER

| Type | Corpus Test | | Corpus Dév. | |
|---|---|---|---|---|
| | Nb. | % | Nb. | % |
| Personne | 1662 | 25.2 | 1689 | 27.2 |
| Lieu | 166 | 2.5 | 155 | 2.5 |
| Organisation | 1001 | 15.1 | 839 | 13.5 |
| GSP | 1794 | 27.2 | 1624 | 26.1 |
| Quantité | 501 | 7.5 | 337 | 5.5 |
| Temps | 1071 | 16.3 | 1245 | 20.1 |
| Produit | 286 | 4.3 | 212 | 3.5 |
| Construction | 125 | 1.9 | 99 | 1.6 |
| Total | 6606 | 100.0 | 6200 | 100.0 |

**TAB. 4.6:** *Distribution des types d'entités dans le corpus de test ESTER. Les entités les plus nombreuses sont les groupes géo-socio-politiques (GSP), les noms de personnes, les références temporelles et les organisations.*

Une tâche expérimentale de reconnaissance des entités nommées a été introduite dans la campagne d'évaluation ESTER. Elle permet de mesurer les performances des systèmes d'extraction d'entités nommées à partir d'un flux de parole. Cette campagne est la seule organisée à l'heure actuelle sur le français parlé. Les types d'entités à reconnaître sont les suivants (leur distribution est détaillée dans la table 4.6) :

1. Personne : humaine, fictive, animal familier ;
2. Lieu : géographique, voie de communication, adresse physique et électronique, numéro de téléphone ;
3. Organisation : politique, commerciale, à but non lucratif ;
4. Groupe géo-socio-politique (GSP) : clan, famille, nation, région administrative ;
5. Quantité : durée, devise, longueur, température, âge, poids, vitesse ;
6. Temps : relatif, absolu, heure ;
7. Produit : œuvre artistique, journal, récompense, véhicule ;
8. Construction : bâtiment, monument.

L'évaluation présente de nombreuses difficultés dues à la nature du média, la diversité des classes et la qualité de l'annotation qui en résulte. D'une part, les malformations du discours oral, comme les hésitations, les reprises ou les confusions, ont un impact sur les entités nommées et doivent être annotées. D'autre part, la classification choisie introduit des ambiguïtés sur le choix de la classe, entre par exemple le temps (il y a 5

[10]Détails sur http://www.alias-i.com/lingpipe/demos/tutorial/ne/read-me.html, visité en septembre 2006.

heures) et la durée (pendant 5 heures), le lieu (place de la Tour Eiffel) et la construction (la Tour Eiffel). Le phénomène est accentué par les métonymies fréquentes qui nécessitent la séparation du rôle et de la nature de l'entité. En effet, un nom de pays utilisé pour représenter son peuple (la France part en vacances) doit-il être annoté comme GSP, ou comme Organisation ? Même si tous ces cas particuliers sont décrits et pris en compte dans le guide d'annotation de la campagne, ils ont néanmoins tendance à abaisser la qualité de l'annotation car les annotateurs ne sont pas toujours cohérents entre eux pour sur l'interprétation du guide. Au niveau des métonymies par exemple, les modèles construits sur le contexte d'utilisation des entités doivent prendre en compte l'ambiguïté constante des classes. Un dernier problème affecte plus particulièrement la qualité de la transcription au niveau des entités nommées. En effet, un décalage temporel de 6 mois différencie les données d'apprentissage de celles de test. Ce décalage implique que les entités dont parle l'actualité ont changé et sont moins bien reconnues par les systèmes de transcription avec le risque non moindre que des noms propres fréquents n'existent pas dans le lexique du reconnaisseur.

| **Données** | **Système** | **Corpus Dév.** | | | **Corpus Test** | | |
|---|---|---|---|---|---|---|---|
| | | SER | $f_1$ | WER | SER | $f_1$ | WER |
| Réf. | $S_{texte}$ | 21 | 0.84 | 0.0 | 27 | 0.79 | 0.0 |
| | $S_{audio}$ | 22 | 0.84 | | 34 | 0.74 | |
| Trans. | $S_{texte}$ | 42 | 0.72 | 21.2 | 55 | 0.63 | 26.4 |
| | $S_{audio}$ | 41 | 0.73 | | 54 | 0.63 | |

**TAB. 4.7:** *Comparatif des performances du système Lingpipe ($S_{texte}$) et de l'approche proposée ($S_{audio}$) sur la tâche d'extraction d'entités nommées ESTER. Les systèmes sont comparés sur les transcriptions de référence (Ref.) et sur les sorties du système de transcription (Trans.), en terme de Slot Error Rate (SER) et $f_1$-mesure ($f_1$). Les performances sont comparées entre le corpus de développement (Dév.) et le corpus de test (Test). L'augmentation du taux d'erreur de mots (WER) entre ces deux corpus implique une forte diminution des performances en détection des entités nommées. Cette diminution est expliquée par une différence de 6 mois entre les corpus qui induit un changement des thèmes d'actualités et donc des entités nommées concernées. On observe toutefois que le système adapté à l'audio est au moins aussi bon que Lingpipe sur les donnée transcrites.*

Les performances du système proposé et de Lingpipe selon diverses conditions d'expérimentation sont proposées dans la table 4.7. Les mesures de performances sont le *Slot Error Rate* (SER) et la $F_1$-mesure. Lingpipe est un système optimisé pour l'annotation de texte et il obtient logiquement de meilleures performances sur la transcription de référence. Par contre, l'approche proposée est meilleure dans le cas où la transcription provient de la chaîne de traitement Speeral, avec un taux d'erreur de mots (WER) de plus de 20%. Néanmoins, l'écart entre l'utilisation des données de référence et la transcription automatique est important : cette relation est de l'ordre de 1.9 points de $f_1$-mesure perdus (et 1.1 points de SER en plus) pour 1% de taux d'erreur de mots supplémentaire. Cette observation peut être comparée à celle de Miller et al. (2000) selon laquelle $\Delta\text{WER} = -0.7\Delta f_1$-mesure sur l'évaluation Hub-4 de NIST (Przybocki et al., 1998). La différence s'explique par l'inadaptation du système spécialisé sur la parole sur des données propres (avec 2.4 points de $f_1$-mesure perdus par point de WER sur le corpus de test). La table 4.8 recense les taux d'erreur (SER) pour chaque type d'en-

tité nommée. Les groupes géo-socio-politiques (GSP) sont bien reconnus et nombreux dans le corpus (en général des noms de pays). Par contre, les catégories moins nombreuses, comme les Produits et les Constructions connaissent des taux d'erreur élevés. Ces performances sont explicables par l'absence de caractères morphologiques spécifiques à ces catégories et la diversité des formes rencontrées. Une comparaison entre l'annotation de la transcription de référence et l'annotation des sorties du système de transcription montre que les entités nommées contenant des noms propres sont plutôt mal reconnues. Cette chute s'explique par une trop forte généralisation sur les erreurs de transcription et la mauvaise modélisation des noms propres dans les modèles de transcription.

Bien que cette évaluation reste une tâche expérimentale[11] d'ESTER, il faut noter que notre approche est la plus performante parmi les trois participants à la tâche d'extraction d'entités nommées. De plus, comparé à une approche ne remettant pas en cause la transcription, il est possible de rechercher explicitement des entités nommées dans le graphe d'hypothèses de transcription lorsque la tâche visée apporte des informations sur le type d'entité recherchée.

| **Type** | **Corpus Dév.** | | **Corpus Test** | |
|---|---|---|---|---|
| | Trans. | Réf. | Trans. | Réf. |
| Personne | 43.9 | 21.7 | 69.1 | 32.3 |
| Lieu | 60.5 | 44.9 | 67.2 | 55.2 |
| Organisation | 46.8 | 32.6 | 67.9 | 50.9 |
| GSP | 26.4 | 9.6 | 36.5 | 11.1 |
| Quantité | 54.8 | 36.7 | 68.2 | 49.7 |
| Temps | 33.6 | 20.3 | 51.9 | 37.9 |
| Produit | 80.5 | 56.5 | 86.3 | 72.8 |
| Construction | 70.6 | 56.9 | 91.9 | 65.6 |

**TAB. 4.8:** *Performances du système proposé sur la tâche d'étiquetage en entités nommées d'ESTER. Les résultats sont donné en Slot Error Rate (SER), sur les données de développement (Dév.) et de test (Test), selon que l'annotation est réalisée sur la transcription de référence (Réf.) ou les sorties du système Speeral (Trans.). On observe que les entités les mieux reconnues sont les GSP et que les entités à base de noms propres subissent le plus gros impact lors de la transcription.*

Toujours selon la table 4.7, il existe une forte différence de performances entre le corpus de développement et le corpus de test. La perte de 5% au niveau du taux d'erreur de mots se traduit par une perte de plus de 10% en SER sur l'annotation. Cette baisse de performances s'explique par un décalage temporel de 6 mois entre les deux corpus. Les changements de thèmes d'actualité et des noms propres les plus fréquents sont les principales explications de ce phénomène. Par exemple, les données de test couvrent la libération de George Malbruno et Christian Chesnaut, otages des forces irakiennes, personnes dont parlent très peu les médias dans le corpus de développement. Afin de pallier ce genre de problème, il est essentiel de créer un modèle de l'actualité correspondant à l'époque du corpus de test et d'introduire ces informations dans les processus

[11] La qualité de l'annotation de référence et le protocole d'évaluation ont évolué jusqu'à la dernière minute.

de transcription et d'annotation.

### 4.2.4 Limites

L'étiquetage des entités nommées d'un flux audio permet de favoriser les zones à forte densité d'informations dans un espace informatif et de présenter à l'utilisateur des indicateurs qualifiés pertinents sur les entités impliquées dans un sujet ou un thème. Néanmoins, sans une résolution complète des coréférences (déterminer que *il*, *le président Américain*, *Mr Bush*, *Georges Bush*, *Georges W. Bush* et *Georges Walker Bush* font tous référence à la même personne), il est impossible de raisonner sur le contenu informatif afin d'en déduire des informations précises sur le fond. De plus, dans une application de recherche d'information, des personnes génériques (les pompiers...) sont peut être aussi importantes que les personnes nommées.

Il faut aussi remarquer que les algorithmes à base d'apprentissage souvent utilisés pour l'annotation de séquences nécessitent une forte qualité des références construites manuellement. En effet, les définitions fines des catégories d'objets à retrouver mènent à des interprétations différentes par chaque annotateur et à des corpus d'autant plus ambigus. Un exemple simple est l'annotation de *La tour Eiffel* comme lieu géographique, bâtiment ou personne (Gustave Eiffel). Bien que le guide d'annotation ESTER essaye de contourner et limiter ces problèmes, leur nombre reste élevé et rend l'évaluation de l'annotation en entités nommées compliquée et laborieuse.

Le modèle choisi dans ces travaux a l'avantage d'être modulaire et de s'appuyer sur des outils existants. Toutefois, on peut imaginer l'instanciation d'autres modèles restant dans le même cadre (automates à états finis), comme les CRF, utilisés dans ces travaux pour trouver des frontières de phrases.

La coopération entre la transcription et l'extraction de descripteurs sémantiques ne s'arrête pas à un espace de recherche commun. Nous avons essayé dans (Favre et al., 2005) de mettre à jour les modèles de langage du système de transcription en utilisant des données externes sélectionnées pour leur recouvrement avec les entités retrouvées. Bien qu'aucune amélioration du WER n'a été observée, cette approche améliore de 10% la quantité d'entités potentiellement extraites (car présentes dans les hypothèses de transcription).

## 4.3 Conclusion

Nous venons d'illustrer l'extraction de descripteurs structurels et sémantiques à travers la segmentation en phrases et la détection d'entités nommées. La section 4.1 a présenté une méthode de segmentation en phrases fondée sur CRF par l'interaction entre des paramètres prosodiques et linguistiques. Cette segmentation admet une $F_1$-mesure de 0.70 dans des conditions réelles sur des émissions radiophoniques en français. La section 4.2, pour sa part, s'est concentrée sur une plus grande collaboration entre les processus de transcription et d'annotation en entités nommées. Une $F_1$-mesure de 0.63

a pu être obtenue sur les mêmes données, sans toutefois améliorer de beaucoup un système optimisé pour le texte.

De nombreux autres descripteurs peuvent être extraits par des méthodes similaires. Toutefois, ces tâches nécessitent une définition rigoureuse pour obtenir des données d'apprentissage les plus cohérentes possibles. L'amélioration de telles méthodes passera alors par une amélioration des algorithmes d'apprentissage impliqués, par une amélioration des paramètres extraits (limitation de la variabilité de la parole), et, surtout, par l'augmentation de la quantité de données d'apprentissage. Le prochain challenge est de s'affranchir de chacune de ces contraintes, en recherchant un modèle unique pour une structuration non-supervisée à toutes les granularités, facilitant l'extraction de concepts représentatifs d'une sémantique à l'échelle humaine.

# Deuxième partie

# Résumé automatique de parole multi-document

# Introduction

Les travaux présentés dans ce document sont focalisés sur l'amélioration de l'accès à un contenu parlé. La principale problématique introduite par un média audio est le temps nécessaire pour écouter et dégager l'information importante parmi de grandes quantités de données audio. Nous proposons d'étudier le résumé automatique de parole pour limiter le temps d'écoute dans l'accès à l'information supportée par de grandes quantités d'émissions radiodiffusées. L'objectif de ces travaux est avant tout la mise en place d'un système complet de résumé automatique de parole fonctionnel.

Le chapitre 2 a proposé un tour d'horizon des méthodes de recherche d'information dans des documents textuels et les adaptations envisagées dans la littérature pour traiter un flux audio contenant de la parole. Le chapitre 3 a présenté l'existant en terme de structuration d'un flux audio. La chaîne de structuration Speeral, développée au LIA, fournit une segmentation en macro-classes acoustiques, une indexation en locuteurs et une transcription orthographique. Nous avons introduit au chapitre 4 des descripteurs supplémentaires nécessaires pour le résumé automatique de parole (frontières de phrases et entités nommées).

Cette nouvelle partie aborde le résumé de parole à partir des données audio structurées jusqu'aux interactions avec l'utilisateur. Tout d'abord, le chapitre 5 propose un modèle de résumé de parole par extraction sensible aux contraintes de l'interactivité. Ce chapitre inclut une étude de la validité de l'hypothèse d'extraction sur la parole, la présentation d'un modèle spécifique aux problématiques visées et son intégration dans un algorithme classique de résumé. Puis, le chapitre 6 se focalise sur une validation de la méthode proposée lors d'une campagne d'évaluation du résumé automatique de texte. Les erreurs introduites par la structuration d'un contenu parlé sont simulées sur les données de cette même campagne. Enfin, le chapitre 7 montre comment les points faibles du résumé par extraction peuvent être contournés à l'aide d'une interface utilisateur adéquate. Les éléments de ce chapitre reposent principalement sur la mise en place d'un démonstrateur complet, à l'aide de la chaîne de structuration et de la méthode de résumé proposée.

# Chapitre 5

# Intégration de contraintes d'interactivité dans le résumé

**Sommaire**



Ce chapitre expose une méthode de résumé par extraction d'un média parlé sous contraintes d'interactivité. La section 5.1 explore la validité d'une méthode de résumé par extraction sur des données radiophoniques. Une borne supérieure de performances Rouge est comparée aux performances d'approches triviales pour le résumé. Puis la section 5.2 définit un modèle de résumé par extraction séparant les traitements dépendant du besoin de l'utilisateur de ceux qui n'en dépendent pas. Les traitements coûteux doivent être indépendants de ce besoin pour satisfaire les contraintes d'interactivité. Le modèle proposé est intégré à *Maximal Marginal Relevance* (MMR) dans la section 5.3. Cette intégration est complétée d'une projection des phrases dans un espace pseudo-sémantique à l'aide de *Latent Semantic Analysis* (LSA). Nous concevons ainsi un système de résumé état de l'art prenant en compte les nouvelles contraintes liées à l'application visée.

## 5.1 Portabilité à un média parlé de l'hypothèse d'extraction pour le résumé

L'approche du résumé automatique de texte par extraction provient d'observations comme celles de Lin et Hovy (2003) et Jing (2002) qu'environ 70% du contenu d'un pa-

nel de résumés textuels écrits à la main (mettant en jeu entre 15 et 50 résumés) est extrait directement depuis les textes d'origine (copie de morceaux de phrases). Nous prenons pour hypothèse que cette observation, réalisée sur un corpus de résumés textuels, est transposable au résumé automatique de parole.

Afin de vérifier cette hypothèse, nous proposons une expérience de résumé automatique fondée sur les données de la campagne ESTER. Ces données ne contiennent pas de références pour l'évaluation du résumé. Par contre, elles sont segmentées manuellement en sections thématiques, dont les étiquettes sont à la fois représentatives du contenu sémantique et du contenu structurel. L'étiquetage permet de différencier les titres du journal radio-diffusé de son contenu[1]. Il est possible de tester grâce à ces données le succès d'une méthode de résumé par extraction pour générer les titres du journal (résumé cible) à partir du corps de celui-ci (documents source).

Rouge (Lin, 2004) est le critère utilisé pour valider la qualité des résumés (titres du journal) générés (voir section 2.2.1). Ce critère est connu comme étant fortement corrélé avec la qualité informative des résumés évaluée par des juges humains. Au cours de cette expérience, les résumés de trois systèmes fictifs de résumé sont comparés aux titres du journal :

1. Sélection au hasard des phrases ;
2. Sélection des $N$-premières phrases ;
3. Sélection de phrases maximisant le score Rouge.

Chacun de ces systèmes utilise les mêmes phrases du corps du journal, transcrites et segmentées manuellement (données de référence ESTER). Le taux de compression choisi correspond au rapport entre la longueur, en mots, du corps du journal et celle des titres visés[2]. Les deux premiers systèmes sont des systèmes triviaux couramment utilisés dans les évaluations du résumé par extraction. Ces mêmes évaluations ont prouvé que l'approche consistant en l'extraction des $N$ premières phrases est difficile à battre dans le cadre d'un résumé mono-document (Over et Yen, 2003). Elle est beaucoup plus facile à surpasser dans le cas de résumés multi-documents. La sélection de phrases maximisant Rouge est obtenue par optimisation gloutonne en choisissant les $N$ phrases de plus grand score individuel Rouge. Cette maximisation est sous-optimale, mais toute autre approche ne serait plus triviale (trouver la vraie borne supérieure est un problème $NP$-complet). La figure 5.1 compare les sorties des trois systèmes aux titres du journal pour un exemple tiré des données.

Une meilleure corrélation entre Rouge et les titres des journaux est obtenue en supprimant les mots outils des résumés hypothèse et référence. Cette suppression n'est pas appliquée dans les campagnes d'évaluation car il est tentant pour un participant de maximiser le nombre de mots à contenu informatif en supprimant les mots outils et en

[1] L'annotation d'origine n'est pas de bonne qualité : environ 35% des données ont été écartées pour créer des paires (titres – corps du journal) propres. La segmentation en sections thématiques n'a pas été remise en cause ; les titres contiennent souvent des données hors-sujet comme les annonce d'émissions.

[2] Les titres font en moyenne 177 mots ; les corps des journaux font en moyenne 3245 mots ; soit un ratio de 6% environ.

**(1) Titres du journal de référence :**
bonsoir à tous les grands titres de cette édition : double attentat suicide à Moscou faisant 18 morts et une cinquantaine de blessés , la police tchétchène retenue , la piste tchétchène plutôt retenue . Golfe : 7 policiers irakiens ont été tués et 45 autres personnes ont été blessées dans une explosion à Ramadi . sport tennis : l'américaine Serena Williams a remporté aujourd'hui le titre du simple dame des cent dix-septième internationaux de Wimbledon de tennis . les développements de ces titres , bonsoir à tous !

**(2) Borne supérieure Rouge :**
et en tennis l'américaine Serena Williams a remporté ce samedi , le titre du simple dame des cent dix-septième internationaux de Wimbledon de tennis sur gazon , troisième levée du grand chelem en battant en finale sa soeur aînée Venus par 2 sets à 1 . en Russie , un double attentat suicide perpétré par des femmes kamikazes est attribué par les autorités russes aux rebelles indépendantistes a fait 18 morts dont les 2 kamikazes et une cinquantaine de blessés lors d'un concert de rock ce samedi , à Moscou . dans le Golfe 7 policiers irakiens et un journaliste britannique ont été tués dans 2 attaques ce samedi en Irak , au lendemain de la diffusion d'une bande sonore attribuée à Saddam Hussein qui appelle à la guerre sainte contre les troupes occupantes .

**(3) N premières phrases :**
en Russie , un double attentat suicide perpétré par des femmes kamikazes est attribué par les autorités russes aux rebelles indépendantistes a fait 18 morts dont les 2 kamikazes et une cinquantaine de blessés lors d'un concert de rock ce samedi , à Moscou . selon le dernier bilan fourni par les autorités . un bilan précédent faisant faisait état de 20 morts le ministre russe de l'intérieur Boris Gryzlov ayant indiqué que 16 personnes étaient mortes sur le coup sans compter les kamikazes .

**(4) Sélection aléatoire de phrases :**
cette rencontre portera essentiellement sur la question des détenus palestiniens qu' Israël pourrait libérer ainsi que sur la suite de la mise en oeuvre de la feuille de route . l' agence Anatolie a annoncé ce soir qu'une partie des militaires turques arrêtés avaient été libérés . en Turquie euh , explosion dans un dans une station euh ,(de) service .

**FIG. 5.1:** *Exemple de résumés d'un journal (20030705_2300_2310_RTM_ELDA). Les titres du journal sont utilisés comme référence (1). Les résumés générés par trois systèmes fictifs sont présentés : la sélection de phrases maximisant Rouge (2),* $R_1 = 0.45$, $R_2 = 0.32$ *; les N premières phrases du corps du journal (3),* $R_1 = 0.26$, $R_2 = 0.08$ *; et une sélection aléatoire de phrases (4)* $R_1 = 0.04$, $R_2 = 0.00$.

faisant l'impasse sur la forme. Dans le cadre de nos expériences, un seul résumé de référence est disponible ; les mots outils sont supprimés afin de limiter l'effet sur le score de la réduction du nombre d'expressions anaphoriques permettant le recouvrement entre des expressions sémantiquement proches.

Les résultats de cette expérience sont présentés dans la table 5.1. Il est intéressant de noter que 39% des mots informatifs des titres du journal ont été retrouvés par la méthode maximisant la mesure Rouge. Ce résultat peut paraître peu élevé, mais il faut noter que la maximisation de Rouge est sous-optimale et que les synonymes ne sont pas considérés par la mesure. De plus, les scores des autres systèmes triviaux sont significativement moins bons quelle que soit la portée de Rouge. Ces résultats démontrent qu'une méthode de résumé par extraction peut être utilisée avec des données parlées de façon similaire à ce qui est fait sur le texte. Ce résultat ne remplace pas une évaluation de méthodes de résumé automatique de parole par extraction par des juges humain. De plus, il faut garder à l'esprit les conclusions de Banko et Vanderwende (2004)

| Système fictif | Rouge-1 | Rouge-2 |
|---|---|---|
| Hasard | 0.12756 | 0.03588 |
| └ *écart-type* | *0.00507* | *0.00342* |
| N-premières | 0.19007 | 0.08516 |
| Borne supérieure | 0.38618 | 0.24212 |

**TAB. 5.1:** *Performance des systèmes fictifs pour générer les titres des journaux ESTER à partir du corps du journal. Le premier système fait une sélection aléatoire de phrases (Hasard), le second utilise les premières phrases du corps du journal (N-premières) et le troisième repose sur la sélection de phrases maximisant les mesures Rouge. La borne supérieure indique qu'une méthode de résumé par extraction peut être utilisée sur des données parlées de façon similaire à ce qui est fait sur le texte. La sélection aléatoire est moyennée sur 50 initialisations différentes du générateur pseudo-aléatoire.*

selon lesquelles les méthodes par extraction doivent être outrepassées pour obtenir des avancées significatives dans le domaine du résumé multi-document.

Nous allons maintenant proposer un modèle général pour le résumé intégrant les contraintes de l'interactivité et d'un média parlé.

## 5.2 Modèle général

Une approche idéale pour le résumé serait de générer tous les résumés possibles et de sélectionner, par rapport au besoin de l'utilisateur, celui qui maximise à la fois la qualité du fond (quantité d'information et non-redondance) et celle de la forme (linguistique et acoustique). Comme il a été vu dans la section 2.2, le résumé par extraction est une approximation prenant pour hypothèse une indépendance et une complétude des phrases pour générer un résumé. Selon la plupart des méthodes, les scores de qualité du résumé peuvent être calculés comme la somme des scores des phrases qui le composent. Hassel et Sjöbergh (2006), par exemple, construisent un résumé textuel par extraction optimal au sens de sa représentativité du contenu d'un document mais sans prendre en compte, ni la forme, ni la redondance du fond. Le nombre de résumés à générer pour obtenir une solution exacte à ce problème est un arrangement de $p$ phrases parmi $n$ phrases possibles, soit $n!/(n-p)!$ résumés différents (par exemple, pour $n = 100$ et $p = 10$, il y a un peu moins de $10^{20}$ résumés). Il faut donc trouver une solution approchée dont le coût en temps soit suffisamment raisonnable pour permettre une interaction avec l'utilisateur.

Dans le contexte d'une interaction avec un utilisateur, le temps nécessaire pour donner une réponse à l'utilisateur est un facteur de qualité primordial. Pour le résumé automatique, ce temps est dépendant de la quantité de données traitées et de la complexité des algorithmes mis en œuvre. Des approximations peuvent parfois être employées pour réduire le temps de traitement d'un algorithme trop complexe. Nous allons présenter un modèle général pour minimiser l'effet de ces approximations et améliorer le temps de réponse du système.

La figure 5.2 montre différentes classes de modèles possibles pour le résumé auto-

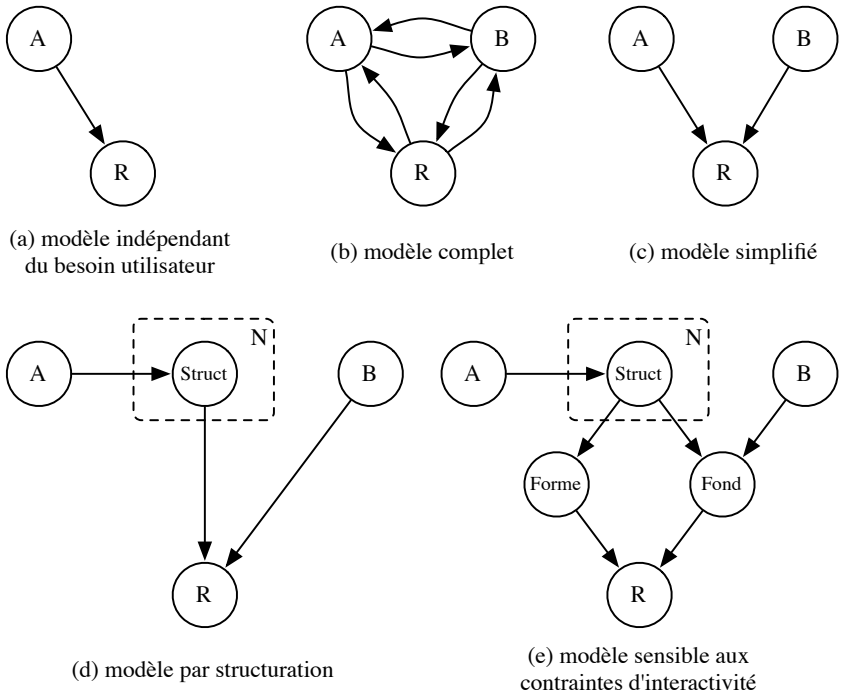



**FIG. 5.2:** *Différentes classes de modèles pour le résumé automatique. A représente l'acoustique, B le besoin utilisateur et R le résumé. Les modèles indépendants du besoin utilisateur déduisent le résumé de l'acoustique (a). Le besoin de l'utilisateur peut être intégré de différentes manières. Le modèle complet (b) prend en compte l'ensemble des dépendances possibles entres les composantes du problème. Le modèle simplifié (c) introduit une hypothèse d'indépendance entre l'acoustique et l'expression du besoin pour faciliter la résolution du problème (modèle classique). Le modèle par structuration (d) ajoute des sous-tâches de structuration pour faciliter l'émergence d'une sémantique dans l'acoustique. Le modèle sensible aux contraintes d'interactivité (e) introduit une séparation entre les paramètres liés à la forme du résumé, indépendants du besoin et ceux liés au fond du résumé, dépendants du besoin. La représentation proposée dans cette figure repose sur le formalisme des modèle graphique : une flèche représente une dépendance conditionnelle.*

matique. Dans cette figure, *A* représente l'acoustique, *R* représente un résumé audio satisfaisant l'utilisateur et *B* représente le besoin de l'utilisateur. Un premier modèle reflète les méthodes ne prenant pas en compte le besoin de l'utilisateur : le résumé est construit uniquement à partir des propriétés intrinsèques de l'acoustique. Pour introduire le besoin dans ce modèle, un modèle complet contenant toutes les interactions possibles entre les trois événements peut être imaginé. Ce modèle fait intervenir des aspects théoriques encore peu explorés, comme une dépendance possible des données et du besoin de l'utilisateur sur le résumé. Bien que très intéressant, le modèle complet n'est malheureusement pas réaliste à l'heure actuelle, surtout pour une application à ressources limitées. La plupart des approches en résumé automatique dépendant du besoin utilisateur sont fondées sur une hypothèse d'indépendance entre l'acoustique et le besoin. Cette hypothèse est illustrée par le modèle simplifié, ou modèle classique en résumé automatique (le résumé dépend de l'acoustique et du besoin). Cette indé-

pendance est discutable mais rend le problème beaucoup plus abordable. Ce modèle implique toutefois la constitution du résumé directement à partir des propriétés de l'acoustique, qui représente une quantité de données trop importante et pas suffisamment riche pour être exploitée correctement. Comment un besoin exprimé en langue naturelle peut-il être confronté directement à des données audio ? Un modèle par structuration emploie des couches de structuration du contenu acoustique avant d'aboutir au résumé. Grâce à ce type de structuration, une sémantique est extraite de l'acoustique et en annote le contenu. Différents éléments de cette structuration ont été étudiés au chapitre 3. Ce type de modélisation représente l'approche la plus intuitive et la plus répandue pour traiter des données audio. Pour rendre ce modèle sensible aux contraintes de l'interactivité, les descripteurs issus de la structuration sont séparés en deux classes, ceux qui auront un impact lié au besoin, et ceux qui peuvent être traités de façon indépendante du besoin. La figure 5.2 matérialise cette idée par la séparation des descripteurs liés à la forme de ceux liés au fond car les premiers sont naturellement moins influencés par le besoin. Néanmoins, cette séparation n'est pas nécessairement limitée à ces deux classes de descripteurs. L'approche par séparation limite l'approximation due aux contraintes d'interactivité à la seule composante liée au besoin. Le reste des traitements peut être mené à l'avance en utilisant des méthodes plus complexes. Pour le résumé par extraction, ceci se traduit par un degré de prédisposition des phrases à l'apparition dans un résumé.

La prochaine section est dédiée à l'intégration du modèle sensible aux contraintes d'interaction dans la méthode de résumé par extraction *Maximal Marginal Releveance* (MMR).

## 5.3 Découplage fond-forme dans Maximal Marginal Relevance

Cette section est organisée en deux sous-parties. Tout d'abord, la formulation de *Maximal Marginal Relevance* est subdivisée en paramètres dépendants du besoin de l'utilisateur (le fond) et de ceux qui peuvent être calculés par avance (la forme). Ensuite, une projection des phrases dans un espace pseudo-sémantique est proposée pour modéliser le fond.

### 5.3.1 Algorithme de sélection de phrases représentatives

La méthode générale de sélection de phrases pour le résumé par extraction est de partitionner l'espace informatif en groupes de phrases sur un thème, ou un événement important, puis de sélectionner une phrase représentative par groupe. Le résumé sera constitué de la juxtaposition des phrases représentant chaque groupe. Des contraintes d'interaction avec l'utilisateur dirigent néanmoins le choix vers une méthode hybride réalisant le partitionnement de l'espace informatif en même temps que la sélection des phrases représentatives. Cette méthode, dénommée *Maximal Marginal Relevance* (MMR), a été proposée par Goldstein et al. (2000) pour déterminer la sélection de phrases candidates dans un résumé par extraction.

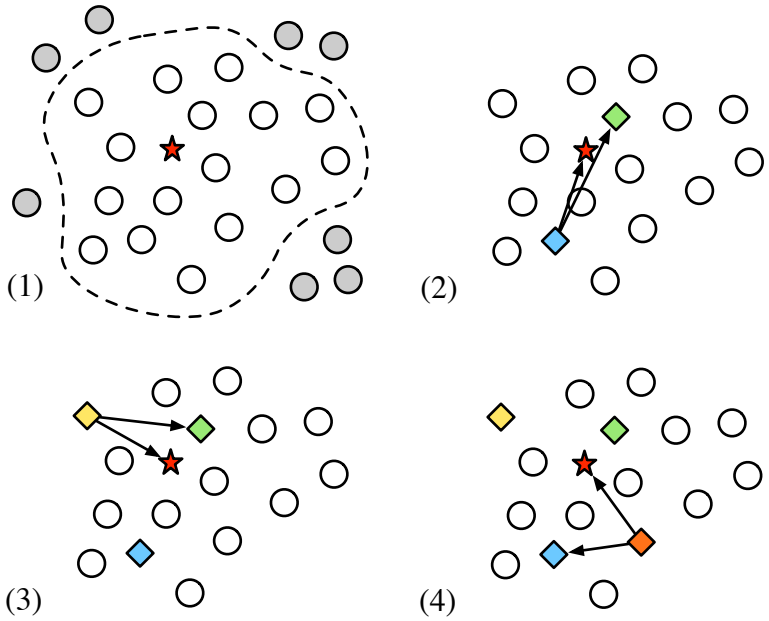



**FIG. 5.3:** *Illustration du fonctionnement de Maximal Marginal Relevance (MMR). La projection du besoin utilisateur est représentée par une étoile, les phrases candidates par des cercles et les phrases sélectionnées par des losanges. La première étape est d'écarter les phrases non pertinentes à l'aide par exemple d'une approche de recherche documentaire (1). La première phrase sélectionnée est celle qui est la plus proche du besoin. Puis, les phrases sont sélectionnées itérativement en fonction de leur distance à la projection du besoin, contrebalancée par leur redondance estimée par leur distance à la phrase déjà sélectionnée la plus proche (2,3 et 4).*

Dans les formulations suivantes, un résumé est constitué à partir d'un ensemble de documents thématiquement cohérents et d'un besoin utilisateur. Les documents sont découpés en phrases qui représentent la matière première de l'algorithme de résumé. À ce niveau, les phrases sont considérées comme indépendantes les unes des autres : leur origine et leur ordre sont ignorés. Une phrase est dénotée ($s_i$) pour la différencier d'une autre phrase ($s_j$). Pour l'instant, aucune hypothèse n'est faite sur le contenu des phrases, il est juste important de pouvoir les comparer deux à deux. Le besoin utilisateur est noté ($b$). La seule hypothèse nécessaire réside dans la possibilité d'évaluer une phrase en fonction de sa réponse au besoin.

MMR est une approximation gloutonne de la résolution du problème d'optimisation consistant à maximiser l'information tout en minimisant la redondance de l'ensemble de phrases sélectionnées pour le résumé. À chaque itération, l'algorithme détermine la phrase ($\hat{s}_k$) la plus proche de l'expression du besoin utilisateur ($b$) tout en étant la plus éloignée des phrases ($s_j$) sélectionnées auparavant. Cette phrase est ajoutée à la sélection et l'algorithme s'arrête lorsqu'une condition est remplie comme par exemple lorsqu'un nombre de phrases $K$, un nombre de mots ou un ratio de compression est atteint. La figure 5.3 illustre ce fonctionnement. Si $n$ est le nombre de phrases à l'origine,

une implémentation efficace de MMR aura une complexité[3] en $O(n^2)$. L'équation 5.1 illustre la fonction de sélection d'une phrase à l'étape $k$.

$$\begin{aligned} mmr_0 &= \emptyset \\ mmr_k &= mmr_{k-1} \cup \{\hat{s}_k\} \\ |mmr_k| &< K \\ \hat{s}_k &= \underset{s_i \notin mmr_{k-1}}{\operatorname{argmax}} \left( \lambda sim_1(s_i, b) - (1-\lambda) \max_{s_j \in mmr_{k-1}} sim_2(s_i, s_j) \right) \end{aligned} \tag{5.1}$$

Dans la formulation originelle de MMR, $sim_1(\cdot)$ et $sim_2(\cdot)$ sont la similarité $cosine(\cdot)$ qui a fait ses preuves en recherche documentaire (voir section 2.1.4). Cependant, n'importe quelle similarité entre phrases est adaptée à ce problème. $\lambda$ est un hyperparamètre[4] devant être ajusté empiriquement en fonction du cadre d'utilisation. Le modèle sensible aux contraintes des interactions utilisateur est introduit en modifiant la façon dont est sélectionnée la meilleure phrase à l'étape $k$ (équation 5.2).

$$\hat{s_k} = \underset{s_i \notin mmr_{k-1}}{\operatorname{argmax}} \left( \lambda_1 \phi(s_i) + \lambda_2 \psi(s_i, b) - \lambda_3 \max_{s_j \in mmr_{k-1}} sim(s_i, s_j) \right) \tag{5.2}$$

Dans l'équation 5.2, $\psi(s, b)$ est le potentiel d'une phrase ($s$) pour le résumé en fonction de paramètres dépendants du besoin ($b$) ; $\phi(s)$ est le potentiel d'une phrase indépendamment du besoin. La limitation de la redondance du résumé est inchangée. Si $\phi(s)$ peut être précalculée et comporter des composantes gourmandes en ressources, $\psi(s, b)$ au contraire doit garantir un résultat rapide pour satisfaire les contraintes de l'interaction. La formulation classique de MMR peut être retrouvée en fixant $\phi(s) = 0$ et $\psi(s, b) = sim(s, b)$.

La fonction $\phi(\cdot)$ représente l'intérêt *a priori* pour une phrase dans le processus de sélection en fonction de ses caractéristiques indépendantes du besoin utilisateur. Le choix de ces composantes est difficile car il s'agit d'un compromis entre le temps de réponse du système et le potentiel d'intégration du besoin utilisateur. En réalité, il existe beaucoup de paramètres pour lesquels un lien direct au besoin utilisateur n'apparaît pas comme primordial. Par exemple, donner la possibilité à l'utilisateur de spécifier la longueur de phrase moyenne du résumé n'est pas indispensable. Il serait plus intéressant d'inférer ces paramètres directement de l'expression du besoin en langue naturelle, qu'ils soient explicités par le discours, ou latents. Les paramètres à prendre en compte sont de trois types (ces paramètres ont été illustrés par les figures 2.6 et 2.7 de la section 2.2) :

[3] Plus précisément, la complexité est de $n \times K$ calculs de la similarité entre deux phrase ou entre une phrase et le besoin, et $3 \times n$ en quantité de mémoire. Ces complexités peuvent être aussi bornées en ignorant de façon arbitraire les phrases répondant le moins au besoin de l'utilisateur.

[4] La nécessité de limiter la redondance dépend du contenu du résumé. Au lieu de faire varier $\lambda$ en fonction du nombre d'itérations de l'algorithme (Murray et al., 2005), nous préférons normaliser la distribution des similarités $sim_1(\cdot)$ et $sim_2(\cdot)$ à chaque itération en leur imposant une moyenne nulle et une variance unitaire (standardisation des distributions).

1. les caractéristiques classiques de la phrase (longueur, position, anaphores non résolues...) ;
2. des mesures de confiance issues de la structuration (probabilité *a posteriori* de la transcription...) ;
3. les caractéristiques intrinsèques à la parole (prosodie, qualité d'élocution, identité du locuteur...).

Utilisées dans $\phi(\cdot)$, ces caractéristiques peuvent être impliquées dans n'importe quelle régression coûteuse mais performante, comme des SVM. Par contre, dans $\psi(\cdot)$, une méthode peu coûteuse comme une combinaison linéaire est indispensable. $\phi(\cdot)$ représente une décision optimale indépendante du besoin (ou dépendante d'un besoin utilisateur moyen), alors que $\psi(\cdot)$ intègre le besoin réel de l'utilisateur.

Dans le cadre de la chaîne de structuration Speeral décrite en section 3.1 et des compléments présentés dans le chapitre 4, nous pouvons introduire plusieurs éléments dans $\phi(\cdot)$. Pour la transcription, une confiance acoustique peut être calculée à partir du rapport entre la probabilité acoustique de la séquence de mots et la probabilité acoustique de la séquence de phonèmes non contrainte par les mots ; une confiance linguistique sera fonction du nombre de fois ou le système a utilisé un repli (séquence de mots non observée dans les données d'apprentissage) dans l'estimation de la probabilité linguistique. D'autres mesures de confiance sont présentées par Mauclair et al. (2006). La qualité de la segmentation en phrases peut être estimée par la probabilité marginalisée de sa frontière de début et de fin. La confiance linguistique dans les entités nommées extraites peut être calculée de la même manière que pour la transcription, mais sera moins bien estimée à cause de la faible quantité de données impliquées dans l'apprentissage des modèles n-grammes d'entités nommées. Une confiance plus représentative sera inférée à partir de la fréquence des entités retrouvées.

### 5.3.2 Projection des phrases dans un espace pseudo-sémantique

La similarité dans l'espace sémantique utilisée par MMR peut être fondée sur la plupart des méthodes de recherche d'information détaillées dans la section 2.1. Nous nous sommes concentrés sur $cosine(\cdot)$, mais afin d'outrepasser les limitations du modèle vectoriel classique (VSM), la base de l'espace sémantique est construite par *Latent Semantic Analysis* (LSA). Nous détaillons cette méthode dans les paragraphes suivants.

L'objectif de LSA est d'obtenir un espace modélisant les affinités contextuelles des mots comme approximation de leurs relations sémantiques. Contrairement à la formulation classique qui projette les requêtes (phrases) dans un espace réduit construit à partir des documents, nous suivons les travaux de Widdows et Peters (2003) et calculons une base unique sur un corpus de grande taille pour pouvoir projeter de nouveaux documents sans avoir à recalculer la base. Cette approche s'apparente au modèle vectoriel généralisé (GVSM). Ainsi, le vecteur représentant une phrase est la somme des vecteurs représentant les mots la composant dans l'espace LSA.

La construction de l'espace LSA demande tout d'abord la création d'une matrice

de cooccurrence entre les mots. Chaque ligne et chaque colonne de cette matrice représentent un mot et la cellule à l'intersection d'une ligne $i$ et d'une colonne $j$ contient le nombre de fois que les mots $i$ et $j$ se sont retrouvés ensemble dans un contexte donné (équation 5.3).

$$w_i^T \rightarrow \begin{matrix} \\ \\ \end{matrix} \begin{matrix} \quad w_j \\ \quad \downarrow \end{matrix} \begin{bmatrix} x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,n} \end{bmatrix} \tag{5.3}$$

La portée de la cooccurrence peut être la phrase, le document, ou un contexte de taille fixe. Par exemple, le tableau de contingence des bi-grammes est construit en limitant le contexte au mot précédent. Généralement, afin de s'affranchir de la nécessité de frontières fixes, une fenêtre glissante de $n$ mots représente le contexte. Bien que la matrice de cooccurrences soit symétrique et creuse, elle demande quand-même une quantité de mémoire en $O(n^2)$, ce qui devient vite exorbitant lorsque le vocabulaire traité est grand. Pour réduire cet effet, il est courant de limiter le nombre de mots observés aux $N$ mots les plus fréquents et de calculer leurs cooccurrences avec non pas l'ensemble du vocabulaire, mais une petite quantité de mots très fréquents et bien choisis, comme les entités nommées, afin de représenter des domaines sémantiques. Dans la suite, le rang de cette matrice va être réduit pour trois principales raisons :

1. la matrice est de trop grande taille pour assurer une faible complexité de calcul dans les traitements de grandes masses de données ;
2. l'estimation des cooccurrences n'est pas fiable par rapport à la distribution réelle (à cause de l'utilisation de synonymes, par exemple) ;
3. les erreurs de transcription bruitent cette matrice (dans le cas où des données textuelles ne sont pas disponibles pour en améliorer l'estimation).

La réduction du rang de la matrice se fait par décomposition en valeurs singulières (*Singular Value Decomposition*, SVD). Le principe est qu'une matrice réelle $X$ peut être factorisée en 3 matrices $U$,$\Sigma$ et $V$ avec $\Sigma$ une matrice diagonale positive, et $U$ et $V$ des matrices orthogonales (équation 5.4).

$$X = U\Sigma V^T \tag{5.4}$$

$\Sigma$ contient les valeurs singulières $\sigma_i$ de $X$, et $U$ et $V$ contiennent les vecteurs singuliers respectifs. Si les $\sigma_i$ sont ordonnés de façon décroissante, le rang de la matrice $\Sigma$ peut être réduit à $k$ en annulant les valeurs singulières de rang supérieur à $k$. Cette réduction correspond à une approximation de $X$ minimisant l'erreur au sens de moindres carrés (équation 5.5). La décomposition en valeurs singulières correspond à la minimisation de la corrélation entre les vecteurs singuliers. La base créée est souvent qualifiée de base « thématique » dans laquelle chaque vecteur représente un thème détecté automatiquement. Les mots sont exprimés dans cette base comme un vecteur de poids des

différents thèmes auxquels ils participent. Dans la pratique, il est difficile d'interpréter les thèmes détectés, mais l'espace créé représente bien les affinités lexicales des mots.

$$\hat{X} = U\Sigma_k V^T \tag{5.5}$$

La base de vecteurs singuliers $U$ est utilisée pour exprimer un contexte $d$ (requête, phrase, paragraphe, document...) en fonction des mots qui le composent. Dans l'équation 5.6, $d$ est un vecteur sur l'ensemble du vocabulaire, fonction de la fréquence des mots dans le contexte observé et $\hat{d}$ est la projection de ce vecteur dans l'espace sémantique.

$$\begin{aligned} d &= (w_0, \cdots, w_n)^T \\ \hat{d} &= \Sigma_k^{-1} U^T d \end{aligned} \tag{5.6}$$

Cette approche a pour principal intérêt de représenter chaque unité informative par un vecteur dans un espace de dimension réduite. Ceci diminue le coût de calcul d'une similarité entre deux éléments. L'espace est considéré comme un modèle du monde (de tout ce qui peut y être instancié) et peut être généré à partir de données externes, disponibles en grande masse. Par contre, par rapport au modèle vectoriel simple utilisé en recherche documentaire, il n'est pas possible de créer un index inversé accélérant — par exemple — le calcul pour trouver l'élément le plus proche d'un élément donné (tous les éléments doivent être parcourus). Comme dans de nombreuses approches, la similarité *cosine* est utilisée pour comparer deux éléments ; ce n'est autre que le cosinus de l'angle entre les deux vecteurs représentant ces éléments :

$$\begin{aligned} cosine(a,b) &= \frac{a \cdot b}{|a||b|}, \\ &= \frac{\sum_i a_i b_i}{\sqrt{\sum_i a_i^2}\sqrt{\sum_i b_i^2}}. \end{aligned}$$

Suivant des travaux de Murray et al. (2005), l'espace sémantique ainsi créé est utilisé pour calculer la similarité entre deux phrases dans l'algorithme de résumé automatique présenté en 5.3.1, sous le nom de MMR-LSA.

## 5.4 Conclusion

Nous avons présenté dans cette section une intégration de contraintes de complexité dans une approche classique de résumé automatique (MMR). Celle-ci se traduit, dans le calcul de l'adéquation de phrase pour le résumé, par la séparation entre les caractéristiques provenant de la forme et celles provenant du fond. La forme est considérée comme indépendante du besoin alors que le fond reste lié à ce dernier. L'adéquation d'une phrase au besoin de l'utilisateur (sur le fond), est calculée après projection des

phrases et du besoin dans un espace pseudo-sémantique (LSA). Nous allons maintenant, dans le chapitre 6, évaluer le système de résumé automatique conçu à partir de ces approches sur des données textuelles, puis simuler sur cette même tâche le type d'erreurs imposé par une structuration automatique du contenu parlé.

# Chapitre 6

# Évaluation indirecte sur le texte

**Sommaire**



Il n'existe pas à notre connaissance de campagne d'évaluation du résumé parlé, même hors des conditions nous intéressant (résumé multi-document, sur le français, portant sur des données radio-diffusées et en association avec une tâche de recherche d'information). Il est tout de même possible d'évaluer indirectement la qualité de l'approche proposée en employant des données d'évaluation portant sur le texte et non sur la parole. Pour cela, une participation conjointe LIA-Thales à la campagne DUC 2006 a permis la validation de l'approche sur des données provenant de journaux, un type de données proche des émissions radio-diffusées (section 6.1). De plus, nous allons simuler l'impact de la structuration automatique d'un contenu parlé sur les données DUC pour avoir une idée de l'évolution des performances du système en conditions dégradées (section 6.2).

## 6.1 La campagne d'évaluation Document Understanding Conference

L'édition 2006 de DUC s'est concentrée sur la stabilisation de la campagne avec une tâche très similaire à l'année précédente afin de vérifier et consolider les acquis[1]. L'ob-

[1]Descriptif de la tâche disponible sur http://www-nlpir.nist.gov/projects/duc/duc2006/tasks.html, visité en janvier 2007.

jectif de l'évaluation est de générer un résumé de 250 mots ou moins (les résumés sont tronqués s'ils sont trop longs, il n'y a pas de bonus à faire plus court), à partir de 25 documents extraits de sources journalistiques et d'une description du besoin utilisateur. Le besoin utilisateur est formulé par un champ *titre* concis et un champ *description* qui liste le type d'information recherchée. En général, le champ *description* contient des sous-besoins du type :

- *Quels sont les causes, les conséquences, les problèmes de (...) ?*
- *Listez les types de (...). Quelles sont leurs particularités ?*
- *Détaillez chronologiquement, et/ou géographiquement, les événements liés à (...).*

Des illustrations de besoin utilisateur sont données dans la table 6.1. Bien que le besoin soit souvent formulé à l'aide de questions, ces dernières ne peuvent pas être traitées comme les questions fermées du type de celles apparaissant dans les tâches de questions-réponses. Ces questions n'ont clairement pas une réponse complète écrite dans un des document, mais elles mettent en jeu des capacités d'abstraction et de raisonnement.

| |
|---|
| **D0617H – Le vol 990 d'EgyptAir** |
| Qu'est-ce qui a causé le crash du vol 990 d'EgyptAir ? Détaillez les éléments de preuves, les théories et les spéculations. |
| **D0629B – Les virus informatiques** |
| Identifiez les virus informatiques ayant eu une propagation mondiale. Détaillez de quelle façon ils se répandent, les systèmes d'exploitation affectés, leurs pays d'origine, et leurs créateurs quand cela est possible. |
| **D0641E – Le réchauffement climatique** |
| Décrivez les théories concernant les causes et effets du réchauffement climatique et les arguments contre ces théories. |

**TAB. *6.1:*** *Exemples de topics DUC 2006 traduits de l'anglais.*

DUC 2006 implique 50 *topics* (définitions de besoin utilisateur, requêtes, sujets ou thèmes) avec leurs 25 documents associés. Il faut noter que les documents fournis pour un *topic* sont pertinents et ne contiennent pas d'information hors-sujet. En terme d'évaluation, les résumés soumis sont notés manuellement sur le fond, la forme et une note globale prenant en compte à la fois fond et forme. Des évaluations automatiques Rouge-2, Rouge-SU4, et *Basic Elements* viennent les compléter grâce à des références produites par NIST, contenant 4 résumés par *topic*. Enfin, l'évaluation *Pyramids* (Nenkova et Passonneau, 2004) est produite par une partie des participants. La tâche telle qu'elle est décrite est très similaire à celle qui a été conduite en 2005. 34 participants ont soumis des résumés lors de l'édition 2006, 20 d'entre eux ont participé à l'évaluation par la méthode *Pyramids*.

### 6.1.1 Descriptif de la soumission LIA-Thales

Notre participation à DUC 2006 est exposée dans cette section ; les autres participants ont décrits leurs systèmes dans les actes de l'atelier de clôture de l'évaluation[2].

#### Principe

La tâche principale de DUC n'a pas évolué entre 2005 et 2006, ceci permet de profiter des données de 2005 pour affiner les paramètres d'un système destiné à l'évaluation en 2006. Cependant, le faible nombre de *topics* (50) impose une certaine prudence dans l'utilisation de ces données. En optimisant trop un système sur 2005, ce dernier risque de réduire ses performances sur l'évaluation 2006 à cause de la différence dans les thèmes traités et les résumés attendus. Cette nécessaire prudence nous conduit à développer une méthode minimisant le risque de sur-apprentissage et améliorant la robustesse de l'estimation des paramètres.

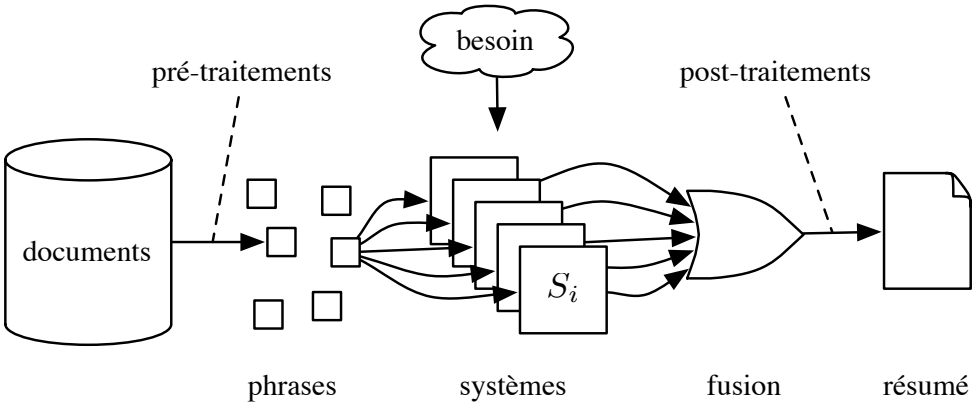



**FIG. 6.1:** *Schéma de fonctionnement du système LIA-Thales pour DUC 2006. Les documents sont découpés en phrases après application de pré-traitements. Puis différents systèmes établissent des listes de priorité d'inclusion des phrases dans le résumé. Les listes sont fusionnées pour trouver une sélection optimale de phrases. Des post-traitements améliorent la forme du résumé final.*

Le principe de la méthode est de fusionner les résumés soumis par plusieurs systèmes de résumé automatique ayant des caractéristiques différentes. Chaque système effectue un résumé par extraction à partir d'une segmentation en phrases commune à tous les systèmes. Seuls les identifiants de phrases sont véhiculés pour réunir les sorties de tous les systèmes dans un espace de représentation commun. La fusion est réalisée dans cet espace et implique à la fois la cohérence des sorties de chaque système et un certain nombre d'heuristiques destinées à améliorer la qualité linguistique des résumés.

[2]Les actes de DUC 2006 sont disponibles sur http://www-nlpir.nist.gov/projects/duc/pubs.html, visité en janvier 2007.

**Les systèmes**

Deux types de systèmes sont utilisés pour la sélection de phrases : des systèmes conçus pour la création de résumés multi-documents par extraction et des systèmes issus de la tâche question-réponse. L'idée est d'essayer de profiter des avantages de la spécialisation de chaque type de système à la fois pour déterminer globalement l'information importante issue de plusieurs documents et pour répondre aux questions précises accompagnant les *topics* DUC. Les 5 systèmes mis en œuvre sont décrits dans (Favre et al., 2006) et représentent un travail de collaboration entre les équipes du laboratoire.

- Le système $S_1$ est fondé sur les travaux présenté en 5.3 (résumé par MMR-LSA). Les phrases sont projetées dans un espace LSA construit à partir d'une matrice de cooccurrences sur le corpus DUC 2006[3]. Dans MMR[4], la fonction $\phi(\cdot)$ est réduite à un retrait des phrases de moins de 10 mots pleins. La fonction $\psi(\cdot)$ correspond $cosine(\cdot)$ dans l'espace LSA. La requête est interpolée[5] avec le centroïde des phrases pour en améliorer la portée. Le contexte des phrases est introduit par une interpolation de chacune d'entre elles avec la précédente[6] dans l'espace LSA ;
- le système $S_2$ a été construit à partir d'une version modifiée de CORTEX (Torres-Moreno et al., 2005) fondé sur différentes caractéristiques des phrases pour en guider la sélection (position, pertinence, redondance, longueur...) et une procédure de décision élaborée ;
- le système $S_3$ utilise une notion d'alignement des phrases avec les concepts exprimés dans le *topic*. Le score final d'une phrase est fonction de cet alignement à plusieurs niveaux morphologiques (stem, lemme, mot...), d'un score de couverture et de la position de la phrase dans le document ;
- le système $S_4$ est le module de recherche de passages pertinents du système de questions-réponses du LIA (Bellot et al., 2003). Le score d'une phrase est influencé par la densité d'apparition des mots du *topic* dans le document ;
- le système $S_5$, décrit dans (Gillard et al., 2005), est le module d'extraction de réponse du système de questions-réponses du LIA. Le score d'une phrase est exprimé à travers la « compacité » d'apparition des mots du *topic* dans un contexte proche autour de celle-ci.

Les systèmes $S_1$ et $S_2$ sont issus de la problématique « résumé automatique » et intègrent une minimisation de la redondance, alors que $S_3$, $S_4$ et $S_5$ répondent aux problématiques questions-réponses et se concentrent sur la précision de la réponse au *topic*. Il est intéressant de noter que $S_4$ et $S_5$ n'ont pas été modifiés pour la tâche DUC. Notamment, leurs paramètres n'ont pas été optimisés grâce aux données d'apprentissage. Cette dernière caractéristique permettra de déterminer si le corpus d'apprentissage apporte de meilleurs résultats sur le corpus de test malgré sa petite taille.

[3]La matrice est construite sur une fenêtre glissante de 30 mots pleins et représente les 60000 mots les plus fréquents par rapport aux 3000 mots les plus fréquents. La réduction de la matrice par SVD est fixée empiriquement à 200 dimensions.

[4]Les paramètres sont $\lambda_1 = 1$, $\lambda_2 = 0.95$, $\lambda_3 = 0.05$.

[5]Facteur d'interpolation avec le centroïde de 0.9.

[6]Facteur d'interpolation avec la phrase précédente de 0.05.

### Traitements linguistiques

**WASHINGTON (AP) –**(1) **The**(2) Department of Housing and Urban Development is taking steps to preserve thousands of federally subsidized housing units for the poor. Housing Secretary Andrew Cuomo said Thursday the department will begin increasing payments to landlords who participate in the program. Many of the payments, set years ago, are below current market levels, and increasing them may entice landlords to stay in the program**, he said**(3). The increased payments will cost $30 million this fiscal year, HUD said...

<s> the Department of Housing and Urban Development is taking steps to preserve thousands of federally subsidized housing units for the poor . **</s>**(4)
<s> housing Secretary Andrew Cuomo said Thursday the department will begin increasing payments to landlords who participate in the program . </s>
<s> many of the payments ,(5) set years ago , are below current market levels , and increasing them may entice landlords to stay in the program </s>
<s> the increased payments will cost $30 million this fiscal year </s>
...

**FIG. 6.2:** *Illustration des pré-traitements appliqués à un document DUC (APW19990428.0245). La marque de l'agence de presse est supprimée (1) ; les majuscules de début de phrase sont transformées en minuscules (2) ; les marques de discours rapporté sont supprimées (3) ; le document est segmenté en phrases (4) ; la ponctuation est séparée des mots (5).*

Les pré-traitements et post-traitements linguistiques améliorent le fond et la forme des résumés produits. Avant la phase de « sélection de phrases », un pré-traitement du texte brut aboutit au découpage en phrases et en mots communs utilisé par les systèmes de résumé (quelques pré-traitements sont illustrés par la figure 6.2). Cette étape est nécessaire afin de normaliser la morphologie des mots et de supprimer les éléments qui pourraient parasiter la modélisation de l'information contenue dans les phrases. Les traitements suivants sont réalisés :

- suppression des marque d'agence de presse (lieu, date et source) ;
- normalisation du vocabulaire selon un lexique propre ;
- découpage en phrases ;
- suppression des majuscules en début de phrase ;
- suppression des titres de personnes (Mr, Mme, Dr...) ;
- formatage des dates et quantités numériques ;
- suppression des formules rhétoriques organisant le discours ;
- nettoyage de la ponctuation (doublons, fins de phrase ...) ;
- suppression des expressions rapportant le discours d'une tierce entité ;
- suppression d'expressions temporelles relatives.

Tous ces traitements sont effectués grâce à un ensemble de règles, de dictionnaires et d'expressions régulières. L'objectif est d'avoir des phrases les plus proches possibles des phrases utilisées dans les résumés. Pour cela, les règles sont écrites de façon à réduire la longueur des phrases et minimiser le « risque linguistique » des pré-traitements. Hors contexte, les formules de construction du discours peuvent dégrader rapidement la cohérence du résumé, par exemple en opposant deux phrases qui ne traitent pas du même sujet. La suppression des expressions liées au discours rapporté (... a dit .., ...

écrit que ...) peut être polémique étant donné que la source de l'information, et donc sa crédibilité, est perdue. Ce type de règle permet toutefois de centrer les phrases sur leur thème principal tout en réduisant leur taille. Finalement, seules les règles qui amènent un résultat correct dans la majorité des cas sont conservées.

**republics**(1) of the former Soviet Union agreed in talks at Nato headquarters in Brussels to enforce reductions in heavy army weapons and aircraft as possible and without renegotiating the 1990 Conventional Arms Forces in Europe treaty(2) bush , responding to a series of Soviet proposals for reducing conventional forces , agreed for the first time to include manpower , helicopters and land based military aircraft in the Conventional Forces in Europe talks in Vienna . **bush , responding to a series of Soviet proposals for reducing conventional forces , agreed for the first time to include manpower , helicopters and land based military aircraft in the Conventional Forces in Europe talks in Vienna .**(3) while strategic nuclear arms will be the main topic of the trip , Baker also is expected during his four day stay in Moscow to provide greater detail about Bush 's new proposal to slash U.S. and Soviet troop strength in Central and Eastern Europe . the 22 nations represented at the Conventional Forces in Europe talks began cleaning up the 200 page text , which sets ceilings on the number of weapons each alliance can hold . **republics of the former Soviet Union agreed in talks at Nato headquarters in Brussels yesterday to enforce reductions in heavy army weapons and aircraft as soon as possible and without renegotiating the 1990 Conventional Arms Forces in Europe treaty , David White writes .**(3)

Republics of the former Soviet Union agreed in talks at Nato headquarters in Brussels to enforce reductions in heavy army weapons and aircraft as possible and without renegotiating the 1990 Conventional Arms Forces in Europe treaty. Bush, responding to a series of Soviet proposals for reducing conventional forces, agreed for the first time to include manpower, helicopters and land based military aircraft in the **Conventional Forces in Europe (CFE)**(4) talks in Vienna. While strategic nuclear arms will be the main topic of the trip, Baker is expected during his four day stay in Moscow to provide greater detail about Bush's new proposal to slash US and Soviet troop strength in Central and Eastern Europe. The 22 nations represented at the **CFE**(5) talks began cleaning up the 200 page text, which sets ceilings on the number of weapons each alliance can hold. **The successor states to the Soviet Union have promised to agree by the end of May on a share out of the weapons cuts to which Moscow committed itself under the 1990 CFE treaty. Gorbachev's announcement that he will propose massive cuts in military manpower on both sides came as part of a new, more detailed Soviet proposal for the current Vienna talks on CFE, US officials said.**(6)

**FIG. 6.3:** *Illustration des post-traitements appliqués à un résumé DUC (D398E). Les débuts de phrase sont recapitalisés (1) ; la ponctuation est normalisée et les phrases sont terminées par des points (2) ; les phrases dupliquées sont supprimées (3) ; les acronymes sont d'abord présentés avec leur forme complète (4), puis remplacés par leur forme réduite (5) ; la réduction du nombre de mots permet d'introduire de nouvelles phrases (6).*

Après formulation d'un résumé, des post-traitements sont appliqués dépendant de l'ordre des phrases ou pouvant malmener les systèmes de sélection de phrases (la figure 6.3 présente un résumé avant et après les post-traitements).

- Réécriture des acronymes ;
- réécriture des noms de personnes ;
- suppression des expressions entre parenthèses ;
- normalisation de la ponctuation pour maximiser le nombre de mots au sens de DUC (éléments séparés par des espaces).

Le principe de réécriture des acronymes et noms de personnes est le suivant : la première occurrence est complète et les suivantes utilisent une forme réduite. Pour les

acronymes, la forme complète détaille la signification de l'acronyme et la forme réduite est l'acronyme lui même. Pour les noms de personne, la forme complète contient le prénom et le nom alors que la forme réduite est limitée au nom de famille seul. Il n'est pas évident de détecter les noms et les acronymes et encore moins d'en faire la résolution lorsque la forme complète n'est pas connue.

Les définitions d'acronymes sont découvertes dans le corpus sous la forme d'une séquence de mots suivie d'un mot entre parenthèses. Les lettres de l'acronyme sont ensuite alignées sur la forme développée à l'aide de quelques heuristiques sur les majuscules, les déterminants et les conjonctions. Le score d'alignement, la fréquence d'occurrence et le nombre de résultats d'une requête jointe sur le moteur de recherche Google permettent d'établir un score de confiance pour ne garder que les résolutions les plus probables. Une expansion en aveugle sur Google ou un corpus plus volumineux est possible, mais comme les acronymes ont souvent plusieurs significations, la méthode permettant de choisir la bonne forme développée doit prendre en compte son contexte d'utilisation afin d'éviter les erreurs grossières comme celle présentée dans la table 6.2.

| |
|---|
| The test can be just as valuable if it discourages athletes from using European Patent Office (EPO) for fear they might ... |

**TAB. 6.2:** *Exemple de résolution erronée de l'acronyme EPO à l'aide de Google. La bonne forme complète est Erythropoietin (EPO).*

La détection de noms de personnes est compliquée car il faut pouvoir différencier des noms de personnes utilisés pour représenter une marque (ou un lieu) de ceux représentants des personnes physiques. La présence de titres, de noms de métier, de prénoms et l'utilisation des formes étendues et réduites dans le corpus permettent d'établir une mesure de confiance dans la construction de la résolution. Toute la difficulté est de détecter par exemple la marque de cigarettes « Philip Morris » bien qu'elle contienne un prénom pour ne pas la remplacer par « Morris ».

| Corpus DUC | Sans | Avec |
|---|---|---|
| 2005 | 260.50 mots | 249.26 mots |
| 2006 | 259.00 mots | 249.22 mots |

**TAB. 6.3:** *Longueur moyenne des résumés avec et sans les traitements linguistiques. Ils permettent de réduire le nombre de mots d'environ 5%, ce qui correspond dans le cadre de la tâche DUC à ajouter une phrase au résumé.*

L'ensemble des traitements linguistiques améliorent la lisibilité de façon significative, bien que ce point reste difficile à mesurer. Ces traitement sont néanmoins très dépendants du type de données. La longueur des résumés est réduite d'environ 5%, soit 10 mots sur les 250 accordés dans la tâche DUC (table 6.3). Ceci représente une phrase supplémentaire en moyenne, apportant potentiellement un gain non négligeable de contenu dans un résumé.

**Fusion**

Les sorties des différents systèmes de sélection de phrases sont représentées dans un même espace de recherche afin de générer un résumé optimal. Le processus de fusion représente les contraintes du problème sous la forme d'un automate à états finis pondéré (*Weighted Finite State Transducer*, WFST) et le meilleur chemin dans cet automate est déterminé par programmation dynamique (Mohri et al., 2002). La construction de l'automate est la suivante :

1. chaque phrase est représentée par un transducteur acceptant ses mots en entrée et générant l'identifiant de la phrase en sortie ;

2. une seule occurrence de chaque phrase est conservée ;

3. les transducteurs sont concaténés en un unique transducteur et chaque phrase est doublée d'une transition « epsilon » pour autoriser n'importe quelle sélection de phrase ;

4. un automate est construit pour représenter un résumé valide d'environ 250 mots. Il contient 251 états et 250 transitions acceptant l'ensemble du vocabulaire, seuls les 20 derniers états sont finaux. Une fois composé avec le transducteur représentant les sélections de phrases possibles, seules les sélections aboutissant à une longueur entre 230 et 250 mots sont conservées.

5. le graphe d'hypothèses est finalement pondéré en utilisant une fonction de coût à plusieurs niveaux. Chaque phrase a un coût pondéré selon le nombre de systèmes pour lesquels la phrase est avant un rang donné et selon son rang maximum dans les sorties des systèmes. Un coût est associé aux états finaux pour avantager les résumés plus longs. Enfin, une pénalité est définie sur des mots spécifiques pour minimiser les anaphores pronominales en début de phrase et les références temporelles relatives.

La recherche du résumé optimal (chemin de coût minimal dans le graphe de fusion) est illustrée par l'équation 6.1. Dans cette équation, $r$ est un résumé, $p$ une phrase et $m$ un mot ; $c_R(\cdot)$ est le coût associé à un résumé ; $c_P(\cdot)$ est le coût associé à une phrase ; $c_M(\cdot)$ est le coût associé à un mot ; $l_{max}$ est la limite en nombre de mots d'un résumé ; $l(\cdot)$ représente sa longueur effective ; $rg_{sys}(p)$ est le rang de la phrase $p$ pour le système $sys$ ; $nb_{sys}\{rg_{sys}(p) < N\}$ représente le nombre de « votes » pour la phrase $p$. Les $\alpha_i$ sont

des hyperparamètres qui doivent être affinés sur un ensemble d'apprentissage[7].

$$\hat{r} = \underset{r}{\operatorname{argmin}}\, c_R(r) \tag{6.1}$$

$$c_R(r) = \alpha_1 \times (l_{max} - l(r)) + \sum_{p_i \in r} c_P(p_i)$$

$$c_P(p) = \alpha_2 \times \underset{sys}{\mathrm{nb}}\{rg_{sys}(p) < N\} + \alpha_3 \times \max_{sys} rg_{sys}(p) + \sum_{m_j \in p} c_M(m_j)$$

$$c_M(m) = \begin{cases} \alpha_4 & \text{si } m \text{ est une expression anaphorique en début de phrase} \\ \alpha_5 & \text{si } m \text{ est une expression anaphorique} \\ 0 & \text{sinon} \end{cases}$$

Les post-traitements appliqués après la fusion ont la particularité de modifier la longueur des phrases, avec pour conséquence que le résumé n'est plus optimal (en moyenne, dans le cas de DUC, une phrase peut y être ajoutée). Il faut réaliser une deuxième passe de fusion afin de prendre en compte les nouvelles longueurs de phrases, en forçant la présence des phrases modifiées sur lesquelles dépendent d'autres phrases du résumé (résolution d'acronymes, de noms de personnes).

#### Ordre chronologique et géographique

La campagne DUC 2005 a montré que l'organisation et la cohérence des résumés devaient être grandement améliorées. Pour cela, les phrases constituant chaque résumé de notre système sont ordonnées en fonction d'un ordre partiel temporel et géographique dépendant des caractéristiques du *topic* traité. Cette méthode repose sur l'observation que la plupart des *topics* DUC traitent d'événements ayant un étalement temporel important (plusieurs années) et impliquant plusieurs régions au niveau mondial.

Afin de déterminer le type d'ordre appliqué à un *topic*, des règles simples permettent de lui donner les étiquettes suivantes : spécifique, général, temporel, et géographique. Par exemple, un *topic* est étiqueté comme général s'il ne contient pas de noms propres ; un *topic* est étiqueté comme géographique s'il contient des mots comme « pays », « mondial », ou « nation » ; un *topic* est étiqueté comme temporel s'il contient des mots comme « événements », ou « dates ».

Chaque document se voit attribué une étiquette temporelle (l'année de publication de la nouvelle) et une étiquette géographique (le pays de publication). Les phrases sont d'abord ordonnées selon les étiquettes de leur document d'origine, puis selon leur position à l'intérieur du document. Les deux sous-ordres sont appliqués séquentiellement en fonction des caractéristiques du *topic* (les règles sont déterminées empiriquement sur le corpus de développement). Le premier ordre permet de former des paragraphes éventuellement introduits par l'étiquette en question. Des exemples sont donnés dans la table 6.4.

[7] Pour DUC 2006, les paramètres sont $N = 30$, $\alpha_1 = 1$, $\alpha_2 = 1000$, $\alpha_3 = 100$, $\alpha_4 = 2$, $\alpha_5 = 1$. Les expressions anaphoriques pénalisantes sont les pronoms et les jours de la semaine.

*Topic : NASA's Galileo Mission*
How successful was NASA's Galileo space probe mission of Jupiter ? What discoveries were made about the **planet*** and its moons ? Include details about **when*** the probe was launched and any troubles it may have encountered.

*Résumé généré :*
**In London, in 1996,** the planet Jupiter is much hotter and windier than previously believed, the latest information from the space probe Galileo has shown.
**In Washington, in 1998,** volcanoes on a moon of Jupiter called Io may be the hottest place in the solar system outside the sun itself. Using instruments orbiting Jupiter on the Galileo spacecraft, researchers calculated the temperature of lava spewing from volcanoes on Io. In addition to the gravitational studies, Galileo has measured magnetic fields surrounding Jupiter and its moons. For about two years the space probe Galileo has been gathering ever more evidence that a large ocean lies hidden beneath the frozen, fractured surface of Jupiter's moon Europa. Photographs taken by the Galileo spacecraft show that Jupiter's moon Io is aglow with colorful light.
**In Pasadena, in 1999,** the Galileo spacecraft halted all non essential activities by going into a safe mode shortly after close approaches to the moon Europa and Jupiter.
**In Los Angeles,** NASA's aging Galileo spacecraft flew within 380 miles of Jupiter's moon Io, exposing the craft to so much radiation that mission controllers feared the probe might not survive. The aging and glitch prone Galileo spacecraft successfully flew within 380 miles of Jupiter's moon Io, overcoming huge doses of radiation and a computer problem just hours before the approach.
**In Pasadena, in 2000,** NASA's aging Galileo spacecraft swooped past Jupiter's frozen moon Europa on Monday but apparently did not experience any computer problems from the planet's intense radiation.

**FIG. *6.4:*** *Exemple de topic (sujet) classé par erreur à la fois comme temporel et géographique (D0638B). Les indices ayant provoqué cette classification sont dénotés par une étoile*. Les paragraphes du résumé sont introduits par le lieu géographique du document d'origine et son année de publication. Malgré leur bon potentiel pour améliorer la structure des résumés, ces indications représentent un risque d'incohérence avec les informations déjà présentes dans les phrases et avec la structure de la phrase (mauvais temps des verbes...).*

Les étiquettes géographiques et temporelles sont détectées au niveau du document et non de la phrase. Cette méthode peut faire émerger des informations contradictoires avec le contenu de la phrase. Pour limiter ce risque, les paragraphes contenant déjà des références temporelles ou géographiques ne sont pas introduits par l'étiquette correspondant au document.

### 6.1.2 Résultats sur DUC 2006

L'édition 2006 a impliqué 34 participants évalués entre eux et par rapport au système trivial NIST générant des résumés à partir des 250 premiers mots du document le plus récent parmi les documents correspondant à un *topic* donné. Le protocole et les mesures d'évaluation ont été décrits dans la section 2.2.1. La figure 6.5 contient les résultats globaux du système LIA-Thales sur l'ensemble des évaluations manuelles et automatiques, accompagnées du classement du système par rapport aux autres systèmes. Notre soumission obtient un très bon classement sur les scores prenant en compte le fond des résumés, compte tenu du peu d'expérience de l'équipe sur DUC (première participation). Par contre, dans l'absolu, on observe que le Contenu et la Qualité générale représentent 50% du score idéal. Au niveau des évaluations automatiques, environ

| |
|---|
| En 1998, paragraphe ordonné géographiquement. |
| En 2001, paragraphe ordonné géographiquement. |
| En 2004, paragraphe ordonné géographiquement. |
| Aux États Unis, paragraphe ordonné temporellement. |
| Au Mexique, paragraphe ordonné temporellement. |
| En Amérique du sud, paragraphe ordonné temporellement. |

**TAB. 6.4:** *Ordre des phrases selon les caractéristiques du topic. Dans le premier cas, l'ordre temporel prévaut alors que dans le second, l'ordre géographique prévaut. Dans les 2 cas, une étiquette introductive est générée.*

| **Évaluation manuelle** | **Score** | **Rang /35** | **Min** | **Max** |
|---|---|---|---|---|
| Qualité linguistique moyenne | 3.57 | 14 | 2.32 | 4.08 |
| *Grammaticalité* | 4.08 | 7 | 1.38 | 4.62 |
| *Non redondance* | 3.84 | 31 | 3.76 | 4.66 |
| *Clarté des références* | 3.42 | 6 | 1.90 | 4.00 |
| *Focalisation* | 3.74 | 13 | 2.50 | 4.28 |
| *Structure et cohérence* | 2.76 | 19 | 1.16 | 3.28 |
| Contenu | 2.80 | 8 | 1.68 | 3.08 |
| Qualité générale | 2.40 | 8 | 1.34 | 2.84 |
| Pyramids | 0.21 | 6/21 | 0.13 | 0.25 |
| **Évaluation automatique** | **Score** | **Rang /35** | **Min** | **Max** |
| Rouge 2 | 0.08700 | 6 | 0.02834 | 0.09558 |
| Rouge SU4 | 0.14522 | 5 | 0.06394 | 0.15529 |
| Basic Elements | 0.04763 | 6 | 0.00459 | 0.05107 |

**TAB. 6.5:** *Résultats du système LIA-Thales sur DUC 2006.*

8% des bigrammes de l'ensemble des résumés de référence apparaissent dans la soumission du système. La qualité linguistique du système est correcte avec un rang de 14 sur 35. Une analyse détaillée fait apparaître que la redondance de surface (utilisation de pronoms à la place de la forme complète) est relativement mauvaise comparée aux autres systèmes. Cela s'explique par la pénalisation des phrases contenant des pronoms dans le processus de fusion. Cependant, une conséquence de ce traitement est l'obtention d'un bon score de clarté des références. Il faut tout de même remarquer que les résumés sont relativement peu focalisés et que malgré les efforts pour les structurer, le score de structure et cohérence reste relativement bas.

Un exemple de résumé généré par le système LIA-Thales est présenté dans la figure 6.6 conjointement à un des résumés de référence. La table 6.8 donne les différents scores d'évaluation associés à ce résumé : ce *topic* a obtenu de bons scores dans la plupart des domaines et représente un bon exemple de conditions de bon fonctionnement du système.

La figure 6.5 montre les performances Rouge 2 et Rouge SU4 (pour DUC 2005 et DUC 2006) des différents sous-systèmes avant fusion ($S_1$ à $S_5$), la fusion des 3 meilleurs

Soumission

Scientists looking for signs of global warming should spend more time scrutinizing Earth's weather circulation patterns, a new study suggests. The findings suggest that global warming blamed on human industry is manifesting itself not just by causing a global rise in temperatures but by favoring certain established weather patterns. While the study does not prove that human induced global warming is what caused the change in frequency of circulation patterns, he said, the change is consistent with it. The theory is that at certain critical altitudes, roughly from 6 to 12 miles, increasing carbon dioxide has the effect of warming the tropics but cooling the polar regions. The data suggest that global warming has caused a slowdown of the Earth at a rate of 0.56 milliseconds a century..

This warming would touch off widespread disruptions in climate and weather and cause the global sea level to rise and flood many places.

In Beijing, chinese scientists believe that global warming, particularly higher winter temperatures, will continue to the first half of the 21 st century.

In San Antonio, while the Global Climate Coalition questions whether global warming is happening, it advocates that companies voluntarily explore and employ new technology to reduce emissions that contribute to global warming.

In Washington, greenhouse gas emissions blamed for global warming may cause the collapse of the West Antarctic Ice Sheet and raise the average global sea level by four to six metres, beginning as as the, a new scientific study predicted recently.

Référence

Global warming is thought to be at least partly caused by emissions of waste industrial gases like carbon dioxide, produced by burning fossil fuels like coal, oil and natural gas. These emissions trap solar radiation and produce a greenhouse effect. Methane and nitrous oxide emissions from agriculture (ruminants and manure) make up 8% of greenhouse gases. Controls on sulfur dioxide emissions reduce a balancing cooling effect.

Global warming already causes more frequent El Nino appearances, receding shorelines, longer warm seasons, and a slower earth spin. It affects habitats and threatens marine life. If emissions are not reduced, average surface temperature will rise 2-6 degrees over the next century, bringing widespread climatic, ecological and economic dislocation. Floods and droughts will increase in frequency and intensity. Melting polar ice will cause rising sea levels and coastal flooding. Malaria will increase. Rates of habitat loss and species extinction will increase. Communities will need to adapt to new conditions.

Skeptics argue that human activities have little influence on climate. Most observed warming is due to natural causes like changes in solar radiation or the circulation of heat-bearing ocean waters. Measurements taken by satellites have found little temperature rise in the upper atmosphere. Computer models are unreliable. Any warming over the next century would be most pronounced in the winter, at night, and in sub-Arctic regions, doing little harm and creating benefits like longer growing seasons and faster plant growth. Industry argues that reducing the use of fossil fuels would cause economic harm to consumers.

**TAB. 6.6:** *Soumission du système et l'un des résumés de référence pour le topic D0641 (Global warming) : certains paragraphes ont été introduits par le lieu géographique concerné. Les quelques erreurs de post-traitements montrent les limites d'une approche à base de règles. Bien que ce résumé soit l'un des meilleurs produits par le système, la qualité de l'abstraction n'est pas encore à la hauteur de ce que produit un expert humain. Un comparatif des résumés produits par les 5 sous-systèmes sur ce topic est disponible en annexe A.*

systèmes ($F1$) et de l'ensemble des systèmes ($F2$). Les mesures d'évaluation permettent de comparer le comportement des sous-systèmes sur les données de développement (2005) et de test (2006). Les paramètres des systèmes $S_1$, $S_2$ et $S_3$ ont été optimisés dans le but de maximiser les scores Rouge sur 2005, contrairement à $S_4$ et $S_5$ qui ont été

Soumission

The custody flap over 6 year old Elian Gonzales could ultimately strengthen U.S. Cuba relations if American officials stand firm and do not succumb to political pressure.

The organizers cited a more positive course in the ongoing custody battle over the 6 year old. Dan Burton, R Ind., to have the boy testify before a House committee and efforts by Elian's relatives to attain custody of the child. As the city's streets settled into an uneasy calm, the battle over six year old Cuban rafter Elian Gonzalez moved to court Friday with the boy's Miami relatives hoping to overturn a decision by federal immigration officials to send him back to his father in Cuba. As Gonzalez was speaking, his relatives in Miami appeared in a county Family Court, with Elian's great uncle, Lazaro Gonzales, asking for temporary custody of the boy. Cuba ELIAN miami demonstrators sang and prayed outside the home where Elian Gonzales is staying as all sides in the custody battle waited for a federal appeals court ruling that could lead to the boy's reunion with his father. Cuba ELIAN miami as both sides in the Elian Gonzales custody case await a federal appeals court ruling, the mayor of Miami flies to Washington to meet with Attorney General Janet Reno. Cuba ELIAN miami pressure mounts on the relatives of Elian Gonzales to turn the boy over to his father as Attorney General Janet Reno reportedly has given approval to take the child by force, if necessary.

**TAB. 6.7:** *Exemple de résumé généré par notre système pour le topic D0647 (Polémique autour de l'enlèvement d'Elian Gonzales). Le résumé a une très faible qualité linguistique a cause d'une mauvaise détection des indications de lieu et source en début de document. Toutefois, les scores automatiques de ce résumé sont élevés, en contradiction avec l'évaluation manuelle, ce qui montre les limites de l'évaluation Rouge.*

utilisés tels quels. Cela se traduit par une différence de performances sur ce jeu de données. Le fait que cette différence soit toujours significative sur le corpus 2006 indique que l'utilisation de données d'apprentissage est bénéfique, même si elles sont peu nombreuses (50 *topics*). Un autre aspect intéressant est que les différentes fusions ($F1$ et $F2$) ont de meilleurs résultats que le meilleur système, que ce soit pour 2005 ou 2006. Enfin, la fusion $F2$ (celle qui a été utilisée pour la soumission) admet de meilleurs résultats sur 2006 par rapport à $F1$ alors que ses résultats étaient moins bons sur 2005. Ceci prouve que l'utilisation d'un plus grand nombre de systèmes aux performances variées agit comme un garde-fou en limitant le sur-apprentissage presque inévitable sur des corpus de petite taille comme ceux de DUC.

Les résultats peuvent être observés en fonction des étiquettes données à chaque *topic*, qui ont un impact sur le choix de l'ordre des phrases et la façon d'introduire les paragraphes. Quatre étiquettes sont retenues selon que le *topic* est *Spécifique*, *Général*, *Géographique* ou *Temporel*. L'étiquette fictive *Inconnu* est ajoutée lorsque le *topic* n'est ni *Géographique*, ni *Temporel*. Comme le montre la table 6.10, les *topics* du type *Géographique* et *Général* obtiennent les meilleures performances. Visiblement, le fait d'introduire les paragraphes par le lieu géographique concerné améliore la contextualisation des phrases dans les résumés. Par contre, l'ajout d'un contexte temporel aux *topics* étiquetés *Temporel* n'apporte rien comparé aux *topics* étiquetés *Inconnu*.

Une étude des intervalles de significativité des évaluations automatiques dénombre les systèmes significativement meilleurs (à 95%) ou moins bon qu'un système donné. La figure 6.9 montre ces résultats pour le système LIA-Thales. Cette étude montre que le système est au niveau des meilleurs systèmes sur Rouge 2 et Rouge SU4.

| **Évaluation manuelle** | **D0641** | **D0647** |
|---|---|---|
| Qualité linguistique moyenne | 4.4 | 1.6 |
| *Grammaticalité* | 4 | 1 |
| *Non-redondance* | 5 | 2 |
| *Clarté des références* | 4 | 1 |
| *Focalisation* | 5 | 3 |
| *Structure et cohérence* | 4 | 1 |
| Contenu | 3 | 1 |
| Qualité générale | 3 | 1 |
| Pyramids | n/a | 0.083 |
| **Évaluation automatique** | **D0641** | **D0647** |
| Rouge 2 | 0.08564 | 0.11011 |
| Rouge SU4 | 0.14662 | 0.16382 |
| Basic Elements | 0.05161 | 0.05284 |

**TAB. 6.8:** *Résultats de l'évaluation pour les topics D0641 et D0647. Le premier est un des meilleurs résumés de la soumission alors que le second est le moins bon. Il faut remarquer que les scores automatiques du topic D0647 ne sont pas en accord avec les scores manuels.*

| **Évaluation automatique** | **score** | **inf.** | **sup.** | **nb.** $>$ | **nb.** $<$ |
|---|---|---|---|---|---|
| Rouge 2 | 0.08700 | 0.00368 | 0.00396 | 2 | 25 |
| Rouge SU4 | 0.14522 | 0.00365 | 0.00358 | 1 | 26 |
| Basic Elements | 0.04763 | 0.00299 | 0.00282 | 2 | 26 |

**TAB. 6.9:** *Évaluations automatiques du système LIA-Thales pour DUC 2006, avec les incertitudes inférieure (inf.) et supérieure (sup.) de chaque score et le nombre de systèmes significativement meilleurs (nb. $>$) et moins bons (nb. $<$) sur chaque score.*

Les résultats du système pour cette campagne d'évaluation montrent que le système $S_1$ (développé dans ces travaux) est au niveau de l'état de l'art et peut éventuellement être utilisé conjointement à d'autres systèmes grâce à un processus de fusion (figure 6.6). Malheureusement, les données d'évaluation sont textuelles et ne reflètent que partiellement les problématiques du résumé audio. Pour cela, la prochaine section est dédiée à la dégradation des données DUC pour simuler une structuration automatique d'un contenu parlé.

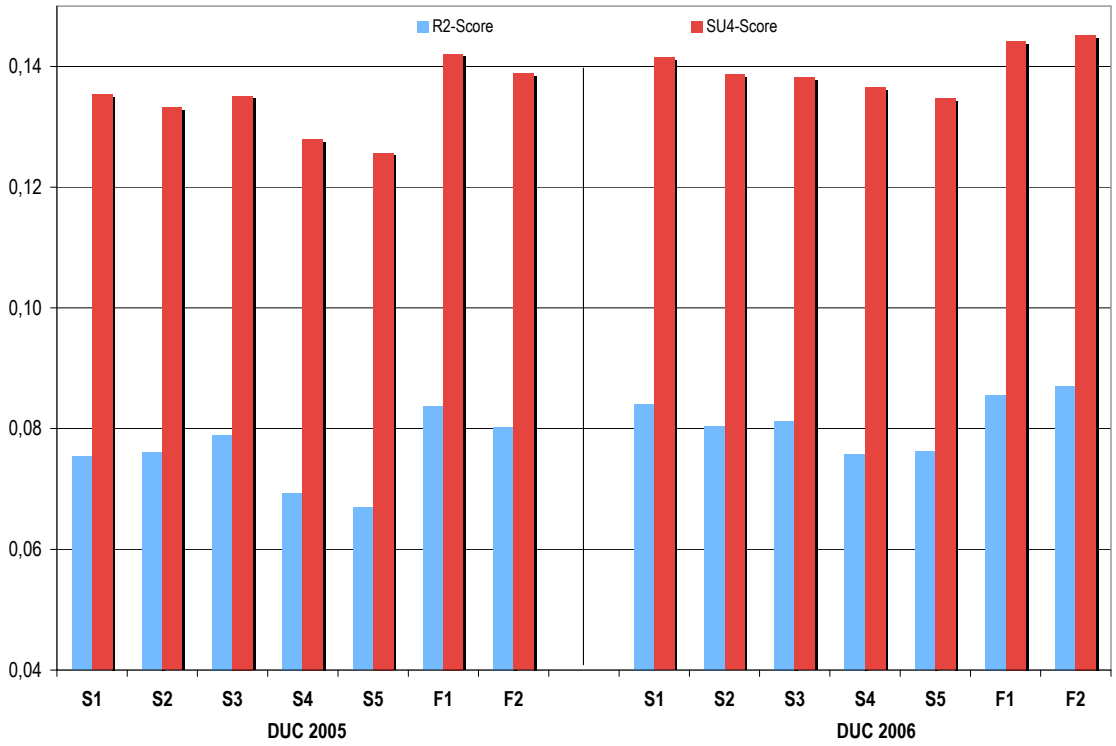



**FIG. 6.5:** *Scores Rouge 2 et Rouge SU4 sur les données de DUC 2005 et DUC 2006 pour les 5 systèmes LIA-Thales ($S_1$ à $S_5$), la fusion des 3 meilleurs (F1) et la fusion des 5 systèmes (F2). L'apprentissage sur le corpus 2005 est bénéfique car les systèmes non optimisés restent significativement moins performants que les autres sur DUC 2006. La fusion limite le sur-apprentissage lorsqu'elle est appliquée sur les 5 systèmes (F2 est meilleure sur DUC 2006).*

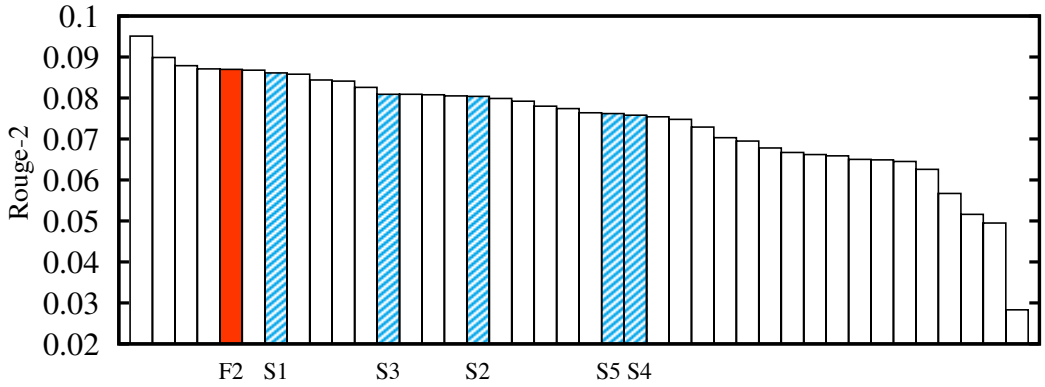



**FIG. 6.6:** *Classement des sous-systèmes $S_1$ à $S_5$ (barres hachées bleues) et de leur fusion $F_2$ (barre pleine rouge) par rapport aux soumissions des autres participants à DUC 2006. Les scores sont exprimés selon la mesure Rouge-2. Le système $S_1$ est à la hauteur des systèmes état de l'art en résumé automatique de texte.*

| **Évaluation manuelle** | **Spécif.** | **Gén.** | **Temp.** | **Géo.** | **Inconnu** |
|---|---|---|---|---|---|
| Qualité linguistique moyenne | 3.39 | **3.58** | 3.34 | **3.77** | 3.48 |
| *Grammaticalité* | 3.90 | **4.30** | 3.80 | 4.00 | **4.25** |
| *Non redondance* | 3.52 | **4.28** | 3.81 | **4.33** | 3.75 |
| *Clarté des références* | **3.59** | 3.19 | **3.69** | 3.33 | 3.28 |
| *Focalisation* | **3.79** | 3.66 | 3.37 | **4.33** | 3.82 |
| *Structure et cohérence* | 2.14 | **2.42** | 2.00 | **2.83** | 2.28 |
| Contenu | 2.55 | **3.09** | **3.06** | 3.00 | 2.57 |
| Qualité générale | 2.38 | **2.48** | 2.37 | **2.66** | 2.39 |
| **Évaluation automatique** | **Spécif.** | **Gén.** | **Temp.** | **Géo.** | **Inconnu** |
| Rouge 2 | **0.090** | 0.082 | 0.077 | **0.095** | 0.091 |
| Rouge SU4 | **0.146** | 0.144 | 0.136 | **0.151** | 0.149 |
| Basic Elements | 0.045 | **0.051** | 0.046 | **0.056** | 0.047 |

**TAB. *6.10:*** *Scores automatiques et manuels sur DUC 2006 en fonction des étiquettes, générées pour chaque topic, qui influencent l'ordre des phrases du résumé. Les étiquettes sont Spécifique (Spécif.), Général (Gén.), Temporel (Temp.), Géographique (Géo.) et Inconnu.*

## 6.2 Simulation de l'impact d'un contenu parlé

Les expériences présentées dans cette section visent à simuler l'impact des différents éléments de structuration automatique d'un contenu parlé sur la tâche DUC. Nous nous attachons à évaluer le fond du résumé par la méthode automatique Rouge-2 tout en ignorant la forme car aucune méthode automatique ne permet de le faire. La chaîne de structuration du contenu audio décrite en 3.1 peut amener divers types d'erreurs :

1. la détection des classes acoustiques peut pousser à ignorer des segments entiers de parole par confusion avec la classe musique, par exemple. Ces segments ne pourront pas être utilisés dans le résumé ;
2. la segmentation en locuteurs a un impact sur la sélection des modèles de transcription et sur la segmentation en phrases. Les erreurs sur les tours de parole réduisent la qualité des frontières de phrase et celles sur les identités ont tendance à faire augmenter le taux d'erreur de mots ;
3. la transcription automatique engendre des erreurs sur le contenu linguistique sous la forme d'insertions, de substitutions et de suppressions de mots. Ces erreurs ont un impact sur l'ensemble des tâches de plus haut niveau, dont le résumé ;
4. l'extraction des entités nommées (comme n'importe quel type de descripteur sémantique) peut affecter une méthode de création du résumé les utilisant fortement ;
5. la segmentation en phrases est cruciale pour le rendu final du résumé audio car la qualité des syntagmes grammaticaux extraits est directement liée à la compréhension du contenu.

Les données traitées pour cette expérience sont celles de DUC 2006. Le système utilisé est le système $S_1$ de la soumission LIA-Thales décrit en 6.1.1. Les post-traitements appliqués ignorent toutefois le processus de fusion et ne changent en aucun cas l'ordre des phrases.

### 6.2.1 Cadre expérimental

L'ensemble des types d'erreurs provoqués par la structuration sont simulés sous la forme d'erreurs touchant la transcription : suppressions, insertions et substitutions du contenu lexical. Les autres types d'erreurs peuvent être inférés à partir de ces erreurs (par exemple, la suppression d'un segment correspond à une suppression en chaîne des mots). Obtenir une dégradation réaliste pour émuler le comportement d'un système de transcription est envisageable (Deng et al., 2003), mais très dépendant de nombreux paramètres liés aux données, à l'implémentation et à la tâche. Les dégradations appliquées dans les expériences suivantes sont uniformes selon l'algorithme de la figure 6.7. Malgré ses défauts, une dégradation uniforme a l'avantage de représenter le cas le plus défavorable pour un système de sélection de phrase car ce dernier ne pourra pas profiter de la variance des dégradations pour choisir des phrases plus « propres ». Des exemples de phrases dégradées sont donnés dans la table 6.11.

**pour chaque** *mot* $\in$ *documents* **faire**
  $p$ =nombre au hasard uniforme sur $\in [0;1]$;
  **si** $p < p_{sup}$ **alors**
    supprimer le mot;
  **sinon si** $p < p_{del} + p_{ins}$ **alors**
    insérer un mot du vocabulaire choisi au hasard;
  **sinon si** $p < p_{del} + p_{ins} + p_{sub}$ **alors**
    substituer le mot avec un mot choisi au hasard;
  **sinon**
    utiliser le mot d'origine;
  **fin**
**fin**

**FIG. 6.7:** *Algorithme pour la génération aléatoire des erreurs selon une probabilité de suppression ($p_{del}$), d'insertion ($p_{ins}$) ou de substitution ($p_{sub}$). Le taux d'erreur de mots obtenu n'est pas exactement de $p_{del} + p_{ins} + p_{sub}$ à cause des effets de bord entre les différents types d'erreurs.*

Pour mieux comprendre l'impact de données parlées, le système MMR-LSA est testé selon deux conditions correspondant à un résumé textuel « lu » par l'utilisateur et à un résumé audio « écouté ». Pour mettre en œuvre la première condition, les données sont dégradées avant d'être injectées dans le système de résumé, dont les sorties sont utilisées pour calculer Rouge-2. Dans la seconde condition, l'utilisateur ne perçoit pas les erreurs de transcription. Cette contrainte est modélisée en remplaçant les phrases des résumés générés pour la condition précédente par les phrases propres d'origine. Ainsi, il est possible de mesurer directement la robustesse de la méthode de sélection de phrases.

| **Type** | **Exemple** |
|---|---|
| Original | Andrew Cuomo said the department will begin increasing payments |
| Insert. | Andrew **lake** Cuomo said the department **change** will begin increasing payments **London** |
| Supp. | Cuomo the department will begin increasing |
| Subst. | Andrew Cuomo **lake** the department **London** begin increasing **change** |
| Tout-type | Andrew Cuomo **lake** the will begin **London** increasing payments |
| supp. EN | said the department will begin increasing payments |
| remp. EN | **lake change** said the department will begin increasing payments |

**TAB. 6.11:** *Exemples de phrases dégradées. Les phrases ayant subit des insertions (Inser.), des suppressions (Suppr.) des substitutions (Subst.) et un mélange de ces trois types d'erreur (Tout type), ont un WER de 33%. Les phrases dont les entités nommées sont supprimées (suppr. EN) ou remplacées aléatoirement (remp. EN), ont un WER de 22%.*

| Type de dégradation | WER | Rouge-2 « lu » | | Rouge-2 « écouté » | |
|---|---|---|---|---|---|
| Aucune | 0.0 | 0.08407 | | 0.08407 | |
| Remplacement des OOV | 1.0 | 0.08255 | -1.8% | 0.08318 | -1.0% |
| └ *écart-type* | | *0.00034* | | *0.00034* | |
| Suppression des OOV | 1.0 | 0.08283 | -1.4% | 0.08279 | -1.5% |
| Remplacement des EN | 10.4 | 0.06741 | -19.8% | 0.08029 | -4.4% |
| └ *écart-type* | | *0.00083* | | *0.00094* | |
| Suppression des EN | 10.4 | 0.07211 | -14.2% | 0.07991 | -4.9% |
| Erreurs aléatoires | 10.0 | 0.07440 | -11.5% | 0.08232 | -2.0% |
| └ *écart-type* | | *0.00118* | | *0.00104* | |

**TAB. 6.12:** *Impact d'une transcription automatique sur les performances de MMR-LSA en dégradant artificiellement les documents de DUC 2006. Rouge-2 est calculé sur les phrases dégradées (condition d'un résumé « lu ») et sur ces mêmes phrases remplacées dans les résumés par leur contenu d'origine (condition d'un résumé « écouté »). Le vocabulaire est limité aux 65000 mots les plus fréquents de Gigaword pour simuler l'utilisation d'un lexique de cette taille. Les mots hors vocabulaire (OOV) sont soit supprimés, soit remplacés par des mots choisis aléatoirement dans le lexique. Les mêmes types de dégradations sont appliqués aux entités nommées (EN). L'impact sur les entités nommées est comparé avec celui de l'introduction d'erreurs aléatoires au même niveau de WER. Les expériences aléatoires sont répétées 50 fois.*

### 6.2.2 Résultats sur les données dégradées

Les documents ont d'abord été dégradés en limitant leur vocabulaire aux mots contenus dans un lexique de système de transcription. Ce lexique est construit en conservant les 65000 mots les plus fréquents d'un grand corpus, Gigaword (Graff, 2003), habituellement utilisé pour créer des modèles de langage. Cela correspond sur les données de DUC 2006 à un taux de mots hors-vocabulaire (OOV) de 1%. Cette limitation du lexique donne lieu à deux dégradations distinctes : le remplacement aléatoire des mots hors-vocabulaire et leur suppression. Ce type de lexique n'est pas forcément représentatif de la différence observée entre les données d'apprentissage et les données de test (essentiellement pour les modèles de transcription, provoquant une forte baisse des performances en test). Dans le cadre de journaux radio-diffusés, les entités nommées sont les plus touchées par cette différence. Nous explorons cet aspect en dégradant uniquement les entités nommées (suppression et remplacement). Ce type de dégradation est comparé à des erreurs aléatoires caractérisées par un taux d'erreur de mots similaire. La table 6.12 donne une idée de l'impact de toutes ces dégradations sur Rouge-2 dans les deux conditions que nous avons envisagées (« lu », « écouté »). Une analyse de ces résultats montre que l'impact sur Rouge-2 est toujours plus fort sur la condition « lu » que sur la condition « écouté ». La limitation du lexique aux mots les plus fréquents implique une baisse relativement faible de Rouge-2 : ce facteur n'est pas limitant mais correspond à des conditions optimales de fonctionnement de la transcription. Par contre, de manière attendue, la suppression des entités nommées réduit fortement la qualité des résumés lus (-14% sur Rouge-2) et a un impact similaire sur les résumés écoutés (-5% sur Rouge-2). Comparé à des erreurs aléatoires, ce type de dégradation est beaucoup plus pénalisant pour le résumé et prouve une nouvelle fois l'intérêt des

entités nommées.

Les figures 6.8 et 6.9 illustrent la variation de Rouge-2 par rapport au taux d'erreur de mots introduit par les différents types de dégradation. Les paramétrages de l'algorithme (fig. 6.7) aboutissent aux dégradations suivantes : uniquement des suppressions, uniquement des insertions, uniquement des substitutions et une distribution uniforme de ces trois classes d'erreurs. Les performances du système sont comparées à un résumé trivial par sélection aléatoire des phrases[8] (indépendante du système) et à un reclassement aléatoire des phrases du système avant sélection[9] (dépendant du système).

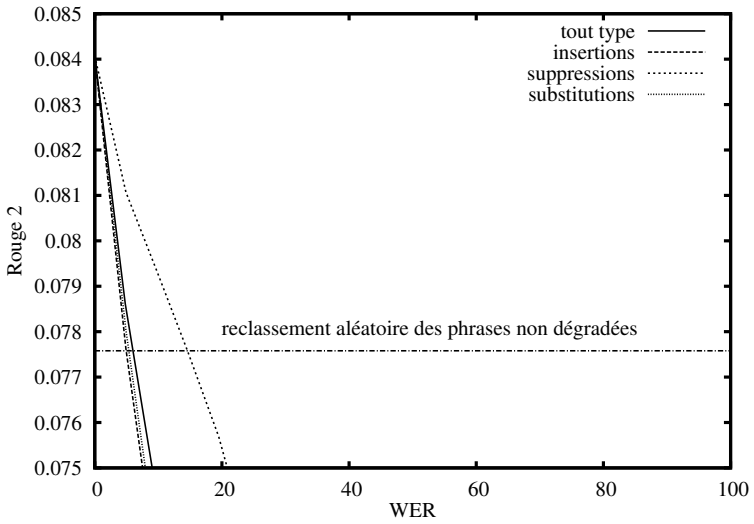



**FIG. 6.8:** *Comparaison des dégradations pour un résumé « lu », pour le système MMR-LSA sur les données DUC 2006. Les erreurs sont des suppressions, des insertions et des substitutions de mots dispensées de façon uniforme pour atteindre divers degrés de WER. Comme les erreurs se répercutent sur le texte du résumé, leur impact est très corrélé à la dégradation pour une mesure fondée sur la distribution des mots comme Rouge-2. Seules les suppressions sont compensées dans une moindre mesure par MMR-LSA. Cette figure est à la même échelle que la figure 6.9 présentant les résultats dans la condition « écouté » (ce choix d'échelle provoque le chevauchement des courbes).*

L'analyse de la figure 6.8 montre que Rouge-2 décroît fortement avec l'augmentation du WER lorsque le résumé est « lu ». Cette observation est prévisible compte tenu de la manière dont fonctionne Rouge. La proportion d'erreur des phrases se retrouve dans les résumés car les erreurs sont uniformes et chacun des *n*-grammes observés par Rouge a autant de chances d'être affecté que les autres. Les performances du reclassement aléatoire sont atteintes pour un WER d'environ 10% et celles de la sélection aléatoire

[8]Cette *baseline* est obtenue à partir des documents d'origine sans pré/post-traitements. Les phrases sont arbitrairement segmentées à chaque occurrence d'un point suivit d'un espace (« . »). Les performances obtenues (Rouge-2 de 0.05576) sont similaire à la *baseline* DUC consistant à créer un résumé à l'aide des 250 premiers mots du document le plus récent (Rouge-2 de 0.0495).

[9]Pré/post-traitements compris, sans appliquer d'autre dégradation (Rouge-2 de 0.07611).

pour un WER d'environ 30%. Seules les suppressions semblent être compensées par le système de résumé, quand l'algorithme choisit, par exemple, des phrases plus longues à l'origine que celles choisies sans dégradation. Ces diverses remarques montrent bien que, malgré les observations (dans des conditions réelles) de taux d'erreur de mots réduits dans les résumés par Christensen et al. (2003) et Murray et al. (2005), le résumé de parole sous forme textuelle nécessite de développer des techniques pour détecter et écarter les phrases mal transcrites.

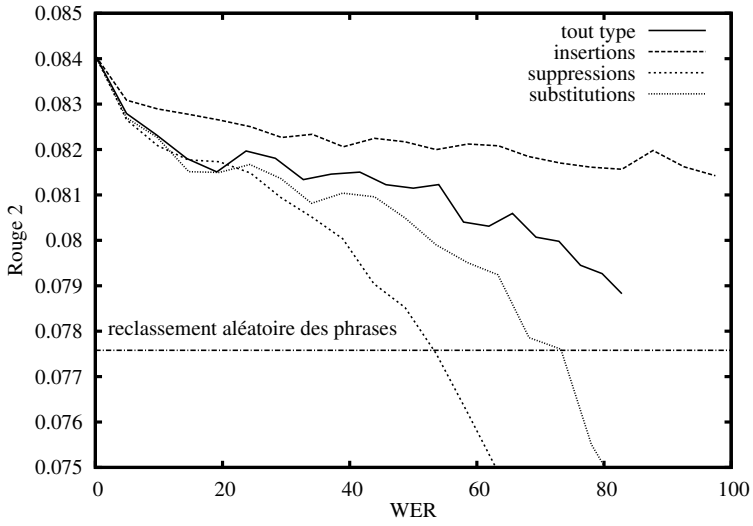



**FIG. 6.9:** *Comparaison des dégradations d'un résumé « écouté », pour le système MMR-LSA sur les données DUC 2006. Contrairement à l'expérience présentée par la figure 6.8, les phrases dégradées sont remplacées dans le résumé par leur version propre pour simuler une écoute de l'audio. Ceci amène à une bonne conservation des performances, même à de forts taux d'erreur, prouvant soit que MMR-LSA est robuste aux erreurs de ce type, soit que Rouge-2 ne reflète pas la qualité des résumés dans de telles conditions.*

La figure 6.9 est plus intéressante car MMR-LSA conserve des scores Rouge-2 élevés lorsque le résumé est « écouté », même sur des données fortement dégradées. Le système est globalement meilleur que les deux *baselines* aléatoires pour un WER inférieur à 50. De plus, les insertions ne sont jamais pénalisantes dans la mesure où le contenu d'origine est toujours présent dans les documents. Une analyse de la variance des résultats pour 50 initialisations différentes du générateur pseudo-aléatoire est présentée dans la figure 6.10. Bien que dans certaines conditions, les données aléatoires augmentent les performances du système (ce qui représente le défaut d'optimisation des paramètres du système sur DUC 2005), l'allure générale des courbes est représentative. Ces observations ne peuvent être expliquées que par une robustesse du système ou un échec de Rouge à évaluer des résumés « écoutés ».

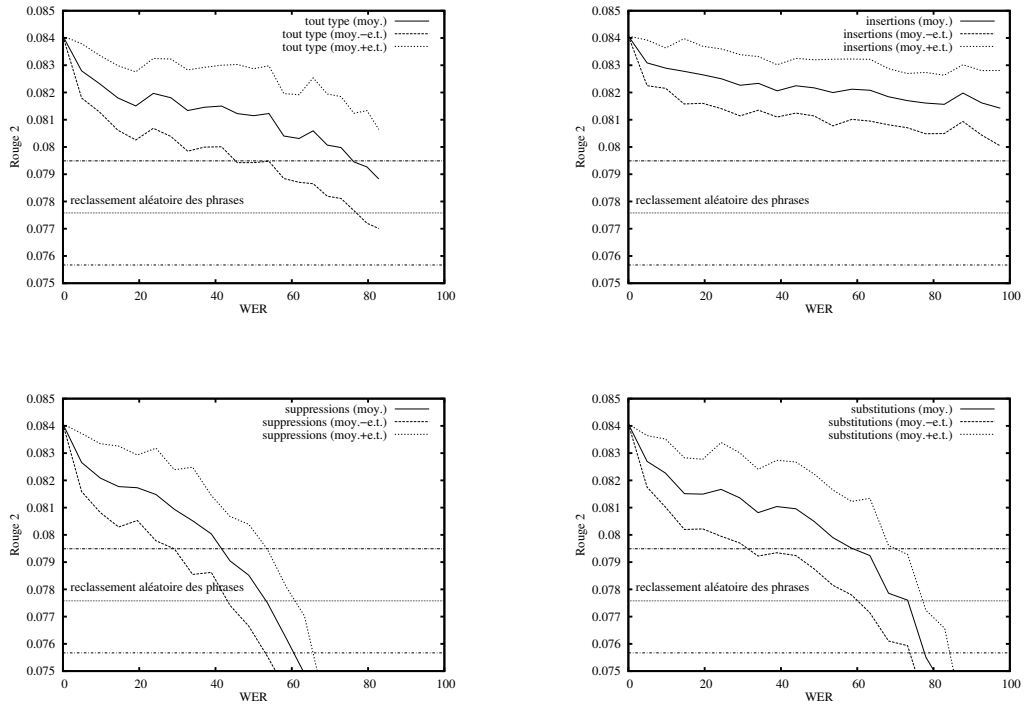



**FIG. *6.10:*** *Détail des courbes de performances Rouge-2 du système sur des données bruitées, dans la condition « résumé écouté ». Les performances sont comparées au reclassement aléatoire des sorties du système représentant une dégradation maximum de la sélection de phrases (Rouge-2 : 0.07611 de moyenne et un écart-type de 0.00191). La moyenne et l'écart-type sont illustrés pour chaque type d'erreurs. Les expériences aléatoires sont répétées 50 fois avec des initialisation différentes du générateur pseudo-aléatoire.*

### 6.2.3 Interprétation des résultats

La robustesse du système est obtenue principalement par l'utilisation du document moyen dans l'expression du besoin utilisateur (les mots les plus fréquents restent fréquents après dégradation). Ainsi, le bruit introduit dans les phrases est également présent dans la requête. Par exemple, pour une dégradation par insertion, le contenu originel des phrases est toujours présent et porteur d'une information plus cohérente que le bruit rajouté. Le reclassement aléatoire des sorties du système montre que les prétraitements et les post-traitements jouent aussi un rôle pour compenser le bruit. La composante la plus déterminante du pré-traitement consiste à écarter les phrases de moins de 10 mots informatifs. La longueur des phrases est prise en compte implicitement de cette façon et il a été prouvé que ce paramètre est déterminant pour le résumé. Le post-traitement contient un garde-fou pour éviter d'insérer des phrases identiques dans le résumé (des phrases quasi-dupliquées ont été introduites par les organisateurs dans DUC 2005 et 2006) : une phrase est écartée si elle n'apporte pas de mots nouveaux au résumé. Cette analyse permet de déduire qu'une grande partie de la robustesse du

système provient des traitements annexes, et que la sélection de phrases en elle-même (MMR-LSA) est bénéfique pour des taux d'erreur inférieurs à 50%.

Au delà de la robustesse du système, l'observation du maintien des performances de MMR-LSA dans des conditions dégradées pose la question de la validité de la mesure Rouge. Cette mesure évalue la qualité du fond d'un résumé par son taux de rappel en *n*-grammes par rapport à un ensemble de résumés de référence. Bien qu'elle soit fortement corrélée avec les évaluations manuelles, les conditions dans lesquelles la mesure n'est plus représentative ne sont pas bien connues. Il serait intéressant de comparer dans les conditions d'un résumé « écouté », les performances Rouge d'une soumission fondée sur des données fortement dégradées et la perception par l'utilisateur de la qualité du contenu. Une autre piste serait de voir à quel point reproduire dans un résumé la distribution des mots dans les documents est une *baseline* performante. Cette dernière, bien que trop élaborée pour être considérée comme une *baseline*, pourrait bien nous amener à reconsidérer la notion de qualité dans les approches statistiques au résumé par extraction.

## 6.3 Conclusion

Nous avons prouvé dans ce chapitre que le système proposé est au niveau des systèmes état de l'art sur une tâche de résumé textuel. Pour cela, la méthode a été évaluée à travers une participation conjointe LIA-Thales à la campagne *Document Understanding Conference* (DUC) 2006. Cette soumission est une fusion de cinq systèmes de sélection de phrases (dont MMR-LSA, décrit dans ces travaux). En plus de cette évaluation ciblant le résumé textuel, nous avons dégradé les données DUC pour simuler les erreurs de la structuration automatique d'un contenu audio. Cette expérience montre que le système proposé est robuste à des erreurs uniformes (le type d'erreur le moins favorable pour un système de résumé) jusqu'à un taux d'erreur mots (WER) d'environ 40%. Les évaluations DUC ont tout-de-même montré que les approches par extraction aboutissaient généralement à une faible qualité de la structure des résumés. L'objectif du prochain chapitre est d'étudier des moyens de contourner cet aspect à l'aide d'interactions utilisateur complémentaires. Ce chapitre sera aussi l'occasion de mettre en valeur la chaîne de traitement complet « de l'audio à l'utilisateur », au sein du démonstrateur développé.

# Chapitre 7

# Interactions complémentaires au résumé parlé

**Sommaire**



Dans les chapitres précédents, une méthode de résumé automatique de parole adaptée à un contexte interactif a été présentée et évaluée de manière indirecte sur des données textuelles. Il s'avère que le gros défaut du résumé par extraction est le manque de structure des informations présentées. En effet, créer un résumé par juxtaposition de phrases retirées de leur contexte, et sans autre lien que leur représentativité thématique, a tendance à générer une réponse dénuée de cohérence. Nous allons tenter d'ébaucher quelques représentations complémentaires au résumé parlé afin de donner à l'utilisateur une idée de la structure des données représentées (section 7.1). Pour cela, le prototype implémenté est décrit en 7.2 et son potentiel est évalué par une enquête, de portée restreinte, auprès d'un panel d'utilisateurs en 7.3. Le lecteur se référera à la section 2.1.7 pour un bref historique des différentes interfaces proposées pour faciliter l'accès à des données parlées.

## 7.1 Frise chronologique interactive

La plupart des échecs des systèmes de recherche d'information provient d'une mauvaise perception du besoin de l'utilisateur exprimé au travers d'une requête. Pour ou-

trepasser ce genre de situation, l'utilisateur n'a d'autre choix que de reformuler son besoin jusqu'à obtenir un succès. En général, il utilise deux types de sources d'information pour cette reformulation : sa connaissance générale du domaine et la structure de la base documentaire traitée. La section 2.1.5 présente quelques méthodes d'expansion de requête pour guider l'utilisateur dans l'expression de son besoin. Ces méthodes sont orientées par les cooccurrences des mots de la requête dans les documents considérés comme pertinents. Cette seule dimension ne permet pas de discriminer les mots proposés en fonction de la structure thématique de l'« espace informatif ». Chuang et Chien (2004) construisent par exemple une hiérarchie thématique à partir des résultats d'un moteur de recherche. La distribution temporelle de l'information est un autre axe fortement structurant dans le cas de nouvelles radio-diffusées.

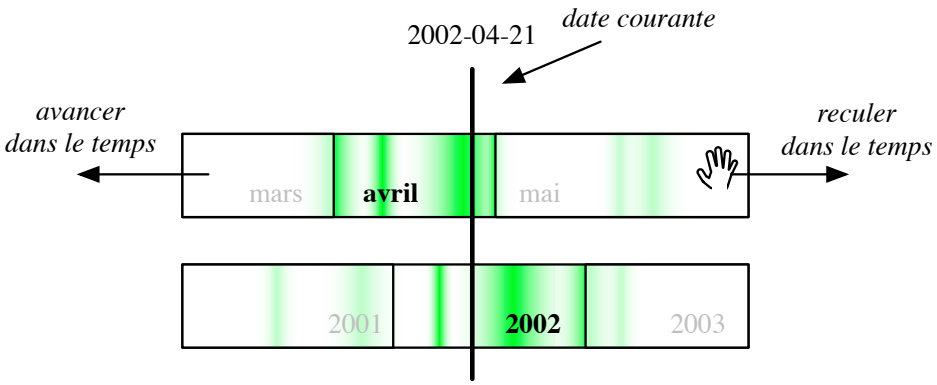



**FIG. 7.1:** *Fonctionnement de la frise chronologique. Le curseur central représente la position de la lecture dans le temps. Chaque échelle est centrée sur ce curseur en fonction de la date courante. L'utilisateur peut faire glisser les échelles pour explorer les résultats dans le temps. La densité de résultats pertinents est représentée sur chaque échelle par un dégradé.*

Nous proposons d'étudier cette distribution à travers une frise chronologique interactive. La figure 7.1 illustre son fonctionnement. L'information potentielle d'une tranche de temps est représentée comme la densité de résultats pertinents dans cette tranche. Au lieu d'avoir une résolution ajustable, les différentes granularités sont discrétisées sous la forme de plusieurs échelles temporelles synchronisées sur l'instant de lecture du flux audio. L'utilisateur peut explorer les différentes échelles en les faisant glisser vers le futur ou le passé de cet instant. L'idée derrière cette répartition en échelles est de créer une correspondance avec un système de quantification du temps naturel pour l'utilisateur : les années, les mois, les jours.... La frise chronologique est implémentée conjointement au résumé de parole dans le prototype présenté par la section suivante.

## 7.2 Description du prototype

L'ensemble des techniques présentées dans ces travaux est implémenté au sein d'un prototype dont l'objectif est de démontrer leur faisabilité technique et d'en effectuer

une analyse globale. La mise en place du démonstrateur complet permet la validation des concepts proposés dans une approche impliquant directement l'utilisateur.

La problématique principale étudiée dans ce document est provoquée par la quantité de données audio à écouter dans le cadre d'une recherche documentaire audio. Si le cas d'un moteur de recherche textuel est considéré, les documents retrouvés sont montrés à l'utilisateur sous forme d'une vue condensée incluant diverses informations jugées utiles pour déterminer rapidement la pertinence d'un document. Ces informations incluent souvent le titre, le thème, l'auteur, la source, ou la date de publication du document. Parfois, des extraits sont ajoutés pour contextualiser les résultats. Cette première représentation permet à l'utilisateur de décider s'il souhaite continuer l'exploration dans la direction d'un certain document. Le système lui délivre alors le document textuel que l'utilisateur commence par observer avant d'en lire une ou plusieurs parties. Très rapidement, en fonction de son expérience, l'utilisateur peut déterminer si un approfondissement s'avère nécessaire ou s'il lui faut passer au document suivant pour obtenir la réponse à son besoin en information. En général, la charte graphique, la structure thématique et une lecture rapide dirigent la décision précédente avec succès et rapidité. Dans le cas de l'audio, l'impossibilité d'avoir un aperçu global d'un document force l'utilisateur à passer beaucoup plus de temps à écouter le contenu. Cette perte de temps peut devenir fortement handicapante dans le cas de bases de données audio de grande taille.

La solution de résumer les résultats du moteur de recherche trouve sa valeur réelle dans l'exploitation de données audio. Nous proposons donc de concevoir un démonstrateur incluant un moteur de recherche interrogeable de la même façon qu'un moteur de recherche textuel, mais dont les résultats sont résumés sous la forme d'extraits audio. Afin d'améliorer la perception de la structure des résultats par l'utilisateur, deux représentations supplémentaires lui sont proposées : des mots-clés évoquant le contenu et une frise chronologique interactive pour naviguer rapidement dans les résultats.

### 7.2.1 Interface utilisateur

La figure 7.2 montre un écran représentatif des éléments du démonstrateur. Tout d'abord, un champ de requête permet d'entrer une requête textuelle et de la soumettre au moteur de recherche. Les résultats retrouvés sont présentés sous trois formes complémentaires : une frise chronologique interactive, une série de mots-clés et une liste d'extraits sélectionnés pour le résumé.

Les segments présentés dans la partie inférieure de la capture-écran ont été sélectionnés parmi les résultats du moteur de recherche en utilisant MMR-LSA[1], la méthode de résumé automatique présentée dans la section 5.3.1. Il est possible d'écouter les extraits, de voir leur transcription et d'explorer l'espace informatif qu'ils représentent en utilisant leur transcription comme requête. La transcription n'est pas montrée directement à l'utilisateur car il a tendance à trop lui faire confiance et à ne pas écouter l'audio

[1]Les données résumées sont les 100 heures d'ESTER. Le modèle LSA a été appris sur 300 millions de mots du journal Le Monde. Les autres paramètres sont identiques à ceux appliqués sur DUC 2006.

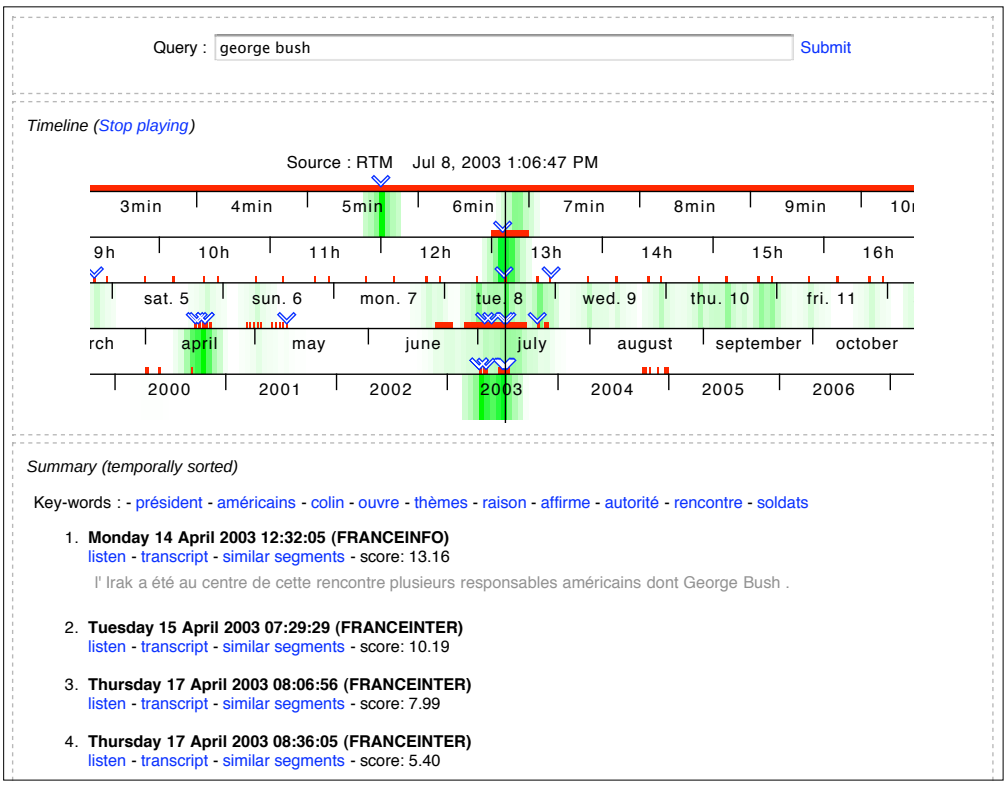



**FIG. 7.2:** *Capture écran de l'interface utilisateur du prototype de moteur de recherche audio incluant de haut en bas : un champ de requête, une frise chronologique interactive, des mots-clés et les extraits audio sélectionnés pour le résumé. La frise chronologique permet de naviguer selon plusieurs échelles temporelles sur lesquelles sont représentés la densité de résultats (dégradé vert), les données disponibles (rectangles rouges) et les extraits sélectionnés pour le résumé (chevrons bleus).*

(Hirschberg et al., 2001). Des informations supplémentaires caractérisent les extraits : leur date de diffusion, leur source et le score de pertinence assigné par le moteur de recherche. Les segments sont ordonnés temporellement pour ne pas briser la chronologie des événements.

Ce premier résumé des résultats est complété par une liste de mots-clés construite par la même méthode (MMR-LSA) mais dans l'espace de représentation des mots et non des phrases. Ces mots-clés sont généralement signifiants mais un filtrage sur les formes syntaxiques permettrait d'en améliorer l'utilité. L'utilisateur peut affiner sa requête en cliquant sur les mots-clés. Cette action les ajoute à la requête et met à jour les résultats.

Une représentation temporelle des résultats et de leur résumé est conçue dans le but d'améliorer la capacité de l'utilisateur à localiser l'information dans le temps. Les résultats du moteur de recherche font appel à deux types de chronologie : la chronologie naturelle des événements (ce qui arrive) et la chronologie de leur apparition dans les thèmes du flux audio (quand on en parle). Afin de minimiser la confusion de l'utilisateur, seul le second type de chronologie est utilisé, mais il serait très intéressant

d'intégrer ces deux chronologies dans une même visualisation bénéfique à l'utilisateur. La frise chronologique du démonstrateur permet de naviguer dans l'ensemble de la base de données audio selon plusieurs échelles temporelles (année, mois, jour, heure, minute) sur lesquelles sont représentés la densité de résultats (dégradé vert), les données disponibles (rectangles rouges) et les extraits sélectionnés pour le résumé (chevrons bleus). Le curseur central représente la date courante de lecture du flux audio. L'utilisateur peut déplacer l'une des échelles afin d'observer l'évolution de la densité de résultats sur une autre échelle, ou de positionner le curseur à une date précise. La densité de résultats est calculée par rapport au score de chaque segment, quantifié sur une durée temporelle fixe (par exemple toutes les 5 minutes pour l'échelle des heures). Les intervalles sont normalisés sur la durée affichée (pour avoir un contraste relatif plus intuitif), et interpolés selon l'équation 7.1.

$$d_k(t_1, t_2) = \sum_{j=-k}^{k} \frac{nb_p\left\{(t_2 - t_1)j < t(p) - t_1 \leq (t_2 - t_1)(j+1)\right\}}{|j| + 1} \tag{7.1}$$

Dans cette équation, $d_k(t_1, t_2)$ est la densité entre un temps $t_1$ et un temps $t_2$, prenant en compte $k$ voisins pour l'interpolation. $nb_p(\cdot)$ est le nombre de points $p$ dans l'intervalle $[t_1; t_2]$ décalé de $j \times (t_2 - t_1)$, $|\cdot|$ est la valeur absolue de son argument, et $t(p)$ représente la date associée à un point.

### 7.2.2 Architecture technique

Le schéma technique général du prototype est décrit dans la figure 7.3. La structuration automatique de la base de donnée audio et son indexation sont effectuées hors ligne. Les étapes de segmentation acoustique, et d'indexation en locuteur, de transcription, de segmentation en phrases, d'extraction d'entités nommées et d'indexation sont opérées en amont car ces traitements sont coûteux en temps et en ressources. L'architecture s'appuie sur ces données pour proposer un service de recherche d'information audio à l'utilisateur. Côté serveur, sont opérés les traitements « métier » comme la recherche d'information, l'extraction de mots-clés et la création de résumés. Côté client, les résultats sont mis en forme de façon indépendante. La diffusion de flux audio doit avoir les mêmes performances que ce soit en local ou à travers un lien réseau, tout en respectant les contraintes temps réel liées au média (délais, gigue...).

Des standards technologiques et des logiciels *open-source* sont utilisés pour faciliter l'intéropérabilité avec les systèmes existants, faciliter le déploiement du démonstrateur et préparer une éventuelle intégration produit. Les technologies utilisées sont détaillées dans la table 7.1.

## 7.3 Enquête utilisateurs

Cette section relate les détails de l'enquête menée auprès d'utilisateurs vis-à-vis du démonstrateur.

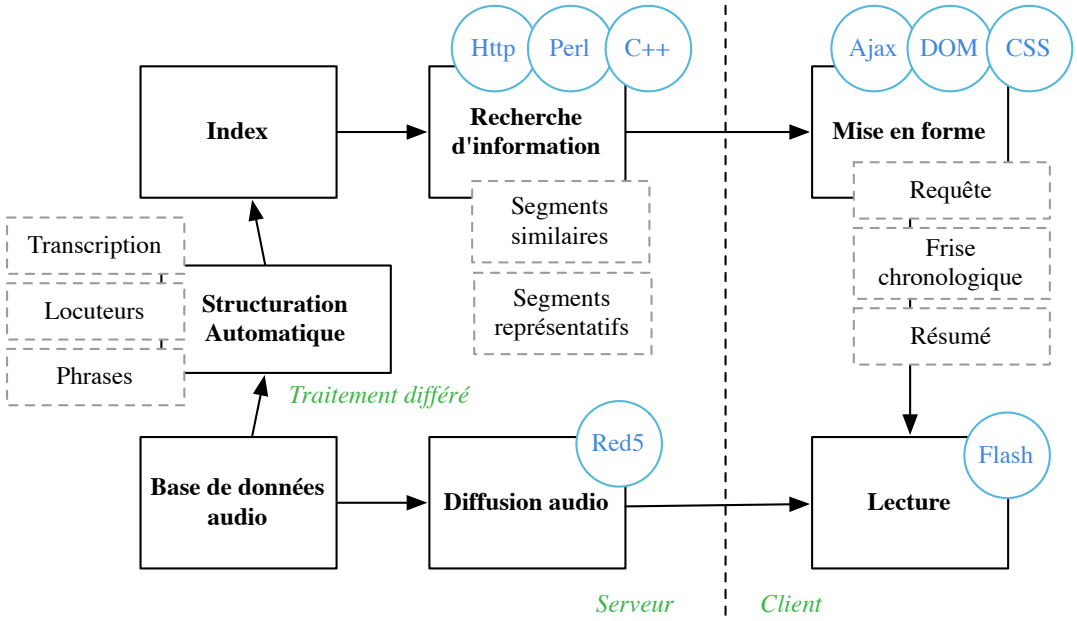



**FIG. 7.3:** *Architecture technique pour la mise en place du prototype. La structuration est réalisée hors ligne ; la recherche d'information, le résumé et la diffusion audio sont proposés sous forme de services et exploités par un client léger fondé sur des standards technologiques.*

### 7.3.1 Principe

En plus de l'évaluation de l'algorithme MMR-LSA sur les données d'une campagne reconnue (DUC 2006), l'interface est évaluée à l'aide lors d'une enquête auprès d'utilisateurs. Cette enquête repose sur un protocole du type « pensez tout haut » lors duquel l'utilisateur doit décrire oralement le maximum de ses réflexions, intentions et actions. Ce protocole a été étudié par Ericsson et Simon (1993) qui le qualifie d'« excellent moyen d'observer le cheminement mental d'un utilisateur ». Ce protocole est largement utilisé dans de nombreux domaines cognitifs ou liés à des interactions avec l'humain comme en ergonomie. Nielsen (1993) souligne qu'il faut choisir les sujets avec soin pour que ce protocole s'avère efficace et qu'un nombre réduit de sujets peut suffire dans le domaine des interfaces logicielles.

Dans le cadre de nos expériences, des sessions de 4 utilisateurs représentatifs ont été enregistrées et analysées. Chaque session dure environ 45 minutes pendant lesquelles le sujet s'exprime sur son expérience, guidé par un interlocuteur connaissant déjà le prototype. Les sessions se passent en plusieurs étapes durant lesquelles l'utilisateur réalise différentes tâches face au démonstrateur :

1. lecture du descriptif de l'expérience présenté en figure 7.4 ;
2. description de son profil en détaillant son expérience de l'outil informatique en général et des moteurs de recherche ;

| | |
|---|---|
| Transcription | Speeral (Outil LIA) |
| Suivi de locuteurs | Alize/LIA_SpkSeg (Outil LIA) |
| Segmentation en phrases | CRF++ |
| Entités nommées | AT&T fsm/grm |
| Serveur HTTP | serveur léger (Perl), support Ajax |
| Moteur de recherche | Modèle vectoriel, serveur léger (Perl/C++) |
| Résumé automatique | MMR-LSA, serveur léger (Perl/C++) |
| Serveur audio | Red5 |
| Lecteur audio | Flash (Ming/Perl) |
| Frise chronologique | Applet (Java) |
| Mise en forme | DOM, CSS et Javascript, (navigateur web) |

**TAB. 7.1:** *Technologies utilisées dans les différents composants du démonstrateur. Red5 est un serveur de diffusion pour du contenu Flash temps réel (http://osflash.org/red5, visité en septembre 2006). Ming est une bibliothèque de génération d'applicatifs Flash (http://ming.sourceforge.net, visité en septembre 2006).*

3. première confrontation visuelle avec l'interface sans l'utiliser, description et intuition sur les fonctions qu'elle peut accomplir ;
4. découverte de l'interface, l'utilisateur peut « jouer » avec l'interface en essayant ses différentes fonctions ;
5. mise en situation en remplissant un besoin en information concret avec une analyse des résultats ;
6. retour sur impressions permettant de collecter l'avis de l'utilisateur sur les points forts et les manques du logiciel.

### 7.3.2 Résultats

Pour chaque phase de l'enquête, les enregistrements ont été analysés. Les points suivants en résument les traits principaux.

– **Profil utilisateur :**

Tous les sujets utilisent le web régulièrement, mais montrent différents niveaux de connaissance de l'informatique théorique. Le genre est distribué uniformément entre les sujets. Ils ont de 17 à 30 ans.

– **Intuitivité *a priori* :**

Les utilisateurs doivent décrire l'interface et imaginer comment elle marche sans interagir avec elle. Plus précisément, ils doivent décrire l'interface comme si leur interlocuteur ne l'avait pas sous les yeux et ne la connaissait pas. Cet exercice permet de faire apparaître les points non intuitifs du logiciel.
Lors de cette phase, les sujets ont tous comparé l'interface à celle d'un moteur de recherche web comme Google en imaginant qu'elle devait fonctionner de la même

## Interface d'accès à une base de données audio.

Ce logiciel permet de trouver et écouter des informations audio provenant de plusieurs radios entre 1998 et 2004. Vous allez participer à une enquête afin de tester la validité du prototype auprès d'utilisateurs. Le cheminement de l'enquête est décrit sous forme de grandes étapes guidées par des questions. Afin de mieux observer votre cheminement mental, vous devrez **penser à voix haute**. Vous serez enregistré(e) pour des analyses ultérieures mais vos informations personnelles ne seront pas diffusées; vous avez un droit de retrait conformément à la loi.

**0) Votre profil.**

Remplissez la feuille concernant vos informations personnelles.

**1) Première approche visuelle.**

Comment imaginez-vous que fonctionne ce logiciel ? Décrivez les éléments que vous voyez. A quoi pensez-vous qu'ils servent ? Qu'attendez-vous de ce logiciel ?

**2) Premier contact.**

Familiarisez-vous avec l'interface en jouant avec ses divers composants. Fonctionne-t-elle comme vous vous l'imaginiez ? Décrivez le fonctionnement de l'interface comme si vous l'expliquiez à quelqu'un qui ne l'a jamais utilisée.

**3) Utilisation effective.**

Choisissez un thème en vous aidant de la liste suivante.

- Fin de la guerre en Irak
- Arrêt de production chez Coca-cola
- Coupe d'Afrique de football
- La crise au Front National
- Le SRAS

Avez-vous des connaissances sur ce thème ? Pourquoi l'avez-vous choisi ? Retrouvez les informations importantes liées à ce thème parmi les données audio. Décrivez votre stratégie. Êtes-vous satisfait des informations trouvées ? Pensez-vous avoir retrouvé toutes les informations importantes liées au thème ?

**4) Sentiment général.**

Qu'est-ce qui vous a plu, déplu ? Que faudrait-il ajouter à ce logiciel ?

**FIG. 7.4:** *Feuillet de questions remis aux utilisateurs.*

façon mais sur des données audio. Malgré l'intuitivité apparente du moteur de recherche, ils ont trouvé que la frise chronologique paraissait obscure et difficile à utiliser. En effet, cette dernière ne correspond pas à une représentation classique et contient des éléments qui ne « s'expliquent pas d'eux même ». De plus, certains éléments liés aux résultats du moteur de recherche ne sont pas affichés avant la première interaction, ce qui ne facilite pas l'interprétation.

– **Découverte du système :**

Les utilisateurs peuvent interagir avec le logiciel et sont encouragés à en décou-

vrir les fonctionnalités en les essayant une par une. Ils n'ont toutefois aucune explication sur le fonctionnement du système.
Durant cette étape, les sujets ont dû inventer des requêtes portant sur des émissions radiophoniques de 1998 à 2004 sans avoir d'autre *a priori* sur les données. Les approches utilisées pour inventer ces requêtes sont très intéressantes et font apparaître une problématique souvent ignorée dans le domaine de la recherche d'information : comment représenter une *vue* informative et concise du contenu d'une base de données spécialisée qui soit utile pour explorer de l'information nouvelle ? Les stratégies mises en œuvre pour inférer le contenu de la base sont, d'une part, d'utiliser des requêtes très génériques et d'analyser les réactions du système et, d'autre part, d'utiliser des requêtes précises sur des événements bien connus appartenant à la période étudiée. La seconde approche est beaucoup plus frustrante car la base de données ne couvre qu'une centaine d'heures et des dates très peu uniformes, donc un petit nombre d'événements.
Au niveau de la frise chronologique, la plupart des utilisateurs ont aimé la façon d'explorer le temps en faisant glisser l'une des échelles temporelles, mais pour des raisons ergonomiques, certains utilisateurs ont pensé que la frise était figée et n'ont même pas essayé d'interagir avec elle. Il est intéressant de noter que la transcription des extraits sélectionnés pour le résumé est cachée à l'utilisateur et qu'il doit cliquer sur un lien pour la faire apparaître. Le but de cette approche était de forcer l'utilisateur à écouter plutôt que de s'appuyer sur sa vision, car Hirschberg et al. (2001) ont observé que les utilisateurs ont tendance à trop avoir confiance dans la transcription et n'écoutent plus l'audio. Ainsi, les utilisateurs ne regardent la transcription que pour valider le comportement du système, dans les cas où il n'est plus cohérent (lorsque les erreurs de transcription ont un impact sur les résultats).

– **Utilisation en situation réelle :**

Après des explications sur le fonctionnement de l'interface, les utilisateurs doivent répondre à des besoins réels en information, sélectionnés dans une liste. Les sujets détaillent dans un premier temps leurs connaissances du thème choisi, les raisons de leur choix et la stratégie qu'ils comptent mettre en œuvre.
Dans cette mise en situation, les utilisateurs ont développé des stratégies variées tirant parti de la frise chronologique, de la liste d'extraits ou une utilisation conjointe de ces deux éléments. Il est intéressant de noter qu'un utilisateur s'est concentré sur la liste de mots-clés et sur les transcriptions, sans écouter l'audio avant d'être sûr d'avoir trouvé l'information recherchée.
Les différentes requêtes proposées aux utilisateurs ont été spécialement conçues pour couvrir un maximum de situations : pour certaines, le moteur de recherche retrouve plus ou moins d'extraits, voire aucun extrait ; d'autres nécessitent une reformulation ; certaines expriment des arguments à l'opposé de ce qui sera retrouvé dans les données ; d'autres sont délibérément à propos d'événements ambigus. Les utilisateurs, confrontés à ces requêtes, ont bien compris les particularités des requêtes proposées et ont été satisfaits par les résultats obtenus.

– **Retour sur expérience et suggestions :**

  À la fin de l'expérience, il est demandé aux utilisateurs de commenter leur session, de donner leur opinion sur le système et éventuellement de suggérer des nouveautés à intégrer au démonstrateur.
  La majorité des suggestions concernent l'ergonomie de l'interface qui n'est pas suffisamment intuitive et nécessite beaucoup de mouvements à la souris. Par exemple, la lecture audio peut être arrêtée soit en cliquant sur le lien en haut de la page, soit en faisant glisser le curseur de la frise temporelle sur une zone non couverte par des données. Un raccourci clavier ou le remplacement du lien « lecture » des extraits issus du résumé faciliteraient l'ergonomie du lecteur audio. Il a aussi été suggéré que des mouvements fluides de la frise chronologique pourraient améliorer la compréhension du processus (lorsqu'un extrait du résumé est sélectionné sur la frise ou sur l'interface). Afin d'améliorer l'intuitivité de la frise chronologique, il faudrait afficher moins d'échelles temporelles différentes (par exemple 2) avec la possibilité d'en changer la précision. Le problème des multiples échelles est qu'elles représentent plusieurs fois la même information. Plus généralement, les utilisateurs auraient apprécié des icones représentatives, des pointeurs de souris et des aides contextuelles leur montrant les possibilités de l'interface.
  La fonctionnalité jugée comme étant la plus intéressante est l'idée de supprimer complètement la liste des extraits du résumé et de représenter leur contenu au sein même de la frise chronologique. Dans cette optique, des mots-clés sont utilisés pour annoter les zones denses en information pertinente. Il faudrait également développer un moyen de représenter la transcription, sachant que l'échelle de représentation du texte par rapport au temps n'est pas la même que celle de l'audio.
  À propos du sentiment général des utilisateurs sur le prototype, ils ont eu des réponses variées et pas toujours positives. Par exemple, une utilisatrice a remarqué qu'elle ne voyait « pas l'intérêt d'une base de données audio constituée de journaux radio-diffusés alors qu'elle peut trouver les mêmes informations au format textuel sur internet, le texte étant tellement plus rapide à lire ». Un autre utilisateur a précisé qu'il avait « peur des moteurs de recherche car ils indexent de plus en plus d'informations liées à notre vie privée qui pourraient être utilisées à mauvais escient par la personne qui les contrôle ».

**Conclusions de l'enquête**

Nous avons voulu expérimenter le résumé automatique de parole dans les conditions réelles et une utilisation conjointe avec d'autres moyens de localiser l'information. Un prototype a été implémenté pour tester l'ensemble des composants dans des conditions réelles (chaîne de structuration, résumé automatique, interface). Ces composants permettent de construire un résumé de parole à partir de données audio entièrement automatiquement. Le prototype a été évalué au moyen d'une enquête préliminaire auprès d'utilisateurs. Cette enquête a révélé qu'une frise chronologique interactive était

un bon complément à la couverture thématique du résumé. Toutefois, il serait intéressant de parfaire cette approche par l'introduction de la chronologie des événements, reliée à celle des données. Une telle perspective nécessite tout de même la résolution des références temporelles pour dater les informations au niveau de la phrase. Un autre point attirant serait une représentation de la granularité thématique sous une forme similaire à celle de la frise chronologique. Couper une arborescence thématique à plusieurs hauteurs de l'arbre peut être une première idée, mais y introduire une interactivité ne sera sans doute pas trivial.

## 7.4 Conclusion

Dans ce chapitre, un prototype de résumé automatique de parole complet a été implémenté pour tester les réactions d'utilisateurs dans des conditions réelles. Ce prototype sert avant tout de support pour tester des représentations conjointes au résumé dans le but de dépasser les limites de l'approche par extraction. Ces derniers travaux représentent une ouverture vers la recherche de méthodes pour l'accès à l'information parlée pour aller plus loin qu'une simple amélioration des méthodes de structuration et de sélection de l'information.

# Chapitre 8

# Conclusion

Depuis près d'un siècle, l'humanité construit une société de l'information, centrée sur le partage des connaissances. Grâce à de nombreuses avancées techniques dans le domaine des communications, chaque individu peut potentiellement accéder au savoir de l'ensemble de la société. Internet, un des principaux vecteurs d'informations et de connaissances, semble jouer le rôle de système nerveux de notre société. Savoir accéder à l'information semble devoir devenir plus important que le Savoir. Toutefois, bien que la communication permette un transport des différents médias (texte, voix, image, vidéo), seule la connaissance écrite est réellement pérenne. En effet, l'accès à l'information n'est possible que par l'intermédiaire d'une recherche dans les descripteurs conceptuels du contenu des données. Alors que pour le texte, ces descripteurs sont directement extraits à partir des mots, les autres médias nécessitent une annotation textuelle (manuelle ou automatique) de leur contenu.

Dans ce document, nous nous sommes focalisés sur la parole, le moyen le plus utilisé pour communiquer. La parole a la caractéristique d'être éphémère : elle disparaît avec l'onde sonore qui la transporte. Les technologies actuelles permettent de transporter et conserver cette onde, et la rendent ainsi exploitable à plus long terme. Dès lors qu'il est possible de transcrire son contenu, elle acquiert les mêmes possibilités d'indexation que le texte. Ce processus a facilité l'apparition de moteurs de recherche sur le contenu parlé de flux audio. Ces derniers génèrent une liste d'extraits pour satisfaire le besoin de l'utilisateur. Cependant, une écoute exhaustive des extraits est essentiel pour déterminer leur pertinence. Cet effet est comparable à la nécessité d'explorer plusieurs pages de résultats dans un moteur de recherche Internet si l'information n'est pas trouvée dans la première page. Ce défaut, déjà présent à l'écrit, est exacerbé par le temps passé à l'écoute d'informations non pertinentes.

Nous avons proposé dans cette étude d'améliorer l'efficacité de l'accès à des bases de données parlées à l'aide d'une approche reposant sur le résumé automatique de parole. Ce concept, proche du résumé textuel, correspond à générer un résumé parlé à partir des données audio répondant au besoin d'un utilisateur. Une telle approche fait apparaître de nombreuses problématiques sur la portabilité des méthodes dédiées au texte, sur le traitement des données parlées et sur la manière de satisfaire l'utilisateur.

Le principal objectif de ce travail était de mettre en place une chaîne de traitement complète, en réutilisant les briques disponibles et en développant les éléments manquants, pour pouvoir étudier l'impact de chaque sous-système sur l'application visée.

Cette conclusion expose d'abord les résultats obtenus, puis détaille les perspectives de recherche majeures découlant de ce travail.

## 8.1 Résultats obtenus

Le chapitre 2 a tout d'abord présenté la recherche d'information à travers ses origines textuelles et ses adaptations à un média oral. Cette description s'est appuyée sur le domaine de la recherche documentaire (section 2.1) et celui du résumé automatique (section 2.2). Il est ressorti de cette étude qu'une bonne structuration était nécessaire pour passer de l'acoustique à une sémantique. Les éléments de structuration suivants ont été abordés dans le chapitre 3 : la segmentation en classes acoustiques, la segmentation en tours de parole, l'indexation en locuteurs et la transcription de la parole. Une implémentation cohérente de ces tâches sous la forme de la chaîne de transcription enrichie du LIA a ensuite été décrite et évaluée sur les données radiophoniques francophones de la campagne ESTER.

La chaîne de structuration a été complétée au chapitre 4 par deux descripteurs sémantiques et structurels majeurs : la segmentation en phrases et l'extraction d'entités nommées. La première est importante pour le résumé par extraction car la qualité des frontières de phrases est un paramètre influençant fortement la qualité de la forme du résumé. La seconde a pour but d'améliorer la projection du contenu dans un espace sémantique en faisant émerger des entités du domaine (dans notre cas : des noms de personnes, d'organisation, de lieux...) peu ambiguës hors contexte.

La solution proposée dans ce document pour la segmentation en phrases (section 4.1) repose sur une modélisation des frontières de phrase par des paramètres acoustiques et prosodiques à l'aide de *Conditional Random Fields* (CRF). Notre approche a obtenu des résultats au niveau de l'état de l'art sur les frontières de phrases ESTER avec une $F_1$-mesure de 0.68. Par rapport aux techniques de segmentation en phrases de la littérature, notre approche constitue une première voie pour modéliser conjointement des événements linguistiques et prosodiques. Cependant, cette approche ne profite pas des valeurs de confiance des processus sous-jacents (transcription, segmentation en locuteurs) et ne permet pas un apprentissage des différentes modalités sur des jeux de données disjoints.

Notre solution pour l'extraction d'entités nommées dans le flux de parole (section 4.2) intègre l'espace d'hypothèses de la transcription pour tenter d'alléger les erreurs qui pourrait être commises dans sa meilleure hypothèse. La méthode proposée repose sur une annotation du graphe d'hypothèses de transcription par des grammaires locales couplées à un modèle probabiliste génératif. Elle a obtenu les meilleurs résultats sur la tâche expérimentale d'annotation en entités nommées de la campagne ESTER. Ce résultat doit cependant être tempéré : Les écarts de performances relevés entre les différentes

approches et systèmes proposés durant ESTER restent cependant peu significatifs.

La seconde partie de ce document (chapitres 5 et suivants) a été consacrée au résumé automatique de parole dans notre cadre applicatif. Le chapitre 5 a présenté les problèmes liés à l'interactivité et à un contenu parlé. Tout d'abord, les contraintes d'interactivité nous ont poussés à modifier le modèle général de résumé automatique, en séparant les paramètres liés au besoin de l'utilisateur de ceux considérés comme indépendants de ce besoin. En effet, l'estimation de la contribution d'une phrase pour le résumé peut être plus gourmande en ressources et calculée en temps différé pour les paramètres indépendants du besoin de l'utilisateur (ou adaptés à un utilisateur moyen). Par contre, les paramètres liés à une demande spécifique requièrent une inférence rapide. Le modèle proposé repose sur l'approche *Maximal Marginal Relevance* (MMR), vue comme une solution tendant à séparer les influences du fond et de la forme (les contraintes liées à la forme n'étant pas capturées par l'expression du besoin). Toutefois, le manque de données dédiées au résumé automatique de parole dans le cadre d'ESTER (cette campagne ne proposait pas de tâche directement liée au résumé automatique), n'a pas permis une évaluation exhaustive de la méthode proposée.

Pour pallier ce problème, le chapitre 6 a proposé une évaluation indirecte du modèle proposé sur des données textuelles, lors de la campagne *Document Understanding Conference* (DUC) 2006. Elle a donné lieu à une collaboration LIA-Thales pour une soumission fusionnant 5 systèmes et obtenant des résultats état de l'art (environ $6^{e}$/34 soumissions, dont deux significativement meilleures). L'approche développée dans ce document a obtenu des performances juste en dessous de la fusion et meilleures que les 4 autres systèmes selon l'évaluation automatique Rouge-2. Obtenir des performances satisfaisantes sur des données textuelles n'offre pas la garantie d'un même niveau de performances sur la parole, essentiellement à cause des erreurs et des imprécisions ajoutées par la chaîne de structuration (découpage en phrases, erreurs de transcription...). En conséquence, nous avons évalué l'impact de tels problèmes sur la qualité des résumés proposés par notre approche en « bruitant » les données textuelles, pour simuler des données provenant d'une chaîne de structuration automatique. Les résultats obtenus ont démontré la robustesse de notre approche à de telles dégradations. Ce point reste cependant à confirmer, à cause de la mesure de performance utilisée (Rouge) et de la nature artificielle des données employées.

Dans le but de prouver la faisabilité d'un système de recherche d'information reposant sur les approches proposées dans ce document, nous avons décrit dans le chapitre 7 une implémentation complète sous la forme d'un démonstrateur. Réaliser un démonstrateur nous permettait également de prendre en compte les facteurs liés à l'interaction utilisateur-système. Cette implémentation a prouvé la faisabilité de la méthode proposée en terme de complexité de développement, de calcul et d'efficacité. Elle a facilité l'étude des différentes composantes en application réelle (données radiophoniques ESTER). Le facteur « interactivité » a été intégré sous la forme d'une présentation originale des résultats correspondant au besoin exprimé par l'utilisateur. Chaque segment audio extrait par le procédé de résumé est placé dans une frise chronologique interactive. L'utilisateur peut aisément naviguer dans les résultats et sélectionner une partie ou l'autre des segments, affinant ainsi l'expression de son besoin. Bien que l'impact de

cet élément soit difficile à évaluer, nous en avons proposé une première étude à l'aide d'une enquête utilisateur sur des aspects ergonomiques.

## 8.2 Perspectives

Nous allons explorer les perspectives de ce travail au regard des challenges listés par Zechner (2003) pour le résumé automatique de parole :

1. « *améliorer la transcription automatique du contenu parlé ;*
2. *intégrer les informations prosodiques dans les méthodes textuelles de résumé automatique ;*
3. *faire converger les méthodes de résumé issues de la recherche d'information, robustes au domaine, et celles issues de l'intelligence artificielle, dépendantes du domaine ;*
4. *rapprocher le résumé automatique et la problématique questions/réponses ;*
5. *évaluer le résumé de parole grâce à des critères adaptés.* »

Le résumé automatique de parole nécessite une amélioration des performances de la transcription automatique (1), et plus généralement de l'ensemble de la chaîne de structuration. L'étude proposée dans ce document est restreinte à un type de contenu radiodiffusé sur lequel la plupart des travaux en transcription ont été réalisés. En changeant de domaine et de conditions (moins bonnes conditions acoustiques, plus de parole spontanée, langue ou thème différents), la qualité de la structuration sera certainement détériorée à cause d'une plus faible quantité de données d'apprentissage et d'une plus grande variabilité des observations. Cette variabilité peut être compensée par la transformation des observations (débruitage), par augmentation de la diversité des situations observées en apprentissage, et par une amélioration des capacités de généralisation des algorithmes de reconnaissance. D'un autre côté, la variabilité des observations peut être amortie par l'utilisation de mesures de confiance entre les tâches de structuration. Ce procédé correspond à une extension des espaces de recherche de la meilleure hypothèse d'une étape à la suivante qui pourrait être généralisée à une fusion de tous les espaces de recherche. Une telle fusion n'est pas triviale car elle demande d'accorder des hypothèses provenant de données d'apprentissage et de cadres théoriques différents (espaces acoustiques ou linguistiques, discrets ou continus, cadre probabiliste...). L'apprentissage supervisé reste très coûteux car il requiert une annotation complète des données : il convient peut-être de se tourner vers des approches non-supervisées ou semi-supervisées comme celle de Haghighi et Klein (2006). Cette approche se fonde sur des exemples peu nombreux dont l'étiquetage est étendu à un corpus par des statistiques globales. Les performances restent bien en dessous des approches sur des corpus totalement étiquetés, mais un annotateur peut observer les résultats de l'étiquetage et réitérer l'apprentissage après ajout de nouveaux exemples représentatifs.

Plus spécifiquement, parmi les éléments ajoutés à la chaîne de traitement, la méthode de détection de frontières de phrases doit être étendue à une méthode de segmentation, au sens propre du terme. En effet, bien que les paramètres de *Conditional*

*Random Fields* dépendent d'une séquence dans son ensemble, le modèle n'est pas capable de prendre en compte la cohérence globale de la phrase, voire l'enchaînement des phrases. Cette limitation est due à l'hypothèse de Markov (localité des dépendances) qui facilite la résolution du problème. Les CRF Semi-Markoviens de Sarawagi et Cohen (2005) remettent en cause cette hypothèse avec pour conséquence une augmentation de la complexité de l'apprentissage et du décodage. Ce genre de modèle sera certainement bénéfique à l'annotation en entités nommées. Il faudra cependant instancier le modèle dans le cadre des automates à états finis ou trouver un autre moyen de prendre en compte le treillis d'hypothèses de transcription.

Nous n'avons pas abordé la détection de l'emphase par analyse prosodique (2) dans ce document car les méthodes d'extraction de prosodie sont relativement peu robustes dans un environnement acoustique contenant un fond sonore. Notre approche repose surtout sur les statistiques de cooccurrences des mots. Toutefois, la formulation présentée pour *Maximal Marginal Relevance* permet d'introduire des informations d'emphase au niveau de la granularité de la phrase. Il serait très intéressant de pouvoir employer cette caractéristique au niveau des mots et surpondérer les mots à forte intonation dans le modèle vectoriel. Néanmoins, ces considérations nécessitent de trouver un bon équilibre entre la contribution des informations prosodiques, des informations de fréquence et la contribution des autres mots de la phrase dans la normalisation du vecteur la représentant. De plus il faut différencier les nombreux mots informatifs non accentués des mots accentués non informatifs. De façon similaire, la méthode proposée dans ces travaux facilite l'intégration d'informations provenant du domaine (3) comme des indices lexicaux ou structurels. Pour compléter ce type d'intégration, il faut se poser la question de l'interaction entre les mesures de confiance issues de la structuration et les patrons linguistiques découverts.

La campagne d'évaluation DUC nous a offert la possibilité de construire un résumé par extraction grâce à la fusion des sorties de systèmes spécialisés dans le résumé automatique et de systèmes orientés questions/réponses (4). Ce rapprochement des deux domaines a été bénéfique (la fusion est meilleure que les systèmes séparés), mais la fusion proposée n'utilise que le pouvoir informatif estimé des phrases et ne prend pas en compte la redondance du résumé. Pour remédier à cette limitation, le problème de sélection des phrases doit être formulé à l'aide d'une fonction objective globale. Les équations 8.1 et 8.2 donnent une éventuelle formulation de ce problème prenant en compte la représentativité d'une phrase par rapport au besoin de l'utilisateur et aux documents d'origine, et l'inférence du contenu d'une phrase du résumé par une autre phrase du résumé. Dans ces équations, $S$ est une sélection de phrases, $s_i$ est une phrase, $b$ est le besoin, $D$ est l'ensemble des documents source.

$$\hat{S} = \underset{S}{\operatorname{argmax}}\, \text{information}(S) - \text{redondance}(S) \tag{8.1}$$

$$= \underset{S}{\operatorname{argmax}} \sum_{s_i \in S} \text{représentativité}(s_i, b, D) - \sum_{s_i \in S} \sum_{s_j \in S} \text{inférence}(s_i, s_j) \tag{8.2}$$

Cette formulation peut être assimilée à un problème de sac-à-dos quadratique en ajoutant la contrainte de longueur du résumé. Une solution exacte à ce problème étant

inaccessible lorsque le nombre de phrases augmente, il conviendra de mettre en place des heuristiques et des approximations pour traiter le problème dans un temps raisonnable.

Les difficultés d'évaluation du résumé automatique sont déjà nombreuses à l'écrit à cause de l'impossibilité de définir un résumé modèle de référence (5). Il n'existe pas à l'heure actuelle d'évaluation du résumé de parole à grande échelle, bien que NIST compte orienter prochainement *Document Understanding Conference* vers la parole. Nous avons proposé quelques critères spécifiques à ce média (Q6-Q11, section 2.2.1, page 47) pour une évaluation manuelle de la forme d'un résumé parlé. Les critères sur le fond sont-ils identiques à ceux utiles au résumé textuel ? Les mesures automatiques d'évaluation comme Rouge sont-elles adaptées à un média parlé ? Nous avons étudié la différenciation entre un résumé « lu » et un résumé « écouté ». Les deux types de résumé ont des applications très différentes et mèneront certainement à des approches différentes. Nous avons simulé l'impact d'un contenu parlé sur des données textuelles pour évaluer ces nouvelles conditions, mais une évaluation sur des données réelles sera nécessaire pour confirmer les résultats obtenus. Rouge, la mesure d'évaluation automatique dédiée à un développement rapide des systèmes de résumé, atteint ses limites, car elle n'offre aucune contrainte sur la forme du résumé et ignore sa redondance. La recherche d'une meilleure mesure d'évaluation pourrait être l'objet d'une prochaine campagne DUC, les participants proposant des systèmes capables de prédire au mieux le comportement des juges.

Un dernier point n'est pas évoqué par Zechner (2003) : les interactions avec l'utilisateur pour construire un résumé. En plus du besoin exprimé sous forme textuel, DUC introduit en 2007 les connaissances présumées de l'utilisateur au travers d'une tâche de mise à jour d'un résumé. Toutefois, ce type d'évaluation n'a pas la dynamique nécessaire pour étudier les interfaces complémentaires au résumé comme la frise chronologique présentée dans nos travaux. L'enquête préliminaire associée à cet élément devra être étendue en focalisant l'évaluation sur la perception de la structure temporelle des informations présentées par les résumés. Cette évaluation peut par exemple prendre la forme d'un questionnaire présenté à des utilisateurs ayant écouté un résumé et d'autres utilisateurs ayant écouté le résumé et ayant eu la possibilité d'interagir avec la frise chronologique. En plus de l'information temporelle, cette frise pourra être améliorée en ajoutant des indicateurs thématiques. Pour cela, la structure thématique devra être représentée conjointement à la structure temporelle tout en offrant les mêmes possibilités de navigation que la frise actuelle.

Finalement, il semble que les travaux sur le résumé automatique de parole doivent s'orienter vers un *modèle complet* ne faisant aucun *a priori* d'indépendance entre chacun des éléments du triplet acoustique, utilisateur et résumé.

# Annexe A

# Résultats DUC détaillés pour le *topic* D0641 (réchauffement climatique)

| | |
|---|---|
| Topic | D0641E |
| Titre | global warming |
| Détails | Describe theories concerning the causes and effects of global warming and arguments against these theories. |

## A.1 Résumés générés par les systèmes $S_1$ à $S_5$ et leur fusion $F_2$

### Résumé $S_1$ (MMR-LSA)

To them, the observed surface warming of about 1 degree over the with an especially sharp rise in the last quarter century is mostly or wholly natural, and there is no significant human influence on global climate. While the Global Climate Coalition questions whether global warming is happening, it advocates that companies voluntarily explore and employ new technology to reduce emissions that contribute to global warming. The changes are caused by a temperature increase, which is one of the effects of global warming. For that reason, slightly higher rates of warming and sea level rise related to warming are expected, according to a climate study written by Wigley and released June 29 by the Pew Center on Global Climate Change in Washington. The findings suggest that global warming blamed on human industry is manifesting itself not just by causing a global rise in temperatures but by favoring certain established weather patterns.

| | |
|---|---|
| Rouge-2 | 0.12444 |
| Rouge-SU4 | 0.17684 |

**Résumé $S_2$ (CORTEX)**

The changes are caused by a temperature increase, which is one of the effects of global warming. Greenhouse gas emissions blamed for global warming may cause the collapse of the West Antarctic Ice Sheet and raise the average global sea level by four to six metres, beginning as as the, a new scientific study predicted recently. While the Global Climate Coalition questions whether global warming is happening, it advocates that companies voluntarily explore and employ new technology to reduce emissions that contribute to global warming. The study believed that because carbon dioxide is one of the primary greenhouse gases contributing to the warming of the planet, the decrease of this greenhouse gas may help slow the global warming. The researchers, at the Max Planck Institute for Meteorology in Hamburg, Germany, used a computerized model of the climate system's general circulation to investigate the effects of global warming on El Nino's frequency.

| | |
|---|---|
| Rouge-2 | 0.07072 |
| Rouge-SU4 | 0.12892 |

**Résumé $S_3$ (alignement)**

Agriculture is estimated to be responsible for eight percent of the total greenhouse gas emissions blamed for global warming. South African Environment Minister Valli Moosa said on Friday that global warming resulted in the rise in incidence of malaria. While the Global Climate Coalition questions whether global warming is happening, it advocates that companies voluntarily explore and employ new technology to reduce emissions that contribute to global warming. Scientists looking for signs of global warming should spend more time scrutinizing Earth's weather circulation patterns, a new study suggests. Chinese scientists believe that global warming, particularly higher winter temperatures, will continue to the first half of the 21 st century. Scientists who argue against global warming often cite nearly 20 years of satellite readings showing slight cooling of the atmosphere two miles up. These effects open up a new way of tracking the progress of global warming without the uncertainties in simple temperature measurements. An American biologist has obtained the clearest evidence so far that global warming is forcing living species to move toward the poles and to higher altitudes. New study suggests that El Nino may help slow the global warming, though it has been blamed for many climate disorders such as floods and droughts. Under way in Buenos Aires is a giant international conference on global warming, a follow up to last December's United Nations sponsored confab in Kyoto, Japan.

| | |
|---|---|
| Rouge-2 | 0.06572 |
| Rouge-SU4 | 0.10855 |

### Résumé $S_4$ (Q/A densité)

Methane and nitrous oxide are more powerful in retention of heat and potential causes for global warming. One ton of methane will have the same warming effect as 56 tons of carbon dioxide over a period of 20 years. North said the effect of global warming in Texas and the projected corresponding 1 to 2 foot rise in the sea level could add additional stress on already tight water supplies. The increase rate of gases that cause the greenhouse effect will accelerate, so much so that the earth's warming trend will continue in the coming dozens of years. The predicted warming is expected to cause widespread climatic, ecological and economic dislocation. McKelvey said the team's ultimate aim was to show the likely effect of future global warming. It is an enormous and expensive task to study a continental ice sheet that is twice as big as Australia. Heat waves triggered off by global warming have caused thousands of additional deaths each year in major cities worldwide, and would lead tens of millions of people to the risk of malaria, the report warns. When Wentz and co author Matthias Schabel took the falling satellite effect into account, they found the lower troposphere was warming 0.13 degrees Fahrenheit per decade. It can disrupt climate around the world, producing extra rain in the southeastern United States and in Peru during the winter, while causing drought in the western Pacific.

| | |
|---|---|
| Rouge-2 | 0.07263 |
| Rouge-SU4 | 0.12170 |

### Résumé $S_5$ (Q/A compacité)

The changes are caused by a temperature increase, which is one of the effects of global warming. Hurricane forecasters doubt extreme weather events such as Mitch show the influence of global warming. I think it's agreed that the globe is about 1 degree Fahrenheit warmer, but as its effect on tropical cyclones, we haven't been able to notice any difference, said forecaster Brian Jarvinen at the U.S. National Hurricane Center in Miami. The theory is that at certain critical altitudes, roughly from 6 to 12 miles, increasing carbon dioxide has the effect of warming the tropics but cooling the polar regions. Methane and nitrous oxide are more powerful in retention of heat and potential causes for global warming. Sea levels would rise as the Antarctic ice sheet melted because of higher temperatures caused by global warming. The theory is that a warmer atmosphere holds more water vapor, increasing the warming more. This warming would touch off widespread disruptions in climate and weather and cause the global sea level to rise and flood many places. North said the effect of global warming in Texas and the projected corresponding 1 to 2 foot rise in the sea level could add additional stress on already tight water supplies. Computer models of the effects of warming point to the appearance of jets of fast moving wind between the troposphere and stratosphere, roughly 12 kilometers above the Earth's surface.

| | |
|---|---|
| Rouge-2 | 0.08556 |
| Rouge-SU4 | 0.13721 |

**Résumé $F_2$ (Fusion des 5 systèmes)**

Scientists looking for signs of global warming should spend more time scrutinizing Earth's weather circulation patterns, a new study suggests. The findings suggest that global warming blamed on human industry is manifesting itself not just by causing a global rise in temperatures but by favoring certain established weather patterns. While the study does not prove that human induced global warming is what caused the change in frequency of circulation patterns, he said, the change is consistent with it. The theory is that at certain critical altitudes, roughly from 6 to 12 miles, increasing carbon dioxide has the effect of warming the tropics but cooling the polar regions. The data suggest that global warming has caused a slowdown of the Earth at a rate of 0.56 milliseconds a century..

This warming would touch off widespread disruptions in climate and weather and cause the global sea level to rise and flood many places.

In Beijing, chinese scientists believe that global warming, particularly higher winter temperatures, will continue to the first half of the 21 st century.

In San Antonio, while the Global Climate Coalition questions whether global warming is happening, it advocates that companies voluntarily explore and employ new technology to reduce emissions that contribute to global warming.

In Washington, greenhouse gas emissions blamed for global warming may cause the collapse of the West Antarctic Ice Sheet and raise the average global sea level by four to six metres, beginning as as the, a new scientific study predicted recently.

| | |
|---|---|
| Rouge-2 | 0.08564 |
| Rouge-SU4 | 0.14662 |

## A.2 Résumés de référence

**Référence D0641E — C**

Global warming is thought to be at least partly caused by emissions of waste industrial gases like carbon dioxide, produced by burning fossil fuels like coal, oil and natural gas. These emissions trap solar radiation and produce a greenhouse effect. Methane and nitrous oxide emissions from agriculture (ruminants and manure) make up 8% of greenhouse gases. Controls on sulfur dioxide emissions reduce a balancing cooling effect.

Global warming already causes more frequent El Nino appearances, receding shorelines, longer warm seasons, and a slower earth spin. It affects habitats and threatens marine life. If emissions are not reduced, average surface temperature will rise 2-6 degrees over the next century, bringing widespread climatic, ecological and economic dislocation. Floods and droughts will increase in frequency and intensity. Melting polar

ice will cause rising sea levels and coastal flooding. Malaria will increase. Rates of habitat loss and species extinction will increase. Communities will need to adapt to new conditions.

Skeptics argue that human activities have little influence on climate. Most observed warming is due to natural causes like changes in solar radiation or the circulation of heat-bearing ocean waters. Measurements taken by satellites have found little temperature rise in the upper atmosphere. Computer models are unreliable. Any warming over the next century would be most pronounced in the winter, at night, and in sub-Arctic regions, doing little harm and creating benefits like longer growing seasons and faster plant growth. Industry argues that reducing the use of fossil fuels would cause economic harm to consumers.

### Référence D0641E — E

As early as 1968 scientists suggested that global warming might cause disintegration of the West Antarctic Ice Sheet. Greenhouse gas emissions created by burning of coal, gas and oil were believed by most atmospheric scientists to cause warming of the Earth's surface which could result in increased frequency and intensity of storms, floods, heat waves, droughts, increase in malaria zones, rise in sea levels, northward movement of some species and extinction of others.

Some scientists, however, argued that there was no real evidence of global warming and others accepted it as a fact but attributed it to natural causes rather than human activity. In 1998 a petition signed by 17,000 U.S. scientists concluded that there is no basis for believing (1) that atmospheric CO2 is causing a dangerous climb in global temperatures, (2) that greater concentrations of CO2 would be harmful, or (3) that human activity leads to global warming in the first place.

By 1999 an intermediate position emerged attributing global warming to a shift in atmospheric circulation patterns that could be caused by either natural influences such as solar radiation or human activity such as CO2 emissions.

By 2000 opponents of programs to cut back greenhouse emissions admitted that there was evidence of global warming but questioned its cause and dire consequences. Proponents of plans to control emissions to a large extent admitted that the size of the human contribution to global warming is not yet known.

### Référence D0641E — F

By 2000 most climate scientists agreed that the Earth's atmosphere warmed about 1 degree Fahrenheit in the 1990s alone. Earlier debates about conflicting evidence for atmospheric warming were resolved when weather satellite data were recalibrated. Also agreed is that atmospheric concentrations of greenhouse gases are rising – 360 parts per million, up from 315 ppm in the late 1950s ; that carbon dioxide (CO2) is nearly 30% higher than before the Industrial Revolution and the highest in the last 420,000 years.

Agriculturally produced methane and nitrous oxide make up 8% of greenhouse gases, small compared to C02. Atmospheric angular momentum correlates with atmospheric temperature rises.

All agree that we don't yet know definitively what is causing global warming. The debate is not sharply polarized, but a broad spectrum of nuanced scientific views, always open to new evidence.

The dominant scientific view, expressed by a U.N. science panel, concluded that increased man-made greenhouse gases, especially CO2 from fossil fuel burning, are significant causes of rapid global warming, which if not reduced, would raise global temperature by 2.3 to 7.2 degrees over the next 100 years. In 2000, the U.S. EPA concluded that fossil fuel burning contributes to global warming.

A significant minority – over 170,000 scientists – have questioned the accuracy of climate models predicting dangerous heat-raising effects of man-made greenhouse gases. They said naturally-produced water vapor is far more significant, and that doubling atmospheric CO2 would add only 1 degree Fahrenheit by 2100 – near the low end of opponents' estimates.

### Référence D0641E — G

There is wide agreement by scientists that the average temperature of the earth's surface has risen some 1.2 degrees over the past century. There is a wide range of opinion, however, as to what is actually causing this global warming and what its direct effects are.

Global warming is the change in climate generally thought to occur mostly from the release of methane and nitrous oxide and especially carbon dioxide gas from agriculture. The result is retention of heat in the atmosphere. However, there is a dominant view that surface warming is at least partly due to human activities ; namely, the emissions of heat-trapping waste industrial gases like carbon dioxide from burning fossil fuels like coal, oil, and natural gas. A UN scientific panel has predicted that unless these greenhouse gas emissions are reduced, the earth's average surface temperature will rise by some 2 to 6 degrees F over the next century. The panel says that the warming would touch off widespread disruptions in climate and weather causing the melting of polar ice packs. This, in turn, would cause widespread flooding and droughts threatening ecosystems that support marine life and numerous animal species. Studies also show that wind changes due to global warming are responsible for a one-third slowing down of the Earth's spin.

Skeptics say that global warming is wholly natural, that there is no significant human influence on global climate, and any future warming may be inconsequential. They do agree, however, that fossil fuel burning should be reduced.

# Annexe B

# Résumé automatique de ce document

Les recherches sur les interfaces d'accès à une base de données audio ont convergé vers l'utilisation de la même métaphore que celle permettant l'accès à des documents textuels, ou des documents indexés par des métadonnées textuelles. Le besoin utilisateur et l'information parlée sont projetés dans un espace sémantique, puis des méthodes de recherche documentaire et de résumé automatique permettent la génération d'un résumé parlé. Les premières approches de la recherche d'information dans un contenu parlé ont d'abord utilisé des techniques similaires à celle développées pour les documents textuels, appliquées à la transcription automatique du flux de parole. Un résumé automatique de parole est constitué à partir d'un flux audio parlé (entrées) et généré sous forme écrite ou parlée (sorties). L'objectif de ces travaux est de faciliter l'accès à l'information audio à l'aide du résumé de parole et les éléments de structuration présentés au chapitre précédent ne sont pas suffisants pour obtenir un résumé de qualité. Ces derniers travaux représentent une ouverture vers la recherche de méthodes pour l'accès à l'information parlée pour aller plus loin qu'une simple amélioration des méthodes de structuration et de sélection de l'information. Nous avons proposé dans cette étude d'améliorer l'efficacité de l'accès à des bases de données parlées à l'aide d'une approche reposant sur le résumé automatique de parole. Toutefois, le manque de données dédiées au résumé automatique de parole dans le cadre d'ESTER (cette campagne ne proposait pas de tâche directement liée au résumé automatique), n'a pas permis une évaluation exhaustive de la méthode proposée.

## B.1 Phrases du résumé remises dans leur contexte

233 elle se limite généralement à l'utilisation de méta-informations (étiquettes) décrivant le contenu audio, générées manuellement au moment de l'*acquisition*. **234 Les recherches sur les interfaces d'accès à une base de données audio ont convergé vers l'utilisation de la même métaphore que celle permettant l'accès à des documents**

**textuels, ou des documents indexés par des métadonnées textuelles.** 235 En effet, le processus habituel est de transcrire le contenu parlé et d'exploiter cette transcription comme un texte classique.

295 Schéma général pour une réduction de l'information parlée à l'aide du résumé automatique. **296 Le besoin utilisateur et l'information parlée sont projetés dans un espace sémantique, puis des méthodes de recherche documentaire et de résumé automatique permettent la génération d'un résumé parlé.** 297 La projection de la parole dans un espace sémantique nécessite une structuration préalable.

599 Extension à la parole **600 Les premières approches de la recherche d'information dans un contenu parlé ont d'abord utilisé des techniques similaires à celle développées pour les documents textuels, appliquées à la transcription automatique du flux de parole.** 601 La recherche documentaire audio (*Spoken Document Retrieval*, SDR) est la première formalisation de la tâche au travers de la campagne TREC 7.

1047 Spécificités de la parole **1048 Un résumé automatique de parole est constitué à partir d'un flux audio parlé (entrées) et généré sous forme écrite ou parlée (sorties).** 1049 La méthode la plus naturelle consiste à profiter des approches développées pour le texte, grâce à une étape de la transcription automatique du contenu parlé.

1374 Par la suite, le chapitre 3 a introduit les différentes tâches de structuration et leur mise en œuvre dans la chaîne de structuration Speeral. **1375 L'objectif de ces travaux est de faciliter l'accès à l'information audio à l'aide du résumé de parole et les éléments de structuration présentés au chapitre précédent ne sont pas suffisants pour obtenir un résumé de qualité.** 1376 Nous nous concentrons maintenant sur la présentation de deux compléments à la structuration pour le résumé automatique de parole.

3395 Ce prototype sert avant tout de support pour tester des représentations conjointes au résumé dans le but d'outrepasser les limites de l'approche par extraction. **3396 Ces derniers travaux représentent une ouverture vers la recherche de méthodes pour l'accès à l'information parlée pour aller plus loin qu'une simple amélioration des méthodes de structuration et de sélection de l'information.** 3397

3415 Ce défaut, déjà présent à l'écrit, est exacerbé par le temps passé à l'écoute d'informations non pertinentes. **3416 Nous avons proposé dans cette étude d'améliorer l'efficacité de l'accès à des bases de données parlées à l'aide d'une approche reposant sur le résumé automatique de parole.** 3417 Ce concept, proche du résumé textuel, correspond à générer un résumé parlé à partir des données audio répondant au besoin d'un utilisateur.

3450 Le modèle proposé repose sur l'approche *Maximal Marginal Relevance* (MMR), vue comme une solution tendant à séparer les influences du fond et de la forme (les contraintes liées à la forme n'étant pas capturées par l'expression du besoin). **3451 Toutefois, le manque de données dédiées au résumé automatique de parole dans le cadre d'ESTER (cette campagne ne proposait pas de tâche directement liée au résumé automatique), n'a pas permis une évaluation exhaustive de la méthode proposée.** 3452 Pour pallier ce problème, le chapitre 6 a proposé une évaluation indirecte du modèle

proposé sur des données textuelles, lors de la campagne *Document Understanding Conference* (DUC) 2006.

## B.2 Information sur le résumé

| Id. | Gain | Longueur | Pleins | Score |
|---|---|---|---|---|
| 3416 | 1.00 | 28 | 14 | 63.14 |
| 234 | 0.80 | 37 | 17 | 56.95 |
| 1375 | 0.71 | 37 | 18 | 49.68 |
| 296 | 0.67 | 29 | 16 | 49.39 |
| 3396 | 0.46 | 33 | 15 | 45.27 |
| 3451 | 0.74 | 38 | 21 | 45.19 |
| 1048 | 0.57 | 23 | 14 | 43.57 |
| 600 | 0.47 | 34 | 17 | 43.39 |

**TAB. *B.1:*** *Caractéristiques des phrases du résumé : l'identifiant de la phrase (Id.), le pourcentage de mots pleins apportés par la phrase (Gain), la longueur de la phrase (Longueur), le nombre de mots pleins de la phrase (Pleins), et sa similarité à la requête (Score).*

| Mot | Poids | Mot | Poids | Mot | Poids |
|---|---|---|---|---|---|
| résumé | 3.00000 | utilisateur | 0.41723 | tâche | 0.23356 |
| donnée | 3.00000 | recherche | 0.36735 | contenu | 0.22902 |
| base | 3.00000 | transcription | 0.35828 | type | 0.21995 |
| efficace | 3.00000 | entité | 0.31973 | table | 0.21542 |
| parole | 3.00000 | figure | 0.31519 | extraction | 0.20408 |
| automatique | 3.00000 | section | 0.28571 | corpus | 0.19955 |
| parlée | 3.00000 | besoin | 0.27891 | performance | 0.19955 |
| audio | 3.00000 | duc | 0.27211 | équation | 0.19048 |
| accès | 3.00000 | approche | 0.26077 | nombre | 0.18821 |
| phrase | 0.73696 | forme | 0.25397 | locuteur | 0.18141 |
| mot | 0.73243 | évaluation | 0.24943 | acoustique | 0.17687 |
| information | 0.57143 | méthode | 0.24943 | nommée | 0.17687 |
| document | 0.51927 | erreur | 0.23810 | segmentation | 0.17460 |
| système | 0.50794 | résultat | 0.23583 | align | 0.16327 |
| modèle | 0.43311 | rouge | 0.23583 | traite | 0.16100 |

**TAB. *B.2:*** *Mots formant le besoin utilisateur pour construire le résumé de ce document (45 premiers sur 556 mots apparaissant plus de 10 fois). Les mots du titre du document sont surpondérés. Les autres mots ont un poids fonction de leur fréquence dans le document.*

# Glossaire

Ce glossaire regroupe les principaux acronymes et les principales notions apparaissant tout au long du document. Les définitions sont focalisées sur les sens utilisés dans ce document.

| | |
|---|---|
| AFCP | *Association Francophone de la Communication Parlée*, organisateur de campagnes d'évaluation |
| Ajax | standard de communication |
| Alize | boîte à outils pour la reconnaissance du locuteur |
| Basic Elements | mesure d'évaluation en résumé automatique, dérivée de Rouge |
| BIO | *Begin Inside Outside*, sous-classes d'annotation de séquences |
| BIR | *Binary Information Retrieval*, modèle de recherche documentaire |
| CLEF | *Cross Language Evaluation Forum*, campagne d'évaluation |
| CORTEX | système de résumé développé au LIA |
| CRF | *Conditional Random Fields*, modèle d'étiquetage de séquences |
| CRF++ | boîte à outils pour l'étiquetage CRF |
| CSS | *Cascading Style Sheet*, standard de définition d'interfaces graphiques |
| DGA | Délégation Générale pour l'Armement, organisateur de campagnes d'évaluation |
| DOM | *Document Object Model*, standard de définition d'interfaces graphiques |
| DUC | *Document Understanding Conference*, campagne d'évaluation |
| ELDA | *European Language resource Distribution Agency*, organisateur de campagnes d'évaluation |
| ESTER | Évaluation des Systèmes de Transcription d'Émissions Radiophoniques, campagne d'évaluation |
| *F*-mesure | mesure d'évaluation d'un résultat de classification |
| Flash | interface graphique opérant sur le client |
| FSM | *Finite State Machine*, automate à états finis |
| GIS | *Generalized Iterative Scaling*, méthode d'apprentissage des modèles à maximum d'entropie |
| GMM | *Gaussian Mixture Model*, modèle probabiliste |
| GPL | *Gnu Public Licence*, licence logicielle |
| GSP | groupe Géo-Socio-Politique, type d'entité nommée |
| GVSM | *Generalized Vector Space Model*, modèle de recherche documentaire |

| | |
|---|---|
| HMM | *Hidden Markov Model*, modèle d'étiquetage de séquence |
| HTTP | *Hyper Text Transfer Protocol*, protocole de communication |
| IDF | *Inverse Document Frequency*, pondération des mots |
| IIS | *Improved Iterative Scaling*, méthode d'apprentissage des modèles à maximum d'entropie |
| LBFGS | *Limited-memory Broyden-Fletcher-Goldfarb-Shanno*, méthode d'apprentissage des modèles à maximum d'entropie |
| LDA | *Latent Dirichlet Allocation*, modèle de recherche documentaire |
| LIA | Laboratoire Informatique d'Avignon |
| LIA_SpkSeg | outil de segmentation en locuteur |
| LIA_SpkDet | outil de vérification du locuteur |
| LPCC | *Linear Predictive Cepstral Coefficient*, paramètres acoustiques |
| LSA | *Latent Semantic Analysis*, modèle de recherche documentaire |
| LSI | *Latent Semantic Indexing*, modèle de recherche documentaire |
| LVCSR | *Large Vocabulary Continuous Speech Recognition*, transcription de la parole |
| MAP | *Mean Average Precision*, mesure d'évaluation en recherche documentaire |
| MMR | *Maximal Marginal Relevance*, modèle de résumé automatique |
| MMR-LSA | système de résumé présenté dans ces travaux |
| MFCC | *Mel Frequency Ceptstrum Coefficents*, paramètres acoustiques |
| NIST | *National Institute of Standards and Technology*, organisateur de campagnes d'évaluations |
| OOV | *Out-Of-Vocabulary*, mot hors-vocabulaire |
| PLP | *Perceptual Linear Predictive*, paramètres acoustiques |
| Red5 | serveur opensource de distribution de contenu multimédia |
| Rouge | mesure d'évaluation de la qualité d'un résumé |
| SDR | *Spoken Document Retrieval*, recherche d'information parlée |
| SEE | *Summarization Evaluation Environment*, interface d'évaluation du résumé automatique |
| SER | *Slot Error Rate*, taux d'erreur en extraction des entités nommées |
| Speeral | moteur de transcription automatique de la parole |
| SVD | *Singular Value Decomposition*, décomposition en valeurs singulières |
| SVM | *Support Vector Machines*, modèle de classification supervisée |
| TDT | *Topic Detection and Tracking*, campagne d'évaluation |
| TREC | *Text REtrieval Conference*, campagne d'évaluation |
| TVSM | *Topic-based Vector Space Model*, modèle de recherche documentaire |
| VSM | *Vector Space Model*, modèle de recherche documentaire |
| WER | *Word Error Rate*, mesure d'évaluation de la transcription automatique |
| WFST | *Weighted Finite State Transducer*, automate à états finis transducteur pondéré |

# Liste des illustrations

# Liste des tableaux

# Bibliographie

(Abdillahi et al., 2006) N. Abdillahi, P. Nocéra, et J.-F. Bonastre, 2006. Towards Automatic Transcription of Somali Language. Dans les actes de *Language Resource and Evaluation Conference (LREC)*.

(Acero et al., 2004) A. Acero, Y.-Y. Wang, et K. Wang, 2004. A Semantically Structured Language Model. Dans les actes de *Special Workshop in Maui (SWIM)*.

(Adams, 1979) D. N. Adams, 1979. *The Hitchhiker's Guide to the Galaxy*. Pan Books.

(Agirre et Edmonds, 2006) E. Agirre et P. Edmonds, 2006. *Word Sense Disambiguation : Algorithms and Applications*. Springer.

(Alfonseca et al., 2004) E. Alfonseca, J. M. Guirao, et A. Moreno-Sandoval, 2004. Description of the UAM system for generating very short summaries at DUC-2004. Dans les actes de *Document Understanding Conference (DUC)*.

(Allan, 2002) J. Allan, 2002. *Topic Detection and Tracking : Event-Based Information Organization*. Kluwer.

(Allauzen, 2003) A. Allauzen, 2003. *Modélisation linguistique pour l'indexation automatique de documents audiovisuels*. Thèse de Doctorat, Université Paris Sud.

(Arons, 1993) B. Arons, 1993. SpeechSkimmer Interactively Skimming Recorded Speech. Dans les actes de *Symposium on User Interface Software and Technology (UIST)*.

(Atrey et al., 2006) P. K. Atrey, N. C. Maddage, et M. S. Kankanhalli, 2006. Audio Based Event Detection for Multimedia Surveillance. Dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

(Baeza-Yates et Ribeiro-Neto, 1999) R. A. Baeza-Yates et B. A. Ribeiro-Neto, 1999. *Modern Information Retrieval*. Addison-Wesley Harlow.

(Banko et Vanderwende, 2004) M. Banko et L. Vanderwende, 2004. Using n-grams to understand the nature of summaries. Dans les actes de *North American chapter of the Association for Computational Linguistics (NAACL)*.

(Bar-Hillel, 1958) Y. Bar-Hillel, 1958. The mechanization of literature searching. *Mechanization of Thought Processes 10*, 4–8.

(Barzilay et al., 2000) R. Barzilay, M. Collins, J. Hirschberg, et S. Whittaker, 2000. The Rules Behind Roles Identifying Speaker Role in Radio Broadcasts. Dans les actes de *Natinal Conference on Artificial Intelligence (AAAI)*.

(Baum et al., 1970) L. E. Baum, T. Petrie, G. Soules, et N. Weiss, 1970. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *Annals of Mathematics and Statistics 41*(1), 164–171.

(Bazzi et Glass, 2000) I. Bazzi et J. Glass, 2000. Modeling Out-Of-Vocabulary Words for Robust Speech Recognition. Dans les actes de *International Conference on Spoken Language Processing (ICSLP)*.

(Becker et Kuropka, 2003) J. Becker et D. Kuropka, 2003. Topic-based vector space model. Dans les actes de *Business Information Systems (BIS)*, 7–12.

(Bellot, 2000) P. Bellot, 2000. *Méthodes de Classification et de Segmentation Locales Non Supervisées pour la Recherche Documentaire*. Thèse de Doctorat, Université d'Avignon.

(Bellot et al., 2003) P. Bellot, E. Crestan, M. El-Bèze, L. Gillard, et C. de Loupy, 2003. Coupling Named Entity Recognition, Vector-Space Model and Knowledge Bases for TREC-11 Question-Answering Track. Dans les actes de *Text REtrieval Conference (TREC)*.

(Bellot et El-Bèze, 2000) P. Bellot et M. El-Bèze, 2000. Classification locale non supervisée pour la recherche documentaire. *Traitement Automatique des Langues (TAL), été 2*(41), 335–365.

(Bender et al., 2003) O. Bender, F. Och, et H. Ney, 2003. Maximum Entropy Models for Named Entity Recognition. Dans les actes de *Conference on Natural Language Learning (CoNLL)*, 148–151.

(Berger et al., 1996) A. L. Berger, S. D. Pietra, et V. J. D. Pietra, 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics 22*(1), 39–71.

(Bergler et al., 2004) S. Bergler, R. Witte, Z. Li, M. Khalifé, Y. Chen, M. Doandes, et A. Andreevskaia, 2004. Multi-ERSS and ERSS 2004. Dans les actes de *Document Understanding Conference (DUC)*.

(Biber, 1991) D. Biber, 1991. *Variation Across Speech and Writing*. Cambridge University Press.

(Bikel et al., 1997) D. Bikel, S. Miller, R. Schwartz, et R. Weischedel, 1997. Nymble : a High-Performance Learning Name-Finder. Dans les actes de *Applied Natural Language Conferences (ANLC)*, 194–201.

(Blair-Goldensohni et al., 2004) S. Blair-Goldensohni, D. Evans, V. Hatzivassiloglou, K. McKeown, A. Nenkova, R. Passonneau, B. Schiffman, A. Schlaikjer, A. Siddharthan, et S. Sieglman, 2004. Columbia University at DUC 2004. Dans les actes de *Document Understanding Conference (DUC)*.

(Blei et al., 2003) D. Blei, A. Ng, et M. Jordan, 2003. Latent dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

(Brandenburg, 1999) K. Brandenburg, 1999. MP3 and AAC explained. Dans les actes de *International Conference on High-Quality Audio Coding*.

(Braschler et Peters, 2004) M. Braschler et C. Peters, 2004. Cross-Language Evaluation Forum : Objectives, Results, Achievements. *Information Retrieval 7*(1), 7–31.

(Brown et al., 2001) E. W. Brown, S. Srinivasan, A. Coden, D. Ponceleon, J. W. Cooper, et A. Amir, 2001. Toward Speech as a Knowledge Resource. *IBM Systems Journal 40*, 985–1001.

(Buckley et al., 1994) C. Buckley, G. Salton, J. Allan, et A. Singhal, 1994. Automatic Query Expansion Using SMART : TREC 3. Dans les actes de *Text REtrieval Conference (TREC)*.

(Byrne et al., 2004) W. Byrne, D. Doermann, M. Franz, S. Gutsman, J. Hajic, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, et W. Zhu, 2004. Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Transactions on Speech and Audio Processing 12*(4), 420–435.

(Béchet et al., 2004) F. Béchet, A. L. Gorin, J. H. Wright, et D. H. Tür, 2004. Detecting and Extracting Named Entities from Spontaneous Speech in a Mixed-Initiative Spoken Dialogue Context How May I Help You. *Speech Communiation, Elsevier 42*, 207–225.

(Callan, 2000) J. Callan, 2000. *Distributed Information Retrieval*, Chapter 5, 127–150. Kluwer Academic Publishers.

(Callan et al., 1992) J. P. Callan, W. B. Croft, et S. M. Harding, 1992. The INQUERY Retrieval System. Dans les actes de *Database and Expert Systems Applications (DEXA)*, 78–83.

(Carmel et al., 2005) D. Carmel, E. Yom-Tov, et I. Soboroff, 2005. SIGIR workshop report : predicting query difficulty-methods and applications. Dans les actes de *ACM Special Interest Group on Information Retrieval (SIGIR)*, 25–28. ACM Press New York, NY, USA.

(Chen et al., 2004) J. Chen, H. Ge, Y. Wu, et S. Jiang, 2004. UNT at TREC 2004 : Question Answering Combining Multiple Evidences. Dans les actes de *Text REtrieval Conference (TREC)*, 695–702.

(Chen, 1999) S. Chen, 1999. A gaussian prior for smoothing maximum entropy models. Technical Report CMUCS 99-108, Carnegie Mellon University.

(Chieu et Ng, 2002) H. Chieu et H. Ng, 2002. Named Entity Recognition : a Maximum Entropy Approach Using Global Information. Dans les actes de *International Conference On Computational Linguistics (ACL/Coling)*, 1–7. Association for Computational Linguistics Morristown, NJ, USA.

(Christensen et al., 2003) H. Christensen, Y. Gotoh, B. Kolluru, et S. Renals, 2003. Are Extractive Text Summarisation Techniques Portable to Broadcast News ? Dans les actes de *Automatic Speech Recognition and Understanding (ASRU)*, 489–494.

(Chuang et Chien, 2004) S.-L. Chuang et L.-F. Chien, 2004. A Practical Web-Based Approach to Generating Topic Hierarchy for Text Segments. Dans les actes de *Conference on Information and Knowledge Management (CIKM)*, New York, NY, USA, 127–136. ACM Press.

(Church et Gale, 1995) K. Church et W. Gale, 1995. Poisson mixtures. *Natural Language Engineering 1*(2), 163–190.

(Coden et al., 2002) A. R. Coden, E. Brown, et S. Srinivasan, 2002. Acm sigir 2001 workshop "information retrieval techniques for speech applications". Dans les actes de *ACM Special Interest Group on Information Retrieval (SIGIR)*, New York, NY, USA, 10–13. ACM Press.

(Collins et Singer, 1999) M. Collins et Y. Singer, 1999. Unsupervised models for named entity classification. Dans les actes de *Empirical Methods in Natural Language Processing (EMNLP)*, 189–196.

(Copeck et Szpakowicz, 2004) T. Copeck et S. Szpakowicz, 2004. Vocabulary Agreement Among Model Summaries and Source Documents. Dans les actes de *Document Understanding Conference (DUC)*.

(Crestani et van Rijsbergen, 1995) F. Crestani et C. van Rijsbergen, 1995. Information Retrieval by Logical Imaging. *Journal of Documentation 51*(1), 3–17.

(Croft, 2000) W. Croft, 2000. Combining approaches to information retrieval. *Advances in Information Retrieval 7*, 1–36.

(Darroch et Ratcliff, 1972) J. Darroch et D. Ratcliff, 1972. Generalized Iterative Scaling for Log-Linear Models. *Annals of Mathematical Statistics 43*(5), 1470–1480.

(Daumé III et Marcu, 2001) H. Daumé III et D. Marcu, 2001. A Noisy-Channel Model for Document Compression. Dans les actes de *International Conference On Computational Linguistics (ACL/Coling)*, Morristown, NJ, USA, 449–456. Association for Computational Linguistics.

(Daumé III et Marcu, 2004) H. Daumé III et D. Marcu, 2004. A Tree-Position Kernel for Document Compression. Dans les actes de *Document Understanding Conference (DUC)*.

(Davis et Mermelstein, 1980) S. B. Davis et P. Mermelstein, 1980. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuous Spoken Sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing 28*, 357–366.

(de Loupy et al., 1998) C. de Loupy, P. Bellot, M. El-Bèze, et P. Marteau, 1998. Query Expansion and Classification of Retrieved Documents. Dans les actes de *Text REtrieval Conference (TREC)*, 382–389.

(De Mori, 1998) R. De Mori, 1998. *Spoken dialogues with computers*. Academic Press.

(Deerwester et al., 1990) S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, et R. Harshman, 1990. Indexing by Latent Semantic Analysis. *Journal of the Society for Information Science 41*(6), 391–407.

(Della Pietra et al., 1997) S. Della Pietra, V. Della Pietra, et J. Lafferty, 1997. Inducing Features of Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence 19*(4), 380–393.

(Dempster et al., 1977) A. Dempster, N. Laird, et D. Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society 39*(1), 1–38.

(Deng et al., 2003) Y. Deng, M. Mahajan, et A. Acero, 2003. Estimating Speech Recognition Error Rate without Acoustic Test Data. Dans les actes de *European Conference on Speech Communication and Technology (EUROSPEECH)*, 929–932.

(Dewey, 1876) M. Dewey, 1876. *A Classification and Subject Index for Cataloguing and Arranging the Books and Pamphlets of a Library*. disponible en ligne[1].

(Donald, 1991) M. Donald, 1991. *Origins of the modern mind : three stages in the evolution of culture and cognition*. Harvard University Press Cambridge, Mass.

(Doran et al., 2004) W. Doran, N. Stokes, E. Newman, J. Dunnion, J. Carthy, et F. Toolan, 2004. News Story Gisting at University College Dublin. Dans les actes de *Document Understanding Conference (DUC)*.

(Dufaux et al., 2000) A. Dufaux, L. Besacier, M. Ansorge, et F. Pellandini, 2000. Automatic Sound Detection and Recognition for Noisy Environment. Dans les actes de *European Signal Processing Conference (EUSIPCO)*.

(Edmundson, 1969) H. Edmundson, 1969. New Methods in Automatic Extracting. *Journal of the ACM (JACM) 16*(2), 264–285.

(Ericsson et Simon, 1993) K. A. Ericsson et H. A. Simon, 1993. *Protocol analysis ; Verbal reports as data*. Bradford books/MIT Press.

(Erkan et Radev, 2004) G. Erkan et D. R. Radev, 2004. The University of Michigan at DUC 2004. Dans les actes de *Document Understanding Conference (DUC)*.

(Falavigna et al., 2002) D. Falavigna, R. Gretter, et G. Riccardi, 2002. Acoustic and Word Lattice Based Algorithms for Confidence Scores. Dans les actes de *International Conference on Spoken Language Processing (ICSLP)*, 1621–1624.

(Favre, 2003) B. Favre, 2003. Indexation Multimédia : Caractérisation du Déséquilibre entre les Modalités Texte et Parole. Mémoire de Master, Université d'Avignon.

[1] http://www.gutenberg.org/etext/12513, visité en novembre 2006.

(Favre et al., 2006) B. Favre, F. Bechet, P. Bellot, F. Boudin, M. El-Beze, L. Gillard, G. Lapalme, et J.-M. Torres-Moreno, 2006. The LIA-Thales summarization system at DUC-2006. Dans les actes de *Document Understanding Conference (DUC)*, 131–138.

(Favre et al., 2005) B. Favre, F. Béchet, et P. Nocéra, 2005. Robust Named Entity Extraction from Large Spoken Archives. Dans les actes de *Human Language Technologies (HTL/EMNLP)*.

(Fleiss, 1971) J. Fleiss, 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin 76*(5), 378–382.

(Floridi, 2005) L. Floridi, 2005. Is Information Meaningful Data ? *Philosophy and Phenomenological Research 70*(2), 351–370.

(Fredouille et al., 2004) C. Fredouille, D. Matrouf, G. Linares, et P. Nocera, 2004. Segmentation en macro-classes acoustiques d'émissions radiophoniques dans le cadre d'ESTER. Dans les actes de *Journées d'Étude sur la Parole (JEP)*, 225–228.

(Furui et al., 2004) S. Furui, T. Kikuchi, Y. Shinnaka, et C. Hori, 2004. Speech-To-Text and Speech-To-Speech Summarization of Spontaneous Speech. *IEEE Transactions on Speech and Audio Processing 12*(4), 401–408.

(Galibert et Rosset, 2005) G. Galibert et S. Rosset, 2005. Ritel : dialogue homme-machine à domaine ouvert. Dans les actes de *Traitement Automatique des Langues (TALN)*, Volume 1, 439.

(Galliano et al., 2005) S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J. Bonastre, et G. Gravier, 2005. The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News. Dans les actes de *European Conference on Speech Communication and Technology (EUROSPEECH)*.

(Garofolo et al., 1999) J. Garofolo, C. Auzanne, et E. Voorhees, 1999. The TREC spoken document retrieval track : A success story. Dans les actes de *Text REtrieval Conference (TREC)*, Volume 8, 16–19.

(Gilbert et Zhong, 2003) J. E. Gilbert et Y. Zhong, 2003. Speech User Interfaces for Information Retrieval. Dans les actes de *Conference on Information and Knowledge Management (CIKM)*, New York, NY, USA, 77–82. ACM Press.

(Gillard et al., 2005) L. Gillard, P. Bellot, et M. El-Bèze, 2005. Le LIA à EQueR. Dans les actes de *Traitement Automatique des Langues (TALN)*, Volume 2, 81–84.

(Godfrey et al., 1992) J. Godfrey, E. Holliman, et J. McDaniel, 1992. SWITCHBOARD : Telephone speech corpus for research and development. Dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 517–520.

(Goldstein et al., 2000) J. Goldstein, V. Mittal, J. Carbonell, et J. Callan, 2000. Creating and Evaluation Multi-Document Sentence Extract Summaries. Dans les actes de *Conference on Information and Knowledge Management (CIKM)*.

(Gong et Liu, 2001) Y. Gong et X. Liu, 2001. Generic Text Summarization using Relevance Measure and Latent Semantic Analysis. Dans les actes de *ACM Special Interest Group on Information Retrieval (SIGIR)*, 19–25. ACM Press New York, NY, USA.

(Gonzalo et al., 1998) J. Gonzalo, F. Verdejo, I. Chugur, et J. Cigarran, 1998. Indexing with WordNet synsets can improve text retrieval. Dans les actes de *International Conference On Computational Linguistics (ACL/Coling)*.

(Goode, 2002) B. Goode, 2002. Voice over Internet protocol (VoIP). *IEEE Standard 90*(9), 1495–1517.

(Gotoh et Renals, 1999) Y. Gotoh et S. Renals, 1999. Statistical Annotation of Named Entities in Spoken Audio. Dans les actes de *ESCA Workshop : Accessing Information in Spoken Audio*, 43–48.

(Gotoh et Renals, 2000) Y. Gotoh et S. Renals, 2000. Sentence Boundary Detection in Broadcast Speech Transcripts. Dans les actes de *Automatic Speech Recognition : Challenges for the new Millennium (ASR)*.

(Graff, 2003) D. Graff, 2003. English Gigaword. Linguistic Data Consortium, No. LDC2003T05[2].

(Grishman, 1998) R. Grishman, 1998. Information Extraction and Speech Recognition. Dans les actes de *DARPA Broadcast News Transcription and Understanding Workshop*.

(Haghighi et Klein, 2006) A. Haghighi et D. Klein, 2006. Prototype-Driven Learning for Sequence Models. Dans les actes de *Human Language Technologies (HTL/EMNLP)*, 320–327.

(Hansen et al., 2004) J. H. Hansen, R. Huang, P. Mangalath, B. Zhou, M. Seadle, et J. R. Deller Jr, 2004. SPEECHFIND : Spoken Document Retrieval for a National Gallery of the Spoken Word. Dans les actes de *Nordic Signal Processing Symposium (NORSIG)*.

(Hassel et Sjöbergh, 2006) M. Hassel et J. Sjöbergh, 2006. Towards Holistic Summarization — Selecting Summaries, Not Sentences. Dans les actes de *Language Resource and Evaluation Conference (LREC)*.

(Haton et al., 2006) J.-P. Haton, C. Cerisara, D. Fohr, Y. Laprie, et K. Smaïli, 2006. *Reconnaissance automatique de la parole*. Dunod.

(Hermansky, 1990) H. Hermansky, 1990. Perceptual linear predictive (PLP) analysis of speech. *Journal of Acoustical Society of America 4*(87), 1738–1752.

(Hirschberg et al., 2001) J. Hirschberg, M. Bacchiani, D. Hindle, P. Isenhour, A. Rosenberg, L. Stark, L. Stead, S. Whittaker, et G. Zamchick, 2001. SCANMail : Browsing and Searching Speech Data by Content. Dans les actes de *European Conference on Speech Communication and Technology (EUROSPEECH)*, 2377–2380.

[2] http://www.ldc.upenn.edu/catalog/, visité en novembre 2006.

(Hirschberg et al., 1999) J. Hirschberg, S. Whittaker, D. Hindle, F. Pereira, et A. Singhal, 1999. Finding information in audio : A new paradigm for audio browsing/retrieval. Dans les actes de *ESCA Workshop : Accessing Information in Spoken Audio*, 117–122.

(Hofmann, 2000) T. Hofmann, 2000. Learning the similarity of documents : An information-geometric approach to document retrieval and categorization. *Advances in Neural Information Processing Systems 12*, 914–920.

(Hori et al., 2003a) C. Hori, T. Hori, et S. Furui, 2003a. Evaluation Method for Automatic Speech Summarization. Dans les actes de *European Conference on Speech Communication and Technology (EUROSPEECH)*.

(Hori et al., 2003c) C. Hori, T. Hori, H. Isozaki, E. Maeda, S. Katagiri, et S. Furui, 2003c. Study on Spoken Interactive Open Domain Question Answering. Dans les actes de *Spontaneous Speech Processing and Recognition (SSPR)*, 111–113.

(Hori et al., 2003b) T. Hori, C. Hori, et Y. Minami, 2003b. Speech Summarization using Weighted Finite-State Transducers. Dans les actes de *European Conference on Speech Communication and Technology (EUROSPEECH)*.

(Horlock et King, 2003) J. Horlock et S. King, 2003. Named Entity Extraction from Word Lattices. Dans les actes de *European Conference on Speech Communication and Technology (EUROSPEECH)*.

(Hovy et al., 2005) E. Hovy, C. Lin, et L. Zhou, 2005. A BE-based Multi-document Summarizer with Sentence Compression. Dans les actes de *Multilingual Summarization Evaluation (MSE)*.

(Huang et al., 2001) X. Huang, A. Acero, et H. W. Hon, 2001. *Spoken Language Processing : A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, Upper Saddle River, N.J.

(Ihadjadene, 2004) M. Ihadjadene, 2004. *Les systèmes de recherche d'informations*. Lavoisier.

(Istrate et al., 2005) D. Istrate, N. Scheffer, C. Fredouille, et J. Bonastre, 2005. Broadcast News Speaker Tracking for ESTER 2005 Campaign. Dans les actes de *European Conference on Speech Communication and Technology (EUROSPEECH)*, 2445–2448.

(Jenhani, 2006) O. Jenhani, 2006. *WebSum : Système de résumé automatique de réponses des moteurs de recherche*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France.

(Jing, 2002) H. Jing, 2002. Using hidden Markov modeling to decompose human-written summaries. *Computational Linguistics 28*(4), 527–543.

(Johnson et al., 2000) S. E. Johnson, P. Jourlin, K. Spärck-Jones, et P. C. Wooland, 2000. Audio Indexing and Retrieval of Complete Broadcast News Shows. Dans les actes de *Recherche d'Information Assistée par Ordinateur (RIAO)*.

(Kando, 2005) N. Kando, 2005. Overview of the Fifth NTCIR Workshop. Dans les actes de *National Institute of Informatics Test Collections for Information Retrieval (NTCIR)*.

(Kanerva et al., 2000) P. Kanerva, J. Kristofersson, et A. Holst, 2000. Random Indexing of Text Samples for Latent Semantic Analysis. Dans les actes de *Annual Conference of the Cognitive Science Society (CogSci)*, Volume 1036.

(Kazama et al., 2002) J. Kazama, T. Makino, Y. Ohta, et J. Tsujii, 2002. Tuning Support Vector Machines for Biomedical Named Entity Recognition. Dans les actes de *Workshop on Natural Language Processing in the Biomedical Domain (ACL)*, 1–8.

(Kim et al., 2004) J. Kim, S. Schwarm, et M. Ostendorf, 2004. Detecting Structural Metadata With Decision Trees and Transformation-Based Learning. Dans les actes de *Human Language Technologies (HLT/NAACL)*, 137–144.

(Kobayashi et Schmandlt, 1997) M. Kobayashi et C. Schmandlt, 1997. Dynamic Soundscape Mapping Time to Space for Audio Browsing. Dans les actes de *MIT project CHI97*.

(Kubala et al., 1998) F. Kubala, R. Schwartz, R. Stone, et R. Weischedel, 1998. Named Entity Extraction from Speech. Dans les actes de *DARPA Broadcast News Transcription and Understanding Workshop*.

(Kudo et al., 2004) T. Kudo, K. Yamamoto, et Y. Matsumoto, 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. Dans les actes de *Empirical Methods in Natural Language Processing (EMNLP)*.

(Kupiec et al., 1995) J. Kupiec, J. Pedersen, et F. Chen, 1995. A Trainable Document Summarizer. Dans les actes de *ACM Special Interest Group on Information Retrieval (SIGIR)*, 68–73. ACM Press New York, NY, USA.

(Kuropka, 2004) D. Kuropka, 2004. Modelle zur Repräsentation natürlichsprachlicher Dokumente–Information-Filtering und-Retrieval mit relationalen Datenbanken. *Advances in Information Systems and Management Science 10*.

(Lacatusu et al., 2006) F. Lacatusu, A. Hickl, K. Roberts, Y. Shi, J. Bensley, B. Rink, P. Wang, et L. Taylor, 2006. LCC's GISTexter at DUC 2006 : Multi-Strategy Multi-Document Summarization. Dans les actes de *Document Understanding Conference (DUC)*.

(Lafferty et al., 2001) J. Lafferty, A. McCallum, et F. Pereira, 2001. Conditional random fields : Probabilistic models for segmenting and labeling sequence data. Dans les actes de *International Conference on Machine Learning*, 282–289.

(Lappin et Leass, 1994) S. Lappin et H. Leass, 1994. An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics 20*(4), 535–561.

(Laukka, 2004) P. Laukka, 2004. *Vocal Expression of Emotion*. Thèse de Doctorat, Uppsala university.

(Lavrenko, 2002) V. Lavrenko, 2002. *A Generative Theory of Relevance*. Thèse de Doctorat, University of Massachusetts.

(Le Meur et al., 2004) C. Le Meur, S. Galliano, et E. Geoffrois, 2004. Guide d'Annotation en Entités Nommées ESTER.

(Lin, 2004) C. Lin, 2004. Rouge : A Package for Automatic Evaluation of Summaries. Dans les actes de *Workshop on Text Summarization Branches Out*, 74–81.

(Lin et Hovy, 2003) C. Lin et E. Hovy, 2003. The potential and limitations of automatic sentence extraction for summarization. Dans les actes de *Human Language Technologies (HLT/NAACL)*, 73–80. Association for Computational Linguistics Morristown, NJ, USA.

(Lindman et al., 1976) H. Lindman, R. Sner, W. Ziegler, J. Jackson, H. Linstone, et M. Turoff, 1976. Analysis of Variance in Complex Experimental Designs. *Technometrics 18*(3), 361–362.

(Liu et Nocedal, 1989) D. Liu et J. Nocedal, 1989. On the Limited Memory BFGS Method for Large Scale Optimization. *Mathematical Programming 45*(1), 503–528.

(Liu et al., 2005) Y. Liu, A. Stolcke, E. Shriberg, et M. Harper, 2005. Using Conditional Random Fields for Sentence Boundary Detection in Speech. Dans les actes de *Association for Computational Lingusitics (ACL)*, Ann Arbor, Michigan, 451–458. ACL.

(Luhn, 1958) H. Luhn, 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development 2*(2), 159–165.

(Lévy et al., 2004) C. Lévy, G. Linarès, P. Nocéra, et J.-F. Bonastre, 2004. Reducing Computational and Memory Cost for Cellular Phone Embedded Speech Recognition System. Dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

(Malouf, 2002) R. Malouf, 2002. A Comparison of Algorithms for Maximum Entropy Parameter Estimation. Dans les actes de *International Conference On Computational Linguistics (ACL/Coling)*, 1–7. Association for Computational Linguistics Morristown, NJ, USA.

(Mani, 2001) I. Mani, 2001. *Automatic Summarization*. Amsterdam/Philadephia : John Benjamins Publishing Company.

(Mani et al., 2002) I. Mani, G. Klein, D. House, L. Hirschman, T. Firmin, et B. Sundheim, 2002. Summac : a Text Summarization Evaluation. *Natural Language Engineering 8*(01), 43–68.

(Manjunath et al., 2002) B. Manjunath, P. Salembier, et T. Sikora, 2002. *Introduction to MPEG-7 : Multimedia Content Description Interface*. Wiley & Sons.

(Markert et Nissim, 2002) K. Markert et M. Nissim, 2002. Metonymy Resolution as a Classification Task. Dans les actes de *Empirical Methods in Natural Language Processing (EMNLP)*, 204–213. Association for Computational Linguistics Morristown, NJ, USA.

(Maskey et Hirschberg, 2006) S. Maskey et J. Hirschberg, 2006. Summarizing Speech Without Text Using Hidden Markov Models. Dans les actes de *Human Language Technologies (HTL/EMNLP)*.

(Mauclair et al., 2006) J. Mauclair, S. Estève, Y. andPetit-Renaud, et P. Deléglise, 2006. Automatic Detection of Well Recognized Words in Automatic Speech Transcription. Dans les actes de *Language Resource and Evaluation Conference (LREC)*.

(McCallum et Li, 2003) A. McCallum et W. Li, 2003. Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. Dans les actes de *Conference on Natural Language Learning (CoNLL)*.

(McKeown et al., 2005) K. McKeown, J. Hirschberg, M. Galley, et S. Maskey, 2005. From Text to Speech Summarization. *IEEE International Conference on Acoustics, Speech, and Signal Processing 5*, 997–1000.

(McKeown et Radev, 1995) K. McKeown et D. R. Radev, 1995. Generating Summaries of Multiple News Articles. Dans les actes de *ACM Special Interest Group on Information Retrieval (SIGIR)*, New York, NY, USA, 74–82. ACM Press.

(Miller, 1995) G. Miller, 1995. WordNet : A Lexical Database for English. *Communications of the ACM 38*, 11–39.

(Miller et al., 2000) S. Miller, H. Fox, L. Ramshaw, et R. Weischedel, 2000. A novel use of statistical parsing to extract information from text. Dans les actes de *North American chapter of the Association for Computational Linguistics (NAACL)*, 226–233.

(Minel, 2004) J.-L. Minel, 2004. *L'évaluation des systèmes de traitement de l'information*, Chapter L'évaluation des systèmes de résumé automatique, 171–184. Éditions Hermès, Paris.

(Mohri et al., 2002) M. Mohri, F. Pereira, et M. Riley, 2002. Weighted Finite-State Transducers in Speech Recognition. *Computer, Speech and Language 16*(1), 69–88.

(Mooers, 1950) C. Mooers, 1950. *The Theory of Digital Handling of Non-numerical Information and Its Implications to Machine Economics.* Zator Co.

(Mori et Sasaki, 2002) T. Mori et T. Sasaki, 2002. Information Gain Ratio meets Maximal Marginal Relevance. Dans les actes de *National Institute of Informatics Test Collections for Information Retrieval (NTCIR)*.

(Mrozinski et al., 2006) J. Mrozinski, E. W. D. Whittaker, P. Chatain, et S. Furui, 2006. Automatic Sentence Segmentation of Speech for Automatic Summarization. Dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Volume 1, 981–984.

(Murray et al., 2005) G. Murray, S. Renals, et J. Carletta, 2005. Extractive summarization of meeting recordings. Dans les actes de *European Conference on Speech Communication and Technology (EUROSPEECH)*.

(Nam et Tewfik, 1999) J. Nam et A. H. Tewfik, 1999. Dynamic Video Summarization and Visualization. Dans les actes de *ACM international conference on Multimedia (MULTIMEDIA)*, New York, NY, USA, 53–56. ACM Press.

(Nenkova et Passonneau, 2004) A. Nenkova et R. Passonneau, 2004. Evaluating Content Selection in Summarization : the Pyramid Method. Dans les actes de *Human Language Technologies (HLT/NAACL)*.

(Neumann et Sacaleanu, 2004) G. Neumann et B. Sacaleanu, 2004. Experiments on Robust NL Question Interpretation and Multi-layered Document Annotation for a Cross-Language Question/Answering System. Dans les actes de *Cross Language Evaluation Forum (CLEF)*. Springer.

(Nielsen, 1993) J. Nielsen, 1993. *Usability Engineering*. AP Professional.

(Nobata et Sekine, 2004) C. Nobata et S. Sekine, 2004. CRL NYU Summarization System at DUC-2004. Dans les actes de *Document Understanding Conference (DUC)*.

(Nocéra et al., 2004) P. Nocéra, C. Fredouille, G. Linares, D. Matrouf, S. Meignier, J.-F. Bonastre, D. Massonié, et F. Béchet, 2004. The LIA's French Broadcast News Transcription System. Dans les actes de *Special Workshop in Maui (SWIM)*.

(Over et Yen, 2003) P. Over et J. Yen, 2003. An Introduction to DUC 2003 : Intrinsic Evaluation of Generic News Text Summarization Systems. Dans les actes de *Document Understanding Conference (DUC)*.

(Paice, 1990) C. D. Paice, 1990. Constructing Literature Abstracts by Computer : Techniques and Prospects. *Information Processing and Management : an International Journal 26*(1), 171–186.

(Ponte et Croft, 1998) J. M. Ponte et W. B. Croft, 1998. A Language Modeling Approach to Information Retrieval. Dans les actes de *ACM Special Interest Group on Information Retrieval (SIGIR)*, 275–281.

(Porter, 1980) M. Porter, 1980. An algorithm for suffix stripping. *Program 14*(3), 130–137.

(Przybocki et al., 1998) M. Przybocki, J. Fiscus, J. Garofolo, et D. Pallett, 1998. HUB-4 Information Extraction Evaluation. Dans les actes de *DARPA Broadcast News Transcription and Understanding Workshop*, 13–18.

(Radev et McKeown, 1998) D. Radev et K. McKeown, 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics 24*(3), 470–500.

(Rahim et Lee, 1996) M. Rahim et C.-H. Lee, 1996. Simultaneous ANN Feature and HMM Recognizer Design Using String-based Minimum Classification Error (MCE) Training. Dans les actes de *International Conference on Spoken Language Processing (ICSLP)*.

(Raina et al., 2005) R. Raina, A. Ng, et C. Manning, 2005. Robust textual inference via learning and abductive reasoning. Dans les actes de *Natinal Conference on Artificial Intelligence (AAAI)*.

(Ribeiro-Neto et Muntz, 1996) B. Ribeiro-Neto et R. Muntz, 1996. A Belief Network Model for IR. Dans les actes de *ACM Special Interest Group on Information Retrieval (SIGIR)*, 253–260.

(Robertson et Spärck-Jones, 1988) S. Robertson et K. Spärck-Jones, 1988. Relevance Weighting of Search Terms. *Taylor Graham Series In Foundations Of Information Science 27*, 143–160.

(Robertson et Walker, 1994) S. Robertson et S. Walker, 1994. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. Dans les actes de *ACM Special Interest Group on Information Retrieval (SIGIR)*, 232–241. Springer-Verlag New York, Inc. New York, NY, USA.

(Salton et al., 1983) G. Salton, E. Fox, et H. Wu, 1983. Extended Boolean Information Retrieval. *Communications of the ACM 26*(11), 1022–1036.

(Salton et al., 1975) G. Salton, A. Wong, et C. S. Yang, 1975. A Vector Space Model for Automatic Indexing. *Communications of the ACM 18*(11), 613–620.

(Sanderson et Shou, 2002) M. Sanderson et X. M. Shou, 2002. Speech and Hand Transcribed Retrieval. Dans les actes de *ACM Special Interest Group on Information Retrieval (SIGIR)*. Springer.

(Sarawagi et Cohen, 2005) S. Sarawagi et W. Cohen, 2005. Semi-Markov Conditional Random Fields for Information Extraction. *Advances in Neural Information Processing Systems 17*, 1185–1192.

(Savoy et Berger, 2005) J. Savoy et P. Berger, 2005. Report on CLEF-2005 Evaluation Campaign : Monolingual, Bilingual, and GIRT Information Retrieval. Dans les actes de *Cross Language Evaluation Forum (CLEF)*.

(Schilder et al., 2006) F. Schilder, A. McCulloh, B. McInnes, et A. Zhou, 2006. TLR at DUC : Tree Similarity. Dans les actes de *Document Understanding Conference (DUC)*.

(Seki et al., 2004) Y. Seki, K. Eguchi, et N. Kando, 2004. User-Focused Multi-Document Summarization with Paragraph Clustering and Sentence-Type Filtering. Dans les actes de *National Institute of Informatics Test Collections for Information Retrieval (NTCIR)*, 459–466.

(Shriberg et al., 2000) E. Shriberg, A. Stolcke, D. Hakkani-Tur, et G. Tur, 2000. Prosody-Based Automatic Segmentation of Speech into Sentences and Topics. *Journal of Speech Communications : Special Issue on Accessing Information in Spoken Audio 32*(1-2), 127–154.

(Singhal et Pereira, 1999) A. Singhal et F. Pereira, 1999. Document Expansion for Speech Retrieval. Dans les actes de *ACM Special Interest Group on Information Retrieval (SIGIR)*, New York, NY, USA, 34–41. ACM Press.

(Smoliar et al., 1996) S. W. Smoliar, J. D. Baker, T. Nakayama, et L. Wilcox, 1996. Multimedia Search : an Authoring Perspective. Dans les actes de *International Workshop on Image Databases and Multimedia Search*, 1–8.

(Spärck-Jones et al., 2000) K. Spärck-Jones, S. Walker, et S. Robertson, 2000. A probabilistic model of information retrieval : development and comparative experiments. *Information Processing and Management : an International Journal 36*(6), 779–808.

(Stenchikova et al., 2006) S. Stenchikova, D. Hakkani-Tur, et G. Tur, 2006. QASR : Spoken Question Answering Using Semantic Role Labeling. Dans les actes de *International Conference on Spoken Language Processing (ICSLP)*.

(Stevenson et Gaizauskas, 2000) M. Stevenson et R. Gaizauskas, 2000. Experiments on Sentence Boundary Detection. Dans les actes de *Applied Natural Language Conferences (ANLC)*, San Francisco, CA, USA, 84–89. Morgan Kaufmann Publishers Inc.

(Strassel, 2003) S. Strassel, 2003. Simple Metadata Annotation Specification Version 5.0. Linguistic Data Consortium, Philadelphia, PA.

(Thong et al., 2000) J.-M. V. Thong, D. Goddeau, A. Litvinova, B. Logan, P. Moreno, et M. Swain, 2000. SpeechBot a Speech Recognition Based Audio Indexing System. Dans les actes de *Recherche d'Information Assistée par Ordinateur (RIAO)*.

(Torres-Moreno et al., 2002) J. Torres-Moreno, P. Velazquez-Morales, et J. Meunier, 2002. Condensés de Textes par des Méthodes Numériques. Dans les actes de *Journées internationnales d'Analyse statistiques des Données Textuelles (JADT)*.

(Torres-Moreno et al., 2005) J. Torres-Moreno, P. Velazquez-Moralez, et J. Meunier, 2005. CORTEX, un Algorithme pour la Condensation Automatique de Textes. Dans les actes de *Association pour la Recherche Cognitive (ARCo)*, Volume 2, 365.

(Van Noord et al., 2000) G. Van Noord, G. Bouma, R. Koeling, et M. Nederhof, 2000. Robust Grammatical Analysis for Spoken Dialogue Systems. *Natural Language Engineering 5*(01), 45–93.

(Vanderwende et al., 2004) L. Vanderwende, M. Banko, et A. Menezes, 2004. Event-Centric Summary Generation. Dans les actes de *Document Understanding Conference (DUC)*.

(Varges et al., 2006) S. Varges, F. Weng, et H. Pon-Barry, 2006. Interactive Question Answering and Constraint Relaxation in Spoken Dialogue Systems. Dans les actes de *Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

(Viterbi, 1967) A. J. Viterbi, 1967. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory 13*(2), 260–269.

(Vlachos et al., 2006) A. Vlachos, C. Gasperin, I. Lewin, et T. Briscoe, 2006. Bootstrapping the Recognition and Anaphoric Linking of Named Entities in Drosophila Articles. Dans les actes de *Pacific Symposium in Biocomputing*.

(Voorhees, 2003) E. Voorhees, 2003. Overview of the TREC 2003 Question Answering Track. Dans les actes de *Text REtrieval Conference (TREC)*.

(Voorhees et Harman, 1999) E. M. Voorhees et D. Harman, 1999. Overview of the Eighth Text REtrieval Conderence (TREC-8). Dans les actes de *Text REtrieval Conference (TREC)*.

(Walls et al., 1999) F. Walls, H. Jin, S. Sista, et R. Schwartz, 1999. Topic detection in broadcast news. Dans les actes de *DARPA Broadcast News Transcription and Understanding Workshop*, 193–198.

(Wang et Acero, 2003) Y.-Y. Wang et A. Acero, 2003. Combination of CFG and N-gram modeling in Semantic Grammar Learning. Dans les actes de *European Conference on Speech Communication and Technology (EUROSPEECH)*.

(Wechsler et al., 1998) M. Wechsler, E. Munteanu, et P. Schauble, 1998. New Techniques for Open Vocabulary Spoken Document Retrieval. Dans les actes de *ACM Special Interest Group on Information Retrieval (SIGIR)*, 20–27. ACM Press.

(Wei et Croft, 2006) X. Wei et W. Croft, 2006. LDA-Based Document Models for Ad-Hoc Retrieval. Dans les actes de *ACM Special Interest Group on Information Retrieval (SIGIR)*, 178–185. ACM Press New York, NY, USA.

(Wheatley, 1879) H. B. Wheatley, 1879. *What is an Index ? A Few Notes on Indexes and Indexers*. London : Longmans, Green & Co.

(Widdows et Peters, 2003) D. Widdows et S. Peters, 2003. Word Vectors and Quantum Logic Experiments with Negation and Disjunction. *Mathematics of Language 8*, 141–54.

(Wilkinson et Hingston, 1991) R. Wilkinson et P. Hingston, 1991. Using the Cosine Measure in a Neural Network for Document Retrieval. Dans les actes de *ACM Special Interest Group on Information Retrieval (SIGIR)*, 202–210. ACM Press New York, NY, USA.

(Witte et Bergler, 2003) R. Witte et S. Bergler, 2003. Fuzzy Coreference Resolution for Summarization. Dans les actes de *International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization (ARQAS)*, 43–50.

(Witten et Bell, 1991) I. Witten et T. Bell, 1991. The Zero-Frequency Problem : Estimating the Probabilities of Novelevents in Adaptive Text Compression. *IEEE Transactions on Information Theory 37*(4), 1085–1094.

(Wong et al., 1985) S. K. M. Wong, W. Ziarko, et P. C. N. Wong, 1985. Generalized Vector Spaces Model in Information Retrieval. Dans les actes de *ACM Special Interest Group on Information Retrieval (SIGIR)*, 18–25.

(Yu et Seide, 2004) P. Yu et F. Seide, 2004. A Hybrid Word/Phoneme-Based Approach for Improved Vocabulary-Independent Search in Spontaneous Speech. Dans les actes de *International Conference on Spoken Language Processing (ICSLP)*.

(Zajic et al., 2004) D. Zajic, B. Dorr, et R. Schwartz, 2004. BBN UMD at DUC-2004 Topiary. Dans les actes de *Document Understanding Conference (DUC)*.

(Zechner, 2001) K. Zechner, 2001. *Automatic Summarization of Spoken Dialogues in Unrestricted Domains*. Thèse de Doctorat, Language Technologies Institute, School of Computer Science, Carnegie Mellon University.

(Zechner, 2002) K. Zechner, 2002. Summarization of Spoken Language-Challenges, Methods, and Prospects. *Speech Technology Expert eZine 6*.

(Zechner, 2003) K. Zechner, 2003. Spoken Language Condensation in the 21st Century. Dans les actes de *European Conference on Speech Communication and Technology (EUROSPEECH)*, 1989–1992.

(Zhu et Penn, 2005) X. Zhu et G. Penn, 2005. Evaluation of Sentence Selection for Speech Summarization. Dans les actes de *Recent Advances in Natural Language Processing (RANLP)*, 39–45.

# Publications Personnelles

(Favre, 2003) B. Favre, 2003. Indexation Multimédia : Caractérisation du Déséquilibre entre les Modalités Texte et Parole. Mémoire de Master, Université d'Avignon.

(Favre et al., 2006) B. Favre, F. Bechet, P. Bellot, F. Boudin, M. El-Beze, L. Gillard, G. Lapalme, et J.-M. Torres-Moreno, 2006. The LIA-Thales summarization system at DUC-2006. Dans les actes de *Document Understanding Conference (DUC)*, 131–138.

(Favre et al., 2004a) B. Favre, P. Bellot, et J.-F. Bonastre, 2004a. Information Retrieval on Mixed Written and Spoken Documents. Dans les actes de *Recherche d'Information Assistée par Ordinateur (RIAO)*, 826–835.

(Favre et al., 2004b) B. Favre, J.-F. Bonastre, et P. Bellot, 2004b. Recherche d'Information dans un Mélange de Documents Écrits et Parlés. Dans les actes de *Journées d'Étude sur la Parole (JEP)*.

(Favre et al., 2006) B. Favre, J.-F. Bonastre, P. Bellot, et F. Capman, 2006. Accès aux Connaissances Orales par le Résumé Automatique. Dans les actes de *Extraction et Gestion des Connaissances (EGC)*.

(Favre et al., 2005a) B. Favre, F. Béchet, et P. Nocéra, 2005a. Mining Broadcast News data : Robust Information Extraction from Word Lattices. Dans les actes de *European Conference on Speech Communication and Technology (EUROSPEECH)*.

(Favre et al., 2005b) B. Favre, F. Béchet, et P. Nocéra, 2005b. Robust Named Entity Extraction from Large Spoken Archives. Dans les actes de *Human Language Technologies (HTL/EMNLP)*.