Grenoble INP – Institut National Polytechnique de Grenoble
École Nationale Supérieure d'Informatique, Mathématiques Appliqueées et Télécommunications
LIG – Laboratoires d'Informatique de Grenoble
Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole

CARLOS EDUARDO RAMISCH

Multi-word terminology extraction for domain-specific documents

Mémoire de Master
Master 2 de Recherche en Infromatique
Option Intelligence Artificielle et Web

Christian BOITET
Advisor

Aline VILLAVICENCIO
Coadvisor

Grenoble, juin 2009

ANNEXE A RÉSUMÉ ÉTENDU

EXTRACTION DE TERMES MULTIMOT POUR DES DOCUMENTS SPÉCIALISÉS

A.1 Introduction

L'un des objectifs majeurs du Traitement Automatique des Langues (TAL) est de rendre la communication entre les êtres humains et les ordinateurs aussi simple que demander un service à un ami. Le langage naturel est une composante essentielle au comportement intelligent, ce qui conduit à la classification du TAL en tant que sous-domaine de l'intelligence artificielle, dont le TAL hérite des nombreux principes et méthodes. Le langage humain est, pour plusieurs raisons, complexe pour la modélisation informatique. Une fraction significative de la connaissance humaine est exprimée par le langage naturel : il est donc un moyen d'échange de connaissances et se situe ainsi à la base de toute communication dans le plan scientifique. Par conséquent, le TAL porte un grand intérêt à ce que l'on appelle le langage spécialisé ou le langage technico-scientifique. Une des caractéristiques de cette langue de spécialité est la richesse de son lexique en structures terminologiques, ces dernières étant l'objet de ce travail. Dans ce contexte, nous croyons que la terminologie peut être acquise automatiquement à partir de corpus spécifiques à des domaines et de techniques d'apprentissage dirigé par les données.

Supposons un utilisateur fictif qui doit trouver des informations sur la traduction automatique basée sur les exemples. Avec un moteur de recherche sur le Web, cette tâche peut être fastidieuse, car non seulement la majorité des résultats seront en anglais, mais il sera également difficile de filtrer la grande quantité de contenu (non pertinent) généré par les utilisateurs. Les moteurs de recherche spécialisés doivent tenir compte de la terminologie du domaine, par exemple les mots *arbre* et *feuille* sont des termes, mais les concepts qu'ils représentent ne sont pas les mêmes en botanique et en informatique. En somme, aujourd'hui, l'utilisateur ayant un besoin d'information spécifique à un domaine exprimé dans sa langue maternelle doit d'abord consulter plusieurs sources d'information auxiliaires pour traduire les mots-clés, et ensuite les documents renvoyés pas la recherche. Non seulement sur le Web mais de façon générale, l'accès multilingue aux textes spécialisés est d'un grand intérêt parce qu'il a des applications dans tous les domaines de la science et de la diffusion des produits. L'étude de méthodes pour la recherche d'informations multilingues n'entre pas dans le contexte de ce travail, mais l'exemple ci-dessus sert à motiver l'intégration des termes du domaine avec la traduction automatique.

Tout d'abord, nous devons définir trois concepts-clés avec lesquels nous travaillerons : *expression multimot*, *terme* scientifique et technique, et *terme multimot*. La définition d'Expression Multimot (EMM) est vaste, puisque ce terme englobe de nombreux phénomènes linguistiques tels que les noms composés (*traffic light*, *feu rouge*), les expressions idiomatiques (*clound number nine*, *sem pé nem cabeça*), les termes composés (*système d'exploitation*, *Datenbanksystem*), etc. Les termes sont, contrairement aux EMM de la langue courante, des entités linguistiques liées au texte technique

et scientifique. Un terme peut être une unité mono-lexicale tandis qu'une EMM est nécessairement composée de plusieurs mots. Dans la Théorie Générale de la Terminologie, il est estimé qu'il existe une relation fonctionnelle et injective (1 :1) entre les termes et les concepts. Cette idée a toutefois été sévèrement critiquée pour être très réductrice et pour ne pas tenir compte de l'habitat naturel des termes : le texte. Les Termes Multimot (TMM) sont définis comme des locutions qui ont le statut de terme. Nous soulignons que les termes multimot ne correspondent pas à la notion de phraséologie du domaine. Un TMM accepte peu de variabilité (morphologique, mais rarement syntaxique), tandis qu'une phraséologie est hautement variable. Alors que le premier représente un concept particulier, il n'est pas rare que la phraséologie soit une structure complexe comprenant plus d'un concept.

La traduction et la terminologie ont de nombreux points communs en ce qui concerne la traduction spécialisée. De même, les TMMs et la traduction automatique (TA) devraient être des domaines proches, malgré la réalité actuelle dans le TAL qui reflète peu cette interdépendance. Ainsi, l'un des objectifs de ce travail est la détection automatique de termes pour l'adaptation des ressources linguistiques informatisées à plusieurs domaines. Nous avons choisi de travailler uniquement avec les termes multimot parce que (a) ils représentent environ 70% de la terminologie d'un domaine, (b) les méthodes en informatique pour la détection des termes multimot et « mono-mots » sont très différentes et (c) les TMMs représentent un défi majeur en linguistique computationnelle.

Definition A.1 *Une Expression Multimot (EMM) est une suite de deux mots ou plus dont la sémantique est non-compositionnelle, c'est-à-dire que le sens du syntagme ne peut pas être totalement compris à travers le sens des mots qui le composent (Sag et al. 2002).*

Definition A.2 *Un terme est une unité lexicale ou multilexicale avec un sens non ambigu quand elle est employée dans un texte spécialisé, de façon à ce que la terminologie est la manifestation linguistique des concepts d'un domaine (Krieger and Finatto 2004).*

Definition A.3 *Un Terme Multimot (TMM) est un terme composé par plus d'un mot (SanJuan et al. 2005, Frantzi et al. 2000).*

A.2 Révision bibliographique

Le travail présenté ici traite de l'intégration de trois sous-domaines du TAL : les expressions multimot, la terminologie et la traduction statistique. Les EMM constituent un défi pour le TAL en raison de leur grande variabilité morphosyntaxique et de leur nature hétérogène. Les techniques pour le traitement des EMM sont classifiées comme étant soit du type « mots-avec-espaces » soit compositionnelles. Quatre tâches sont liées au traitement des EMMs : l'identification de leur occurence dans les textes, l'interprétation de leur structure (syntaxe), leur classification ou construction de groupements (sémantique) et leur application à d'autres systèmes de TAL. Il est possible d'identifier les candidats à EMM à travers des motifs morphologiques de surface (Villavicencio et al. 2007, Manning and Schütze 1999) ou par des métodes plus ou moins profondes qui tiennent compte des phénomènes syntaxiques (Seretan 2008, Baldwin 2005). Le filtrage des candidats peut se faire par des mesures statistiques (Evert and Krenn 2005, Villavicencio et al. 2007) souvent complétées par des fréquences obtenues dans le Web (Keller et al. 2002, Zhang et al. 2006). Les approches symboliques emploient généralement des thésaurus et des dictionnaires de synonymes (Pearce 2001). Aussi l'apprentissage artificiel suppervisé peut se montrer effectif pour la construction de modèles d'identification de EMMs (Pecina 2008). L'interprétation de la structure interne des EMMs dépend de leurs types spécifique, par exemple, Ramisch et al. (2008) classent les verbes composés en anglais par rapport à leur idiomaticité, et Nakov and Hearst (2005) travaillent sur la structure d'imbrication des noms composés. Par

rapport à la classification des EMMs, nous citons le travail de Lapata (2002), qui traite en particulier des nominalisations.

Une technique couramment utilisée pour l'extraction de termes est l'utilisation de motifs morphologiques fréquents (Justeson and Katz 1995). Frantzi et al. (2000) proposent une technique plus fine pour le classement des candidats, pendant que Smadja (1993) utilise l'information mutuelle des mots afin de détecter des collocations. Par rapport à la terminologie biomédicale du corpus Genia, SanJuan et al. (2005) effectuent une classification sémantique automatique qui est ensuite comparée à l'ontologie du domaine.

La traduction automatique (TA) a évolué au fil des dernières années vers des méthodes collaboratives au niveau opérationnel ou statistiques. Les premières sont fondées sur le triplet automaticité, couverture et précision, permettant à l'utilisateur de choisir quels aspects sont plus importants pour leur application (Bey et al. 2006, Boitet et al. 2008). D'autre part, les systèmes de traduction statistique apprennent les probabilités de traduction à partir d'un corpus parallèle (Brown et al. 1993). L'alignement au niveau des mots est obtenu à travers l'algorithme d'apprentissage non supervisé EM. Selon Koehn et al. (2003), les systèmes de TA statistiques peuvent être de deux types : basés sur les mots ou basés sur les syntagmes (séquences de mots). Dans des domaines spécialisés, des problèmes peuvent se produire lorsque le corpus d'entraînement est différent du texte à traduire, comme le montre Baldwin et al. (2004) sur un sustème d'analyse syntaxique. L'identification simultannée des TMMs et des alignements au niveau des mots permet d'extraire à la fois les unités lexicales multimot et leurs traductions (Melamed 1997, Caseli et al. 2009).

A.3 Extraction de la terminologie biomédicale

Le corpus Genia est un ensemble de 2.000 résumés dans le domaine des sciences biomédicales sur les mots-clés « humain », « cellules sanguines » et « facteur de transcription » (Ohta et al. 2002). Il contient 18,5K phrases et 490,7K tokens, annotés avec la classe morphosyntaxique et avec la terminologie, comme les noms de maladies (*tumeur fibroblastique*) et de cellules (*lymphocytes t primaires*). En moyenne, les phrases contiennent 3 TMMs, chacun avec une moyenne de 3 tokens. Cela montre que la terminologie est omniprésente dans ce corpus. Les termes du corpus Genia ont tendance à être profondément imbriqués. Les expériences décrites ci-dessous ont été implémentées à travers une série de scripts en Python, awk et bash disponibles à l'adresse `www.inf.ufrgs.br/~ceramisch/tcc.zip`.

Tout d'abord, les données du corpus ont été prétraitées (SanJuan et al. 2005) : toutes les phrases ont été converties en minuscules et une version simplifiée du corpus en XML a été générée. Dans cette version, l'annotation des termes est gardée, non pas pour être utilisée pendant l'extraction, mais afin de servir comme base de comparaison lors de l'évaluation. D'une part, un *candidat à TMM* est extrait du corpus s'il obéit à certaines normes ou règles pré-établies, indépendamment de l'annotation. D'autre part, la *terminologie standard* est l'ensemble des termes annotés qui serviront comme base de comparaison pour l'évaluation automatique des candidats. La terminologie standard de Genia possède 28.340 TMM distincts. Prenons l'exemple ci-dessous (la classe du mot est indiquée par un slash, les termes sont entre crochets) :

Example A.1 `[ [ alpha/N a/UKN ] -gene/A product/N ]`

Remarquez que l'annotation terminologique casse les mots dans les tirets et dans les slashs. Cela est nuisible puisque à l'une des moitiés du mot est affectée une classe inconnue (UKN). Comme ce phénomène arrive dans 17% des phrases du corpus, nous avons décidé d'inclure le token extérieur dans l'expression, de façon à améliorer la qualité de l'annotation. Le résultat peut illustré par

l'exemple ci-dessous. Avec cette modification, seuelement 81 phrases sont éliminées pour avoir des mots avec une classe inconnue :

Example A.2 `[ [ alpha/N a-gene/A ] product/N ]`

Afin de prendre en compte les déclinaisons de nombre, il a fallu réduire chaque réalisation d'un mot à sa forme au singulier. Un lemmatiseur généraliste pourrait ramener le nom *binding* au verbe *bind*, tandis qu'une simple règle d'élimination de la lettre « s » à la fin des mots aboutirait à des lemmes comme *virus* → *viru** ou *viruses* → *viruse**. Par conséquent, nous avons développé un lemmatiseur simple basé sur des règles et sur des fréquences du Web, dont le fonctionnement est décrit en détail dans l'annexe D.

La dernière étape de prétraitement concerne une caractéristique propre aux textes biomédicaux, qui ont tendance à insérer des acronymes entre parenthèses après les termes récurrents. Lorsque ces parenthèses se présentent à l'intérieur d'un TMM, ils génèrent du bruit pour les filtres, comme dans l'exemple *[[granulocyte-macrophage colony-stimulating factor] (gm-csf) promoter]*. Environ 3% des termes contiennent ces parenthèses, et pour supprimer uniquement celles contenant des acronymes, nous avons utilisé un algorithme qui combine des règles et une procédure de correspondance par sous-chaîne de longueur maximale, décrite dans l'annexe D. Ainsi, nous sommes capable d'identifier les acronymes dans le corpus et, ensuite, éliminer ceux qui apparaissent entre parenthèses. Cette liste d'acronymes (par exemple *AD – alzheimer 's disease*, *5LOX – 5 lipoxygenase*, *AD – adenovirus*) est un produit supplémentaire à l'extraction des TMMs.

Une fois que le corpus a été prétraité de manière uniforme, il est nécessaire d'extraire les candidats à TMM. Ainsi, avant tout il faut le diviser en deux parties : *Genia-test* (100 résumés, 895 phrases) et *Genia-train* (le reste). La première partie est prévue pour une utilisation future au cours de la phase d'évaluation. La dernière partie est utilisée pour effectuer non seulement l'apprentissage des motifs morphologiques mais aussi des poids donnés à chaque mesure d'association durant l'étape de filtrage des candidats, postérieure à l'extraction.

L'extraction de motifs morphologiques est effectuée de la façon suivante : (1) en nous basant sur l'annotation des termes du corpus, nous avons extrait la terminologie standard de *Genia-train*, (2) nous avons sélectionné les 118 motifs qui apparaissent plus de 10 fois, (3) parmi ces derniers, nous avons gardé seulement les 57 motifs dont la précision (proportion de candidats positifs) est supérieure à 30% et (4) nous avons extrait du corpus l'enseble des n-grammes qui corespondent à ces motifs. Les valeurs des seuils de fréquence et de précision ont été obtenues à partir de l'observation empirique des données. Les motifs sélectionnés comprennent des séquences de noms et d'adjectifs (*N-N*, *A-N*, *A-N-N*), mais aussi de mots étrangers (*FW-FW*), et des verbes (*N-V-N*). Le résultat de cette étape est une liste avec 60.585 candidats à être des TMM, dont 22.507 sont positifs (31,52%).

Après avoir extrait les candidats, nous procédons alors à leur filtrage. Les filtres sont basés sur des Mesures d'Association (MA) statistiques (Evert and Krenn 2005, Ramisch et al. 2008, Pecina 2008). La distribution de probabilités du vocabulaire de la langue est zipfienne et comporte un phénomène connu sous le nom de « longue queue ». Par conséquent, les fréquences deviennent très faibles, ce qui implique une basse fiabilité pour les MA. Ainsi, nous avons utilisé également les fréquences obtenues à partir du World Wide Web (via une API de recherche de *Yahoo!*), que nous considérons ici comme un corpus de langue générale dorénavant dénommé *Web*. Pour estimer la taille N du corpus, nous avons utilisé une estimation grossière[1], qui est simplement un facteur d'échelle linéaire identique pour tous les candidats. La liste d'acronymes est utilisée ici pour augmenter la fréquence des mots qui y participent à chacune de leurs occurences. Pour chaque candidat $w_1 \dots w_n$ et pour chaque corpus de taille N, la fréquence simple $f(w_1 \dots w_n)$ et les fréquences marginales $f(w_1) \dots f(w_n)$ sont utilisées pour le calcul des MA ci-dessous :

[1] 50 milliards de pages, selon `http://www.worldwidewebsize.com/`

$$\texttt{prob} = \frac{f(w_1 \dots w_n)}{N} \qquad \texttt{PMI} = log_2 \frac{f(w_1 \dots w_n)}{N^{n-1} f_\emptyset(w_1 \dots w_n)}$$

$$\texttt{t} = \frac{f(w_1 \dots w_n) - N^{n-1} f_\emptyset(w_1 \dots w_n)}{\sqrt{f(w_1 \dots w_n)}} \qquad \texttt{Dice} = \frac{n * f(w_1 \dots w_n)}{\sum_{i=1}^{n} f(w_i)}$$

Ici, l'hypothèse nulle suppose l'indépendance entre les mots, c'est-à-dire que leur co-occurence se produit par hasard et que leur fréquence correspond à $f_\emptyset(w_1 \dots w_n) = (f(w_1) \dots f(w_n))/N^{n-1}$. Les mesures qui utilisent le tableau de contingence, même si elles sont plus robustes, ne sont pas utilisées ici, parce que, si n est arbitraire, il est non seulement difficile de produire un hypercube de fréquences, mais aussi le bruit et les incohérences (surtout dans le corpus *Web*) empêchent en pratique son utilisation. Plusieurs travaux discutent le choix de la meilleure MA, mais notre méthode consiste à utiliser l'apprentissage artificiel pour cette tâche. Nous séparons donc les candidats à filtrer en deux catégories : `MWT` et `NonMWT`. Les algorithmes utilisés incluent les arbres de décision (J48), les classificateurs bayésians par méthodes statistiques, les réseaux de neurones artificiels (MLP et *Voted Perceptron*) et les classificateurs numériques qui agissent sur la frontière de décision, c'est-à-dire les machines à vecteur de support (SVM).

Une fois que les termes multimot ont été extraits du corpus, nous proposons des moyens pour les intégrer la boîte à outils Moses de TA statistique. Le système a été entraîné sur le corpus Europarl avec la paire de langues anglais → portugais, pour laquelle nous disposons de locuteurs natifs aptes à évaluer les résultats. Le corpus Europarl a été aligné au niveau des phrases, filtré, la séparation des mots et la casse ont été homogénéisées, puis nous avons séparé deux sous-ensembles pour le développement et pour les tests. Ensuite, nous avons généré un modèle de langue, les alignements mot à mot et la table de traduction, pour finalement apprendre, par minimisation du taux d'erreur, les coefficients de pondération de chaque attribut du modèle.

Il existe plusieurs techniques proposées dans la littérature pour séparer et recombiner des mots composés en allemand ou en suédois (Nießen and Ney 2004, Nießen et al. 2000). Nous utilisons une méthode similaire pour les termes en anglais : considérons, à titre d'exemple, la phrase d'entrée *ras-related gtp-binding proteins and leukocyte signal transduction*. La première approche (wws) consiste à concaténer les mots qui intègrent un terme, par exemple, *ras-related#gtp-binding#proteins and leukocyte signal#transduction*. La deuxième méthode (head) est de remplacer le terme par son noyau, supposé être le dernier nom du n-gramme, par exemple, *proteins and leukocyte transduction*. La terminologie standard a également été utilisée pour évaluer la séparation et la recombination de termes dans Moses. La technique de remplacement du noyau simplifie le texte original pour faciliter le fonctionnement du système de TA, de manière que le texte original entraîne le texte simplifié, qui est alors traduit. Sémantiquement, l'information véhiculée par la traduction contient le sens de la phrase originale, mais ne lui est pas équivalent.

A.4 Évaluation

La première évaluation concerne le prétraitement, et est donc effectuée sur la totalité du corpus. Nous appellerons ici « normalisation » le processus de lemmatisation et homogéneisation de la séparation de mots. Cette étape est suivie par l'Élimintation d'Acronymes (EA). L'évaluation manuelle sur un ensemble de mots de l'algorithme de transformation pluriel-singulier montre qu'il donne un résultat correct sur 99% des mots. Dans le tableau A.1, nous pouvons vérifier que, vis-à-vis de la configuration sans normalisation, la terminologie standard est 3,89% fois plus petite, alors que la liste des candidats est réduite de 5,35%. Cela montre que les candidats et termes similaires (par exemple, où l'un est le pluriel de l'autre) sont désormais traités en tant qu'entité unique représentant un même

TAB. A.1 – Évaluation de la normalisation et de l'Élimination d'Acronymes (EA).

	Sans norm.	**Sans EA**	**Avec norm. et EA**
TMMs de la term. standard	29.513	28.601	28.340
Candidats à TMM	66.817	61.401	63.210
Précision	15,56%	37,53%	37,24%
Rappel	35,22%	80,58%	83,06%
F-measure	21,58%	51,20%	51,42%
Fréquence moyenne des top-100	189,98	205,05	205,38
Hapax legomena	51.169 (77%)	46.330 (75%)	47.770 (76%)

concept. Malgré son aspect réductionniste (le contexte est ignoré), cette approche donne des résultats pratiques intéressants : le rappel augmente significativement avec la normalisation, de 35% à 83% ; la précision est doublée, passant de 15% à 37%. De plus, les fréquences des mots sont moins faibles puisque les candidats similaires sont regroupés et leurs nombre d'occurrences additionnées. Le nombre de hapax legomena (qui apparaissent une seule fois) diminue, mais correspond encore aux deux tiers des candidats. D'autres heuristiques de normalisation pourraient être utilisées, comme par exemple l'unification des déclinaisons de genre ou des conjugaisons verbales. Il est néanmoins important de souligner que tout traitement de ce type est fortement dépendant de la langue et du domaine en question.

Tojours dans le même tableau, nous présentons la configuration sans EA. Ici, la terminologie standard est aussi plus petite, mais la réduction n'est pas aussi importante que nous l'espérions (la liste d'acronymes a 1.300 entrées), probablement parce que la plupart des acronymes apparaissent après les TMMs, et non pas au milieu d'eux. Plus de candidats sont extraits, ce qui augmente le rappel de 2,5% mais diminue légèrement la précision. Il semble aussi que l'EA aide à générer des données moins rares. Les acronymes sont également utilisés pour augmenter la fréquence des mots qu'ils remplacent. Par conséquent, le nombre d'occurrences marginal moyen augmente de 37 à 47. Cependant, l'importance de l'étape d'élimination des acronymes est remise en question, car les améliorations provoquées sont peu visibles sur le résultat final.

TAB. A.2 – Précision et rappel des motifs de TMM.

Motifs morphologiques	**Précision**	**Rappel**
J&K	**29,58%**	66,71%
$f > 10$	16,50%	**92,57%**
$f > 10, p > .3$	31,52%	89,43%

Une fois que le prétraitement est terminé, nous voulons évaluer l'étape de sélection des motifs morphologiques des candidats. Quand nous comparons les motifs utilisés par Justeson and Katz (1995) (J&K), ceux qui ont plus de 10 occurrences et ceux qui ont été sélectionnés ($f > 10, p > .3$), nous pouvons voir que, d'une part, J&K ont un bon rappel mais une précision très faible sur ce corpus (où les TMMs ont tendence à être longs et imbriqués), et que, d'autre part, les motifs les plus fréquents ont un rappel élevé tandis que la précision est légèrement supérieure à la moitié de J&K.

Cela nous mène aux motifs sélectionnés, qui offrent un bon équilibre, avec la plus grande précision parmi les ensembles de motifs étudiés et un rappel légèrement inférieur à celui de l'ensemble $f > 10$. Parmi les motifs sélectionnés, nous soulignons (a) des problèmes par rapport aux conjonctions comme dans *young and aged subject*, qui est en fait composé de deux TMMs indépendants,

TAB. A.3 – Précision moyenne (%) des MA statistiques.

	Genia				*Web*				
	prob	*PMI*	*t*	*Dice*	*prob*	*PMI*	*t*	*Dice*	**Base**
train	46,16	40,08	52,90	38,84	35,77	41,02	45,13	37,85	37,15
test	38,83	17,51	31,36	22,78	41,49	20,52	43,21	43,74	24,16

(b) la présence de motifs contenant des chiffres (*interleukin 2*) et (c) l'absence du motif N-P-N, qui est considéré par J&K comme un bon indicateur de TMM. Il est intéressant de vérifier que, quand la terminologie standard de l'ensemble d'entraînement est utilisée pour évaluer l'ensemble de test, les motifs sélectionnent environ deux fois plus de termes que l'annotation existante. Ainsi, le rappel atteint 200% alors que, pour chaque motif, la précision oscille entre 65% et 95%. Ces résultats montrent qu'une annotation plus soigneuse pourrait améliorer les valeurs absolues de l'évaluation.

Le tableau A.3 montre la précision moyenne de chacune des mesures d'association ainsi que la précision de la ligne de base (proportion initiale de candidats positifs) pour les fragments d'entraînement et de test. D'après ces valeurs, malgré la supposition fausse que les mots ont une distribution normale, le test t a une bonne performance sur les deux parties du corpus, ce qui est en accord avec les résultats d'Evert and Krenn (2005), Manning and Schütze (1999). La probabilité simple (prob) semble également un bon indicateur de TMM puisque sa précision moyenne est élevée. Parmi les mesures évaluées, PMI est celle qui présente la plus faible performance, étant même inférieure à la ligne de base sur *Genia-test*, contrairement à ce qui est indiqué dans Smadja (1993). Cependant, Evert and Krenn (2005) trouvent le même résultat et affirment que le test t et le rapport de vraisemblance logarithmique (mesure qui utilise le tableau de contingence) ont souvent les meilleures performances. Un algorithme simple de sélection d'attributs confirme que les mesures les plus pertinentes sont $prob_{Genia}$, t_{Genia} et t_{Web}. Quelques incohérences ont été observées dans le comportement des MA sur le corpus *Web*, puisque les mesures semblent mauvaises sur la partie *Genia-train* tandis qu'elles sont bonnes sur *Genia-test*.

L'analyse du vocabulaire du corpus *Web* (figure A.1) montre qu'il existe une explication possible pour ce comportement apparemment arbitraire des mesures sur le *Web*. L'API de *Yahoo!* retourne, pour les 735 mots les plus fréquents, un nombre constant d'occurrences (2.147.483.647), de sorte qu'il est impossible de faire la différence entre des mots comme *the* et *green* (qui apparaissent respectivement 22.446 et 1 fois dans le corpus Genia). Cette situation pathologique, résultat d'une ap-

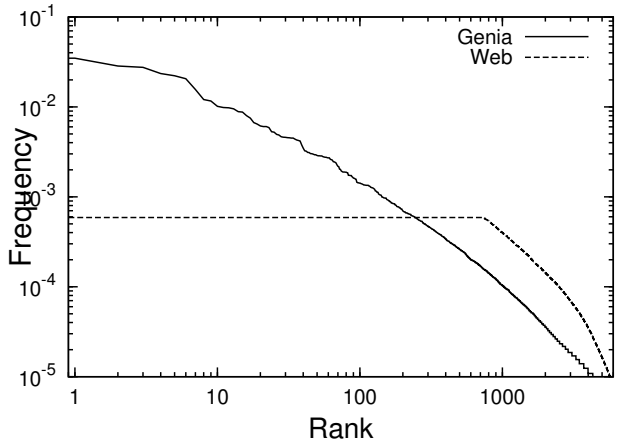

FIG. A.1 – Histogramme du vocabulaire de *Genia-train* e du *Web*.

TAB. A.4 – Performance des algorithmes d'apprentissage.

	J48	**Bayes(TAN)**	**Bayes(SA)**	**MLP**	**VP**	**SVM(pol.)**	**SVM(rad.)**
Précision	47.9%	48.6%	52.4%	27.3%	36.3%	34.9%	**52.6%**
Rappel	19.7%	32.1%	28.9%	**0.5%**	**63.8%**	60.1%	**35.4%**

TAB. A.5 – Évaluation comparative de la méthode SVM avec plusieurs seuils et l'algorithme Xtract.

	Genia-test				*Genia-train*
	Aucun seuil	$f_{Genia} > 1$	$f_{Genia} > 5$	**Xtract**	**Xtract**
nb. TMMs extraits	763	739	174	70	1.558
Précision	52,56%	56,83%	74,14%	70%	66,81%
Rappel	19,96%	20,91%	6,42%	2,44%	3,84%

proximation grossière de la part du moteur de recherche, génère la courbe « cassée » visible dans le graphique, certainement nuisible aux mesures d'association. Ainsi, contrairement à ce que suggèrent Zhang et al. (2006), le coefficient de corrélation de Pearson entre les corpus est de seulement 0,12 alors que le coéfficient τ de Kendall est de 0,21. Dans l'avenir, nous souhaitons étudier une façon d'intégrer les points forts des deux corpus, puisque Genia sous-estime la probabilité des mots peu fréquents, tandis que *Web* le fait pour les mots les plus fréquents.

La performance des algorithmes d'apprentissage artificiel sur le corpus *Genia-test* est comparée dans le tableau A.4. Seul la classe `MWT` est considérée parce que, comme les données ne sont pas équilibrées, la classe `NonMWT` a une F-mesure supérieure à 80% pour tous les algorithmes. L'algorithme J48 génère un modèle de données facile à interpréter (un arbre avec 484 nœuds) qui classifie correctement 75% des instances. Les attributs les plus proches de la racine (les plus pertinents) sont les valeurs *t* dans les deux corpus, tandis que l'attribut qui semble être le moins important est le motif morphologique, généralement dans le nœud qui précède les feuilles.

Le problème de cet algorithme (et bien d'autres) est qu'il considère que les deux classes ont le même poids. Dans ce sens, l'apprentissage bayésien est plus efficace, surtout quand la recherche par arbre de dispersion minimal (TAN) est utilisée au lieu de la recherche par recuit simulé (SA).

La couverture de ces deux méthodes reste décevante parce qu'elles ont le même problème que l'algorithme J48, i.e. elles donnent un poids égal aux deux classes. Le réseau de neurones multicouche (MLP) avec 33 neurones dans la couche caché a une performance très basse : seulement 5% des candidats positifs sont capturés.

L'algorithme *Voted Perceptron (VP)* a une performance surprenante, avec un très bon rappel sur les candidats positifs. Les causes de ce résultat restent à étudier.

Les meilleurs résultats ont été obtenus avec les machine à vecteur de support (SVM) de noyau polynomial et radial. Chacun tend à donner plus d'importance à l'un des aspects : le premier a un bon rappel alors que le dernier a une précision de 52,6%. Le choix entre un noyau polynomial ou radial dépend de l'application. Par exemple, la construction d'un dictionnaire est très onéreuse, de sorte qu'une haute précision diminue la quantité d'effort manuel à fournir. Cependant, une application de TAL peut éventuellement vivre avec des données bruitées à condition que le rappel soit suffisamment grand. Dans lesétapes suivantes, nous avons choisi un SVM à noyau radial, qui renvoie sur l'ensemble de test une liste de 763 instances positives dont 362 sont fausses. Au cas où nous n'aurions pas filtré les données avec un SVM, la proportion de candidats positifs serait de seulement 24%.

Malgré l'impression que l'on a, que les valeurs absolues de précision et de rappel sont basses.

Quand nous comparons ces résultats avec le système conventionnel d'extraction d'unités multimot Xtract, nous pouvons voir que la méthode suggérée ici est supérieure. Dans le tableau A.5, nous comparons la performance du classificateur SVM avec plusieurs valeurs de seuil de fréquence sur la liste de candidats. La configuration où nous éliminons les hapax legomena ($f_{Genia} > 1$) fait monter la précision et le rappel, ce qui montre que ces candidats doivent effectivement être enlevés de la liste. Cependant, un seuil haut ($f_{Genia} > 5$) fait rapidement chuter le rappel. Comme notre méthode utilise les informations présentes dans *Genia-train* pour extraire les TMMs de *Genia-test*, nous la comparons avec Xtract sur l'ensemble de test, et aussi avec Xtract sur l'ensemble d'entraînement, plus grand donc avec des fréquences plus fiables et moins faibles. Néanmoins, dans les deux cas, Xtract est très conservateur et génère une liste de TMMs avec une précision assez haute mais un rappel extrêmement bas. Nous observons que, même dans la configuration la plus restrictive, le SVM obtient un résultat supérieur à celui obtenu par Xtract, prouvant ainsi que la méthode d'extraction de TMMs proposée dans ce travail a un gain réel sur un algorithme standard.

Finalement, pour des raisons pragmatiques de temps de traitement et d'espace de stockage, il n'a pas été possible d'achever l'évaluation de l'intégration TMM-TA. L'idée de cette étape est de fournir aux évaluateurs (des locuteurs natifs) une liste de traductions produites par les différentes méthodes d'intégration (wws et head avec TMMs extraits et terminologie standard). Les traductions seront alors évaluées par leur fluidité, car leur adéquation sera certainement plus basse, pourvu que la méthode effectue une simplification sur l'entrée. Une interface permettant la visualisation des termes dans l'original et la collaboration dans leur traduction sera également créée et soumise à des techniques standard d'évaluation d'interfaces.

A.5 Conclusions

Le travail présenté ici offre une évaluation systématique des techniques d'extraction terminologique sur un corpus biomédical. D'abord, nous avons discuté les hypothèses et les notions de base d'expression et terme multimot, ainsi que leur caractérisation à travers des théories linquistiques. Plusieurs travaux traitent les expressions multimot avec un type et dans une langue déterminés. La plupart des tâches et des méthodes discutées ont comme base le bigramme, et sur ce point nous croyons que la méthode proposée ici est supérieure, puisque nous sommes capable de traiter les TMMs avec une longueur n quelconque.

En somme, notre méthode est composée de quatre étapes : prétraitement, extraction de candidats à travers des motifs morphologiques sélectionnés, filtrage des candidats à travers une combination de mesures d'association et d'algorithmes d'apprentissage artificiel, et en dernier lieu, intégration de la terminologie avec les systèmes de TA statistiques. Le tableau A.6 liste quelque expressions extraites de *Genia-test* avec le SVM $n_{Genia} > 5$. La plus grande partie des termes sont des noms composés avec prédominance des bigrammes. Les exemples négatifs semblent soit appartenir à un terme plus long, soit contenir un terme plus court ; probablement une évaluation fine avec des classes comme `Contains` et `PartOf` pourrait vérifier cette hypothèse.

Il est important de souligner le fait que la terminologie extraite ne veut pas, en aucun moment, remplacer ou améliorer l'annotation existant a priori dans le corpus. L'objectif de ce travail n'était pas la génération d'une nouvelle terminologie pour le corpus Genia, parce que celui-ci comporte déjà une telle structure. Au contraire, l'annotation des termes a été utilisée pour atteindre notre objectif initial qui était d'évaluer (de façon automatisée) les méthodes d'extraction d'unités multimot dans un domaine spécifique.

Dans l'avenir, nous voulons évidemment conclure l'évaluation suggérée ici par l'intégration entre terminologie et traduction automatique, afin d'établir la validité de la méthodologie que nous proposons. De plus, nous voulons étudier de nouvelles façons de combiner les fréquences obtenues dans des

TAB. A.6 – Exemples de TMMs extraits de *Genia-test* (« * » représente une erreur).

TMM extrait	**Motif**	**TMM extrait**	**Motif**
t cell	N-N	**thromboxane receptor***	N-N
nf kappa b	N-N	peripheral blood	A-N
kappa b	N-N	nf kappa b	N-N-N
cell line	N-N	human t	A-N
transcription factor	N-N	receptor alpha	N-N
i kappa b	N-N-N	monocytic cell	A-N
i kappa*	N-N	binding activity	N-N
kappa b alpha*	N-N-N	stat protein	N-N
b alpha*	N-N	**receptor gene***	N-N
i kappa b alpha	N-N-N-N	**rap1 protein***	

corpus de nature différente. Un des points faibles de l'évaluation que nous avons réalisé est qu'elle est fondée sur un seul corpus d'un domaine unique, donc l'objectif des travaux à venir est aussi étendre ces résultats à d'autres langues et domaines.

Le modèle et les attributs de filtrage que nous avons discutés sont totalement indépendants de la langue et du domaine. Ainsi, seules les étapes de prétraitement et de sélection de motifs doivent être adaptées à un nouveau contexte d'application. Il sera donc possible d'adapter facilement les outils de TAL pour traiter le lexique spécialisé des différents domaines, de manière à aider à la diffusion des connaissances.

REFERENCES

R. Harald Baayen. *Word Frequency Distributions*, volume 18 of *Text, Speech and Language Technology*. Springer, 2001. ISBN 978-0-7923-7017-8.

Timothy Baldwin. Deep lexical acquisition of verb-particle constructions. *Computer Speech & Language Special Issue on Multiword Expressions*, 19(4):398–414, 2005.

Timothy Baldwin, Emily M. Bender, Dan Flickinger, Ara Kim, and Stephan Oepen. Road-testing the English Resource Grammar over the British National Corpus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, May 2004. European Language Resources Association.

Youcef Bey, Christian Boitet, and Kyo Kageura. The TRANSBey prototype: an on- line collaborative wiki-based CAT environment for volunteer translators. In *Third LREC International Workshop on Language Resources for Translation Work, Research & Training (LR4Trans-III)*, pages 49–54, Genoa, Italy, May 2006.

Ergun Biçici and Marc Dymetman. Dynamic translation memory: Using statistical machine translation to improve translation memory fuzzy matches. In *Proceedings of the 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2008)*, Haifa, Israel, February 2008. Springer.

Christian Boitet, Valérie Bellynck, Mathieu Mangeot, and Carlos Ramisch. Towards higher quality internal and outside multilingualization of web sites. In Pushpak Bhattacharyya, editor, *Summer Workshop on Ontology, NLP, Personalization and IE/IR – ONII-08*, Bombay, India, 2008.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2), 1993. ISSN 0891-2017.

Helena Medeiros Caseli, Carlos Ramisch, Maria das Graças Volpe Nunes, and Aline Villavicencio. Alignment-based extraction of multiword expressions. *Language Resources & Evaluation Special Issue on Multiword Expressions*, 2009.

Maurice de Kunder. Geschatte grootte van het geïndexeerde world wide web. Master's thesis, Universiteit van Tilburg, 2007.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1): 1–38, 1977.

Jianyong Duan, Ruzhan Lu, Weilin Wu, Yi Hu, and Yan Tian. A bio-inspired approach for multi-word expression extraction. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL 2006)*, pages 176–182, Sidney, Australia, July 2006. Association for Computational Linguistics.

Stefan Evert and Brigitte Krenn. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech & Language Special Issue on Multiword Expressions*, 19(4): 450–466, 2005.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130, 2000.

Roxana Girju, Dan Moldovan, Marta Tatu, and Daniel Antohe. On the semantics of noun compounds. *Computer Speech & Language Special Issue on Multiword Expressions*, 19(4):479–496, 2005.

Gregory Grefenstette. The World Wide Web as a resource for example-based machine translation tasks. In *Proceedings of the Twenty-First International Conference on Translating and the Computer*, London, UK, November 1999. ASLIB.

Dan Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, Cambridge, UK, 1997. ISBN 0-521-58519-8.

Cong-Phap Huynh, Christian Boitet, and Hervé Blanchon. SECTra_w.1 : an online collaborative system for evaluating, post-editing and presenting MT translation corpora. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 2008. European Language Resources Association.

Ray Jackendoff. Twistin' the night away. *Language*, 73:534–59, 1997.

John S. Justeson and Slava M. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27, 1995.

Frank Keller and Mirella Lapata. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484, 2003. ISSN 0891-2017.

Frank Keller, Maria Lapata, and Olga Ourioupina. Using the Web to overcome data sparseness. In Jan Hajič and Yuji Matsumoto, editors, *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 230–237, Philadelphia, USA, July 2002. Association for Computational Linguistics.

Adam Kilgarriff and Gregory Grefenstette. Introduction to the special issue on web as corpus. *Computational Linguistics*, 29(3), 2003. ISSN 0891-2017.

Jin-Dong Kim, Tomoko Ohta, Yuka Teteisi, and Jun'ichi Tsujii. GENIA ontology. Technical report, Tsujii Laboratory, University of Tokyo, 2006.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL 2003)*, pages 48–54, Edmonton, Canada, 2003. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 177–180, Prague, Czech Republic, July 2007. Association for Computational Linguistics.

Maria da Graça Krieger and Maria José Bocorny Finatto. *Introdução à Terminologia: teoria & prática*. Editora Contexto, 2004. ISBN 85-7244-258-8.

Mirella Lapata. The disambiguation of nominalizations. *Computational Linguistics*, 28(3):357–388, 2002. ISSN 0891-2017.

Adam Lopez. Statistical machine translation. *ACM Computing Surveys*, 40(3):1–49, 2008. ISSN 0360-0300.

Christopher D. Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, USA, 1999. ISBN 0-262-13360-1.

I. Dan Melamed. Automatic discovery of non-compositional compounds in parallel data. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP-2)*, Brown University, USA, August 1997. Association for Computational Linguistics.

Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, USA, 1997.

Preslav Nakov and Marti Hearst. Search engine statistics beyond the n-gram: Application to noun compound bracketing. In Ido Dagan and Dan Gildea, editors, *Proceedings of the Ninth Conference on Natural Language Learning (CoNLL-2005)*, University of Michigan, USA, June 2005. Association for Computational Linguistics.

Mark E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46: 323–351, 2005. ISSN 0010-7514.

Sonja Nießen and Hermann Ney. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, 30(2):181–204, 2004. ISSN 0891-2017.

Sonja Nießen, Franz Joseph Och, Gregor Leusch, and Hermann Ney. An evaluation tool for machine translation: Fast evaluation for machine translation research. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, pages 39–45, Athens, Greece, May 2000. European Language Resources Association.

Tomoko Ohta, Yuka Tateisi, and Jin-Dong Kim. The GENIA corpus: an annotated research abstract corpus in molecular biology domain. In *Proceedings of the Second Human Language Technology Conference (HLT 2002)*, pages 82–86, San Diego, USA, March 2002. Morgan Kaufmann Publishers.

Darren Pearce. Synonymy in collocation extraction. In *WordNet and Other Lexical Resources: Applications, Extensions and Customizations (NAACL 2001 Workshop)*, pages 41–46, June 2001.

Darren Pearce. A comparative evaluation of collocation extraction techniques. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, Spain, May 2002. European Language Resources Association.

Pavel Pecina. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, Marrakech, Morocco, June 2008.

Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, Anselmo Peñas, Vivien Petras, and Diana Santos, editors. *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, Revised Selected Papers*, volume 5152 of *Lecture Notes in Computer Science*, Budapest, Hungary, 2008. Springer. ISBN 978-3-540-85759-4.

Grzegorz Protaziuk, Marzena Kryszkiewicz, Henryk Rybinski, and Alexandre Delteil. Discovering compound and proper nouns. In *Proceedings of the RSEISP'07 International Conference on Rough Sets and Emerging Intelligent Systems Paradigms*, pages 505–515, 2007.

Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc, editors. *The PASCAL Recognising Textual Entailment Challenge*, volume 3944, 2005. Springer. ISBN 978-3-540-33427-9.

Carlos Ramisch, Paulo Schreiner, Marco Idiart, and Aline Villavicencio. An evaluation of methods for the extraction of multiword expressions. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 50–53, Marrakech, Morocco, June 2008.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico, February 2002. Springer.

Eric SanJuan, James Dowdall, Fidelia Ibekwe-SanJuan, and Fabio Rinaldi. A symbolic approach to automatic multiword term structuring. *Computer Speech & Language Special Issue on Multiword Expressions*, 19(4):524–542, 2005.

Violeta Seretan. *Collocation extraction based on syntactic parsing*. PhD thesis, University of Geneva, 2008.

Frank A. Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177, 1993. ISSN 0891-2017.

Aline Villavicencio. The availability of verb-particle constructions in lexical resources: How much is enough? *Computer Speech & Language Special Issue on Multiword Expressions*, 19(4):415–432, 2005.

Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In Jason Eisner, editor, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 1034–1043, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Yi Zhang, Valia Kordoni, Aline Villavicencio, and Marco Idiart. Automated multiword expression prediction for grammar engineering. In *Proceedings of the COLING/ACL Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 36–44, Sidney, Australia, July 2006. Association for Computational Linguistics.

Xiaohua Zhou, Xiaodan Zhang, and Xiaohua Hu. Dragon toolkit: Incorporating auto-learned semantic knowledge into large-scale text retrieval and mining. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence - ICTAI 2007*, volume 2, pages 197–201, Washington, USA, 2007. IEEE Computer Society.