Du traitement des langues en recherche d'information et vice versa

Vincent Claveau

UMR IRISA

UNIVERSITÉ DE RENNES 1

Habilitation à diriger des recherches

L'UNIVERSITÉ DE RENNES 1

École Doctorale N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : Informatique

Vincent Claveau

Du traitement des langues en recherche d'information et vice versa

Présentée et soutenue à Rennes, le 10 janvier 2020
Unité de recherche : IRISA

Composition du Jury :

Présidente :	Adeline Nazarenko	Professeur des universités, Univ. Paris-Nord
Rapporteurs :	Catherine Berrut	Professeur des universités, Univ. Grenoble, France
	Philippe Langlais	Professeur titulaire, Univ. de Montréal, Canada
	Jacques Savoy	Professeur titulaire, Univ. Neuchâtel, Suisse
Examinateurs :	Patrice Bellot	Professeur des universités, Univ. Aix-Marseille
	Olivier Dameron	Maître de conférences, Univ. Rennes 1

version du : 15 janvier 2020

Avant-propos

En premier lieu, je tiens à remercier le jury d'avoir accepté la charge de juger du travail rapporté dans les pages qui suivent, malgré les courts délais et leur agenda que je sais chargé.

Ce manuscrit d'habilitation ne porte en auteur que mon nom, mais il doit à un grand nombre de personnes. J'ai en effet eu la chance d'avoir de nombreuses collaborations avec des chercheuses et chercheurs de grande qualité, au sein de mes équipes de recherche successives (TexMex, de 2005 à 2014, puis LinkMedia depuis 2014, toutes deux au sein de l'IRISA, Rennes), mais surtout dans des laboratoires extérieurs. Ce mémoire est écrit sous la forme d'une collection de travaux choisis et d'une revisite d'articles rédigés sur la base de ces collaborations. Il convient donc de remercier ces collègues à deux titres : d'une part, pour leur collaboration fructueuse à l'époque, et d'autre part pour leur apport, différé dans le temps, à ce manuscrit. Je tâche de les citer à chaque fois que nécessaire et je leur prie de m'excuser tout oubli que je pourrais avoir fait. Je reviens en conclusion sur la prégnance de ces collaborations sur mon parcours de recherche.

Je souhaite également remercier plusieurs collègues, devenus au fil des missions et des conférences des amis. C'est à mon sens la réalisation la plus notable de ma carrière, même si elle se mesure plus en discussions – souvent maltées – qu'en publications. Merci Haïfa, Natalia, Karen, Max, Patrick...

Depuis ma prise de fonction, j'ai aussi eu la chance d'être accompagné par plusieurs doctorants, travaillant le plus souvent en co-encadrement avec des collègues. Il s'agit de Fabienne Moreau, Pierre Tirilly, Ali-Reza Ebadat, Abir Ncibi, Grégoire Jadi, Cédric Maigrot, ainsi que Clément Dalloux et François Torregrossa dont les thèses sont en cours. Ces jeunes chercheurs ont travaillé sur des sujets connexes à mes thèmes de recherche ; ils ont été particulièrement enrichissants et scientifiquement stimulants en me permettant de me confronter à des domaines parfois éloignés de mon champ d'expertise (*clustering* multimodal, vision par ordinateur...). Je ne présente pas directement leurs travaux dans ce manuscrit, préférant me concentrer sur des réalisations plus personnelles, mais il est évident que le manuscrit porte aussi leur empreinte, soit directement, par leur participation à des travaux menés conjointement, soit indirectement, par le biais des connaissances

acquises en les encadrant.

Plusieurs collègues expérimentés ont également été d'une grande bienveillance avec moi, me donnant des conseils sur le « métier », pas toujours suivis – je l'avoue – mais toujours écoutés. Ainsi, merci de nouveau à Catherine Berrut de m'avoir tiré les oreilles à chaque conférence CORIA pour ne pas avoir encore passé mon HDR. Il faudra désormais trouver d'autres oreilles... Merci à Béatrice Daille pour ses sympathiques discussions abordant indistinctement prises de responsabilités et restaurants étoilés. Merci à Pierre Zweigenbaum pour nos discussions scientifiques, et pour le modèle de tempérance et d'excellence qu'il donne. Merci à Philippe Blache de m'avoir mis le pied à l'étrier du GdR MaDICS, et de ses éclairages sur les arcanes du CNRS. J'ai une pensée très particulière pour Christine Collet, son franc-parler, sa bonne humeur, ses clins d'œil quand elle se commandait un dessert qu'elle me donnerait. À force de ses exhortations à me « débarrasser » de l'HDR, j'avais promis de soutenir avant la fin nos mandats à la direction du GdR...

Enfin, les mots manquent pour exprimer toute la gratitude que j'ai pour ma compagne. Son support, nos discussions scientifiques ou non, sa rigueur et sa bienveillance font que le travail qui est rapporté dans ce manuscrit porte sa marque à plus d'un titre. En clin d'œil pour nos enfants et leur amour des bandes dessinées, certaines illustrations, sortant un peu du cadre académique habituel, leur permettront d'avoir eux aussi leur marque dans les pages qui vont suivre.

Sommaire

Introduction 9

I TAL pour la RI : liens entre mots 15

1 Quelques apports du TAL en RI 16
1.1 Positionnement de nos travaux . . . 16
1.1.1 Morphologie . . . 17
1.1.2 Syntaxe . . . 18
1.1.3 Sémantique et connaissances . . . 20
1.2 Contribution à la prise en compte de la morphologie en RI . . . 21
1.2.1 Contexte et objectif . . . 21
1.2.2 Approche proposée . . . 22
1.2.3 Résultats . . . 23
1.3 Contribution à la translittération pour la RI . . . 24
1.3.1 Contexte et objectif . . . 24
1.3.2 Approche proposée . . . 25
1.3.3 Résultats . . . 25
1.4 Contributions à la prise en compte de relations sémantiques pour la RI . . . 26
1.4.1 Contexte et objectif . . . 26
1.4.2 Approche proposée . . . 26
1.4.3 Résultats . . . 27
1.5 Conclusion . . . 28

2 Morphologie compositionnelle 31
2.1 Travaux connexes . . . 33
2.2 Décomposition par alignement . . . 35
2.2.1 EM Alignment . . . 35
2.2.2 Normalisation morphologique automatique . . . 37
2.3 Evaluation de l'alignement . . . 38

2.3.1 Données et vérité terrain . . . 38
2.3.2 Résultats d'alignement . . . 39
2.4 Analyse morpho-sémantique pour la recherche d'information . . . 40
2.4.1 Graphes morpho-sémantiques . . . 41
2.4.2 Représentation morphémique pour la RI . . . 42
2.5 Expériences de RI biomédicale . . . 44
2.5.1 Contexte expérimental . . . 44
2.5.2 Résultats . . . 44
2.6 Conclusion et perspectives . . . 46

II RI pour le TAL : du bon usage de la similarité 48

3 Quelques apports de la RI en TAL 49
3.1 Positionnement de nos travaux . . . 49
3.1.1 Représentation . . . 50
3.1.2 Classification et RI . . . 51
3.1.3 La RI pour évaluer le TAL . . . 52
3.2 Similarité et fouille de texte . . . 52
3.2.1 Appariement avec un moteur de recherche . . . 53
3.2.2 Classification par k-ppv . . . 53
3.3 Similarités RI plus complexes . . . 54
3.3.1 Similarités entre sacs de sacs de mots . . . 54
3.3.2 Similarité de second ordre . . . 56
3.4 Conclusion . . . 59

4 Analyse distributionnelle par et pour la RI 61
4.1 Introduction . . . 62
4.2 État de l'art . . . 63
4.2.1 Construction de thésaurus distributionnels . . . 63
4.2.2 Évaluation des thésaurus distributionnels . . . 64
4.3 Modèles de RI pour l'analyse distributionnelle . . . 65
4.3.1 Principes et matériel . . . 65
4.3.2 Test des modèles de RI . . . 67
4.3.3 Test des modèles de réduction de dimension et d'*embedding* . . . 69

4.3.4 Analyse par fréquence . . . 71
4.3.5 Limites de l'analogie avec la RI . . . 73
4.4 Évaluation dans un cadre de RI . . . 75
4.4.1 Contexte expérimental . . . 75
4.4.2 Résultats d'extension . . . 76
4.5 Évaluation intrinsèque vs. évaluation extrinsèque . . . 77
4.5.1 Mise en regard des précisions intrinsèque et extrinsèque . . . 78
4.5.2 Faux positifs et bonnes extensions . . . 79
4.6 Conclusion . . . 80

5 Bilan et discussion **85**
5.1 Retour sur nos travaux . . . 85
5.1.1 À la croisée des domaines . . . 85
5.1.2 À l'épreuve du temps . . . 86
5.2 De la distinction entre RI et TAL . . . 87
5.2.1 Au-delà de la recherche documentaire . . . 88
5.2.2 RI translingue . . . 88

6 Évolution et perspectives **90**
6.1 Quelques pistes de recherche . . . 90
6.1.1 Apprentissage artificiel . . . 90
6.1.2 Multimodalité . . . 92
6.1.3 Éthique et sécurité . . . 93
6.2 Regard sur le passé : un rapprochement lent . . . 94
6.3 Regard vers le futur : des révolutions à venir ? . . . 96

A Curriculum **98**
A.1 Position et responsabilités actuelles . . . 98
A.2 Positions et responsabilités précédentes . . . 98
A.3 Encadrements . . . 99
A.4 Projets de recherche majeurs . . . 99
A.5 Bourses, prix . . . 100
A.6 Vulgarisation, presse . . . 100
A.7 Animation de la recherche . . . 101
A.8 Enseignement . . . 101

Bibliographie **103**

Introduction

La recherche d'information (RI) et le traitement automatique des langues (TAL) ont beaucoup en commun : ils traitent tous deux de langues naturelles et de textes. Il serait donc naturel qu'il y ait beaucoup d'interactions entre les communautés RI et TAL. Pourtant, à quelques exceptions notables près, elles ont longtemps évolué indépendamment, développant chacune un corpus de techniques et de connaissances propres, avec peu de contacts fructueux.

Des domaines distincts...

Du point de vue de la RI, les concepts et outils du TAL sont considérés avec défiance, comme le traduit la citation célèbre de Spärck-Jones (1999) :
« *It is not clear, [either] that NLP is required for some tasks that are closely related to ordinary retrieval.* »
En plus de leur éventuelle inutilité, les processus de TAL sont souvent vus comme trop coûteux pour être appliqués aux gigantesques ensembles de textes [1] qu'affectionnent les chercheurs en RI.

Du point de vue du TAL, les outils et concepts de la RI sont relativement mal connus, du moins pour ce qui est de leurs développements récents. Comme nous l'avions souligné dans (Claveau 2012), on réduit souvent les techniques de RI en la pondération TF-IDF. Ainsi, dans les actes de la conférence TALN de 2007 à 2011, la pondération TF-IDF était utilisée dans trente et un articles, souvent comme étant « état de l'art », alors que la pondération plus avancée Okapi-BM25 (Robertson et al. 1998), véritable standard de la RI à l'époque, ne l'était que dans quatre articles. Cet état de fait met bien en évidence le manque du rapprochement entre les deux communautés de chercheurs.

1. Par exemple, la collection ClueWeb contient plus d'un milliard de pages Web dans dix langues différentes.

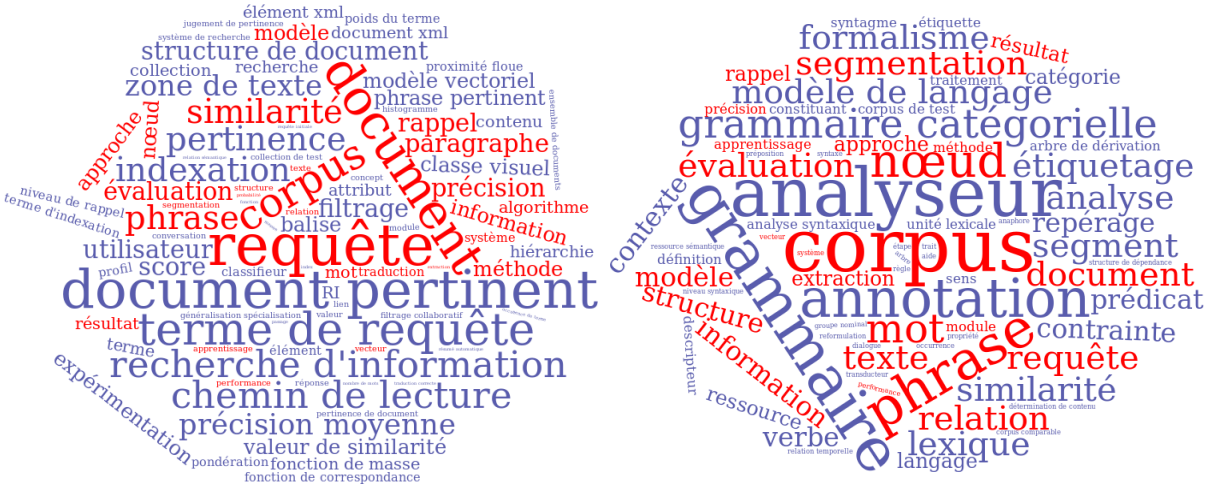


FIGURE 1 – Cent premiers termes extraits des actes de CORIA 2004 et 2005 (à gauche), et de TALN 2004 et 2005 (à droite). La taille des termes est fonction de leur représentativité ; les termes en commun sont en rouge.

...de moins en moins discernables

Ce rapprochement des deux communautés semble cependant inexorable. Il est le fait, d'une part, d'une évolution des besoins : à l'ère du « big data », il faut savoir à la fois gérer la quantité et faire sens des textes. D'autre part, la convergence des techniques des sciences des données, notamment l'apprentissage artificiel et les approches numériques, offre un socle commun en terme d'outillage et de culture. Les grandes conférences de chacun de ces domaines (ACL, SIGIR...) sont désormais peu discernables sur le plan des techniques mises-en-avant (plongements de mots, réseaux de neurones récurrents, etc. nous y revenons en conclusion). Ce rapprochement, entamé, se constate aussi à l'échelle des conférences francophones du domaine : TALN et CORIA. Nous avons employé notre outil d'extraction de terminologie TermEx [2] sur les actes des conférences de 2004 et 2005 d'une part, et sur ceux de 2018 et 2019 d'autre part. Les cent termes jugés par TermEx les plus représentatifs sont ici illustrés par des nuages, et les termes en commun aux deux conférences sont en rouge.

Les termes partagés par les auteurs de TALN et CORIA il y a quinze ans portent sur le matériau commun (corpus, texte, phrase...) et sur le vocabulaire scientifique général

2. TermEx, accessible en web-service repose sur un mélange de techniques traditionnelle du TALN (étiquetage en parties-du-discours par CRF, extraction par patrons morpho-syntaxiques), et de RI (pondération des candidats termes et comparaison au langage général). Il a été employé dans plusieurs de nos projets et vendu sous licence par le CNRS à des entreprises pour des besoins d'indexation de documents.

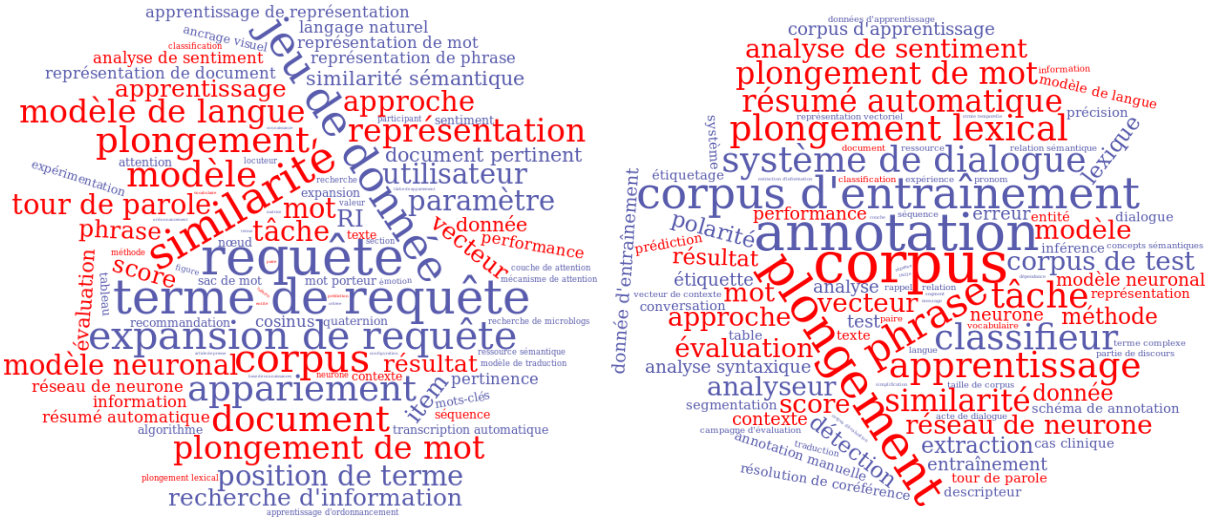


FIGURE 2 – Cent premiers termes extraits des actes de CORIA 2018 et 2019 (à gauche), et de TALN 2018 et 2019 (à droite). La taille des termes est fonction de leur représentativité ; les termes en commun sont en rouge.

(résultat, évaluation, précision...). La situation en 2018-2019 est bien moins contrastée, avec la présence de termes communs portant sur certaines tâches (résumé automatique, analyse de sentiments...) mais surtout sur les méthodes désormais partagées par les deux communauté (apprentissage, plongement lexical, réseau de neurones...).

Contexte et clés de lecture

C'est ce rapprochement entre TAL et RI, pour peu que l'on veuille les distinguer, qui est le fil conducteur principal de ce manuscrit. Au travers de la présentation d'une partie de nos travaux, nous montrons les allers-retours, les convergences, les synergies, qu'il peut y avoir entre ces deux domaines. Ce document n'offre donc pas une vue exhaustive de nos travaux, ni en largeur (tous n'y sont pas présentés), ni en profondeur [3] (tous les détails techniques n'y sont pas reportés). Ce document n'est pas non plus une revue de l'état-de-l'art, mais pour situer les travaux présentés dans un contexte plus large, nous proposons dans les chapitres 1 et 3 deux brefs panoramas des interactions entre TAL et RI.

Les travaux que nous présentons ont eu lieu après ma prise de poste à l'IRISA, en 2005, dans l'équipe de recherche TexMex ensuite renouvelée en LinkMedia. Centrées sur l'analyse de documents multimédias, ces équipes ont réuni des experts en vision par ordinateur,

3. Toutefois, outre les renvois bibliographiques, nous avons inséré lorsque nécessaire des liens cliquables qui permettront au e-lecteur intéressé de trouver des compléments sur certains points abordés.

en base de données, en audio/parole, et donc en TAL et RI. Ce contexte a évidemment joué un rôle important sur l'orientation de nos travaux. Nous avons ainsi bénéficié d'un environnement stimulant par la diversité des profils scientifiques des membres de l'équipe, mais revers de la médaille, le traitement du matériau écrit n'est donc qu'une composante portée par peu de personnes. Cette force de frappe, réduite comparée à des équipes entièrement dédiée au TAL, nous a conduit notamment à développer des approches ne nécessitant pas de préparation de données trop coûteuse ou impliquant trop de personnes, ou des développements de multiples modules basés sur des expertises linguistiques complexes, absentes de l'équipe. La constitution de données annotées a cependant été abordée en l'absence complète de données adaptées (commentaires sportifs (Fort et Claveau 2012), données cliniques en français (Grabar, Claveau et al. 2018 ; Dalloux et al. 2019)...), mais cela a constitué une activité marginale, faite au travers de collaborations. Les approches par apprentissage et le réemploi de données existantes ont donc été au centre de notre activité.

Plan

Ce manuscrit s'appuie grandement, pour sa structure, quelques éléments d'état de l'art et de discussions, sur la préface du numéro TAL et RI (Claveau et Nie 2016) co-écrite avec Jian-Yun Nie (Université de Montréal). Il est organisé en deux parties qui se répondent : le TAL pour la RI, et la RI pour le TAL. Chacune est divisée en deux chapitres, l'un rappelant et positionnant plusieurs de nos contributions de manière assez brève, et l'autre chapitre décrivant plus en détail un travail prototypique, reprenant pour cela une ou plusieurs de nos publications passées. Dans le chapitre 5, nous revenons sur plusieurs aspects communs aux travaux présentés, mais aussi sur la pertinence de la distinction faite ici entre TAL et RI. Nous proposons ensuite dans le chapitre 6 des réflexions plus prospectives en détaillant quelques pistes de recherche semblant prometteuses, mais aussi des considérations plus générales sur l'évolution de notre domaine de recherche. Enfin, l'annexe A présente quelques informations plus factuelles sur notre parcours de recherche.


QUE PENSES-TU DE MON PLAN ?..
JE N'EN PENSE RIEN. JE N'AI RIEN COMPRIS!

Première partie

TAL pour la RI : liens entre mots

Chapitre 1

Quelques apports du TAL en RI

La flexion pose des problèmes aux latinisants... et aux systèmes de RI.

Comme nous l'avons souligné en introduction, le TAL peut apporter une meilleure compréhension des textes et donc améliorer leur indexation et leur recherche. Cela se décline sur les différents niveaux d'analyse de la langue (morphologique, syntaxique, sémantique). Dans nos recherches passées, nous nous sommes intéressé à plusieurs de ces aspects, parfois appliqués à des domaines spécifiques. Ils ont pour point commun de chercher à découvrir des relations entre les mots ou les termes d'indexation, mais exploitent des indices différents pour mettre au jour ces relations. Les travaux résumés en section 1.2 s'appuient ainsi sur des indices morphologiques, ceux en section 1.3 exploitent les régularités de translittération et de traduction, et ceux décrits en section 1.4 étudient les relations entre les termes d'indexation eux-mêmes.

1.1 Positionnement de nos travaux

L'importance de bien prendre en compte la complexité de la langue semble une évidence pour les tâches relevant de la RI : si le contenu langagier d'un document ou d'une requête est bien compris, les performances de la recherche ne peuvent être que meilleures. Deux phénomènes touchent particulièrement les systèmes de RI : l'**ambiguïté** (deux énoncés identiques ont, en contexte, des sens différents) et le **paraphrasage** (deux énoncés différents, en surface, ont un même sens). Le premier va ainsi avoir tendance à dégrader

la précision, et le second le rappel du système. Pourtant, comme nous l'évoquions précédemment, les systèmes de RI classiques représentent le texte de manière très pauvre avec les sacs de mots, et calculent des similarités sur la base de comptages au sein de ces sacs de mots. Plusieurs travaux ont donc exploré les effets d'une meilleure prise en compte des phénomènes connus de la langue, en appliquant des méthodes ou des outils usuels du TAL. Un état de l'art encore d'actualité peut être trouvé dans (Moreau et Sébillot 2005a) ; nous ne présentons dans cette section que les grandes problématiques abordées par ces travaux en les distinguant selon les phénomènes considérés : morphologiques, syntaxiques ou sémantiques.

1.1.1 Morphologie

Les problèmes posés par la variation morphologique dans les systèmes de RI ont été notés très tôt. En effet, dans la représentation sac de mots, si aucun traitement n'est fait, deux formes graphiques différentes, même légèrement (marque du pluriel par exemple), vont correspondre à deux entrées différentes. Pour autant, ces variations de forme marquent des différences sémantiques qui ne sont pas toujours pertinentes pour la tâche de RI ; il s'agit donc d'un problème relevant du paraphrasage. Par exemple, un utilisateur cherchant des documents avec une requête '*dog*' pourra accepter des documents contenant '*dogs*'. Ces variations ont pour conséquence qu'une requête contenant une forme ne sera pas appariée à un document ne contenant que l'autre forme. Et plus généralement, les pondérations de ces formes au sein d'un document, fondées sur des comptages, seront sous-évaluées.

Pour traiter ce problème, des outils de racinisation (*stemming*) ont été développés en RI opérant par conflation, c'est-à-dire regroupant les variantes sous une forme standardisée qui sera celle gardée dans le sac de mots. Ces outils diffèrent des techniques de lemmatisation issues du TAL, à la fois dans leur fonctionnement, et dans leur couverture des variations morphologiques traitées. Les raciniseurs sont le plus souvent des algorithmes reposant sur des heuristiques (suppression des préfixes et suffixes connus, suppression des diacritiques...) et traitent non seulement les cas de flexion, mais aussi certains cas de dérivation (changement de partie du discours et/ou modification légère du sens, comme *chanteur/chanter, louer/location, compiler/décompiler*) (Lovins 1968 ; Porter 1980 ; Savoy 1993). La flexion est la relation existant entre deux mots que seuls distinguent le nombre, le genre, le temps, personne et mode pour les verbes, ou le cas pour les langues à déclinaisons (allemand, polonais, russe, et bien sûr latin comme illustré en tête de chapitre...).

Les différents mots liés par la flexion sont appelés formes fléchies ; parmi celles-ci on choisit souvent un unique représentant, le lemme. Par exemple, pour les verbes en français, le lemme est la forme infinitive, pour les adjectifs, le masculin singulier... En morphologie dérivationnelle, deux mots reliés morphologiquement possèdent une racine commune, et diffèrent par leurs affixes (principalement des préfixes comme re- dans reconstruire ou des suffixes comme -eur dans constructeur, mais aussi des infixes dans certaines langues). Cette dérivation s'accompagne alors d'une modification légère de sens (comme dans faire ↔ défaire) et/ou de catégories grammaticales (décider ↔ décision).

La comparaison de la racinisation avec des outils de lemmatisation montre des résultats variables selon les langues et les tâches de RI, comme nous et d'autres auteurs ont pu le montrer (Moreau, Claveau et Sébillot 2007 ; Savoy 2002). Si la racinisation est très largement utilisée en RI, la prise en compte d'autres phénomènes morphologiques est plus rare. Certains chercheurs (Gaussier 1999 ; Moreau, Claveau et Sébillot 2007) ont néanmoins proposé des approches permettant de traiter plus finement les cas de dérivation (et aussi de flexion), dans un grand nombre de langues, et sans besoin de connaissances ou de ressources externes, en s'appuyant, par exemple, sur des analogies formelles (voir également la section 1.2).

Dans certains domaines de spécialité, les phénomènes morphologiques peuvent être plus riches et sont autant de freins à l'accès à l'information. Ainsi, dans le domaine biomédical, les cas de composition sont fréquents dans la construction des termes spécialisés. Leur prise en compte permet ainsi de noter que `stomachalgie`, `gastrodynie` et `maux d'estomac` sont synonymes, ou de mettre en correspondance une requête sur le `foie` avec un document portant sur l'`hépatite` (Claveau et Kijak 2013). Nous y reviendrons dans le chapitre suivant.

Notons enfin que certaines langues nécessitent également des modules morphologiques dédiés traitant par exemple de l'agglutination (allemand, turc...) (Haddad et Bechikh Ali 2014), de la voyellation (par exemple pour la RI arabe) (Grefenstette et al. 2005) et de segmentation (chinois, japonais) (Peng et al. 2002).

1.1.2 Syntaxe

Avec la représentation sac de mots, la construction syntaxique des textes est complètement ignorée. D'une part, il n'est plus possible de distinguer « `La victoire de la France sur l'Italie` » de « `la victoire de l'Italie sur la France` » ou « `une pomme est tombée à terre` » de « `une pomme de terre est tombée` ». Il s'agit donc

d'un problème d'ambiguïté de la représentation des textes, que plusieurs travaux ont bien entendu cherché à résoudre, ou amoindrir, en conservant des informations syntaxiques plus ou moins riches. D'autre part, des énoncés de même sens peuvent se trouver sous des formes différentes ; considérons par exemple (Strzalkowski et al. 1999) les groupes de mots `information retrieval, retrieval of information, retrieve more information` et `information that is retrieved`... Il s'agit dans ce cas d'un problème de paraphrasage, qui peut être résolu si l'on est capable de ramener ces formes à la relation `retrieve+information` où `retrieve` est l'élément tête et `information` son modifieur.

Selon les travaux, les phénomènes syntaxico-sémantiques considérés (et la façon de les nommer) varient. Il peut s'agir d'expressions multi-mots, de syntagmes, de phrasèmes, de mots composés, de collocations, etc., avec des propriétés également variables : expressions plus ou moins figées, continues ou discontinues, compositionnelles [1] (`sténose aortique`) ou non (`pomme de terre, greenhouse gas effect`). Ces expressions sont détectées avec des méthodes qui peuvent reposer sur des indices purement numériques (par exemple, calculs de *Pointwise Mutual Information* (Acosta et al. 2011)), sur des analyses de surface (patrons de parties-du-discours), sur des analyses syntaxiques complètes, ou bien encore sur l'utilisation de ressources externes (par exemple des lexiques et/ou des outils d'annotation comme l'UMLS et l'analyseur MMTx (Aronson et Lang 2010) pour le domaine biomédical (Shen et Nie 2015), ou des logs de moteurs de recherche généralistes (Chapelle et Chang 2011)).

Une fois détectés, ces liens syntaxiques peuvent être utilisés dans le système de RI de différentes façons. Les expressions multi-mots peuvent servir à étendre les requêtes. Elles peuvent aussi être considérées dès l'indexation de manière figée, comme étant un seul mot ; dans notre exemple précédent `retrieve_information` sera considéré comme un terme d'indexation. Cette solution est facile à mettre en œuvre puisqu'elle ne remet pas en cause l'architecture des systèmes de RI reposant sur la représentation sac de mots. Mais des travaux ont également montré qu'il était possible d'inclure une information syntaxique riche dans la représentation des documents, et donc dans le calcul de similarité (Maisonnasse et al. 2008 ; Jianfeng Gao, Nie et al. 2004), intégrant ainsi dans le même temps les problèmes de l'ambiguïté et du paraphrasage. Ces travaux demandent alors une connaissance fine des processus de RI mais permettent une représentation plus riche du contenu des documents.

1. La compositionnalité est la propriété de pouvoir construire le sens d'un énoncé à partir des sens mots qui le composent.

Enfin, plus récemment, certaines architectures de réseaux de neurones ont permis de s'affranchir de cette représentation sac-de-mots, et ont donc permis une prise en compte d'information plus syntaxique. C'est le cas des réseaux convolutifs qui permettent de capturer des dépendances locales (typiquement de l'ordre de 3 à 4 mots) et des réseaux récurrents de type LSTM (Hochreiter et Schmidhuber 1997) dans lesquels les séquences de mots sont entièrement prises en compte.

1.1.3 Sémantique et connaissances

Bien entendu, une meilleure prise en compte de la sémantique a été envisagée très tôt en RI. Dans une certaine mesure, celle-ci a visé à résoudre les problèmes d'ambiguïté, par exemple en identifiant le sens précis d'un mot-forme (Zhong et Ng 2012), mais la majorité des travaux se sont intéressés au problème du paraphrasage. Dans ce dernier cas, les outils ou ressources utilisés ont pour but d'enrichir la description sac de mots avec des énoncés équivalents pour faciliter l'appariement entre la requête et les documents pertinents. Comme précédemment, les deux questions qui se posent sont celle de l'obtention de ces informations sémantiques et celle de leur intégration dans le système de RI.

Les lexiques sémantiques externes ont largement été utilisés. Bien que les premières expériences rapportaient des résultats négatifs [2], ces ressources apportent des gains parfois importants si elles sont adaptées aux documents et si elles sont bien intégrées au calcul de similarité. Les outils de sémantique du TAL, par exemple l'analyse distributionnelle, ont aussi montré de très bons résultats (Besançon et al. 1999 ; Claveau et Kijak 2015), avec l'avantage de pouvoir être appliqués sur les documents à indexer, et donc adaptés au domaine traité. La révolution des plongements de mots et des représentations continues a bien sûr offert une nouvelle vue sur ces problèmes de représentation du sens et leur exploitation. Nous montrons dans le chapitre 4 comment ces représentations, reposant sur les mêmes concepts que l'analyse distributionnelle, peuvent être avantageusement utilisées en RI. Bien connues en RI, les techniques de type Latent Semantic Indexing (LSI (Deerwester et al. 1990)) ou Latent Dirichlet Allocation (LDA, (Hoffman et al. 2010)) abordent également ce problème sémantique en proposant une représentation du document non plus dans un espace de mots, mais dans un espace de « concepts » (en fait, des combinaisons linéaires de mots apparaissant souvent dans les mêmes documents).

2. C'est le cas notamment des travaux de Voorhees avec WordNet (Voorhees 1994), qui, bien que réfutés par de nombreuses autres études par la suite, restent largement cités comme preuve de l'inutilité des processus TAL de sémantique lexicale en RI.

Ces ressources sémantiques, existantes ou calculées, sont souvent simplement adjointes au système de RI sans modification profonde de son fonctionnement. Cela peut être par extension de requête, ou, plus rarement, en modifiant la représentation du document (extension des documents, c'est-à-dire ajout des synonymes dès la phase d'indexation des documents, ou représentation par *synsets*), et parfois durant la phase de calcul de similarité (par exemple, par des techniques de *back-off* dans les modèles de langues), ou enfin par des techniques plus intégrées, comme des réseaux de neurones siamois permettant de tirer parti naturellement des plongements de mots et des capacités de représentations séquentielles des LSTM (voir plus haut).

Pour aller plus loin que la sémantique lexicale, l'inclusion de connaissances riches et structurées sur le monde et le raisonnement à partir de ces connaissances sont également étudiés. Ce domaine, appelé RI sémantique, mêle donc fortement la RI, le TAL mais aussi des problématiques propres à l'ingénierie des connaissances (Zargayouna et al. 2015 ; Grau et al. 2015).

1.2 Contribution à la prise en compte de la morphologie en RI

Dans de nombreuses langues, certains mots partagent une proximité graphique et sémantique ; on parle de relations morpho-sémantiques. Ainsi en français, le verbe transformer est lié à transformes, transforme, transformateur, transformation... Des termes qui dérivent du même lemme ou de la même racine présupposent donc généralement un sens proche. En ce sens, la variation morphologique constitue un type particulier de variation sémantique qu'il est intéressant de capturer, notamment en RI.

1.2.1 Contexte et objectif

Comme nous l'avons vu en section 1.1.1, on distingue usuellement plusieurs types de relations morphologiques (Mel'čuk 2000) ; celles qui nous intéressent dans ce travail sont la flexion et la dérivation. Pour beaucoup de langues, il existe des outils (raciniseurs, lemmatiseurs) et pour certaines des bases de connaissances morphologiques, mais notre objectif dans ce travail (Moreau et Claveau 2006 ; Moreau, Claveau et Sébillot 2007) était d'en dépasser plusieurs limites : c'est-à-dire d'avoir une approche traitant à la fois la flexion et la dérivation, ne nécessitant pas de connaissances expertes, entièrement

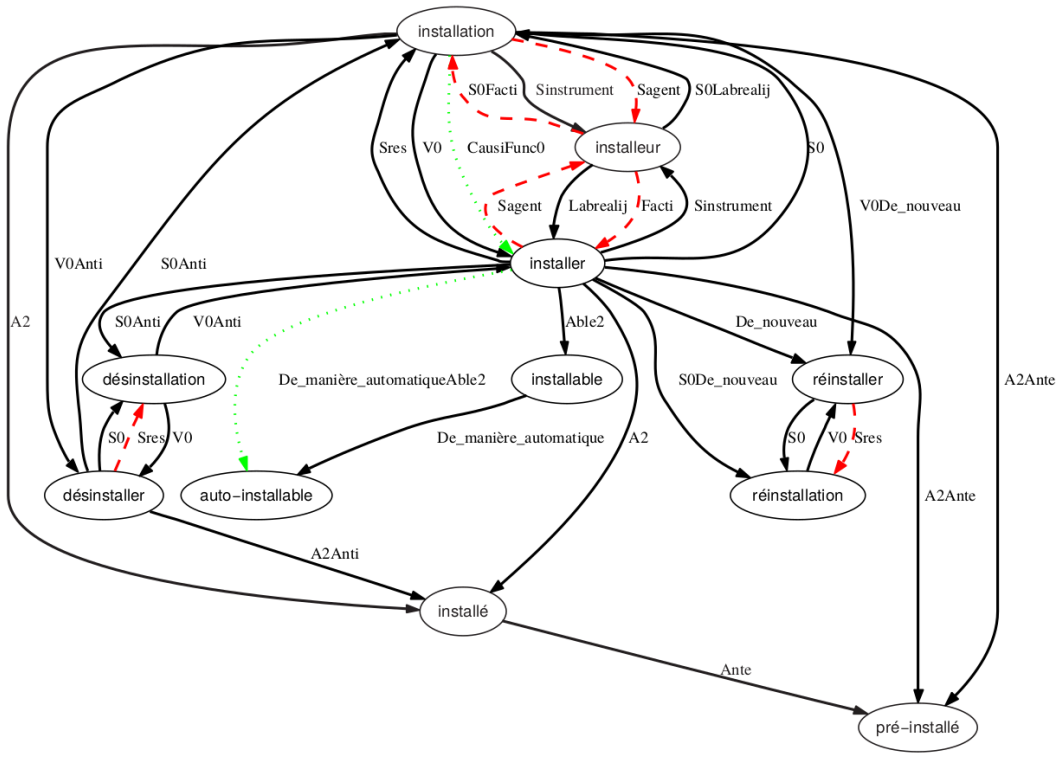


FIGURE 1.1 – Graphe des relations sémantiques de la famille morphologique du terme installer, en informatique, construit par analogie à partir d'exemples de paires de termes et de leurs fonctions lexicales. Les liens en rouge sont ceux erronés, et les verts sont ceux manquants.

automatique, et pouvant donc s'appliquer facilement à un grand nombre de langues. Ces relations morphologiques, une fois repérées, sont utilisées simplement pour étendre des requêtes dans un système de recherche d'information.

1.2.2 Approche proposée

L'approche que nous avons adoptée pour acquérir les variantes morphologiques des mots contenus dans les requêtes s'appuie sur une technique que nous avions développée initialement à des fins terminologiques (CLAVEAU et L'HOMME 2005b ; CLAVEAU et L'HOMME 2005a). Elle nous permettait de détecter automatiquement des liens sémantiques (représentées en l'occurrence par les fonctions lexicales issues de la théorie Sens-Texte) et donc de structurer des connaissances terminologiques. Cela est illustré en figure 1.1 sur les termes dérivés de installer dans le domaine de spécialité de l'informatique.

Le principe de cette technique d'acquisition morphologique est relativement simple

et s'appuie sur la construction d'analogies. En toute généralité, une analogie peut être représentée formellement par la proposition A : B ≐ C : D, qui signifie « A est à B ce que C est à D » ; c'est-à-dire que le couple A-B est en analogie avec le couple C-D. L'analogie a été largement utilisée en TAL, notamment pour la traduction et la translittération (Langlais et al. 2009 ; Langlais 2013), et son utilisation en morphologie, assez évidente, a déjà fait l'objet de plusieurs travaux (Hathout 2001 ; Lepage 2003, *inter alia*) : par exemple, si l'on postule l'analogie

connecteur : connecter ≐ éditeur : éditer

et si l'on sait par ailleurs que connecteur et connecter partagent un lien morpho-sémantique, on peut alors supposer qu'il en est de même pour éditeur et éditer.

Le préalable essentiel à l'utilisation effective de l'apprentissage par analogie est la définition de la notion de similarité qui permet de statuer que deux paires de propositions – dans notre cas deux couples de mots – sont en analogie. La notion de similarité que nous avons proposée est simple mais adaptée aux nombreuses autres langues dans lesquelles la flexion et la dérivation sont principalement obtenues par préfixation et suffixation. Elle considère que deux paires de mots m_1-m_2 et m_3-m_4 sont en analogie si on peut trouver un schéma de réécriture des préfixes et suffixes identique permettant de passer de m_1 à m_2 et de m_3 à m_4.

Notre processus de détection de variantes morphologiques consiste ainsi à vérifier si un couple de mots inconnu m_3-m_4 est en analogie avec un ou plusieurs exemples de couples connus m_1,m_2 (par exemple désinstaller - réinstallation). Par exemple, nous pouvons déterminer que déshydrater : réhydratation ≐ désinstaller : réinstallation (c'est à dire m_3-m_4 est en analogie avec le couple-exemple m_1,m_2).

1.2.3 Résultats

À l'inverse de nombreux travaux existants à l'époque, la technique proposée présente la particularité de ne nécessiter aucune ressource ni connaissance externe, et d'être ainsi applicable à une grande variété de langues. Les évaluations de cette approche pour étendre des requêtes ont été réalisées sur des collections de RI couvrant six langues européennes, et comparées à différents outils existants (*stemmer*, lemmatiseur). Les résultats attestent de l'intérêt de la méthode puisqu'une amélioration significative des performances des système de recherche d'information est constatée dans tous les cas, sur toutes les langues, notamment sur la MAP et les précisions à rangs élevés (P@x avec x > 50).

1.3 Contribution à la translittération pour la RI

L'approche précédente a montré l'intérêt de manipulations simples sur les chaînes de caractères pour faire le lien, au sein d'une langue, entre des mots d'une même famille morphologique. C'est cette même philosophie que nous avons utilisée pour la translittération, c'est-à-dire pour faire le lien entre certains mots en relation de traduction (Claveau 2009a).

1.3.1 Contexte et objectif

Ce travail s'inscrit dans le domaine biomédical, dans lequel les problématiques de recherche et de traitement d'informations textuelles sont particulièrement importantes. Par exemple, la base PubMed regroupe actuellement 30 millions de publications scientifiques dans le domaine médical et fait face à plusieurs millions de requêtes par jour. Dans la plupart de ces bases documentaires médicales, chacun des documents est indexé à l'aide de terminologies de référence, notamment le thésaurus MeSH® (*Medical Subject Heading*) développé par la U.S. National Library of Medicine (NLM). Par ailleurs, la prédominance de l'anglais dans ces terminologies de référence rend cruciale la mise en place de stratégies multilingues pour faciliter l'accès à ces bases pour les non-anglophones. Des terminologies biomédicales multilingues existent, mais elles sont mises en défaut par l'évolution rapide des connaissances, et donc des termes du domaine, et le manque de ressources pour certaines langues.

L'objectif de ce travail était donc de produire des traductions de termes simples (*i.e.* composés d'un seul mot) du domaine biomédical d'une langue source dans une langue cible. Ce travail repose sur deux hypothèses majeures :

1. dans le domaine biomédical, les termes équivalents entre deux langues sont souvent morphologiquement proches ;
2. les différences entre termes de chaque langue sont régulières et peuvent être apprises automatiquement.

Ces deux hypothèses tirent parti du fait que les termes biomédicaux sont construits sur les mêmes racines grecques et latines, et leurs dérivations très régulières (*e.g.* pour le couple français-anglais ophtalmorragie/ophthalmorrhagia, ophtalmoplastie/ophthalmoplasty, leucorragie/leukorrhagia). La technique décrite et les expériences rapportées ici portent sur la traduction entre différentes langues (anglais, espagnol, français, russe…) ; l'accent sera toutefois

mis sur la traduction d'une langue source quelconque vers l'anglais, qui correspond au cas le plus intéressant pour la RI interlingue dans ce domaine.

1.3.2 Approche proposée

Notre approche s'appuie sur une technique d'apprentissage artificiel que nous avons développée. Elle nous permet d'inférer un système de traduction à partir de couples de termes langue source-langue cible en relation de traduction et morphologiquement proches. C'est ce système qui, étant donnés en entrée des termes dans la langue source – dans notre cas, une requête – doit ensuite permettre de produire les termes correspondants dans la langue source. Plus précisément, notre approche repose sur l'apprentissage de règles de réécriture (que l'on peut aussi voir comme des règles de translittération). Ces règles, apprises à partir de listes de paires bilingues de termes du domaine, sont de la forme : $\langle input\ string \rangle \rightarrow \langle output\ string \rangle$.

Pour traduire un terme inconnu dans la langue d'entrée, toutes les règles apprises applicables à ce terme (*i.e.* les règles dont la prémisse correspond à une sous-chaîne du terme) sont effectivement appliquées. Dans le cas de règles concurrentes, toutes les possibilités sont générées. À ce niveau, pour un terme, un grand nombre de traductions candidates sont produites. Pour choisir parmi ces possibilités la traduction la plus probable, nous avons utilisé une approche simple basée sur les modèles de langue (Charniak 1993).

1.3.3 Résultats

L'évaluation que nous avons menée a montré que notre approche offre des performances de traduction variables selon les langues mais d'assez bon niveau pour être utilisée dans un contexte de RI interlingue. À partir de la collection OHSUMED telle qu'utilisée dans TREC-9, nous avons évalué les résultats qu'obtiendrait un système de RI utilisant notre technique pour traduire des requêtes en français, italien, portugais, espagnol ou russe vers l'anglais. Les résultats ont clairement montré l'intérêt de notre approche mais ont aussi mis en évidence une large marge de progression possible.

1.4 Contributions à la prise en compte de relations sémantiques pour la RI

La prise en compte de relations sémantiques pour la RI a fait l'objet de nos tout premiers travaux, en thèse (Claveau 2003). Ils s'inscrivaient dans un cadre linguistique bien défini, celui du lexique génératif (Pustejovsky 1995), mais s'appliquaient à la langue générale. Nous n'y revenons pas dans ce manuscrit. Dans des domaines spécialisés, avec des contextes de recherche d'information particuliers, d'autres relations, plus spécifiques, sont intéressantes, comme dans le domaine médical sur lequel nous revenons ci-dessous.

1.4.1 Contexte et objectif

Comme nous l'avons expliqué dans la section précédente,le thésaurus MeSH® (Medical Subject Headings) est l'outil de prédilection pour indexer la littérature dans le domaine biomédical. Cette indexation est en langage contrôlé, par opposition à l'indexation plein-texte, est principalement manuelle et évidemment très coûteuse. Ainsi, chaque document référencé dans MEDLINE est associé à en moyenne une quinzaine de mots-clés représentant les concepts abordés par les auteurs parmi les quelque 25 000 du thésaurus (*e.g. aphasia*, *patient care*, *hand*...). Le cas échéant, ces mots-clés doivent ensuite être précisés à l'aide des 80 qualificatifs (*e.g. surgery*, *pharmacology*). Pour chaque mot-clé, le MeSH définit un sous-ensemble de qualificatifs « affiliables », de sorte que seules certaines paires peuvent être formées. Par exemple, les paires *aphasia/metabolism* ou *hand/surgery* sont autorisées, mais pas *hand/metabolism*.

La complexité de la tâche d'indexation et le volume croissant de données à traiter fait de l'automatisation de l'indexation une nécessité. En particulier, en ce qui concerne l'indexation MeSH, un progrès notable repose sur l'extraction par des outils automatiques de paires mot-clé / qualificatif et non seulement des mots-clés isolés. L'objectif du travail présenté ici était donc d'inférer automatiquement de nouvelles règles d'indexation impliquant des qualificatifs.

1.4.2 Approche proposée

Pour ce faire, nous adoptons une approche originale en utilisant une méthode d'apprentissage symbolique particulière, la programmation logique inductive (PLI) (Névéol et al. 2008). La PLI est une technique d'apprentissage symbolique supervisée permettant

d'inférer des règles, exprimées sous forme de clauses logiques (clauses Prolog), à partir d'exemples, eux aussi décrits en Prolog (Muggleton et Raedt 1994).

Nous avons adapté les principes de la PLI aux spécificités des données MEDLINE afin d'obtenir des règles pertinentes permettant de traiter efficacement une grande quantité de données. Ces règles, que nous cherchons à inférer, sont du type :

```
Si un terme de l'arborescence « Anatomy » ainsi qu'un « Carboxylic Acids » sont
recommandés pour l'indexation, alors la paire « [Carboxylic Acids]/pharmacology »
doit également être recommandée.
```

1.4.3 Résultats

Nous avons entraîné notre technique sur des documents existants, et donc annotés par des experts. Les performances des règles PLI ont ensuite été évaluées sur quelques qualificatifs pour lesquelles des règles manuelles avaient été proposées par des experts de la NLM. Cela a permis de juger de l'intérêt de notre approche automatique (meilleur rappel que les humains, et baisse très légère en précision). Ces dernières sont beaucoup plus ciblées, ce qui produit de très bons taux de précision mais des rappels généralement faibles, inférieurs non seulement à la PLI mais aussi à la *baseline* (sauf dans le cas de *metabolism*).

La PLI a donc permis d'obtenir rapidement un grand nombre de règles. En les examinant, il apparaît parfois que des règles un peu différentes obtenant des performances proches semblent sémantiquement meilleures pour un expert du domaine. De même, certaines règles PLI peuvent permettre à un expert de créer manuellement une nouvelle règle qui n'aurait pas pu être inférée à partir du corpus d'entraînement. Ainsi, il pourrait être possible d'optimiser la production de règles d'indexation en envisageant une relecture des règles PLI par un expert afin d'améliorer la lisibilité et les performances des règles tout en minimisant le temps passé à la production des règles.

Enfin, ces résultats sont d'autant plus remarquables qu'ils n'exploitent que des régularités implicites des descriptions et non le contenu de l'article et permettent ainsi d'envisager l'utilisation d'autres stratégies de recommandations de qualificatifs basées cette fois sur le corps de l'article à indexer.

1.5 Conclusion

Le survol de nos travaux passés fait dans ce chapitre appelle plusieurs commentaires. Le premier et principal est bien sûr qu'il y a un véritable intérêt à incorporer des connaissances issues de traitements de la langue dans les systèmes de recherche d'informations. Même si les améliorations constatées sont parfois d'ampleur limitée, elles sont tout de même régulières. Dans nos travaux, ces connaissances sont principalement sémantiques : la morphologie sert à trouver des flexions ou des dérivations, qui ont pour propriété de conserver ou d'altérer très légèrement le sens (Mel'čuk 2000), la translittération sert à trouver des équivalents dans une autre langue, notre travail sur l'indexation MeSH exploite les régularités des associations paradigmatiques des termes d'indexation.

Le deuxième point à souligner est que ces travaux concernent la langue générale ou des langues de spécialité, en l'occurrence ici, du domaine biomédical. Certains phénomènes à prendre en compte sont transverses, mais d'autres sont propres à chaque domaine comme nous le verrons avec les termes morphologiquement complexes traités dans le chapitre suivant. Si plusieurs de nos travaux passés se sont positionnés dans le domaine biomédical, c'est qu'il réunit plusieurs caractéristiques intéressantes pour la recherche :

- les besoins applicatifs sont importants et les impacts sociétaux très directs. Les projets FiGTeM et BigClin [3] que nous avons portés récemment s'intéressaient par exemple à l'épidémiologie, la pharmacovigilance, le recrutement de patient pour des essais cliniques, nécessitant ainsi des compétences en extraction d'information et en indexation de document.
- des données textuelles en grande quantité existent. Il s'agit bien sûr des données cliniques (même si on peut regretter la difficulté d'accès à des données réelles en France), mais aussi la littérature scientifique et médicale, les textes de formation, les réseaux sociaux et forums de discussion spécialisés...
- des ressources additionnelles sont également disponibles (thésaurus multilingue, ontologie, classifications).

Enfin, les différents travaux présentés illustrent aussi notre choix récurrent d'utiliser des approches par apprentissage, et autant que possible exploitant peu de connaissances (comme nos travaux reposant uniquement sur les chaînes de caractères) ou alors des connaissances déjà existantes (par exemple, l'UMLS). Les méthodes de fouille ou d'apprentissage utilisées sont dédiées, soit en adaptant des cadres existants (comme la Pro-

3. La description du projet BigClin financé par l'ANR via le LabEx CominLabs est disponible à l'adresse suivante https://bigclin.cominlabs.u-bretagneloire.fr/

grammation Logique Inductive), soit en développant complètement la méthode (comme pour la translittération). De ce fait, les approches proposées ne sont pas dépendantes de la langue mais de la disponibilité de données d'entraînement ; quand cela était possible, nous avons montré leurs bonnes performances sur plusieurs langues (ou paires de langues pour la translittération). On peut également noter que l'ensemble des approches présentées sont principalement symboliques : nous exploitons les lettres, attributs discrets, ou les termes MeSH, sur lesquels nous inférons des règles. Cela contraste avec les approches d'apprentissage statistique plus en vogue, mais permet une interprétation des résultats utile dans certaines circonstances, notamment lorsqu'il est nécessaire que le classifieur obtenu soit appréhendable par l'utilisateur final ou par un expert. Nous reviendrons sur la pérennité de ces approches dans la conclusion de ce manuscrit.

Le chapitre suivant présente un travail synthétisant différents aspects de nos travaux présentés dans ce chapitre : il porte sur la morphologie dans le domaine biomédical pour lequel nous avons développé une approche originale exploitant de nouveau le multilinguisme de l'UMLS. Nous présenterons dans le chapitre 4 un travail portant cette fois sur la sémantique, au travers de l'analyse distributionnel et des plongements de mots, et de leur utilité en RI.

Chapitre 2

Morphologie compositionnelle

Au grand dam de Panoramix, les racines gréco-latines sont particulièrement présentes dans les termes médicaux.

Les relations de flexion et de dérivation à l'étude dans le chapitre précédent ne sont pas les seuls liens morphologiques intéressants pour la RI. La composition permet aussi d'accéder au sens des mots par l'étude de leur forme. Comme nous l'indiquions précédemment, la composition est un phénomène relativement rare dans la langue générale mais très courant dans certains domaines de spécialité. C'est le cas notamment du domaine biomédical, dans lequel les termes morphologiquement complexes sont très courants et créés continuellement. Comme nous l'avons vu dans le chapitre précédent, ces termes spécialisés (par exemple le MeSH) sont la clé de l'accès à l'information dans ce domaine dans lequel les bases de documents sont très consultées (PubMed). La maîtrise de la complexité morphologique de ces termes biomédicaux est donc essentielle pour la recherche d'information biomédicale.

Dans ce chapitre, nous présentons un travail initié en 2010 avec Ewa Kijak (Univ. Rennes 1) proposant une technique d'analyse morphologique des termes bio-médicaux

(Claveau et Kijak 2011) et son utilisation dans une problématique de recherche d'information. Elle permet de décomposer les termes en composants morphologiques, les morphes[1], souvent issus du grec et du latin (n'en déplaise à Panoramix), et d'associer à ces derniers des informations sémantiques. Ces décompositions et les informations sémantiques associées sont ensuite exploitées pour permettre une indexation souple des documents dans un système de recherche d'information. Au contraire des travaux existants dans ce domaine (Deléger et al. 2008 ; Markó, Schulz et Han 2005), qui sont résolument basés sur l'expertise humaine, notre technique de décomposition est automatique et exploite les données existantes dans le domaine biomédical. Elle est donc de fait non supervisée (nous revenons sur ce point en conclusion de ce chapitre).

L'idée au cœur de notre technique de décomposition est d'exploiter l'aspect multilingue des bases terminologiques qui existent dans le domaine biomédical. Nous utilisons le japonais comme langue pivot, et plus précisément les termes écrits en kanjis (caractères chinois utilisés en japonais), pour mener la décomposition des termes dans une autre langue à l'aide d'une technique d'alignement (Claveau et Kijak 2011). En plus d'aider à la décomposition des termes, les kanjis servent aussi à étiqueter les morphes et fournissent ainsi une sorte de représentation sémantique. Ainsi, le terme anglais *photochimiotherapy* correspond au terme japonais par 光化学法 ; l'alignement de ces deux termes mène à la décomposition suivante : *photo* ↔ 光 ('lumière' en japonnais), *chimio* ↔ 化学 ('chimie'), *therapie* ↔ 法 ('thérapie', 'traitement'). Comme cet exemple l'illustre, chaque morphe est associé à des kanjis qui peuvent servir de descripteurs plus pertinents que les termes eux-mêmes pour indexer des documents (trop rares, les termes complexes ne permettent pas un bon rappel). Nous montrons en particulier dans cet article comment les correspondances construites entre morphes et kanjis peuvent être exploitées de différentes façons pour améliorer les résultats d'un système de recherche d'information.

L'analyse morphologique et l'indexation des documents qu'elle permet dépendent donc de l'étape essentielle d'alignement entre morphes et kanjis. Celle-ci est mise en œuvre par une technique originale, adaptée aux données traitées, reposant sur un algorithme *Forward-Backward* et par de l'apprentissage analogique. Après une description des travaux connexes en section 2.1, nous présentons cette technique d'alignement et ses résultats en terme de décomposition morphologique en sections 2.2 et 2.3. L'utilisation des produits de ces décompositions dans un contexte de recherche d'information est détaillée en section 2.4.

1. Dans ce mémoire, nous distinguons les morphes, signes linguistiques élémentaires (segments), des morphèmes, classes d'équivalences de morphes partageant un signifié identique et des signifiants proches (Mel'čuk 2006).

Les résultats obtenus sur une collection biomédicale sont présentés dans la section 2.5.

2.1 Travaux connexes

La morphologie a été étudiée dans de nombreux cadres (lexicologie, terminologie, recherche d'information...). Comme nous l'avons expliqué précédemment, elle tient un rôle particulier dans le domaine biomédical où les termes sont issus d'opérations morphologiques complexes mais régulières et productives. Malheureusement, au contraire des terminologies, très largement développées dans le domaine biomédical, il n'existe pas de bases de données de morphes enrichis avec des informations sémantiques, complètes et maintenues. Par ailleurs, la décomposition d'un terme en morphes, qui permettrait de tirer parti d'une telle base, reste aussi un problème. Notons enfin que les outils communément employés en RI pour tenir compte de la morphologie, notamment la racinisation, sont inadaptés à la complexité des constructions morphologiques à traiter.

En ce qui concerne l'utilisation de la morphologie comme outil d'analyse de termes ou de mots, il convient de distinguer deux visions du problème. Dans la vision lexématique, des relations sémantiques entre mots sont détectées en s'appuyant sur leur forme, mais sans besoin de décomposition (Grabar et Zweigenbaum 2002 ; Claveau et L'Homme 2005c, par exemple). À l'opposé de cet emploi implicite de la morphologie, la vision morphémique repose sur une décomposition en morphe en préalable à tout traitement. Beaucoup de travaux se placent dans ce cadre, qui est aussi celui adopté dans nos travaux. Les approches existantes sont soit fortement manuelles (Deléger et al. 2008 ; Markó, Schulz et Han 2005), ou plus automatiques. Dans ce dernier cas, les morphes sont souvent détectés comme des séquences de lettres répétées dans les mots d'un lexique (Kurimo, Virpioja et al. 2010). Mais de telles techniques n'associent aucune information sémantique aux morphes détectés. À notre connaissance, aucune étude avant notre approche n'a exploré l'utilisation d'une langue pivot pour mener l'analyse morphologique.

D'un point de vue technique, l'utilisation d'une terminologie bilingue est à rapprocher des travaux en translittération, notamment du Katakan ou de l'Arabe (Tsuji et al. 2002 ; Knight et Graehl 1998, par exemple), ou même de traduction et de phonétisation. Le passage par la phonétique, souvent utilisé dans ce type de travaux, n'a pas de sens dans notre cas, de par la nature de l'écriture par kanji et par le fait que les termes en japonais et ceux en français ou anglais n'ont aucune raison d'être phonétiquement proches. Dans ce cadre, citons aussi les travaux de (Morin et Daille 2010) qui proposent de

mettre en correspondance des termes en kanjis avec des termes en français en utilisant des règles s'appuyant sur des indices morphologiques. Cependant, dans ce cas encore, les règles doivent être constituées à la main par un expert. De plus, ce travail ne s'intéresse qu'à un cas bien précis d'opération morphologique issue de la dérivation, et n'est pas adapté à traiter de la composition, l'opération morphologique régissant les composés néo-classiques. Rappelons enfin que des méthodes de traduction de termes biomédicaux considérant les termes comme de simples séquences de lettres ont été proposées (Claveau 2009b). Même si le but est différent, ces méthodes partagent logiquement certaines similarités avec l'approche présentée dans cet article. En effet, elles reposent sur des alignements des termes au niveau des lettres. Cela est effectué le plus souvent avec des algorithmes d'alignement 1-1, c'est-à-dire uniquement capables d'aligner une lettre (ou un caractère vide) du terme de la langue source avec un caractère du terme de la langue cible. Cependant, d'autres travaux récents sur la phonétisation ont souligné l'intérêt d'utiliser des algorithmes d'alignement *many-to-many* (Jiampojamarn et al. 2007). C'est ce type d'algorithme qui sert de base à notre système de décomposition présenté dans la section suivante.

Enfin, en ce qui concerne les traitements morphologiques en recherche d'information, la littérature est très riche. Le lecteur intéressé peut se reporter à (Moreau et Sébillot 2005b) pour un panorama très complet. Si les résultats constatés dans ces études dépendent de nombreux facteurs (langue, outil morphologique, taille de la collection, domaine...), un consensus se dégage pour ce qui est des traitements simples comme la racinisation (*stemming*). Les outils de racinisation, simples et disponibles, permettent en effet d'améliorer les résultats d'un système de RI dans la plupart des cas. La lemmatisation, plus rare en RI, montre aussi de bons résultats. Il est important de noter que les seuls phénomènes morphologiques pris en compte par ces outils sont la flexion et la dérivation. La composition leur reste inaccessible puisqu'ils travaillent principalement sur les suffixes des mots. Plus récemment, les techniques d'analyse morphologique développées dans le cadre de MorphoChallenge ont été appliquées à des problèmes de RI (Kurimo, Creutz et al. 2009). Les auteurs ont là-aussi constaté un gain pour quelques langues, notamment le finnois, très compositionnel, mais les résultats sur l'anglais sont largement moins bons qu'en utilisant un simple *stemming* de Porter.

2.2 Décomposition par alignement

Comme nous l'avons expliqué, notre technique de décomposition repose sur l'alignement avec des termes d'une langue pivot. Cette approche fait donc l'hypothèse que les termes en kanjis ont des constructions parallèles à celles des termes dans la langue étudiée. Cette hypothèse peut paraître forte, mais les résultats donnés ci-après montrent que cette hypothèse est raisonnable dans la plupart des cas.

La technique d'alignement que nous utilisons repose sur un algorithme *Expectation-Maximization* (EM) (Jiampojamarn et al. 2007, pour un exemple d'utilisation), rappelé dans la sous-section suivante. La seconde sous-section présente la modification apportée à cet algorithme pour prendre en compte au mieux les spécificités morphologiques de nos données (Claveau et Kijak 2011).

2.2.1 EM Alignment

L'algorithme d'alignement est relativement standard : il s'agit d'un algorithme *Baum-Welch* étendu pour pouvoir gérer des sous-séquences de symboles et non seulement des alignements 1-1. Les longueurs maximales des sous-séquences dans la langue 1 et dans la langue 2 sont données en paramètres et notées $maxX$ et $maxY$ ci-après. Dans notre cas, l'algorithme prend en entrée des termes dans la langue étudiée (anglais ou français) avec leur traduction en kanji. Ces paires sont issues de bases terminologiques multilingues, en particulier de l'UMLS.

Pour chaque paire de terme (x^T, y^V) à aligner (T et V sont les longueurs des termes en caractères/kanjis), l'algorithme EM (algorithme 1) procède de la manière suivante. La phase d'*Expectation* recense les comptes partiels de tous les alignements possibles entre sous-séquences de kanjis et de caractères. Ces comptes sont conservés dans la table γ ; ils sont ensuite utilisés dans la phase de *Maximization* pour estimer les probabilités d'alignement (table δ associant à chaque traduction une probabilité).

La phase d'*Expectation* implémente une approche *forward-backward* (algorithme 2) : elle estime les probabilités *forward* notées α et *backward* notées β. À chaque position t, v d'une paire de terme, $\alpha_{t,v}$ est la somme des probabilités de tous les alignements possibles du début des termes jusqu'à ces positions (x_1^t, y_1^v), calculés à partir des probabilités d'alignement courantes contenues dans δ (voir algorithme 4). De manière similaire, $\beta_{t,v}$ est calculé en considérant la fin des termes (x_t^T, y_v^V). Ces probabilités α et β sont ensuite utilisées pour re-estimer les comptes de γ. Dans sa version originale, la phase de *Maximi-*

zation (algorithme 3) consiste simplement à calculer les probabilités δ en normalisant les comptes dans γ.

Algorithm 1 *EM Algorithm*

Input : liste de paires (x^T , y^V) , $maxX$, $maxY$
while δ est modifié **do**
 initialisation de γ à 0
 for all paire (x^T , y^V) **do**
  γ = Expectation(x^T , y^V , $maxX$, $maxY$, γ)
 δ = Maximization(γ)
return δ

Algorithm 2 *Expectation*

Input : (x^T, y^V) , $maxX$, $maxY$, γ
α := Forward-many2many(x^T , y^V , $maxX$, $maxY$)
β := Backward-many2many(x^T , y^V , $maxX$, $maxY$)
if $\alpha_{T,V} > 0$ **then**
 for $t = 1...T$ **do**
  for $v = 1...V$ **do**
   for $i = 1...maxX$ t.q. $t - i \geq 0$ **do**
    for $j = 1...maxY$ t.q. $v - j \geq 0$ **do**
     $\gamma(x^t_{t-i+1}, y^v_{v-j+1})$ += $\frac{\alpha_{t-i,v-j}\delta(x^t_{t-i+1},y^v_{v-j+1})\beta_{t,v}}{\alpha^{T,V}}$
return γ

Algorithm 3 *Maximization*

Input : γ
for all sous-séquence a t.q. $\gamma(a, \cdot) > 0$ **do**
 for all sous-séquence b t.q. $\gamma(a, b) > 0$ **do**
  $\delta(a, b) = \frac{\gamma(a,b)}{\Sigma_x \gamma(a,x)}$
return δ

Algorithm 4 *Forward-many2many*

Input : (x^T , y^V) , $maxX$, $maxY$
$\alpha_{0,0} := 1$
for $t = 0...T$ **do**
 for $v = 0...V$ **do**
  if $(t > 0 \vee v > 0)$ **then**
   $\alpha_{t,v} = 0$
  if $(v > 0 \wedge t > 0)$ **then**
   for $i = 1...maxX$ t.q. $t - i \geq 0$ **do**
    for $j = 1...maxY$ t.q. $v - j \geq 0$ **do**
     $\alpha_{t,v}$ += $\delta(x^t_{t-i+1}, y^v_{v-j+1})\alpha_{t-i,v-j}$
return α

Ce processus EM est répété tant que les probabilités dans δ changent. Une fois la convergence atteinte, les alignements sont finalement produits comme ceux maximisant $\alpha(T, V)$. En plus de ces alignements, nous conservons également les probabilités d'alignement des sous-mots collectées dans δ, qui nous sont utiles pour des traitements de certains termes en RI (cf. section 2.4.2).

Cette technique est assez proche de celle utilisée en traduction artificielle mais quelques différences peuvent être soulignées. Cette approche ne permet pas de gérer la distorsion, c'est-à-dire le réordonnement de morphes. En revanche, la gestion de la fertilité, c'est-à-dire la possibilité d'avoir des sous-chaînes vides, n'est pas présentée ici par souci de place mais peut être prise en compte simplement avec cet algorithme.

2.2.2 Normalisation morphologique automatique

La phase de *Maximization* calcule simplement les probabilités de traduire une sous-chaîne de kanjis en une sous-séquence de lettres. Les particularités de nos données, et plus précisément la variation morphologique, y est mal prise en compte, ce qui conduit à dégrader les résultats. Par exemple, pour le kanji 菌 ('*bacteria*'), la table δ peut recenser plusieurs traductions : *bactérie*, ou bien *bactério* (comme dans *bactério/lyse*), ou encore *bactéri* (dans *myco/bactéri/ose*), chacune avec une certaine probabilité. Cette dispersion des probabilités pour les morphes d'un même morphème est préjudiciable.

L'adaptation que nous avons proposé a pour but de rendre la phase de *Maximization* capable de gérer ces différentes variantes, et donc de grouper les différents morphes en un unique morphème. Pour ce faire, nous ré-utilisons le raisonnement analogique présenté dans le chapitre précédent. Nous nous appuyons sur ces derniers travaux pour formali-

ser notre problème de normalisation des morphes. Dans ce cadre, une analogie serait : dermato : dermo $\doteq$ hémato : hémo. Si l'on sait en plus que *dermato* et *dermo* sont deux morphes d'un même morphème, on peut inférer que c'est aussi le cas pour *hémato* et *hémo*. Nous ne disposons pas de tels exemples, mais il est possible d'utiliser une technique d'amorçage. Celle-ci consiste à considérer que deux morphes partageant une grande sous-chaîne commune et connus dans γ comme traductions très probables d'une même sous-chaîne de kanjis sont considérés comme des exemples. Ces amorces sont générées à chaque itération. À partir de ces amorces, les règles de réécriture de préfixation et suffixation sont construites et permettent de trouver d'autres morphes en analogie (au contraire des paires d'amorçage, celles-ci peuvent ne partager qu'une petite sous-chaîne commune). Plus une règle s'applique souvent pour les amorces, plus elle peut être considérée comme fiable. On conserve donc les règles les plus fiables à chaque itération. Le processus est donc entièrement automatique.

Tous les morphes détectés en analogie avec les amorces sont considérés comme appartenant au même morphème. Il est donc maintenant possible d'estimer les probabilités δ en prenant en compte les différentes variantes des morphes. Cette nouvelle version de la *Maximization* assure que tous les morphes supposés appartenir au même morphème aient des probabilités égales et renforcées.

2.3 Evaluation de l'alignement

Nous évaluons la performance de l'alignement en terme de précision : l'alignement d'une paire de termes est considérée correcte si tous ses composants sont correctement découpés et alignés (ce serait l'équivalent du *sentence error rate* en traduction).

2.3.1 Données et vérité terrain

Les données utilisées dans nos expériences sont issues du MetaThesaurus de l'UMLS (Tuttle et al. 1990). Le MetaThésaurus groupe plusieurs terminologies dans différents langages et associe à chaque terme un *identifieur conceptuel unique* (CUI). Les CUI sont indépendants des langues et permettent donc d'extraire facilement des listes de termes dans la langue souhaitée avec leurs équivalents en japonais. Dans notre cas, nous nous sommes intéressés au français et à l'anglais. Dans ces deux cas, nous ne considérons dans l'UMLS que les termes japonais écrit en kanjis et pour le français ou l'anglais, que les

termes simples, c'est-à-dire composés d'un seul mot. Un marqueur de fin de terme (';') est ajouté à ces derniers pour distinguer les suffixes.

Ce sont finalement 14 000 paires de termes anglais-kanjis et 8 000 paires français-kanjis qui sont constituées. Parmi ces paires, 1 600 paires pour le français et 500 pour l'anglais, décomposées et alignées à la main, servent de vérité terrain pour évaluer notre approche.

2.3.2 Résultats d'alignement

Pour chaque paire, l'algorithme EM indique les probabilités de l'alignement proposé. Il nous est donc possible de ne considérer que les alignements dont la probabilité est supérieure à un certain seuil. En faisant varier ce seuil, on peut ainsi calculer une précision en fonction du nombre de termes alignés (nombre de termes dont le score est supérieur au seuil). Les figures 2.1 et 2.2 présentent respectivement les résultats obtenus sur les paires de test pour le français et l'anglais. L'influence de la normalisation morphémique par analogie est illustrée en faisant apparaître les résultats d'alignement avec et sans cette modification de l'algorithme. À des fins de comparaisons, nous rapportons aussi les résultats de GIZA++ (Och et Ney 2003), un algorithme d'alignement de référence dans le domaine de la traduction artificielle. Les différents modèles IBM et paramètres associés disponibles dans GIZA++ ont été testés ; les courbes affichées correspondent aux meilleurs résultats obtenus (IBM modèle 4 sans distorsion).

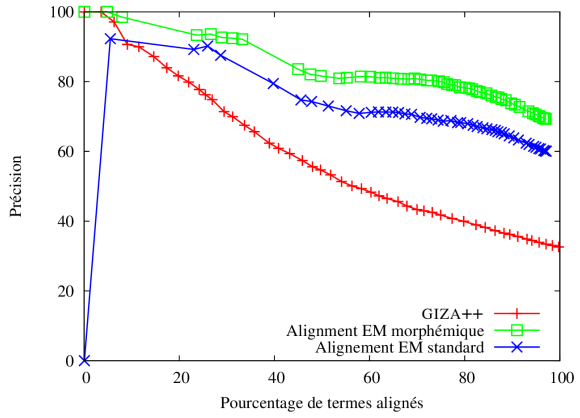


FIGURE 2.1 – Précision de l'alignement français-kanji selon le nombre de paires alignées

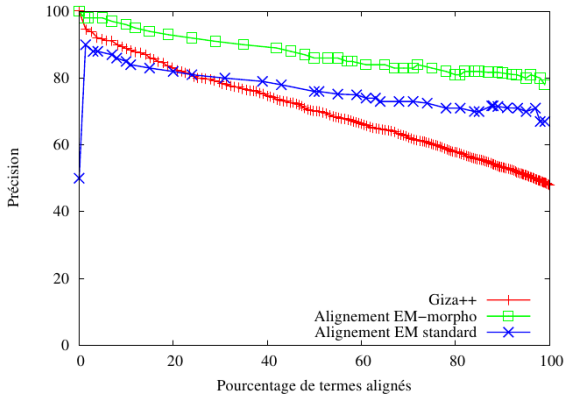


FIGURE 2.2 – Précision de l'alignement anglais-kanji selon le nombre de paires alignées

On note dans ces figures les très bons scores de notre approche, avec dans le pire des cas

(c'est-à-dire quand tous les alignements sont conservés), 70 % de précision pour le français et 80 % pour l'anglais. Comme attendu, l'intérêt de la normalisation morphémique apparaît clairement avec un gain constant d'environ 10 %. La normalisation a aussi un autre intérêt calculatoire puisqu'elle réduit le nombre nécessaire d'itérations de l'algorithme EM et permet donc une convergence plus rapide.

Un examen manuel des résultats montre sans surprise que la plupart des erreurs sont causées par la mise en défaut de notre hypothèse de départ : certaines paires ne se décomposent pas de la même façon en kanjis et dans l'autre langue considérée. Par exemple, le terme français *anxiolytiques* se traduit en japonais par une suite de kanjis signifiant littéralement 'médicament pour la dépression'. Certaines erreurs sont aussi causées par le fait qu'au moins un des deux termes n'est pas un composé néo-classique, comme par exemple *méninges* (alors que sa traduction est bien une composition : 膜 signifie 'membrane du cerveau'). D'autres erreurs sont causées par un manque de données d'entraînement : certains morphes ou séquence de kanjis n'apparaissent qu'une fois dans les données d'entraînement, ou bien toujours combinés avec les mêmes autres morphes, ce qui rend les comptes effectués par l'algorithme peu fiables.

2.4 Analyse morpho-sémantique pour la recherche d'information

Comme nous l'avons dit précédemment, la recherche d'information biomédicale a des caractéristiques particulières dues à l'utilisation des termes spécialisés. À ce titre, l'intérêt de prendre en compte des informations morphologiques riches sur ces termes a déjà été montré, mais uniquement en utilisant des ressources développées manuellement (Markó, Schulz, Medelyan et al. 2005). Dans cette section, nous explorons les différentes utilisations de notre technique de décomposition automatique dans un cadre de RI, sur des documents biomédicaux en anglais. Nous présentons tout d'abord les différentes informations qu'il est possible d'extraire des analyses morphologiques produites par alignement, puis nous indiquons la mise en œuvre adoptée pour inclure ces informations au sein d'un système de RI.

2.4.1 Graphes morpho-sémantiques

Une fois l'alignement effectué, il est possible d'étudier les correspondances récurrentes entre morphes et kanjis dans les données finalement alignées. Plus un morphe est aligné souvent avec une séquence de kanjis, plus le lien sémantique entre eux est sûr. Tous ces liens peuvent être utilisés pour construire un graphe dont les nœuds sont les kanjis ou les morphes ou même les morphèmes (morphes groupés par analogie lors de la phase de maximisation), et les liens représentent donc les correspondances trouvées, pondérées par leur nombre. La figure 2.3 montre un exemple jouet d'un tel graphe anglais-kanji. La taille des arcs est proportionnelle à la force du lien.

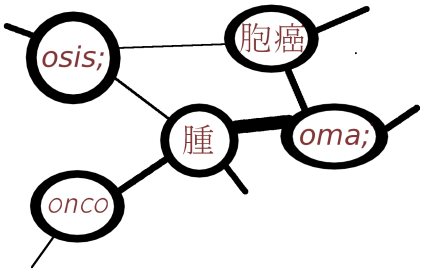


FIGURE 2.3 – Graphe morphème-kanji

Cette représentation nous permet de mettre en lumière les différents types de relations entre morphèmes. Cela se fait simplement en explorant les voisinages des morphèmes avec un algorithme de type Folk-Fulkerson (HEINEMAN et al. 2008) : chaque nœud reçoit une certaine quantité d'énergie qu'il propage, proportionnellement au poids des arcs, à ses voisins ; on peut ainsi lister les nœuds ayant été atteints, et la quantité d'énergie reçue illustre la proximité avec le nœud de départ. Par exemple, la figure 2.4 montre, sous forme de nuages de tags, les morphèmes les plus proches du morphème *ome*, un suffixe de termes liés au cancer. La taille et la couleur des nœuds illustrent la proximité. Les nœuds proches sont supposés conceptuellement liés et doivent être des synonymes ou des quasi-synonymes de *ome*. Il est intéressant de voir que non seulement des suffixes sont trouvés mais aussi des préfixes comme *onco*.

L'alignement et la segmentation produits par notre algorithme rendent également possible l'étude des cooccurrences des morphèmes anglais (ou français) entre eux. Il est par exemple possible d'étudier les affinités de premier ordre, c'est-à-dire quels sont les morphèmes fréquemment associés ensemble. Plus intéressant, on peut également étudier les affinités de second ordre, c'est-à-dire les morphèmes partageant les mêmes morphèmes cooccurrents (mêmes contextes). Les affinités de second-ordre doivent nous permettre de

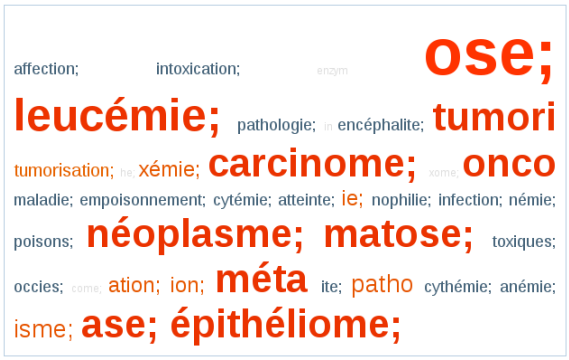


FIGURE 2.4 – Nuage d'affinités du 1er ordre du suffixe *ome*

FIGURE 2.5 – Nuage d'affinités du 2nd ordre du morphème *gastro*

grouper les morphèmes par paradigme. Par exemple, le nuage de mots en figure 2.5 montre les morphèmes associés avec *gastro* (estomac) selon l'affinité de second ordre. On constate comme attendu que ces voisins identifient pour la plupart des organes, et les plus proches désignent des organes les plus proches de l'estomac.

Ces informations de différentes natures permettent d'identifier des relations entre morphèmes et donc entre termes. L'évaluation directe des groupements produits n'est pas possible faute de référence, mais ce sont ces informations qu'on se propose d'utiliser dans un système de RI, offrant ainsi une évaluation indirecte de la pertinence des analyses morphologiques produites.

2.4.2 Représentation morphémique pour la RI

Pour intégrer les informations morphologiques dans le système de RI, nous adoptons une représentation simple : les documents et les requêtes sont considérés comme des sacs de morphèmes et de mots. Les morphèmes sont ceux obtenus par décomposition des termes biomédicaux ou ceux obtenus par affinités de second ordre avec ces premiers. Le but est bien sûr de pouvoir apparier une requête contenant un terme comme *stomachalgia* avec un document contenant *gastrodynia*.

Lors de l'indexation de la collection, les termes sont donc décomposés. Deux cas peuvent se présenter : soit le terme est un terme apparaissant dans les paires d'alignement, soit non. Dans le premier cas, nous récupérons sa décomposition telle que produite

par l'alignement. Dans le second cas, nous exploitons les probabilités collectées dans la table δ pour générer la traduction la plus probable. Pour ce faire, nous utilisons une approche très simple : les probabilités de traductions des morphes dans δ sont utilisées dans un algorithme de Viterbi pour générer la traduction en kanji de probabilité maximale. Nous n'utilisons pas de modèle de langue. Il est important de noter que cette traduction produit en même temps la décomposition voulue du terme initial en assignant à chacun des morphes sa traduction en kanjis. Ce processus de traduction correspond donc bien à l'analyse morpho-sémantique d'un terme inconnu, c'est-à-dire absent des paires utilisées pour l'alignement. Dans chacun de ces deux cas, nous utilisons aussi un autre produit de l'alignement : il s'agit des règles de réécriture (cf. section 2.2.2). Collectées à la dernière itération de l'algorithme EM, celles-ci permettent de détecter les morphes appartenant au même morphème. Elles nous permettent de mettre en correspondance une requête contenant *hemo* avec un document contenant *haemo*, *hemato* ou encore *emia*;.

À partir de ces décompositions, quatre systèmes d'indexation ont été testés. Tous reposent sur un système vectoriel utilisant la pondération BM-25 d'Okapi (Robertson et al. 1998) avec les valeurs standard pour les paramètres b, k_1, k_3.

1. Le premier est basé morphème ; il considère simplement les morphèmes issus de la décomposition des termes rencontrés dans les documents et les requêtes comme des mots à indexer. La pondération du morphème tient compte de la probabilité de décomposition ; elle est définie comme le produit de cette probabilité avec le poids du morphème tel que fourni par le modèle de RI utilisé.
2. Le second est basé kanjis ; les termes sont là-aussi décomposés, mais ce sont les kanjis proches qui sont utilisés comme descripteurs. Ces kanjis proches sont ceux identifiés dans le voisinage des morphes issus de la décomposition.
3. Le troisième système reprend la représentation en morphème du premier système mais étend les requêtes avec les affinités de premier ordre de leurs morphèmes. Les extensions sont pondérées selon leur proximité dans le graphe et par le poids du morphème qu'elles étendent.
4. Le dernier système est identique au troisième mais utilise les affinités de second-ordre pour étendre les requêtes.

À des fins de comparaison, nous utilisons également un système état de l'art avec une indexation classique des documents, racinisation de Porter et une pondération avec BM-25 ; cela constitue notre *baseline*.

	baseline (BM-25 + stemming)	Système basé morphème	Système basé kanji
MAP	29.93	33.94 (+13.4 %)	32.76 (+9.5 %)
R-prec	35.28	39.64 (+12.3 %)	38.59 (+9.4 %)
P@5	69.87	73.45 (+5.1 %)	71.70 (+2.6 %)
P@10	67.99	71.31 (+4.9 %)	*69.65 (+2.4 %)*
P@50	52.98	56.90 (+7.4 %)	55.24 (+4.3 %)
P@100	40.86	44.56 (+9.1 %)	43.39 (+6.2 %)
P@500	15.11	17.21 (+13.9 %)	16.92 (+12 %)
P@1000	8.72	10.10 (+15.86 %)	9.95 (+14.2 %)

TABLE 2.1 – Performances des différents systèmes de RI sur la collection OHSUMED, avec les requêtes TREC. Les différences avec le système *baseline* jugées non statistiquement significatives sont en italiques.

2.5 Expériences de RI biomédicale

2.5.1 Contexte expérimental

Pour les expériences rapportées ci-après, nous utilisons le jeu de données construit pour la tâche de filtrage de la conférence TREC-9. Ce jeu de donnée s'appuie lui-même sur la collection de document OHSUMED, qui est composée de 350,000 résumés d'articles scientifiques extraits de MEDLINE. Dans TREC-9, 4 000 requêtes de filtrage et leur jugement de pertinence ont été développés. Ces requêtes sont composées de plusieurs champs : le sujet, qui est un terme MeSH, et une définition de ce terme. Bien que développé initialement pour le filtrage, nous utilisons ce jeu de donnée comme une collection standard de RI, et ne considérons que le champ sujet pour former nos requêtes.

Les performances des différents systèmes sont évaluées à l'aide des mesures standard : nous calculons la précision sur les 5, 10,... 1000 premiers documents (P@x), la MAP et la R-précision. Pour vérifier la significativité des différences constatées entre les systèmes, nous effectuons un test statistique de Wilcoxon ($p = 0.05$) (HULL 1993).

2.5.2 Résultats

Le tableau 2.1 présente les résultats du système *baseline* et des deux premiers systèmes basés sur la représentation en morphèmes et en kanjis.

Le système basé morphème, reposant donc simplement sur la décomposition des termes

	baseline (BM-25 + stemming)	Système avec affinités 1er ordre	Système avec affinités 2nd ordre
MAP	29.93	34.40 (+14.9 %)	28.74 (-3.9 %)
IAP	31.74	36.63 (+15.4 %)	30.80 (-2.9 %)
R-prec	35.28	39.92 (+13.2 %)	34.38 (-2.6 %)
P@5	69.87	71.76 (+2.7 %)	*68.65 (-1.7 %)*
P@10	67.99	70.46 (+3.6 %)	*66.20 (-2.6 %)*
P@50	52.98	56.30 (+6.7 %)	50.50 (-4.68 %)
P@100	40.86	44.69 (+9.4 %)	39.07 (-4.38 %)
P@500	15.11	17.98 (+18.9 %)	*15.01 (-0.64 %)*
P@1000	8.72	10.56 (+21.1 %)	*8.77 +0.66 %)*

TABLE 2.2 – Performances des différents systèmes de RI sur la collection OHSUMED, avec les requêtes TREC. Les différences avec le système *baseline* jugées non statistiquement significatives sont en italique.

et le regroupement en morphème, obtient de très bons résultats avec un gain en MAP de 13 %. Comme attendu, la décomposition améliore plus particulièrement les performances en fin de liste (P@100 et supérieurs) puisqu'elle permet de ramener des documents même s'ils ne contiennent pas les termes de la requête. Le système basé kanji obtient des performances assez semblables. Le gain espéré de passer à une représentation plus générique que les morphèmes n'est pas réalisé. Il semble que cette représentation soit trop générique pour certaines requêtes et apporte peu d'information supplémentaire par rapport au morphème pour d'autres. Par ailleurs, aucune sélection n'est faite sur les morphèmes à traduire ou non, et certains kanjis trouvés en traduction ont des propriétés (fréquence documentaire) différentes du morphème de départ puisqu'ils peuvent se trouver comme traduction de différents morphèmes. Une technique de pondération tenant compte des fréquences documentaires initiales semble une perspective importante pour développer ce type de système.

Le tableau 2.2 illustre les résultats des deux derniers systèmes proposés reposant sur l'extension de requêtes. Les systèmes à base d'extension ont des résultats plus contrastés. L'extension à l'aide d'affinités du premier ordre donne de très bons résultats, avec une précision un peu dégradée en début de liste mais un rappel amélioré. En revanche, les affinités de second ordre produisent des résultats largement moins bons qu'avec la décomposition morphologique seule. Il semble que ces affinités soient trop éloignées sémantiquement et ramènent trop de documents jugés non pertinents.

2.6 Conclusion et perspectives

Bien que la morphologie soit étudiée de longue date en RI, et malgré la disponibilité d'outils simples comme les *stemmers*, la morphologie reste un enjeu d'importance, au même titre que d'autres phénomènes linguistiques pouvant sembler plus complexes à modéliser. Comme nous l'avons vu, les outils actuels ne sont pas suffisant lorsqu'ils sont confrontés à des opérations morphologiques compliquées comme celles ayant cours dans le domaine médical. À ce titre, nos résultats s'inscrivent dans la continuité de ceux rapportés par d'autres sur l'intérêt de prendre en compte ces phénomènes (Markó, Schulz et Han 2005 ; Deléger et al. 2008), mais à notre connaissance, ce furent les premiers à proposer un processus entièrement automatique, sans intervention humaine, et directement applicable à de nombreuses langues. Bien sûr, il repose sur la disponibilité de terminologies multilingues, mais celles-ci, au contraire de bases de connaissances morphologiques, sont largement disponibles. L'approche repose sur un processus d'apprentissage supervisé, mais les exemples étant dérivés de ressources existantes (bien que construites avec d'autres objectifs), l'ensemble de notre approche est non supervisée de fait.

Ce travail a bien sûr quelques limites, que l'on peut voir comme autant de perspectives ouvertes. Tout d'abord d'un point de vue technique, il serait intéressant de produire des décompositions de termes non plus linéaires, mais hiérarchisées : par exemple, *gastroentérite* serait analysé en *[gastro | entér] ite*. Nous pensons que ces décompositions nous permettraient, dans un cadre de RI, de pondérer plus efficacement les morphes et de choisir plus facilement ceux pouvant être étendus par des morphes liés sémantiquement. Pour ce faire, on peut là encore imaginer exploiter les kanjis dont certains ont une fonction syntaxique connue (certains sont par exemple des prédicats attendant un agent ou un objet). Outre ces considérations syntaxiques au sein des termes, il est aussi possible d'exploiter les liens sémantiques entre kanjis, facilement récupérables à partir de dictionnaires généralistes japonais, pour aider à établir les liens sémantiques entre morphes.

Enfin, le domaine biomédical est aussi très riche en termes complexes (composés de plusieurs mots-formes). Une adaptation de cette approche d'analyse morphologique pouvant s'appliquer à ces termes serait pertinente, mais la difficulté est alors de gérer les différents ordonnancements des mots composant le terme, et donc d'autoriser la distorsion. Pour la RI, l'enjeu est cependant important puisque ces termes sont connus pour les nombreuses variations qu'ils peuvent subir, ce qui empêche la mise en correspondance des documents et des requêtes contenant des variantes différentes d'un même terme.

Dans notre parcours de recherche, ce travail a une position notable ; il illustre en effet plusieurs de nos centres d'intérêt :

— il relève du domaine biomédical, qui est, nous l'avons dit dans le chapitre précédent, un terrain de jeu plein d'intérêt pour le TAL et la RI ;
— il propose une approche dédiée à la fois symbolique (analogie formelle pour la normalisation morphémique) et statistique (algorithme *forward-backward*), reposant sur des concepts connus mais dont la combinaison est, à notre connaissance, originale ;
— l'approche se veut peu supervisée grâce au détournement de données existantes (traduction japonaises de l'UMLS).

Bien entendu, depuis la réalisation de ce travail, d'autres approches ont vu le jour, notamment pour la prise en compte de termes proches dans un contexte de RI. Ces problèmes sont désormais le plus souvent abordés par des plongements dans lesquels la proximité géométrique doit traduire une proximité sémantique. Ainsi certains algorithmes de construction de plongements exploitent des n-grams de caractères (Bojanowski et al. 2016), mais les phénomènes morphologiques pris en compte par ces approches relèvent plutôt de la flexion et de la dérivation et ne sont pas adaptées pour la composition telle que vue dans ce chapitre. D'autres travaux sur les plongements permettent d'avoir une représentation des CUI (Beam et al. 2018), et donc indépendante de la langue et de la forme. Une telle approche est particulièrement intéressante, pour peu que l'on soit capable d'extraire les CUI des textes, c'est-à-dire d'assigner un CUI à une occurrence d'un terme. Des outils existent pour l'anglais (Aronson et Lang 2010), mais les autres langues en sont dépourvues. Enfin, plus généralement, la prise en compte de similarités morphologiques dans des processus de TAL est aussi faite désormais par des réseaux de neurones travaillant directement au niveau du caractères (souvent en supplément d'une couche d'entrée exploitant des plongement de mots). Cette approche donne selon les tâches des résultats intéressants, mais n'offre pas le niveau de finesse et d'interprétabilité de l'approche proposée ici.

Deuxième partie

RI pour le TAL : du bon usage de la similarité

Chapitre 3

Quelques apports de la RI en TAL

Les techniques de recherche d'information, plus modernes que celles d'Obélix, peuvent servir à des nombreuses autres tâches.

En regard du chapitre 1, celui-ci présente quelques apports personnels concernant l'utilisation de concepts issus de la RI pour des tâches notées comme relevant du TAL. Nous y faisons quelques brèves descriptions de travaux passés (sections 3.2, 3.3.1, 3.3). À travers eux, on montre ainsi que la définition d'une mesure de similarité, adaptée à la tâche mais s'inspirant toujours des techniques de RI, permet de résoudre des problèmes parfois complexes. Nous donnons dans la section suivante quelques définitions utiles qui permettent de situer ces travaux.

3.1 Positionnement de nos travaux

La figure 3.1 schématise le fonctionnement d'un moteur de recherche : les documents sont indexés, c'est-à-dire transformés sous une forme appréhendable par la machine (voir sous-section suivante), les requêtes subissent un traitement similaire, et un module permet de calculer la similarité entre la représentation de la requête et celle du document. Ce schéma, présent dans tous les cours d'initiation à la RI, illustre bien le rôle central de ces deux composants : la représentation des documents (et requêtes) au sein d'un système de

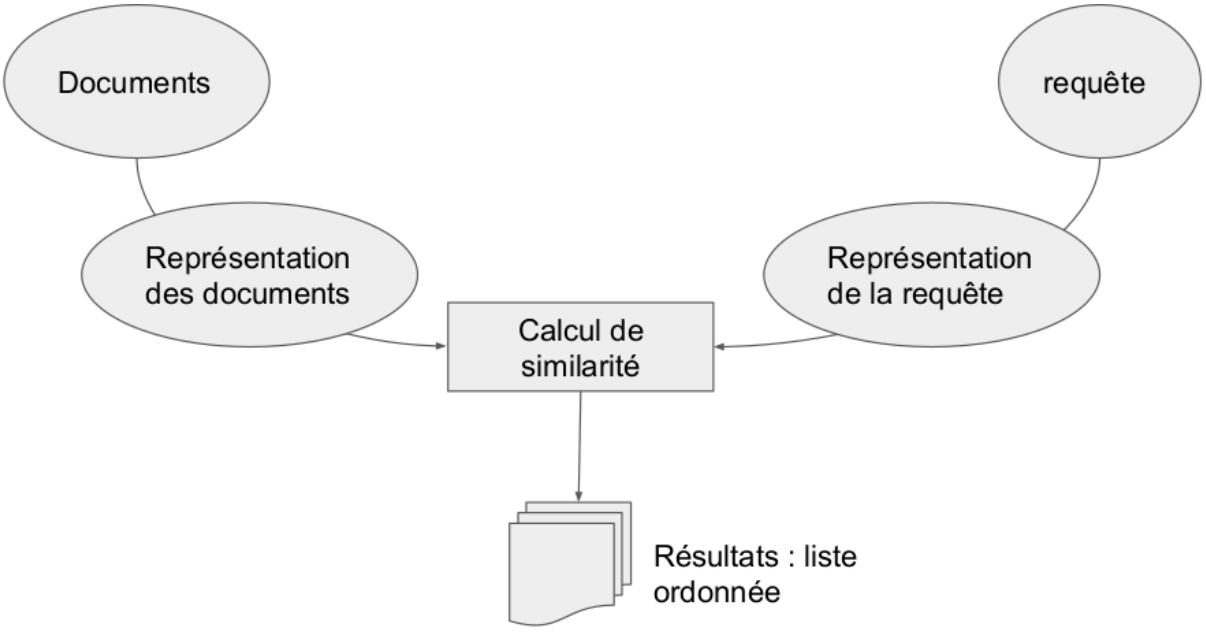


FIGURE 3.1 – Vue simplifiée d'un système classique de recherche d'information.

RI et le calcul de similarité entre ces représentations. Les nombreuses techniques développées pour mettre en œuvre ces deux composants ont été évidemment exploitées au-delà de la RI, notamment dans certaines tâches du TAL.

3.1.1 Représentation

Longtemps tournée vers la recherche dans des grandes bases de documents, la RI a développé différentes approches pour offrir une représentation des textes répondant à certaines contraintes. Elle doit notamment :

— être compacte, mais néanmoins décrire au mieux le contenu du texte ;
— être calculable avec une faible complexité temporelle et mémoire ;
— ne nécessiter aucune ressource externe, ou des ressources facilement disponibles ;
— permettre un calcul de similarité rapide et adapté à la tâche.

À ce titre, la représentation la plus communément employée en RI est celle dite en sac-de-mots : les documents sont considérés comme un multi-ensemble de mots-formes (formes graphiques), non structuré, sans information sur la séquentialité des mots dans le texte.

Usuellement, on calcule pour chaque mot présent dans le document une valeur reflétant

son importance comme descripteur du document. Cette étape de pondération a généré un très grand nombre de travaux dans la communauté RI ; de nombreuses formules de pondération ont été proposées, étudiées et comparées. La plus célèbre d'entre elles est le TF-IDF ; bien que proposée dans les années 70, elle reste, souvent à tort, celle utilisée majoritairement dans les travaux du TAL comme nous l'évoquions dans l'introduction.

Enfin, ces mots et leur valeur associée sont utilisés pour représenter le document dans son ensemble. Là encore, plusieurs modélisations ont été proposées, comme les bien connus modèle de langues (Jay M. Ponte et W. Bruce Croft 1998) et modèle vectoriel (Salton et McGill 1983). Ce dernier est largement employé en RI mais également dans beaucoup d'applications du TAL où il sert à représenter les textes ou des segments de textes. Cette représentation vectorielle est plus particulièrement utilisée en entrée de classifieurs (SVM, forêts aléatoires, bayésien naïf, réseaux de neurones) ; elle est donc au cœur des travaux s'interprétant comme une tâche de classification (détection de spam, assignation de polarité ou de sentiment, attribution d'auteur...).

3.1.2 Classification et RI

Au-delà de la représentation vectorielle vue précédemment, les liens entre la RI et la classification sont plus profonds. En effet, la recherche de documents à partir d'une requête peut être vue comme une tâche de classification, et les moteurs de recherche comme des classifieurs : étant donné une requête, le moteur doit classer les documents entre pertinents et non pertinents (ou les classer du plus pertinent au moins pertinent), et il le fait sur la base de la proximité entre la représentation du document et celle de la requête. Grâce à cette capacité à calculer des proximités entre des textes, les moteurs de recherche peuvent par ailleurs être utilisés directement comme des classifieurs de types k-plus-proches voisins. Ils ont donc été logiquement employés comme tels dans différentes tâches pouvant être mises en œuvre par de la classification de documents (analyse de sentiment, recommandation par le contenu, détection de spam...).

Il faut également noter que cette utilisation des systèmes de RI comme classifieurs peut porter sur d'autres choses que sur des documents bien formés. Dans plusieurs travaux, le texte considéré n'est pas forcément un document. Par exemple, dans Claveau, Kijak et Ferret (2014a), nous montrons que les moteurs de recherche peuvent être utilisés pour comparer des ensembles de contextes de mots dans le but de construire des lexiques distributionnels (cf. chapitre suivant). En segmentation thématique, des systèmes de RI peuvent aussi servir à détecter les ruptures thématiques (Hearst 1997 ; Claveau et

Lefèvre 2015) entre des portions successives d'un flux de textes.

3.1.3 La RI pour évaluer le TAL

Un des apports de la RI pour le TAL, indirect mais important, concerne l'évaluation. La communauté RI a très tôt commencé à formaliser des cadres d'évaluation quantitative (dont les mesures d'évaluation, comme le rappel, la précision, mais aussi les tests de significativité statistique), développer des collections de données, organiser des compétitions[1]. L'idée centrale pour une évaluation objective est que les différents systèmes et méthodes doivent se comparer sur les mêmes données, et vis-à-vis d'une vérité terrain (*ground truth*). Pour permettre la reproductibilité et la comparabilité, la communauté RI a par ailleurs privilégié l'accessibilité des données, ouvertes ou avec un coût d'acquisition faible, et la mise à disposition des scripts d'évaluation.

Outre l'exemple de ces bonnes pratiques en matière d'évaluation quantitative, le cadre de la recherche documentaire a aussi largement été exploité pour servir à l'évaluation indirecte, par la tâche, de différents processus de TAL. Il s'agit dans ce cas d'étudier l'impact d'un outil du TAL sur un moteur de recherche en comparant les résultats avec et sans l'outil de TAL, ou de comparer les résultats obtenus avec plusieurs de ces outils. C'est notamment ce cadre que nous adoptons dans le chapitre suivant pour évaluer la qualité de lexiques distributionnels.

3.2 Similarité et fouille de texte

Comme nous l'avons souligné en section 3.1.2, les représentations des textes utilisées en RI (typiquement, sacs-de-mots pondérés) peuvent être utilisées en entrée d'un classifieur. De plus, les mesures de similarités entre textes définies dans le domaine de la RI peuvent elles-mêmes servir directement à construire des classifieurs simples de type k-plus-proches-voisins (kppv par la suite). C'est une approche que nous avons utilisées à de nombreuses reprises pour diverses tâches de classification et fouille de texte, notamment lors de participations à des challenges de fouille de texte, comme DeFT (Défi Fouille de Textes). Dans certains cas, la tâche se définie directement comme un problème de recherche documentaire (section 3.2.1), dans d'autres, il nous faut définir le classifieur kppv de manière plus appropriée (section 3.2.2).

1. La première compétition TREC a été organisée en 1992.

3.2.1 Appariement avec un moteur de recherche

Certaines tâches de TAL s'interprètent très directement comme des problèmes de recherche documentaire, avec un texte tenant le rôle d'une requête pour lequel on recherche des documents ou des textes pertinents dans une base. C'est notamment le cas des problèmes d'appariement dans lesquels pour un seul document est recherché pour une requête. Dans tous ces cas, l'emploi d'un système de RI "sur étagère" peut tenir lieu de système, avec des performances intéressantes selon les tâches.

Ainsi, lors de l'édition 2011 de DeFT (Grouin et al. 2011), l'une des tâches consistait à apparier des résumés d'articles scientifiques avec le corps des articles. En considérant simplement les résumés comme des requêtes et les articles comme des documents, et avec un système de RI Okapi-BM25 (Robertson et al. 1998), nous avons pu trouver les documents (articles) proches. Pour robustifier cet appariement, nous l'avons combiné avec l'appariement dual : les articles sont considérés comme des requêtes et le système de RI ramène les résumés pertinents. Cette approche, très simple, a donné d'excellents résultats, avec une précision de 95 %, mais surtout une mise-en-œuvre légère, accessible à tous et rapide. Ce même principe a été appliqué pour constituer une *baseline* pour l'édition 2019 de DeFT que nous avons organisés (Grabar, Grouin et al. 2019) dans laquelle l'une des tâches consistait à apparier des description de cas cliniques avec la discussion (expertise) s'y rapportant.

Les textes manipulés sont parfois plus courts, se rapprochant de requêtes standard de RI, ce qui ne modifie en rien nos mises-en-œuvre. Par exemple, c'est une approche similaire qui a été utilisée pour l'édition 2012 de DeFT dans laquelle on cherchait à attribuer des mots-clés à des articles scientifiques. Dans l'une des pistes, les mots-clés étaient fournis, et nous ont donc servi de requête. Là encore, de très bons résultats ont été obtenus, surpassant les résultats d'approches plus complexes (Paroubek et al. 2012). Parfois, les textes sont non pas des documents, mais de simples syntagmes de quelques mots, comme pour la tâche de liage d'entité (*entity linking*) de BioNLP-ST2013 (Claveau 2013).

3.2.2 Classification par k-ppv

Outre les problèmes d'appariement, certaines tâches de classification en TAL peuvent également bénéficier des systèmes de RI. Comme nous l'avons expliqué précédemment, une des façons les plus simples est d'adopter une approche par k-plus-proches-voisins, dans laquelle la recherche des voisins reposent sur une similarité usuelle en RI (par exemple,

Okapi-BM25). Ce type d'approche par kppv est très simple à mettre en œuvre, ce qui est un point important pour ces challenges dans lesquels les délais sont très contraints. De plus, l'approche est assez peu sensible à la quantité de données d'apprentissage, notamment parce qu'il y a très peu de paramètres à optimiser. Enfin, la classification par kppv est moins sensible que d'autres approches lorsque les éléments d'une même classe sont répartis dans l'espace de représentation.

Ainsi, dans DeFT2011 (Raymond et Claveau 2011), l'une des tâches était de dater à l'année près des documents de presse OCRisés sur les 145 dernières années. Bien entendu, différents articles d'une même année peuvent aborder des sujets différents et donc ne partager que peu de vocabulaire commun. Ces articles peuvent donc se trouver à des endroits très éloignés dans l'espace de représentation, et cela nécessite d'employer des méthodes d'apprentissage capables de gérer ces cas. Avec des approches par modèles (un modèle est appris par année, par exemple avec des SVM, utilisés par plusieurs autres participants de DeFT), cela se fait souvent au prix d'une complexité importante (par exemple pour les SVM, emploi d'un noyau non linéaire, impliquant des temps de calcul importants et des besoins de grandes quantités d'exemples...).

Nous avons décliné cette approche par kppv avec des similarités issues de la RI (Okapi-Bm25, modèle de langue (J. M. Ponte et W. B. Croft 1998)...) sur de nombreuses autres tâches. Les objets textuels à classer sont parfois des documents entiers, mais aussi des textes courts comme des tweets ; c'était par exemple le cas pour DeFT 2015 et 2017 (Vukotić et al. 2015 ; Claveau et Raymond 2017) pour faire de l'analyse de sentiments, ou dans MediaEval2016 (Cédric Maigrot et al. 2016) pour détecter des fake news.

3.3 Similarités RI plus complexes

Pour certains travaux, pour répondre à des besoins plus spécifiques, nous avons été amenés à proposer des variantes des représentations sacs-de-mots et des similarités RI. Nous présentons deux de nos travaux dans lesquels les objets textuels manipulés et comparés sont représentés par de telles variantes.

3.3.1 Similarités entre sacs de sacs de mots

La représentation d'un texte complet par un unique sac-de-mots n'est pas toujours adapté. En effet, l'absence de prise en compte de la séquence, déjà problématique pour de

courts textes, devient encore plus brutale lorsque l'on traite de longs textes, abordant des sujets différents. Pour amoindrir ces effets, tout en conservant l'aspect pratique, compact et efficace (pour les calculs de similarité de type cosinus par exemple) de cette représentation, nous avons exploré l'utilisation de représentations en sacs-de-sacs-de-mots (Ebadat et al. 2012a ; Claveau 2014).

Dans cette représentation, le principe est de segmenter le document en phrases (ou paragraphe ou autre unité), chacune étant représentée classiquement par un sac de mot. Le document est ainsi représenté par l'ensemble des sacs-de-mots de ses phrases. Il faut bien noter que le nombre de sacs-de-mots est donc variable d'un document à un autre.

Le calcul de similarité entre deux documents nécessite alors de comparer deux ensembles, le plus souvent en combinant les mesures de similarité usuelle entre chaque paires possibles de sacs-de-mots, notées $\delta(\cdot,\cdot)$ dans la figure 3.2.

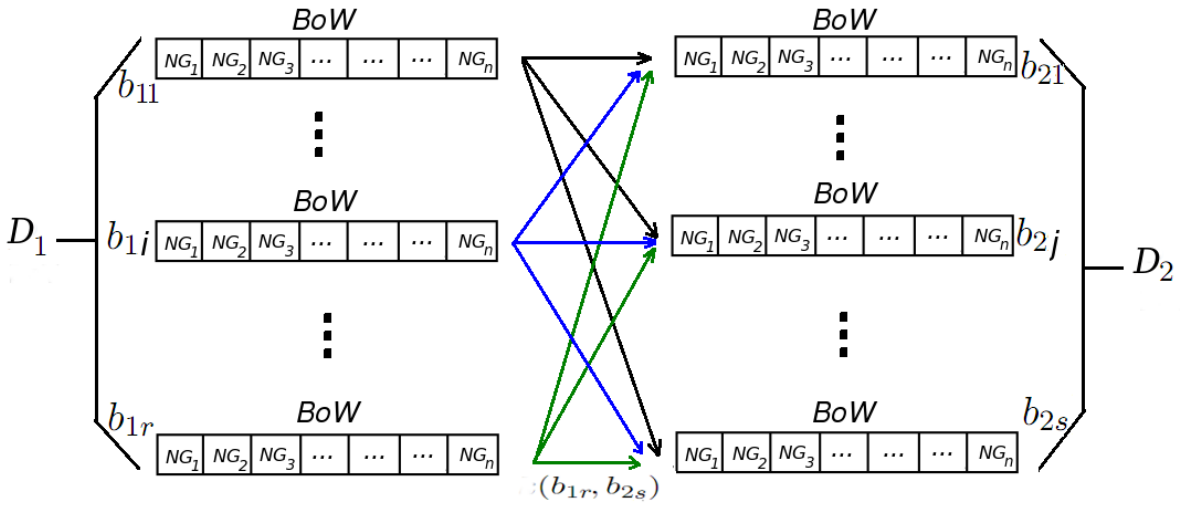


Figure 3.2 – Calcul de similarité entre deux sacs-de-sacs-de-mots.

Deux façons simples pour ce faire ont été proposées par Haussler (1999) :

$$\text{Sum-Sum}(D,Q) = \sum_i \sum_j \delta(b_{Q,i}, b_{D,j})$$

$$\text{Max-Max}(D,Q) = \max_i \max_j \delta(b_{Q,i}, b_{D,j})$$

Une mesure plus générale a été proposée par Gosselin et al. (2007) :

$$\text{PowerScalar}(D,Q) = \sqrt[q]{\sum_i \sum_j \delta(b_{Q,i}, b_{D,j})^q}$$

Le paramètre $q \in [0, \infty[$ donne plus ou moins d'importance aux valeurs hautes et contrôle donc le pouvoir discriminant des similarités mineures. Cette mesure recoupe les deux précédentes pour $q = 1$ (Sum-Sum) et $q \to \infty$ (Max-Max). En nous inspirant des mesures d'agrégation développées en logique floue (Detyniecki 2000), nous avons étudié les propriétés théoriques de ces différentes similarités (Claveau 2014).

Par ailleurs, la plupart des fonctions de pondérations de RI font intervenir des valeurs calculées sur le document (IDF, longueur du document DL...). Le passage à une unité plus petite pose la question du calcul de ces valeurs. Là encore, nous avons proposé plusieurs stratégies : dans un cas, les variables problématiques (IDF, DL...) sont calculées classiquement sur le document, et dans l'autre cas, sur le sous-document (phrase par exemple ; le DL est alors la longueur du sous-document considéré, l'IDF est la fréquence document inverse dans l'ensemble des sous-documents de la collection).

Nous avons utilisé ces approches dans un cadre classique de recherche d'information (*ibid.*) en constatant qu'elles permettent d'obtenir des gains mineurs par rapport aux représentations sac-de-mots classiques. Mais plus intéressant, nous avons montré que ces représentations et similarités étaient bien adaptées à des tâches de découverte d'information. Ainsi, dans le travail de (Ebadat et al. 2012b), nous avons montré que les sacs-de-sacs-de-mots permettaient de représenter des entités nommées pour une tâche de *clustering* : chaque contexte de chaque occurrence d'une entité nommée est représentée indépendamment par un sac de mots et l'entité est représentée par l'ensemble de ses sacs.

3.3.2 Similarité de second ordre

Nous avons également proposé une autre revisite de la représentation RI classique en sac-de-mots et similarités. Elle était motivée par la recherche de représentations plus sémantiques, dépassant la similarité basée sur uniquement sur les mots. En effet, la représentation classique en RI ne permet pas d'apparier une requête contenant vélo avec un document contenant bicyclette. Dans un cadre de RI, nous avons donc proposé une technique simple pour transformer n'importe quel processus d'appariement requête-document fournissant un score (typiquement RSV) en un problème de calcul de distance entre vecteurs. Ces vecteurs sont construits par similarité (avec une fonction RSV usuelle) du document

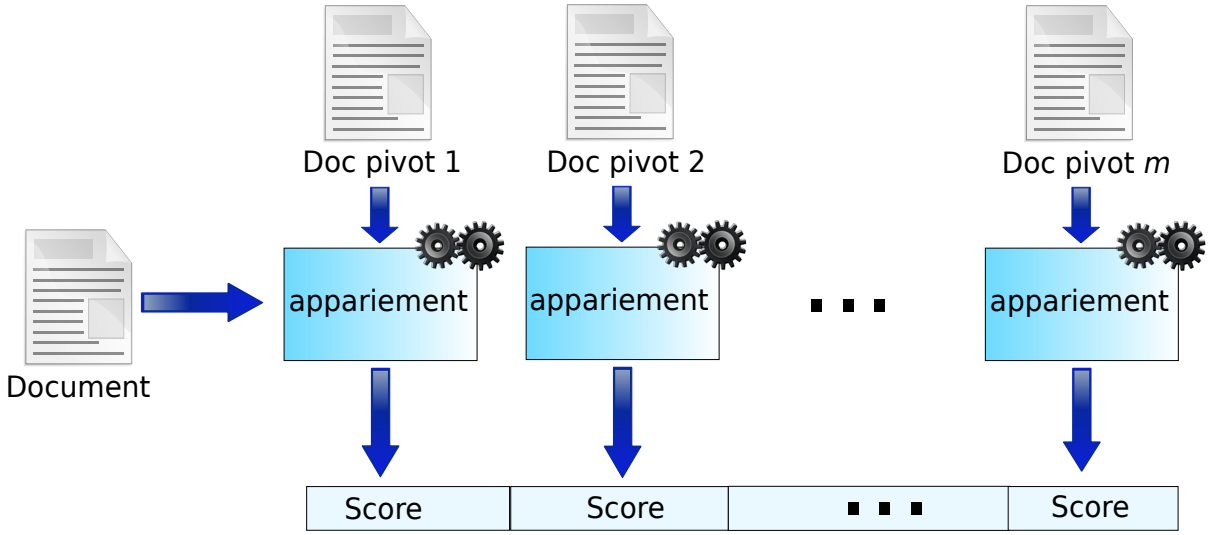


FIGURE 3.3 – Processus de construction des vecteurs de second ordre. Chaque composante du vecteur est le score RSV entre le document et un document-pivot.

avec des documents-pivots (voir figure 3.3). Plus précisément, la $i$ème composante du vecteur représentant un document d est définie par : RSV$(d, pivot_i)$ Les vecteurs résultats se veulent des représentations plus sémantiques des documents, et la similarité entre eux, est alors dite de second ordre. L'intuition est que deux documents sont proches l'un de l'autre s'ils sont proches (ou lointains) des mêmes documents-pivots. Il est donc possible qu'ils soient jugés proches, même s'ils ne partagent pas de mots en commun, comme cela peut se faire avec des approche de type LSI (DEERWESTER et al. 1990 ; HOFMANN 1999).

Nous avons testé l'utilité de ces similarités de second ordre pour la RI (CLAVEAU, TAVENARD et al. 2010). Le gain pour ce type de tâche est faible, et uniquement présent lorsque de nombreux documents sont à retourner (gain de rappel). Mais comme précédemment, l'intérêt de ce travail a été révélé par une tâche de TAL, la segmentation thématique. Dans ce travail (CLAVEAU et LEFÈVRE 2015), nous nous sommes intéressés à la segmentation thématique d'émissions télévisées à partir de la transcription automatique de leur bande-son.

L'un des principes de segmentation thématique les plus communs est de rechercher des frontières thématiques en comparant, à chaque frontière potentielle (fin de phrase par exemple) si ce qui précède est similaire, en terme de vocabulaire, à ce qui suit ; si la réponse est non, on considère qu'il y a changement de thème. Évidemment, plusieurs auteurs ont utilisé les mesures classiques de RI mettre-en-œuvre cette comparaison entre

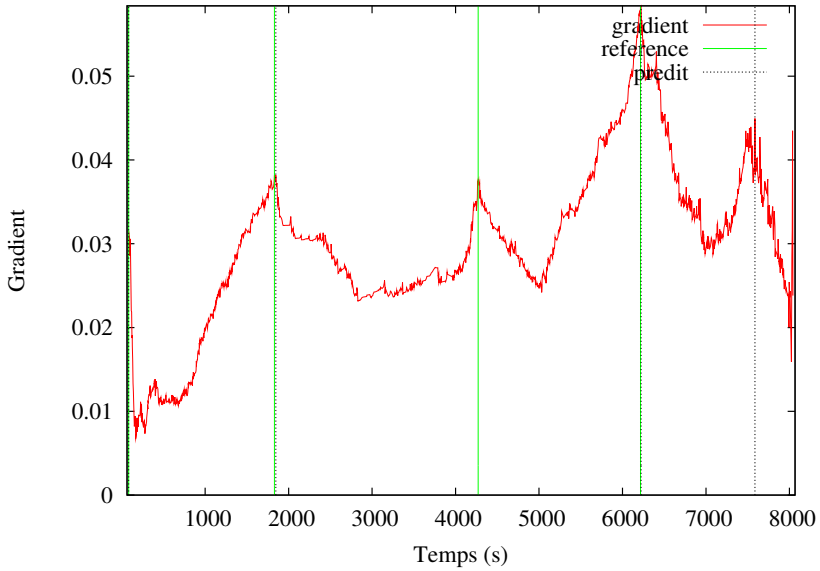


FIGURE 3.4 – Détermination des frontières thématiques par Ligne de Partage des Eaux sur un gradient obtenu par similarité de second ordre (en pratique ici, une dissimilarité, pour une bonne application de la LPE)

ce qui précède et ce qui suit (HEARST 1997, inter alia). En pratique, une mesure de similarité est calculée à chaque frontière potentielle, et il convient donc de déterminer à quel endroit positionner une frontière ; cela était fait par heuristique dans les travaux précédent (quand la similarité tombe sous un certain seuil par exemple).

Dans ce contexte, notre contribution à la segmentation thématique a été double. D'une part, nous avons proposé d'utiliser une technique issue de la segmentation d'image, la Ligne de Partage des Eaux (LPE) ou *watershed* (VINCENT et SOILLE 1991 ; SOILLE 2003), pour détecter les frontières automatiquement (voir figure 3.4). Plus intéressant pour notre propos, nous avons étudié l'impact de différentes mesures de RI, et également de notre similarité de second ordre, pour ce problème. Nous avons ainsi montré les mesures de RI de type Okapi-BM25 apportait déjà un gain important par rapport à l'état de l'art, mais que la similarité de second ordre obtenait des résultats significativement supérieurs. L'analyse des résultats a montré que ces bons résultats étaient dus à la capacité de cette similarité à rapprocher des segments partageant parfois peu de mots communs mais partageant bien une thématique commune, capturée par leur similarité avec des documents-pivots communs.

3.4 Conclusion

Comme nous l'avons illustré dans ce chapitre, les représentations typiques des textes et des calculs de similarités développées pour la RI trouvent des applications au-delà, notamment pour tout un ensemble de tâches du TAL. Évidemment pour des tâches s'exprimant comme un problème d'appariement, l'utilisation de moteurs de recherche est directe, et nous l'avons vu, souvent très performante. Pour toutes les tâches relevant de la classification de texte, il est là-encore très simple d'utiliser un moteur de RI au cœur d'une approche k-plus-proches-voisins. Nos travaux dans ces deux cadres ont été nombreux, notamment au travers de la participation à des challenges de fouille de textes, et ont toujours démontré la pertinence de ces approches conceptuellement simples. Outre les bonnes performances et la simplicité de mise-en-œuvre, ces techniques ont également l'avantage d'être facilement interprétables : on peut aisément expliquer la classification dans un kppv en examinant les documents voisins ayant mené à ce résultat.

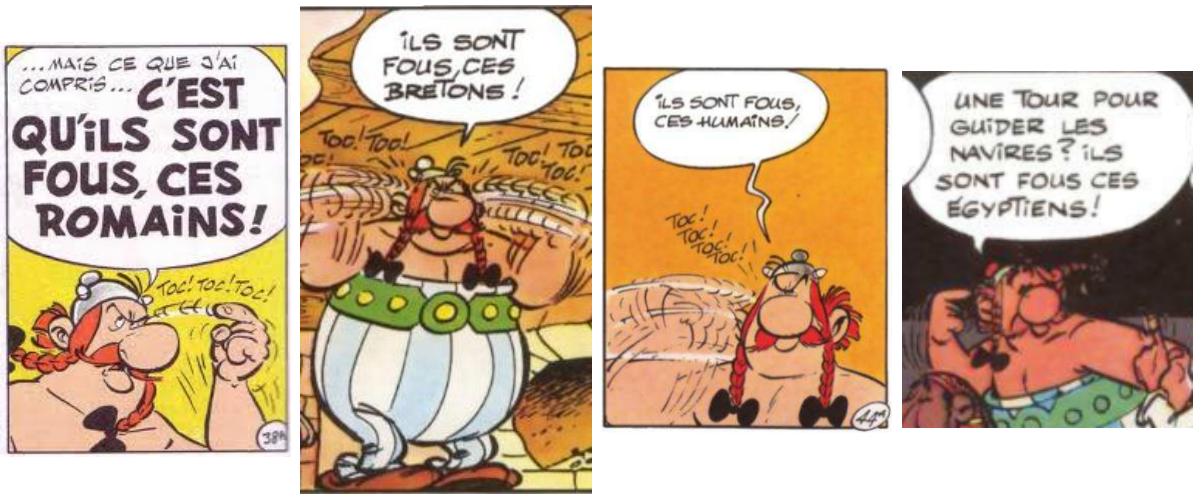
Nous avons conscience que nos contributions scientifiques sont relativement mineures puisqu'il s'agit d'employer des techniques existantes, avec peu de modifications. Cependant, elles ont le mérite de souligner la performance des ces techniques, et surtout de les faire connaître. C'est ce qui avait motivé l'article "prise de position" à la conférence TALN intitulé *Vectorisation, Okapi et calcul de similarité pour le TAL : pour oublier enfin le TF-IDF* (Claveau 2012) dans lequel nous faisions, chiffres à l'appui, le constat du peu de connaissance des publiants à TALN des schémas de pondérations état-de-l'art en RI.

En dehors de ces réutilisations directes d'outils de RI, nous avons également contribué de manière plus significative à la définition de nouvelles représentations de texte ou similarités, comme illustré avec les sac-de-sac-de-mots et les similarités de second ordre. Dans ces travaux, la démarche commune a été de vérifier la viabilité de ces approches dans un cadre classique de recherche documentaire puis d'employer ces approches sur des tâches de TAL (*clustering* d'entités nommées, segmentation thématique). Dans les deux cas, les gains en RI sont relativement mineurs, mais ces approches se sont révélées très performantes pour les autres applications.

Comme nous l'avons souligné en section 3.1.3, la RI a aussi un grand intérêt pour évaluer certains processus de TAL. C'est aussi un cadre que nous avons adopté dans plusieurs de nos travaux depuis longtemps (Claveau 2003). Nous y revenons plus en détail dans le chapitre suivant au travers de travaux récents, faisant bien le lien entre RI et TAL.

Chapitre 4

Analyse distributionnelle par et pour la RI



Selon l'hypothèse distributionnelle, romains, humains, bretons, égyptiens, etc. sont proches puisqu'ils partagent des contextes similaires.

Dans ce chapitre, nous présentons un travail effectué ces dernières années en collaboration avec Ewa Kijak (IRISA-Univ Rennes 1) et Olivier Ferret (CEA LIST) (Claveau, Kijak et Ferret 2014b ; Claveau et Kijak 2016b ; Claveau et Kijak 2016a). Son intérêt, dans le contexte de ce manuscrit, est d'illustrer parfaitement les multiples allers-retours que nous faisons entre tâches typiques du TAL et de la RI. En effet, nous mettons ici les techniques de RI au service de la construction de lexiques distributionnels, eux-mêmes évalués au travers d'une tâche de RI (par expansion de requêtes).

4.1 Introduction

La sémantique distributionnelle a pour objet de construire des thésaurus (ou lexiques) automatiquement à partir de corpus de textes. Pour une entrée donnée (*ie.* un mot donné), ces thésaurus recensent des mots sémantiquement proches en s'appuyant sur l'hypothèse qu'ils partagent une distribution similaire au mot d'entrée. En pratique, cette hypothèse distributionnelle est mise en œuvre simplement : deux mots seront considérés proches s'ils partagent des contextes similaires. Ces contextes sont typiquement les mots cooccurrant dans une fenêtre restreinte autour du mot examiné, ou les mots liés syntaxiquement à celui-ci.

L'évaluation de ces thésaurus reste un point crucial pour juger de la qualité des méthodes de construction employées. Une approche communément utilisée est de comparer le thésaurus produit à un ou plusieurs lexiques de référence. Cette évaluation, qualifiée d'intrinsèque, a pour avantage d'être directe et simple puisqu'elle permet d'estimer la qualité et la complétude du thésaurus produit. Cependant, elle repose sur des lexiques de référence dont la complétude, la qualité, ou tout simplement la disponibilité pour le domaine traité ne sont pas assurés.

Dans cet chapitre, nous examinons ces deux aspects – la construction et l'évaluation des thésaurus distributionnels – sous l'angle de la recherche d'information, utilisée à la fois comme technique et comme usage. Concernant la construction, nous montrons dans ce chapitre comment il est possible d'explorer une approche RI de la construction des thésaurus en examinant l'intérêt de différents modèles classiques de RI, en les comparant à l'état de l'art (Claveau, Kijak et Ferret 2014a). Nous testons également les modèles de type word2vec (Mikolov et al. 2013) qui ont fait l'objet de beaucoup de recherches ces dernières années. Concernant l'évaluation, nous avons proposé une évaluation extrinsèque des thésaurus produits dans une tâche de RI classique. Cela nous permet de mettre en regard ces résultats avec ceux obtenus par évaluation intrinsèque et donc de juger de la pertinence de ces scénarios d'évaluation.

Après un état de l'art (section suivante), le chapitre aborde ces deux contributions successivement : les aspects relatifs à la construction des thésaurus sont présentés en section 4.3, ceux portant sur l'évaluation par RI sont en section 4.4. Nous présentons enfin quelques conclusions et perspectives sur ce travail dans la dernière section.

4.2 État de l'art

4.2.1 Construction de thésaurus distributionnels

La construction de thésaurus distributionnels a fait l'objet de nombreuses études depuis les travaux pionniers de Grefenstette (1994) et Lin (1998). Toutes reposent sur l'hypothèse distributionnelle de Firth (1957) que l'on résume par sa formule célèbre : *You should know a word by the company it keeps*. On considère que chaque mot est caractérisé sémantiquement par l'ensemble des contextes dans lesquels il apparaît. Pour un mot en entrée d'un thésaurus, des mots partageant des similarités de contextes sont proposés ; on les appelle voisins sémantiques par la suite. La nature du lien sémantique entre une entrée et ses voisins est variable ; ils peuvent être des synonymes de l'entrée, des hyperonymes, des hyponymes ou d'autres types de liens sémantiques (Budanitsky et Hirst 2006 ; Adam et al. 2013, pour une discussion). Ces liens sémantiques, même s'ils sont très divers, sont néanmoins utiles pour de nombreuses applications liées au Traitement Automatique des Langues. Cela explique que ce champ de recherche soit encore très actif, avec des contributions portant sur différents aspects liés à la construction de ces thésaurus.

Tout d'abord, différentes pistes sur ce qui peut être considéré comme contexte distributionnel ont été explorées. On distingue ainsi les contextes graphiques des contextes syntaxiques. Les premiers sont simplement les mots apparaissant autour du mot étudié. Les seconds sont les mots recteurs ou dépendants syntaxiques du mot examiné. La seconde approche est souvent considérée comme plus précise, mais elle repose bien sûr sur une analyse syntaxique préalable qui n'est pas toujours disponible et peut même être source d'erreurs.

Les connexions entre sémantique distributionnelle et RI sont nombreuses. Plusieurs chercheurs ont, par exemple, utilisé des moteurs de recherche pour collecter des informations de co-occurrences ou des contextes sur le web (P.D. Turney 2001 ; Bollegala et al. 2007 ; Sahami et Heilman 2006 ; Ruiz-Casado et al. 2005). Les représentations vectorielles des contextes sont également souvent utilisées de différentes manières (P. Turney et Pantel 2010), mais sans lien avec les systèmes de pondérations et les fonctions de pertinence classiques de la RI (à l'exception de (Vechtomova et Robertson 2012) dans un cadre un peu différent de similarité entre entités nommées). Plusieurs travaux se sont pourtant penchés sur la pondération des contextes pour obtenir de meilleurs voisins. Par exemple, Broda et al. (2009) a proposé de ne pas considérer directement les poids des contextes, mais les rangs pour s'affranchir de l'influence des fonctions de pondérations.

D'autres ont proposé des méthodes d'amorçage (*bootstrap*) pour modifier les poids des contextes d'un mot en prenant déjà en compte ses voisins sémantiques (Zhitomirsky-Geffet et Dagan 2009 ; Yamamoto et Asakura 2010). Par ailleurs, beaucoup de travaux se sont basés sur le fait que la représentation "traditionnelle" des contextes distributionnels est très creuse et redondante, comme l'a illustré Hagiwara et al. (2006). Dans ce contexte, plusieurs méthodes de réduction de la dimension ont été testées : depuis l'analyse sémantique latente (T. K. Landauer et S. T. Dumais 1997 ; Padó et Lapata 2007 ; T. Van de Cruys et al. 2011), jusqu'au *Random Indexing* (Sahlgren 2001), en passant par la factorisation par matrices non négatives (Tim Van de Cruys 2010).

Nous avons pour notre part simplement proposés d'identifier plus complètement le processus de recherche de voisins distributionnels comme un problème de recherche documentaire classique (Claveau, Kijak et Ferret 2014a). L'ensemble des contextes d'un mot peut en effet être représenté comme un document ou une requête, ce qui permet de trouver facilement les mots proches, ou plus exactement les ensembles de contextes proches. Bien que partageant de nombreux points communs avec des travaux de l'état de l'art, cette façon simple de poser le problème de la construction des thésaurus distributionnels offre des pistes intéressantes et un outillage facilement accessible. C'est cette approche que nous reprenons dans le cadre de ce chapitre ; nous la décrivons plus en détail dans la section 4.3.1.

4.2.2 Évaluation des thésaurus distributionnels

Comme nous l'avons dit précédemment, l'évaluation des thésaurus produits se fait soit de manière intrinsèque, en les comparant à une ressource de référence, soit de manière extrinsèque, au travers de leur utilisation dans une tâche précise.

Dans le cas de l'évaluation intrinsèque, il faut disposer de lexiques de référence. Il est alors simple de calculer rappel, précision ou toute autre mesure de qualité du lexique produit. Cette approche a été utilisée pour évaluer de nombreux travaux. Parmi les lexiques régulièrement utilisés comme références, on peut citer WordSim 353 (Gabrilovich et Markovitch 2007), ou ceux utilisés par Ferret (2013) qui exploitent des ressources plus larges, à savoir les synonymes de WordNet 3.0 (Miller 1990) et le thésaurus Moby (Ward 1996). Ce sont ces deux derniers lexiques que nous utilisons nous aussi pour l'évaluation intrinsèque ; voir ci-après pour une présentation. D'autres ressources ne sont pas directement des lexiques, mais des jeux de données permettant une évaluation directe, comme le jeu de synonymes du TOEFL (T. Landauer et S. Dumais 1997) ou

l'ensemble de relations sémantiques BLESS (Baroni et Lenci 2011).

L'évaluation directe séduit par sa simplicité, mais pose la question de l'adéquation des lexiques utilisés comme références. Plusieurs recherches ont donc proposé des évaluations indirectes au travers d'une tâche. La plus connue est la tâche de substitution lexicale mise en œuvre à SemEval 2007 (McCarthy et Navigli 2009). Étant donné un mot dans une phrase, le but est de remplacer ce mot par un de ses voisins supposés et de vérifier que cela n'altère pas le sens de la phrase. Les résultats obtenus sont ensuite comparés aux substitutions proposées par des humains. Cette tâche va donc privilégier les synonymes exacts au détriment des autres types de relations sémantiques. L'évaluation de thésaurus distributionnels par des tâches de RI n'a pas, à notre connaissance, été explorée. Bien sûr, l'utilisation d'informations que l'on peut qualifier de distributionnelles dans un cadre de RI a fait l'objet de plusieurs travaux (Besançon et al. 1999 ; Billhardt et al. 2002) qui se prolongent de nos jours par les travaux sur les représentations lexicales apprises par réseaux de neurones (Huang et al. 2012 ; Mikolov et al. 2013). Il s'agit dans tous les cas de tirer parti des similarités de contextes entre mots pour améliorer la représentation des documents et/ou la fonction de pertinence RSV. Cependant, ces travaux ne dissocient pas le processus de création du thésaurus distributionnel du processus de RI, ce qui rend impossible l'évaluation de l'apport des informations distributionnelles seules. Dans notre cas, l'évaluation extrinsèque par RI que nous proposons (cf. section 4.4) repose simplement sur l'utilisation des voisins sémantiques pour étendre des requêtes ; le reste du système de recherche d'information est standard. Cela doit nous permettre de juger au mieux de la qualité des thésaurus produits.

4.3 Modèles de RI pour l'analyse distributionnelle

4.3.1 Principes et matériel

Comme nous l'avons expliqué en introduction, le problème de la construction d'un lexique distributionnel peut être vu comme un problème de recherche de documents similaires et peut donc être mis en œuvre avec des techniques de RI. Dans ce cadre, pour un mot donné, ses contextes dans un corpus sont collectés et rassemblés. C'est cet ensemble de contextes qui forme un document. Construire une entrée du lexique, c'est-à-dire trouver les mots proches au sens distributionnel d'un mot w_i, revient alors à trouver les documents (contextes) proches du document représentant les contextes de w_i.

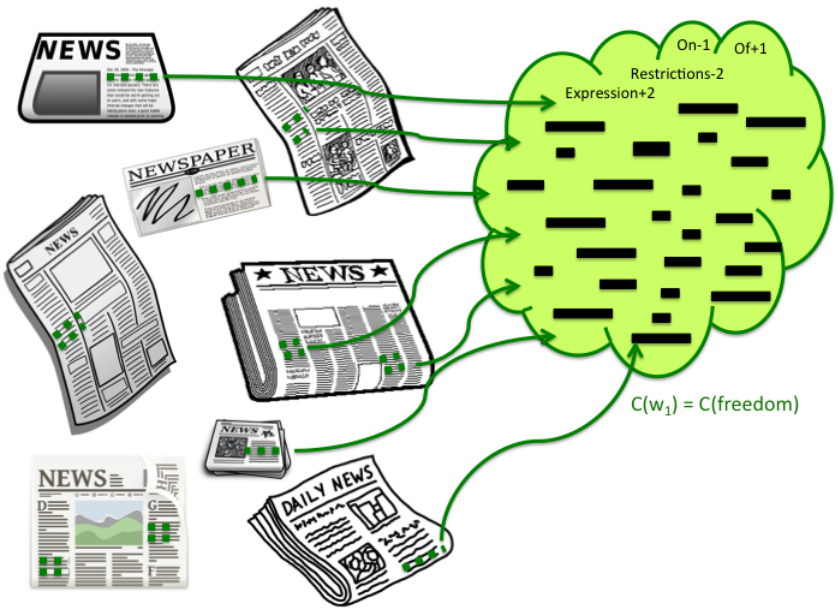


FIGURE 4.1 – Illustration de la collecte des contextes des mots (ici, *freedom*) et de leur représentation comme des documents en RI (sacs de mots, ou sacs de contextes)

Les données que nous utilisons pour nos expériences de construction sont celles utilisées dans plusieurs travaux. Cela va nous permettre de comparer nos résultats à ceux publiés. Le corpus utilisé pour collecter les contextes est le corpus AQUAINT-2 ; il est composé de d'articles de presse en anglais et compte 380 millions de mots. Parmi eux, les mots que nous considérons pour entrées de notre lexique sont les noms communs apparaissant au moins 10 fois dans le corpus, soit 25 000 noms différents. Les contextes de toutes les occurrences de ces mots sont donc collectés ; dans les expériences rapportées ci-dessous, le contexte d'un mot est composé des deux mots à droite et deux mots à gauche du nom visé, en gardant leur position. Par exemple, dans l'extrait : "*... all forms of restriction on freedom of expression, threats ...*", les mots *restriction-2*, *on-1*, *of+1*, *expression+2* sont ajoutés à l'ensemble des contextes *freedom*. La figure 4.1 illustre l'ensemble de cette démarche.

Comme nous l'avons évoqué précédemment, nous utilisons conjointement WordNet (WN) et Moby pour l'évaluation intrinsèque des thésaurus produits. Ces deux ressources offrent des caractéristiques complémentaires : WN recense des liens sémantiques forts (synonymes ou quasi-synonymes) alors que Moby recense une plus grande variété de liens (hyperonymes, méronymes, co-hyponymie...). Une description détaillée des liens considérés

Méthode	MAP	R-Prec	P@1	P@5	P@10	P@50	P@100
Ferret 2013 *base*	5,6	7,7	22,5	14,1	10,8	5,3	3,8
Ferret 2013 *best rerank*	6,1	8,4	24,8	15,4	11,7	5,7	3,8
Ferret 2014 *synt*	7,9	10,7	29,4	**18,9**	**14,6**	**7,3**	**5,2**
Hellinger	2,45	2,89	9,73	6,28	5,31	4,12	3,30
TF-IDF	5,40	7,28	21,73	13,74	9,59	5,17	3,49
TF-IDF ajusté	7,09	9,02	24,68	15,13	11,55	5,96	4,31
Okapi-BM25	6,72	8,41	24,82	14,65	10,85	5,16	3,66
Okapi-BM25 ajusté	**8,97**	**10,94**	**31,05**	18,44	13,76	6,46	4,54
LM Dirichlet $\mu = 25$	6,52	7,56	23,46	11,88	8,16	2,99	1,89
LM Dirichlet $\mu = 250$	6,56	7,43	23,08	12,31	8,17	2,77	1,73
LM Dirichlet $\mu = 2500$	5,83	6,77	23,28	12,06	8,00	2,98	1,81
LM Hiemstra $\lambda = 0,45$	5,41	6,79	25,09	12,07	8,17	3,05	1,90
LM Hiemstra $\lambda = 0,65$	8,10	8,98	27,06	13,35	9,25	3,41	2,13
LM Hiemstra $\lambda = 0,85$	7,06	7,88	25,28	12,44	8,41	3,04	1,89
LM Hiemstra $\lambda = 0,95$	6,49	7,64	27,21	13,62	9,17	3,28	2,06

TABLE 4.1 – Performances des modèles de RI pour la construction des thésaurus distributionnels sur la référence WN+Moby

par ces ressources est donnée par FERRET (2013) ou CLAVEAU, KIJAK et FERRET (2014a). Ainsi, WN propose en moyenne 3 voisins pour 10 473 des noms du corpus AQUAINT-2 et Moby 50 voisins en moyenne pour 9 216 noms. Combinées, ces deux ressources couvrent 12 243 noms du corpus avec 38 voisins en moyenne. Le nombre de noms dans les listes de référence et la variété des relations sémantiques considérées font de ces données un jeu d'évaluation très complet par rapport à d'autres *benchmarks* parfois utilisés tels que WordSim 353 (GABRILOVICH et MARKOVITCH 2007).

4.3.2 Test des modèles de RI

Le tableau 4.1 présente les résultats obtenus par différents systèmes de construction de thésaurus, appliqués au corpus AQUAINT-2. Les mesures de performances utilisées pour comparer les thésaurus produits à la référence WordNet+Moby sont classiquement la précision à différents seuils (P@x), la MAP et la R-précision, moyennés sur les 12 243 noms de la référence WN+Moby et exprimés en pourcentage.

Nous testons notamment des systèmes probabilistes par modèles de langues (notés LM) dans lesquels la pertinence d'une requête est évaluée en fonction de la probabilité

d'apparition des mots selon un modèle appris pour chaque document ; voir équation 4.1.

$$\mathrm{RSV}_{\mathrm{LM}}(Q, D) = \prod_{t \in Q} P(t|\mathcal{M}_D)^{qtf} \tag{4.1}$$

Les notations sont les suivantes : Q est une requête, D un document, t un terme d'indexation, $\mathcal{M}_D$ un modèle de langue du document, et qtf est le nombre d'occurrences de t dans Q.

Pour estimer la probabilité d'apparition d'un terme, deux variantes très usuelles sont testées correspondant à deux techniques de lissage, toutes deux utilisant les probabilités d'apparition des mots dans toute la collection pour mieux estimer celles des documents. Il s'agit d'une part du lissage de Dirichlet que nous testons avec différentes valeurs du paramètre μ ; voir équation 4.2.

$$P(t|\mathcal{M}_D) = \frac{tf + \mu * P(t, C)}{dl(D) + \mu} \tag{4.2}$$

$dl(D)$ est la longueur du document D et $P(t, C)$ la probabilité d'apparition de t dans l'ensemble de la collection (typiquement sa fréquence dans la collection), tf est le nombre d'occurrences de t dans D. Et d'autre part, nous testons le lissage à la Hiemstra (ou lissage de Jelinek-Mercer) avec différentes valeurs de λ ; voir équation 4.3.

$$P(t|\mathcal{M}_D) = (1 - \lambda)\frac{tf}{dl(D)} + \lambda P(t|C) \tag{4.3}$$

Ces modèles sont mis en œuvre en utilisant le système de RI Indri [1].

Nous rapportons également les résultats des systèmes déjà présentés dans (Claveau, Kijak et Ferret 2014a), qui reposent sur la similarité d'Hellinger (Escoffier 1978 ; Domengès et Volle 1979), un TF-IDF/cosinus, et Okapi-BM25 (Robertson et al. 1998) (avec les valeurs par défaut $k_1 = 2$, $k_3 = 1000$ et $b = 0,75$).

Nous proposons une version dite ajustée de la similarité Okapi-BM25, dans laquelle l'influence de la taille du document est renforcée, en prenant $b = 1$, et en mettant l'IDF au carré pour donner plus d'importance aux mots de contexte plus discriminants. Nous appliquons également cette stratégie pour obtenir une version ajustée du TF-IDF/cosinus en prenant l'IDF au carré.

Ces modèles de RI, très classiques, ne sont pas détaillés plus avant ici ; le lecteur intéressé trouvera les notions et détails utiles dans les références citées ou des ouvrages

1. Indri est disponible à `http://www.lemurproject.org/` .

généralistes (Manning et al. 2008 ; Boughanem et Savoy 2008, par exemple).

À des fins de comparaison, nous rapportons les résultats obtenus dans les mêmes conditions expérimentales avec une approche état de l'art notée *base* exploitant une similarité cosinus avec une pondération par information mutuelle (Ferret 2013), une version avec apprentissage (*rerank*) pour réordonnancer les voisins (*ibid.*), et une version (*synt*) reposant non plus sur des contextes graphiques, mais syntaxiques (Ferret 2014).

On observe tout d'abord la difficulté de la tâche puisque dans tous les cas, les précisions relevées sont très faibles selon cette évaluation intrinsèque. La comparaison avec les lexiques de référence conduit donc à une conclusion très sévère quant à la qualité supposée des thésaurus produits. On note tout de même que certains modèles de RI fonctionnent particulièrement bien par rapport à l'état de l'art, comme les modèles basés sur Okapi, ou les modèles de langues.

4.3.3 Test des modèles de réduction de dimension et d'*embedding*

Les plongements de mots (*word embedding*) dans des espaces vectoriels ont connu un regain d'activité ces dernières années avec l'avènement de nouveaux modèles neuronaux. Parmi les différentes approches proposées, les travaux de (Mikolov et al. 2013) reposant sur le concept de *skip-gram* pour représenter le contexte font référence. Ces travaux ont donné naissance à word2vec, un outil permettant de représenter les mots comme des vecteurs denses dans un espace de faible dimension (typiquement $\mathbb{R}^{200}$). Nous rapportons dans le tableau 4.2 le résultat de tels modèles (notés W2V) sur notre tâche dans les mêmes conditions expérimentales que précédemment, avec différents paramètres (nombre de dimension et taille de la fenêtre de contexte considérée). Nous indiquons également le résultat d'un modèle fourni par Google [2] appris sur un corpus d'actualités (Google News) de 100 milliards de mots. Même si ce corpus est différent, la comparaison est néanmoins intéressante, puisqu'il s'agit d'un modèle utilisé directement dans de nombreuses applications, entraîné sur un corpus de même genre mais de taille bien plus conséquente.

Nous testons également des techniques de réduction de dimensions plus classiques, à savoir *Latent Semantic Indexing* (LSI) (Deerwester et al. 1990), *Latent Dirichlet Allocation* (LDA) (Hoffman et al. 2010) et *Random Projections* (RP) (Bingham et Mannila 2001), avec différents nombres de dimensions. Toutes ces approches (W2V, LSI,

2. Ce modèle est disponible à l'URL suivante : `https://code.google.com/p/word2vec/` .

Méthode	MAP	R-Prec	P@1	P@5	P@10	P@50	P@100
Ferret 2013 *base*	5,6	7,7	22,5	14,1	10,8	5,3	3,8
Ferret 2013 *best rerank*	6,1	8,4	24,8	15,4	11,7	5,7	3,8
Ferret 2014 *synt*	**7,9**	**10,7**	**29,4**	**18,9**	**14,6**	**7,3**	**5,2**
LSI dim=50	1,62	2,86	5,00	4,12	3,76	2,78	2,35
LSI dim=500	4,37	6,27	16,00	10,76	8,78	4,61	3,45
LSI dim=1000	5,06	6,87	21,09	13,20	9,96	5,39	4,02
LSI dim=2000	5,11	6,86	23,11	14,34	10,78	5,12	3,72
LDA dim=500	0,60	1,25	2,17	2,21	1,90	1,29	1,13
RP dim=500	5,66	6,48	27,3	12,85	8,67	3,04	1,86
RP dim=2000	5,90	7,04	27,13	13,71	8,94	3,21	1,96
W2V dim=50 w=5	2,89	3,89	13,48	7,36	5,44	2,58	1,82
W2V dim=100 w=5	3,65	4,84	18,49	9,62	7,04	3,16	2,17
W2V dim=200 w=5	3,92	5,44	22,18	11,39	8,32	3,61	2,59
W2V dim=300 w=5	5,25	6,25	18,67	10,72	7,73	3,49	2,38
W2V dim=400 w=5	5,06	6,43	20,37	11,44	8,29	3,66	2,50
W2V dim=50 w=9	3,12	4,11	13,11	7,80	5,68	2,59	1,87
W2V dim=100 w=9	4,14	5,55	17,18	9,25	6,79	3,21	2,21
W2V dim=200 w=9	4,42	5,60	17,69	10,71	7,47	3,40	2,32
W2V dim=300 w=9	4,07	5,53	20,50	11,13	8,02	3,62	2,52
W2V dim=400 w=9	4,39	5,51	17,81	9,95	7,43	3,24	2,21
W2V Google news	5,82	7,51	13,28	11,60	8,94	3,93	2,54

TABLE 4.2 – Performances des modèles de RI pour la construction des thésaurus distributionnels sur la référence WN+Moby

LDA, RP) sont implémentées en utilisant la bibliothèque Python GenSim [3] (ŘEHŮŘEK et SOJKA 2010).

Les résultats obtenus pour l'ensemble de ces méthodes apparaissent comme faibles au regard de ceux vus en sous-section précédente. Les modèles W2V, très utilisés, sont notamment inférieurs à l'état de l'art et même à certaines méthodes de réduction classiques (LSI). Une piste d'explication pourrait être la difficulté de la tâche d'apprentissage rapportée aux nombre de mots, comme le suggèrent les résultats W2V légèrement meilleurs obtenus avec le corpus Google News, qui est 250 fois plus gros qu'AQUAINT-2. Il faut cependant noter que même avec cette quantité de mots, les résultats obtenus sont à peine du niveau de l'état de l'art et restent largement inférieurs aux modèles RI vus précédemment.

Les autres techniques de réduction de dimension donnent des résultats d'autant plus limités que le nombre de dimensions considérées est petit. Ce résultat faible est en ligne avec certaines conclusions de travaux précédents (Tim VAN DE CRUYS 2010). Le fait

3. GenSim est disponible à https://radimrehurek.com/gensim/ .

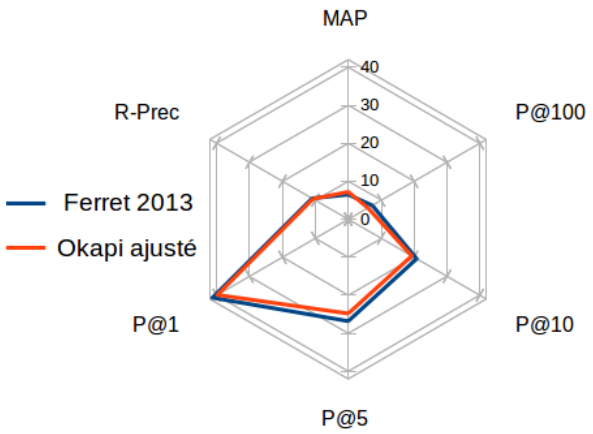


FIGURE 4.2 – Performances pour la construction des thésaurus distributionnels sur la référence WN+Moby pour les mots de fréquence élevée (> 1 000).

d'agréger en une seule dimension des mots différents est donc préjudiciable pour bien distinguer les voisins sémantiques. Autrement dit, l'apparition de certains mots de contexte bien précis est un indicateur fort pour juger de la proximité sémantique des mots. Cela est d'ailleurs confirmé par le fait qu'au sein d'une même famille de modèle de RI (sous-section précédente), les paramétrages menant aux meilleurs résultats sont ceux qui donnent plus de poids aux mots discriminants : IDF au carré pour Okapi, faible lissage pour les modèles de langue (μ et λ relativement petits).

4.3.4 Analyse par fréquence

Certains auteurs ont remarqué que la fréquence des mots dont on essaie de trouver les voisins a une grande influence sur la qualité finale (FERRET 2013). Plus ils sont fréquents, plus on a de contextes pour les décrire et meilleurs sont les résultats avec les méthodes « état de l'art ». On se propose donc de vérifier si l'emploi de méthodes issues de la RI amène la même observation. Pour cela, on reprend le cadre expérimental précédent et le modèle Okapi ajusté, mais on distingue les résultats selon la fréquence des mots-entrées : les mots ayant les plus hautes fréquences (>1000), ceux avec les fréquence les plus basses (<100) et le tiers restant avec des fréquences moyennes. Ces résultats sont présentés dans les figures 4.2 à 4.4. Là encore, nous indiquons les résultats état de l'art de (*ibid.*) pour comparaison.

Il apparaît que l'approche par RI a un comportement bien plus stable selon les fré-

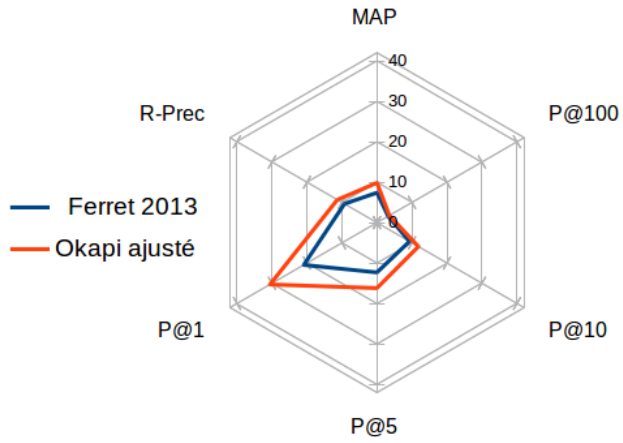


FIGURE 4.3 – Performances pour la construction des thésaurus distributionnels sur la référence WN+Moby pour les mots de fréquence moyenne ($< 1\,000$ et > 100)

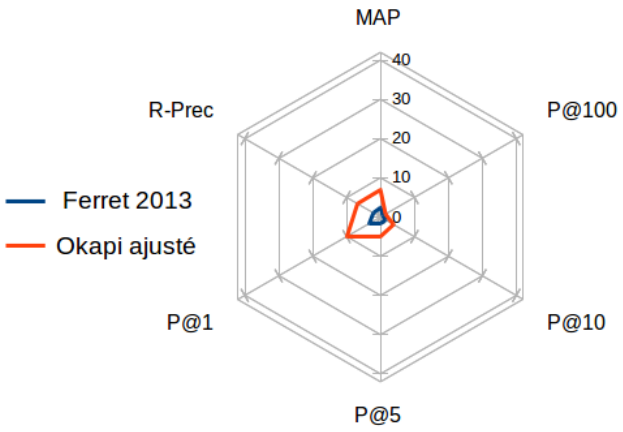


FIGURE 4.4 – Performances pour la construction des thésaurus distributionnels sur la référence WN+Moby pour les mots de fréquence faible (< 100)

quences que le système état de l'art de (FERRET 2013). En particulier, l'approche RI assure des résultats de bonne qualité pour les mots faiblement fréquents. La fréquence des mots étant directement liée à la taille des ensembles de contextes, cela indique l'importance de la normalisation en fonction de la taille des documents dans l'approche RI.

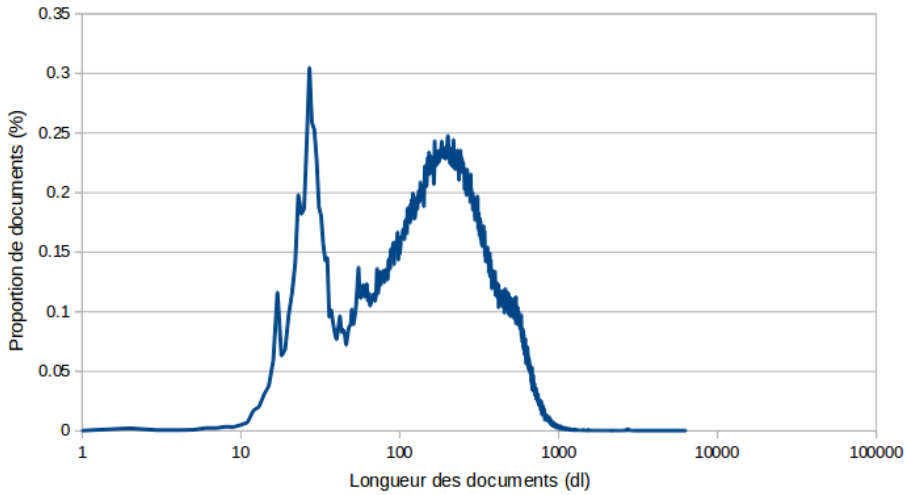


FIGURE 4.5 – Distribution des tailles des documents dans le cadre standard ; échelle log.

4.3.5 Limites de l'analogie avec la RI

L'analogie entre recherche de document similaire et recherche de voisins distributionnels apporte de très bons résultats, mais il convient cependant de pointer certaines limites de cette analogie. En effet, les ensembles de contextes, qui sont considérés comme des documents, ont des propriétés sensiblement différentes des documents réels. Pour illustrer cela, nous produisons respectivement en figures 4.5 et 4.6 la distribution des tailles de documents standard (ce sont ceux du corpus AQUAINT-2, c'est-à-dire des articles de journaux) et de celles des ensembles de contextes. On y observe un éventail de taille beaucoup plus important dans le cas des ensembles de contextes. Il semble donc important dans les fonctions de similarités utilisées de prendre en compte ce besoin de normalisation accrue selon la longueur des documents.

La distribution des mots est également assez différente de ce que l'on trouve dans une vraie collection de documents. Cela est illustré en figures 4.7 et 4.8 dans lesquelles on donne la distribution des fréquences documentaires (DF), en se comparant là encore avec le corpus AQUAINT-2 original. Les mots apparaissent en général dans beaucoup plus de contextes que dans le cas de vrais documents. Par exemple, le nombre de mots apparaissant dans 1 document sur 10 000 (DF=0.0001) est près de 100 fois plus élevé que pour de vrais documents. Comme nous l'avions vérifié expérimentalement, ce phénomène milite pour une prise en compte spécifique de cette distribution dans les modèles (à travers les lissages dans les modèles de langue ou l'IDF dans les modèles vectoriels par exemple,

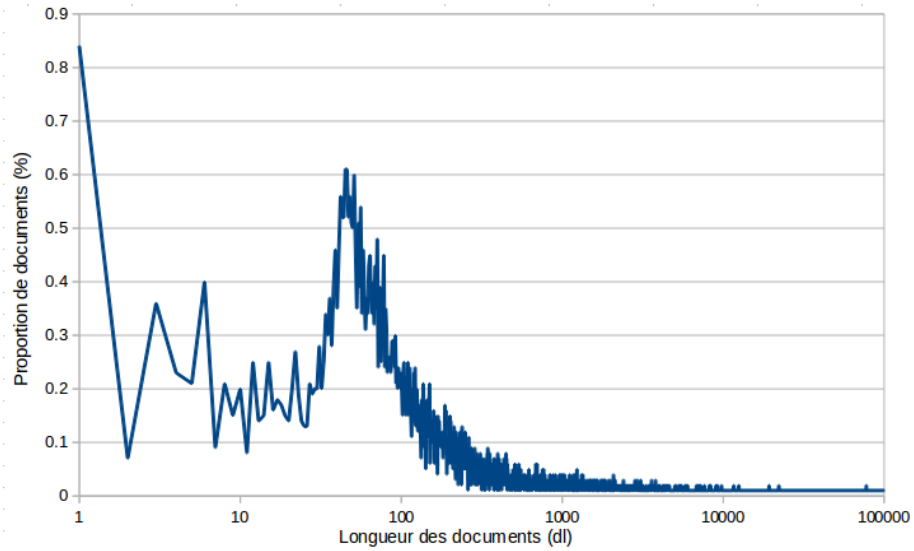


FIGURE 4.6 – Distribution des tailles des documents dans le cadre des ensembles de contextes ; échelle log.

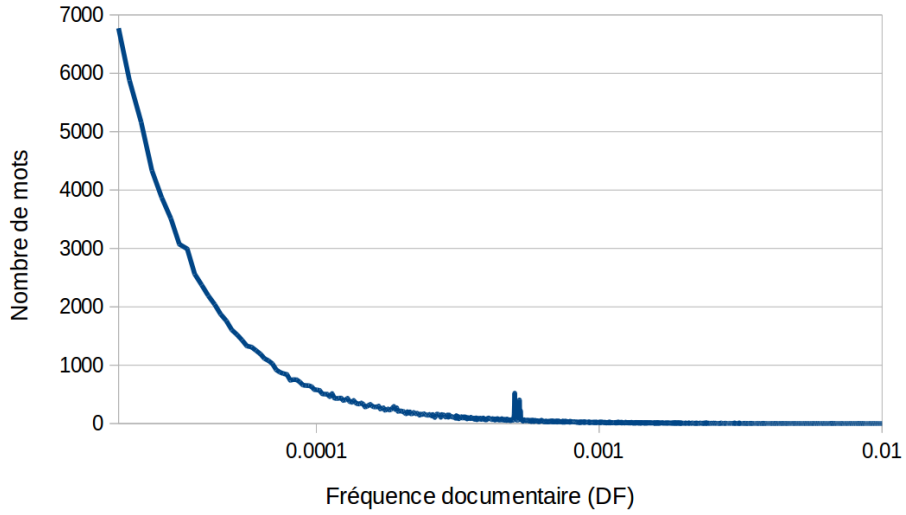


FIGURE 4.7 – Distribution des fréquences documentaires (DF) dans le cadre standard ; échelle log.

ou à travers de nouveaux schémas de pondérations).

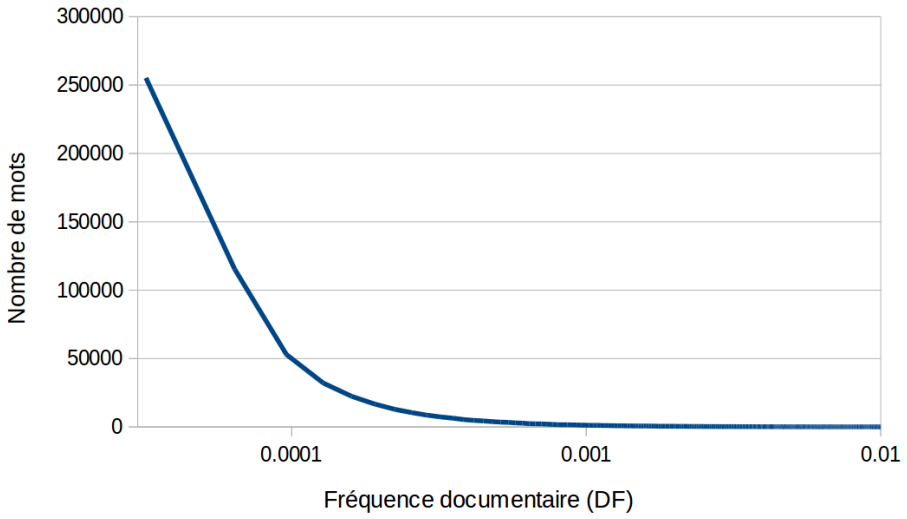


FIGURE 4.8 – Distribution des fréquences documentaires (DF) dans le cadre des ensembles de contextes ; échelle log.

4.4 Évaluation dans un cadre de RI

Pour évaluer l'apport des thésaurus distributionnels dans une tâche classique de RI, nous nous plaçons dans un cadre d'extension de requêtes. Pour chaque nom de la requête, les mots associés dans le thésaurus distributionnel sont ajoutés à celle-ci. Nous décrivons ci-dessous notre contexte expérimental, puis les résultats obtenus. Nous proposons ensuite de mettre en regard les résultats obtenus par cette évaluation indirecte avec les résultats de l'évaluation intrinsèque que nous avons utilisée précédemment.

4.4.1 Contexte expérimental

La collection de RI que nous utilisons est celle développée pour le projet Tipster et utilisée dans le cadre de TREC. Elle contient plus de 170 000 documents et cinquante requêtes. Ces requêtes sont composées de plusieurs champs (la requête à proprement parler, un champ narratif détaillant les critères de pertinence) ; dans les expériences rapportées ci-dessous, nous n'utilisons que le champ requête. Cette collection est particulièrement adaptée puisqu'elle est composée de documents en anglais de même nature que le corpus AQUAINT-2 (articles du *Wall Street Journal*) à partir duquel le thésaurus distributionnel a été construit.

Le système de recherche d'information que nous utilisons est Indri (METZLER et W.

Croft 2004 ; Strohman et al. 2005), connu pour offrir des performances état de l'art. Ce système probabiliste implémente une combinaison de modèle de langue (J. M. Ponte et W. B. Croft 1998), tel que vu précédemment, et de réseaux d'inférence (Turtle et W. Croft 1991) permettant d'utiliser des opérateurs tels que ET OU... Dans les expériences rapportées ci-dessous, nous l'utilisons avec des réglages standard, à savoir un lissage de Dirichlet ($\mu = 2500$). Dans notre cas, ce système de RI offre l'avantage de disposer d'un langage de requête complexe qui nous permet d'inclure les mots du thésaurus distributionnel en exploitant au mieux le modèle par réseau d'inférence à l'aide de l'opérateur dédié '`#syn`' qui permet d'agréger les comptes des mots considérés comme synonymes (voir la documentation d'Indri pour plus de détails). Pour supprimer les effets de flexions (pluriel) sur les résultats, les formes pluriel et singulier des noms de la requêtes sont ajoutées, que ce soit dans les requêtes non étendues avec les synonymes ou celles étendues par les voisins sémantiques.

Les performances pour cette tâche de RI sont également classiquement mesurées en précision à différents seuils (P@x), R-prec, MAP. L'évaluation du lexique consiste donc en la comparaison des résultats obtenus avec ou sans extension, que nous mesurons en gain relatif de précision, de MAP... Nous indiquons également la moyenne des gains d'AP par requête, notée AvgGainAP (à ne pas confondre avec le gain de MAP, qui est le gain calculé sur les moyennes des AP par requête). Les résultats non statistiquement significatifs (Wilcoxon et t-test avec $p < 0,05$) sont en italiques.

4.4.2 Résultats d'extension

Le tableau 4.3 présente les gains de performance obtenus en étendant les requêtes avec les mots collectés dans les thésaurus. Nous choisissons le lexique ayant obtenu les meilleurs résultats : celui construit avec la méthode Okapi ajustée. Puisque ce lexique ordonne les voisins par proximité avec le mot-entrée, on teste différents scénarios : pour chaque mot (nom) de la requête, s'il apparaît dans le thésaurus, on ne garde que ses 5, 10 ou 50 plus proches voisins. Sur les cinquante requêtes, cela concerne 136 noms. À des fins de comparaison, on indique aussi les résultats obtenus en étendant avec les lexiques de référence WN seul et WN+Moby. Voici un exemple de requête, avec sa forme non-étendue et sa forme étendue (Okapi ajusté top 5) utilisant les opérateurs de réseau d'inférence d'Indri :

— requête : `coping with overcrowded prisons`

— forme normale : `#combine( coping with overcrowded #syn( prisons prison ) )`

Extension	MAP	Avg GainAP	R-Prec	P@5	P@10	P@50	P@100
Sans	21,78	-	30,93	92,80	89,40	79,60	70,48
avec WN	+12,44	**+36,3**	+7,01	**+4,31**	**+7,16**	+7,60	+10,87
avec WN+M	+11,00	+28,33	+7,78	*+3,02*	+5,37	+6,53	+9,17
avec Okapi-BM25 ajusté top 5	+13,14	+29,99	**+11,17**	+3,45	+5,15	**+9,40**	**+12,43**
avec Okapi-BM25 ajusté top 10	**+13,80**	+24,36	+9,58	*+2,16*	+4,03	+5,58	+8,26
avec Okapi-BM25 ajusté top 50	+10,02	+17,99	+8,82	+3,45	+3,36	+3,72	+5,36

TABLE 4.3 – Gains relatifs de performance (%) par extension de requête selon le lexique utilisé

— forme étendue : `#combine( coping with overcrowded #syn( prisons prison inmate inmates jail jails detention detentions prisoner prisoners detainee detainees ) )`

On note tout d'abord que quel que soit le lexique utilisé, l'extension de requête apporte un gain significatif de performance. Comme beaucoup de travaux depuis, cela contredit au passage les conclusions de (VOORHEES 1994) sur l'absence d'intérêt à utiliser WN pour étendre des requêtes. Le fait le plus notable est cependant les excellentes performances (MAP) du lexique construit automatiquement, qui dépassent même celles des lexiques de référence. Alors que sa précision sur les 10 premiers voisins a été évaluée à moins de 14 % en section 4.3, ce lexique produit des extensions obtenant le meilleur gain en MAP. La moyenne des gains d'AP (AvgGainAP) apporte également des informations intéressantes : celle-ci est maximale avec WN, qui offre donc une amélioration stable (c'est-à-dire une amélioration concernant beaucoup de requêtes) grâce au fait qu'il ajoute à la requête principalement des voisins très proches sémantiquement (synonymes exacts), sans « prise de risque ». Cette stabilité diminue avec les autres lexiques, et est la plus basse avec les extensions par les 50 plus proches voisins du lexique généré par le modèle Okapi ajusté. Comme la MAP reste globalement bonne, cela indique que seules certaines requêtes bénéficient d'un gain absolu important.

4.5 Évaluation intrinsèque vs. évaluation extrinsèque

Les résultats de l'expérience précédente soulèvent des questions sur la cohérence entre les résultats de l'évaluation intrinsèque et ceux de l'évaluation extrinsèque. Le gain de précision entre deux méthodes de construction de thésaurus, même s'il est jugé statistiquement significatif, est-il sensible en RI ? Dans cette section, nous tentons de répondre à cette question en examinant les différences entre évaluation intrinsèque et extrinsèque

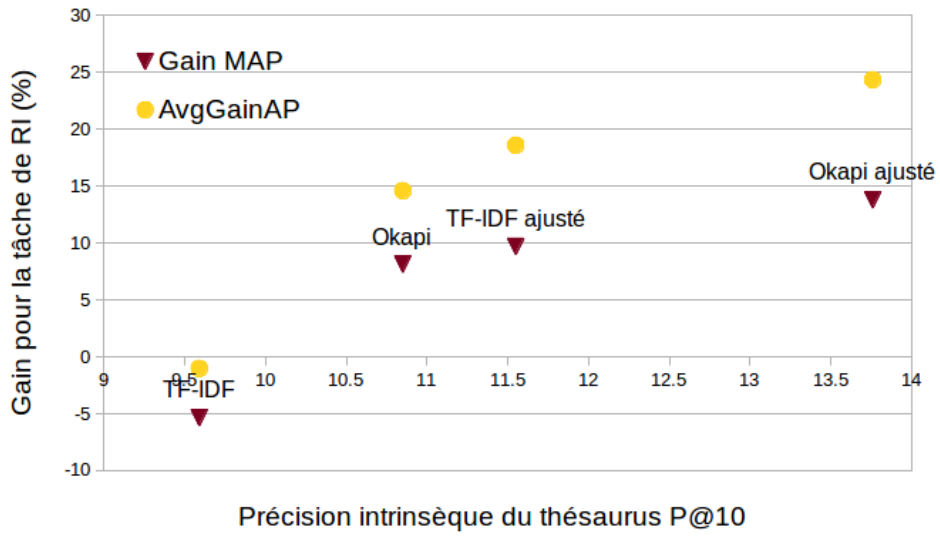


FIGURE 4.9 – Gain en MAP et AvgGainAP de différents modèles selon leur précision @10 lors de l'évaluation intrinsèque

au travers de quelques expériences complémentaires.

4.5.1 Mise en regard des précisions intrinsèque et extrinsèque

Pour cela on complète les résultats précédents avec la figure 4.9 qui rapporte les résultats de différents modèles de RI sur la tâche d'extension (avec les 10 premiers voisins) selon leur P@10 de l'évaluation directe. Il en ressort que la précision mesurée avec l'évaluation directe est liée aux gains mesurés dans la mesure où l'ordre est bien respecté : la meilleure P@10 à l'évaluation directe obtient le meilleur gain de MAP à la tâche de RI, etc. Mais la corrélation n'est pas linéaire comme on pourrait s'y attendre. En outre, des différences statistiquement significatives lors de l'évaluation directe (comme entre TF-IDF ajusté et Okapi ajusté) ne se traduisent pas forcément par des différences statistiquement significatives à la tâche d'extension. Parmi les faux positifs de l'évaluation directe (mots détectés comme proches mais absents dans les lexiques de référence), certains semblent plus ou moins néfastes pour étendre les requêtes.

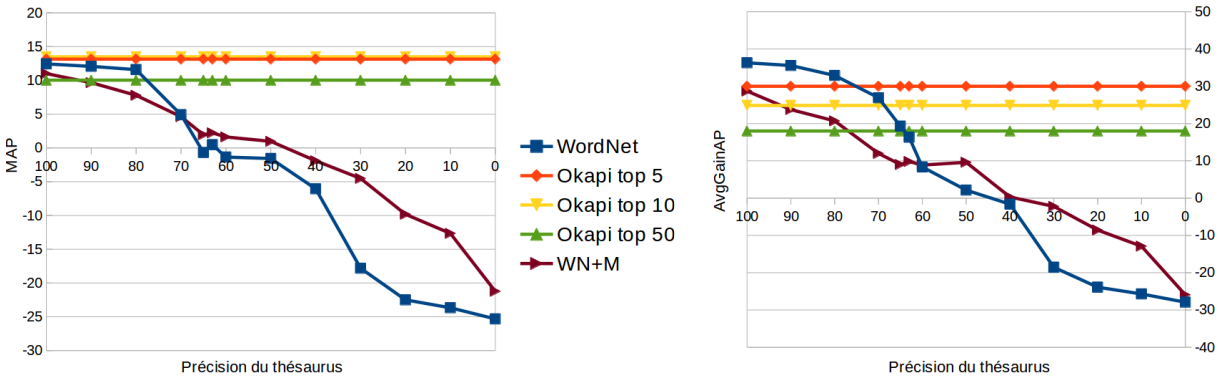


FIGURE 4.10 – Gain en MAP (gauche) et AvgGainAP (droite) selon la précision contrôlée artificiellement des thésaurus utilisés pour étendre les requêtes

4.5.2 Faux positifs et bonnes extensions

Il est alors intéressant d'examiner plus précisément l'effet de ces faux positifs. On examine de nouveau l'évolution des performances sur la tâche de RI en fonction de la qualité des listes de voisins utilisées pour étendre les requêtes, mais cette fois-ci, des listes de voisins plus ou moins bruitées sont générées à partir des thésaurus de référence en remplaçant des voisins par des mots choisis aléatoirement dans le vocabulaire. On peut ainsi produire des listes de voisins avec une précision variable et contrôlée, dont on évalue les performances pour étendre les requêtes comme précédemment. La figure 4.10 montre l'évolution de la MAP et de l'AvgGainAP en faisant varier ainsi la précision des listes de voisins données par les références WN seul et WN+Moby. Une précision du thésaurus de 20 % signifie donc que chaque requête est étendue avec la liste de ses voisins fournie par WN seul ou WN+Moby, dans laquelle 20 % des vrais voisins sont remplacés aléatoirement par d'autres mots. On indique pour comparaison les scores obtenus avec les top 5, 10 et 50 du lexique Okapi ajusté.

Comme attendu, les deux mesures de performance chutent lorsque la précision des listes diminue. Il faut une précision controlée (intrinsèque) des listes inférieure à 50 % pour rendre les gains de performance nuls sur la tâche de RI, et en deçà, les extensions de requête dégradent les résultats. Il y a donc bien une corrélation entre la précision des listes mesurée par évaluation directe et les performances pour l'extension de requête, du moins lorsque que les faux positifs sont pris au hasard. Mais dans le cas du lexique que nous avons généré, les performances obtenues sont comparables à des listes de précision intrinsèque entre 70 et 100 % selon les cas, alors que la précision mesurée par évaluation

Extension avec Okapi-BM25 ajusté	MAP	Avg GainAP	R-Prec	P@5	P@10	P@50	P@100
top 10 sauf WN	+11,80	+21,60	+8,37	*+2,16*	+3,58	+5,08	+6,87
top 10 sauf WN+M	+9,36	+19,22	+6,41	*+3,02*	+3,36	*+3,17*	+5,73

TABLE 4.4 – Gains relatifs de performance (%) par extension de requête avec les voisins jugés faux positifs

intrinsèque variait entre 10 et 20 %. Plus que la sévérité de l'évaluation intrinsèque, cela souligne la faiblesse de la démarche qui repose sur des références incomplètes : certains voisins, jugés comme faux positifs car non listés par les références sont en réalité de bons candidats.

Pour illustrer ce dernier point, nous rapportons dans le tableau 4.4 les performances obtenues par le lexique Okapi ajusté en étendant de nouveau les requêtes avec les 10 premiers voisins de chaque nom, mais en excluant ceux qui sont listés comme voisins dans WN ou WN+Moby. Autrement dit, on ne garde que les voisins jugés comme faux positifs par l'évaluation intrinsèque. Il apparaît clairement que ces faux positifs sont bien liés sémantiquement à l'entrée. Pour le mot `prison` de la requête précédente, parmi les 10 premiers voisins, ceux absents de WN+Moby sont : `sentence, abuse, detainee, guard, custody, defendant, inmate, prisoner`. Ils semblent effectivement bien liés sémantiquement à *prison*.

4.6 Conclusion

Dans cet chapitre, nous avons présenté en détail l'utilisation de la recherche d'information à la fois pour construire et pour évaluer des thésaurus distributionnels. Nous avons d'une part utilisé les modèles de similarités développés en RI sur les contextes des mots, ce qui nous permet, pour un mot donné, de trouver ceux partageant une similarité contextuelle, et donc sémantique. D'autre part, la recherche d'information, à travers la tâche classique de recherche de documents par requête, nous offre un cadre applicatif permettant une évaluation indirecte des thésaurus produits.

De ces travaux, deux conclusions majeures se dégagent. Nous avons démontré le bienfondé de l'approche RI pour la construction des thésaurus sémantiques. Nous avons en particulier montré l'importance de la prise en compte des mots discriminants dans différents modèles (au travers de pondérations spécifiques pour l'IDF ou par le lissage).

Nous avons également souligné l'avantage des modèles RI par rapport aux méthodes classiques en particulier sur les mots avec peu d'occurrences, mais aussi sur les modèles de type WORD2VEC qui font actuellement l'objet de beaucoup de travaux. Mais nous avons également souligné les limites de l'analogie entre RI et sémantique distributionnelle : les ensembles de contextes ont des propriétés statistiques (taille, fréquence d'apparition des mots...) très différentes de « vrais » documents. Cela milite pour l'établissement de fonctions de pondération et de pertinence adaptées à cette réalité et ouvre donc des voies d'amélioration possibles. D'autres perspectives sur ce point concernent l'utilisation de techniques récentes de RI pour la construction des thésaurus (*learning to rank*, représentations continues...).

L'autre conclusion majeure de ce travail porte sur la fiabilité de l'évaluation intrinsèque. En montrant que les thésaurus obtenus offrent des résultats d'expansion de requête au moins aussi bons que les listes de référence servant à l'évaluation intrinsèque, nous remettons en perspective beaucoup de conclusions de travaux précédents qui ne se basait que sur l'évaluation intrinsèque pour comparer des méthodes d'analyse distributionnelle. Les faibles résultats obtenus aux évaluations intrinsèques ne se traduisent pas dans notre cadre applicatif d'extension de requête. Il convient bien sûr de nuancer cette conclusion, qui ne porte ici que sur le cadre applicatif de la recherche documentaire : la tâche et la mise en œuvre que nous utilisons (avec les opérateurs de croyances d'Indri) permettent d'avoir des liens sémantiques relativement distants dans les listes de voisins servant d'extensions sans que cela ne dégrade trop les résultats. Cette conclusion importante nous a cependant poussé étudier les performances de nos analyses distributionnels et de différents plongements sur une plus grande variété de tâches (substitution lexicale notamment) et à étudier l'influence de paramètres (par exemple, taille de fenêtre de voisinage) pour chacune de ces tâches (Claveau et Kijak 2016a). Le but de ce travail subséquent est ainsi de mesurer la corrélation entre les scores d'évaluation intrinsèque et extrinsèque dans différentes tâches pour mieux aider à choisir les méthodes de construction (et leurs paramètres) les plus adaptées selon la tâche finale visée.

Enfin, le travail présenté dans ce chapitre a également servi de base à la définition d'une mesure d'indiscriminabilité s'appliquant aussi bien aux plongements mots qu'aux représentations vectorielles des documents (Claveau 2016 ; Claveau 2018). Cette mesure, qui s'appuie sur la dimensionnalité intrinsèque (Houle et al. 2012 ; Amsaleg et al. 2015), permet de juger de la qualité de la représentation du mot ou du document en fonction de la distribution de son voisinage. Elle permet ainsi d'anticiper, dans un contexte

d'analyse distributionnelle ou de plongement, qu'un mot aura une représentation vectorielle peu fiable, menant à des voisins sémantiquement peu pertinents, ou dans un contexte de RI, qu'une requête ramènera des documents risquant d'être peu pertinents.

Bilan et perspectives

Chapitre 5

Bilan et discussion

5.1 Retour sur nos travaux

5.1.1 À la croisée des domaines

Comme nous le précisions dès l'introduction, nos travaux, évoqués ou non dans ce manuscrit, sont le fruit d'échanges avec des collègues, au sein de mon équipe de recherche, et pour une grande part, en dehors. Le nuage en figure 5.1 les présente proportionnellement au nombre de publications partagées. Ces collaborations ont souvent été l'occasion d'explorer des contextes applicatifs différents, allant de l'indexation de la littérature biomédicale à la terminologie des commentaires de football en passant par l'analyse de sentiment dans les réseaux sociaux. De cette variété, il est difficile de situer une spécificité forte à nos travaux si ce n'est celle d'emprunter et d'adapter des concepts et outils développés dans certaines communautés (RI ou TAL bien sûr, mais aussi vision par ordinateur, logique floue, indexation en grandes dimensions, probabilités...) pour les apporter dans une autre.

Depuis quelques années, nous prolongeons cette activité de fertilisation croisée par notre implication dans des structures de promotion de la recherche (ARIA, GdR CNRS MaDICS) et dans l'organisation d'événements permettant des rencontres en dehors des

FIGURE 5.1 – Nuage de co-auteurs, sur les publications de 2000 à 2018

silos scientifiques habituels :

— journalisme et informatique en 2016 et en 2017
— Big Data et IA
— conférence RI et TAL
— revue RIDoWS sur le TAL, la RI, le multimedia et le Web sémantique
— etc.

5.1.2 À l'épreuve du temps

Beaucoup de nos travaux ont cherché à découvrir des liens sémantiques entre des mots, à extraire des informations spécifiques des textes ou à classer des textes. Nous nous sommes appuyé pour cela sur des indices morphologiques ou distributionnels, des représentations vectorielles, exploités ensuite par des techniques de fouille des données variées (apprentissage statistique ou symbolique, modélisation de séquence, analogie...) que nous avons le plus souvent adaptés aux spécificités du problème traité.

Ces outils de fouille ont largement évolué, et nos travaux les plus anciens seraient certainement faits différemment aujourd'hui. Par exemple, l'identification de liens sémantiques

entre des mots sont désormais largement vus comme des problèmes de plongements lexicaux. Cette approche prend en compte tout à la fois les liens de nature distributionnelle, mais aussi morphologique. Elle a aussi l'avantage de proposer une proximité graduelle (score de proximité, souvent basée sur le cosinus entre les représentations vectorielles), plus adaptée aux systèmes stochastiques, qui diffère du tout-ou-rien de notre approche discrète par analogie de forme (section 1.2). Plus largement, la vision symbolique adoptée dans nombre de nos travaux est désormais remplacée par une vision algébrique ou géométrique dans laquelle les mots ou phrases sont représentées dans des espaces vectoriels. Les problèmes d'extraction d'information et de classification sont quant à eux largement abordés avec des outils issus des réseaux de neurones exploitant ces représentations.

Certaines de nos préoccupations d'alors sont cependant toujours d'actualité, comme celles concernant l'économie vis-à-vis des données d'entraînement. Elles nous ont conduits à développer certaines stratégies faiblement ou non supervisées (Claveau et Ncibi 2013 ; Claveau et Ncibi 2014 ; Claveau et Kijak 2017). Cela constitue encore une piste de recherche très importante (cf. section 6.1.1).

Enfin, on peut constater au cours du temps la disparition de modélisations linguistiques dans nos travaux, en dehors des concepts élémentaires (mot-forme, lemme, morphe/morphème...). Alors que nos premiers travaux s'ancraient respectivement dans la théorie du lexique génératif (Pustejovsky 1995) et de la théorie sens-texte (Mel'čuk 1996) pour définir les liens sémantiques intéressants à extraire et leurs propriétés, tous nos travaux suivants s'en sont émancipés. Ce mouvement n'est pas propre à notre parcours ; les tâches de traitement automatique des langues sont désormais définies en extension (au travers d'exemples) plutôt qu'en intention (au travers d'une définition formelle), et dans le meilleur des cas, au travers d'une mesure opérationnelle de performance (par exemple, une f-mesure calculée sur un jeu de test). Le TAL a rejoint en cela le domaine de la RI, dans lequel les travaux sont traditionnellement très largement guidés par l'évaluation quantitative.

5.2 De la distinction entre RI et TAL

La distinction faite entre techniques et problématiques de RI et celles de TAL, utile pour structurer ce manuscrit, est bien sûr plus floue en réalité. D'une part, beaucoup de travaux en RI ne s'intéressent pas au matériau textuel ; on étudie notamment les problèmes de stockage et d'indexation causés par les volumes de données manipulées (structures

d'indexation, calcul distribué, accès concurrent...), les aspects liés à la prise en compte de l'utilisateur (modélisation des préférences, ergonomie des interfaces, métaphores de visualisation...), les points liés spécifiquement au cas du web (hyperliens et graphe du web, spam, réputation de sites...), etc. D'autre part, les problématiques d'accès à l'information contenue dans des documents textuels ne se résument pas à la seule tâche de recherche documentaire à l'aide d'une requête. Ainsi, d'autres applications sont aussi le point de rencontre entre TAL et RI. Nous en présentons quelques-unes ci-dessous, en nous attardant sur le cas emblématique de la RI translingue.

5.2.1 Au-delà de la recherche documentaire

La recherche de documents à partir d'une requête est l'application prototypique de la RI, mais l'accès à l'information peut prendre la forme d'autres applications. Il s'agit, par exemple, de l'extraction d'information (voir (Ananiadou et al. 2013) pour un état de l'art), du résumé automatique mono- ou multidocument (Kundi et al. 2014), ou encore des systèmes de questions-réponses déjà discutés.

Les problèmes de représentation du contenu textuel, de calcul de similarité et la prise en compte des phénomènes linguistiques que nous avons évoqués se posent également dans ces applications. D'autres pistes sont plus spécifiques, comme par exemple la résolution d'anaphores, très utilisée dans ces applications (Vicedo et Ferrandez 2000 ; Steinberger et al. 2005) alors qu'elle ne l'est pas en recherche documentaire, ou encore la génération de texte, utile dans certaines tâches de résumé...

5.2.2 RI translingue

La RI translingue est née du besoin qu'un utilisateur peut vouloir retrouver des documents écrits en une autre langue que sa propre langue (langue de requête). Par rapport à la RI monolingue, la RI translingue ajoute la dimension de traduction : on doit traduire la requête (langue source) dans la langue des documents (langue cible), ou traduire les documents dans le sens inverse (Nie 2010). Cette phase de traduction étant un des problèmes centraux en TAL, on retrouve en RI translingue une forme de synergie entre TAL et RI. Il s'agit bien d'une synergie et non pas d'une simple combinaison car, en plus de l'utilisation des systèmes de traduction automatique (TA) comme boîte noire pour produire une traduction, il y a aussi beaucoup de tentatives pour adapter la phase de traduction à la tâche de RI. En effet, du fait de la déséquentialisation imposée par la

représentation sac-de-mot, la traduction d'une requête (ou d'un document) n'a pas pour but de produire un texte compréhensible par un être humain (comme dans le cas de TA en général). Cette observation a amené des chercheurs à utiliser seulement des modèles de traduction statistique (modèles IBM) (Kraaij et al. 2003 ; J. Gao et al. 2006) ou même un dictionnaire bilingue (Pirkola et al. 2001 ; Levow et al. 2005) dans la traduction d'une requête, en ignorant le modèle de langue, élément très utile en TA pour la production d'une phrase légitime en langue cible. Il faut noter qu'il s'agit bien d'une synergie : le modèle de traduction est souvent bien intégré dans un modèle de recherche (Kraaij et al. 2003), et non pas utilisé dans une étape indépendante.

Notons que la RI translingue a aussi eu un impact sur la RI monolingue : la traduction a été utilisée pour générer des paraphrases de la requête ou du document (Berger et Lafferty 1999). Ce modèle est maintenant largement répandu dans les moteurs de recherche : on peut considérer une requête et le titre d'un document sur lequel un utilisateur a cliqué comme une paire de textes parallèles. Les modèles de traduction entraînés sur ces données peuvent aider à améliorer la recherche (Jianfeng Gao, He et al. 2010).

Chapitre 6

Évolution et perspectives

6.1 Quelques pistes de recherche

Il est usuel de terminer ce type de manuscrit avec quelques pistes de recherche semblant prometteuses. C'est un exercice délicat auquel le chercheur CNRS a l'habitude de se plier tous les cinq ans, et qui peut donc constater l'écart entre ses prédictions et la réalité cinq ans plus tard. Les cartes sont souvent rebattues par des contingences matérielles (appels à projet, financements acceptés ou refusés, recrutements heureux ou malheureux), les rencontres avec des collègues, et les petites révolutions scientifiques du domaine qui viennent fermer des portes et en ouvrir d'autres. Ne doutons pas que ces éléments viendront nous détourner des quelques axes de recherche que nous présentons ci-dessous.

6.1.1 Apprentissage artificiel

Depuis les années 1990 puis intensivement dans les années 2000, la plupart des tâches du TAL, et désormais de la RI, ont été définies sous le paradigme de l'apprentissage artificiel. Ce travail de conceptualisation (par exemple, appréhender un problème de détection de relations en un problème d'étiquetage de séquence) et de construction de données

(corpus annotés) est, de notre point de vue, la véritable révolution du domaine. Les générations successives de méthodes d'apprentissage (symbolique ou statistiques comme les arbres de décision ou l'apprentissage bayésien, les modèles à noyaux et SVM, les méthodes à base de graphes de type HMM, MaxEnt ou CRF, et bien sûr, récemment, apprentissage profond) n'en sont que la conséquence logique.

De ce fait, beaucoup de perspectives de TAL portent désormais non pas sur les tâches mais sur les techniques permettant de les effectuer. Ces pistes de recherche sont très largement partagées par toute la communauté des sciences des données qui utilisent les mêmes outils d'apprentissage. Nous en évoquons quelques unes ci-dessous.

La question de la quantité de données annotés ou non est un sujet de recherche ancien, mais rendu très actuel du fait de l'emploi de modèles d'apprentissage contenant énormément de paramètres, et nécessitant donc a priori beaucoup de données d'entraînement. L'étude d'approches semi-supervisées ou non-supervisée, de supervision lointaine, de transfert prennent donc tout leur sens dans ce contexte. Nous avons abordé modestement cette question dans le cas particulier de l'apprentissage actif (*active learning*) de modèles graphiques (CRF) en adoptant l'idée d'essayer de maximiser la diversité des données plutôt que leur nombre (Claveau et Kijak 2017). Évidemment, les multiples contextes et multiples techniques à considérer rendent cette piste extrêmement vaste.

L'explicabilité est également une piste largement mise en avant dans de nombreux projets de recherche actuellement. Le fonctionnement dit « boîte noire » de certains classifieurs (par exemple, les réseaux de neurones) et la distribution des applications dans le monde réel justifie ce besoin. Il faut noter que ce besoin est largement lié à l'acceptabilité de ces technologies. Un utilisateur de moteur de recherche acceptera plus facilement un document ne répondant pas à son besoin d'information s'il constate que des mots de sa requête y apparaissent. Avec des modèles de RI fondés sur des plongements de mots, de la similarité de second ordre (voir section 3.3), ou des techniques de type LSI, l'interprétation des résultats est rendue plus difficile et donc potentiellement moins acceptable.

Comme beaucoup d'auteurs, nous avons déjà souligné que les réseaux de neurones, notamment au travers des mécanismes d'attention (Claveau et Raymond 2017) permettaient d'offrir des clés de compréhension des résultats, par exemple en surlignant dans la séquence d'entrée (phrase, tweet, etc.) les portions responsables de la classification finale. L'utilisation d'architectures intégrant plus directement une composante dédiée à l'interprétabilité (par exemple au sein d'un GAN - Generative Adversarial Network (Goodfellow et al. 2014)) pourrait être étudiée. Outre ces approches, intimement liées au classifieur

effectuant la tâche considérée, il est important de souligner que l'interprétabilité peut aussi consister à fournir une explication plausible et compréhensible, a posteriori, même si ce n'est pas elle qui a effectivement conduit à la décision finale. Il est ainsi envisageable d'utiliser un autre classifieur, plus facilement interprétable par l'utilisateur, auquel on fournit en entrée les données et, à qui on impose en contrainte de prédire la même décision que le classifieur opaque. L'incorporation de ces contraintes pose d'intéressants problèmes d'apprentissage.

6.1.2 Multimodalité

La multimodalité est une piste de recherche à la fois ancienne et plus que jamais d'actualité. La langue, et plus précisément le matériau écrit dont il a été question dans ce manuscrit, n'est qu'une petite part de notre système d'interaction avec le monde, et la prise en compte conjointe d'autres modalités de communication, tel l'oral ou l'image, soulève d'intéressants problèmes, notamment de représentation conjointe. Nous y avons été sensibilisé au travers de notre environnement de recherche au sein des équipes TexMex puis LinkMedia à l'IRISA, mais surtout au travers d'encadrements de thèses. La thèse de Pierre Tirilly (Tirilly 2010) portait ainsi sur l'indexation texte-image, pour laquelle nous avions, entre autres, appliqué des principes de TAL (modèles de langues) pour décrire des images. Plus récemment, dans la thèse de Cédric Maigrot (Cédric Maigrot 2019) nous avons réexploré ces liens textes-images dans le contexte très particulier de la détection de fausses informations (*fake news*) dans lesquels il s'agissait de détecter et d'intégrer tous les indices issus du texte (style, contenu, source) et de l'image (traces de manipulation).

Portées là-encore par les réseaux de neurones, les évolutions des méthodes de représentations des textes et des images dans des espaces conjoints (Vukotic 2017, inter alia) permettent d'attaquer de nombreuses tâches multimodales. Parmi celles-ci, le *visual question answering* requiert à la fois une compréhension fine du texte, de l'image et des connaissances sur le monde, que ce soit dans le domaine général (Antol et al. 2015) ou dans un domaine de spécialité (Hasan et al. 2018). L'analyse et la détection des *deepfake* nécessite également une analyse conjointe des indices venant des contenus langagier et visuel des vidéos suspectes et une comparaison à l'existant ; cette application qui prolonge la thèse de C. Maigrot nous semble à ce titre particulièrement propice au mariage de la RI, du TAL et de la vision par ordinateur.

6.1.3 Éthique et sécurité

La diffusion des technologies de la langue dans la société soulève des enjeux d'éthique, de vie privé et de sécurité. Ces enjeux sont pour le chercheur autant de problèmes scientifiques intéressants.

Ainsi, le respect de la **vie privée** est une considération à laquelle le grand public est, tout à fait à propos, de plus en plus sensibilisé. En effet, les interactions d'un utilisateur avec un système de TAL ou de RI qu'il ne contrôle pas (typiquement un moteur de recherche sur le Web) ouvre des possibilités de profilage très précis. Dans de tels contextes, l'enjeu est de fournir des stratégie permettant une utilisation optimale des services offerts par ces systèmes tout en assurant une obfuscation de profil. Ce soucis de vie privée peut aussi concerner les documents confiés à ces systèmes : dans des cas spécifiques, on ne veut pas qu'un système soit capable de reconstruire les documents à partir des représentations que l'on lui fournit. Par exemple, un moteur travaillant sur des textes sensibles peut n'avoir accès qu'à leur représentation (sac-de-mot pondéré, modèle de langue ou représentations par LSTM...). Quelle sont alors les capacités de reconstructions des textes pour chacune de ces représentations ? L'adaptation de cadres théoriques, comme la *differential privacy*, à ces applications est un enjeu de recherche en plein essor (Yang et Zhang 2017).

La détection des **biais sociétaux** dans les données d'apprentissage et la limitation de leurs effet sur les classifieurs est également une considération éthique importante. Ces biais sont par exemple ceux concernant le genre ou l'appartenance ethnique des personnes apparaissant dans les textes. Un système de traduction, entraîné sur des données réelles peut ainsi traduire systématiquement *nurse* par *infirmière*. Bien que valide d'un point de vue statistique sur ses données d'apprentissage et d'évaluation, cette traduction perpétue un biais non acceptable d'un point de vue sociétal. Ces biais ont déjà été constatés grâce aux analogies construites à partir des plongements de mots (Bolukbasi et al. 2016), et récemment des stratégies visant à limiter l'impact de ces biais sur l'apprentissage des classifieurs ont été proposées dans un cadre statistique (Besse et al. 2018). Réduire l'impact de ces biais dans les systèmes de recherche d'information ou de recommandation est également un sujet de recherche sur lequel la communauté RI est en train de se positionner (Boratto et al. 2020). Les pistes de recherche incluent notamment la définition de ces biais (définition opérationnelle, mesures d'évaluation), la façon dont sont constitués les jeux de données en préalable de leur utilisation (sources possibles de biais, équilibrage, pré-traitements...), les techniques pour la découverte de ces biais et bien sûr les mesures de correction.

Les aspects de **sécurité** sont également importants à considérer dès lors que les systèmes ont vocation à sortir de l'environnement bienveillant des laboratoires. On peut distinguer deux types d'attaques sur ces systèmes selon qu'elles ont lieu pendant la construction du classifieur, ou lors de son utilisation. Dans le premier cas, il s'agit par exemple de savoir s'il est possible de construire des textes qui, inclus dans un corpus d'entraînement, perturberont l'apprentissage d'une manière déterminée. Et le cas échéant, la question duale est de savoir comment se prémunir de ce type d'attaque. Pour un moteur de recherche, on voudra par exemple avoir des assurances théoriques qu'un attaquant ne puisse pas polluer la base de document de manière à avoir ses documents toujours en tête des réponses, ou au contraire, toujours cachés. Pour l'extraction d'information, on examinera la possibilité pour un attaquant connaissant le système d'extraction (le classifieur et/ou les données d'apprentissage) de cacher l'information recherchée. De manière similaire, pour des tâches de classification de textes (filtrage de spam, détection de propos haineux...), un attaquant peut-il produire des textes trompant systématiquement le système ? Pour ces différentes tâches, le cadre générique de l'apprentissage antagoniste (*adversarial machine learning*) semble particulièrement adapté, mais sa déclinaison pour les technologies de la langue et la manipulation de séquences n'est pas directe.

6.2 Regard sur le passé : un rapprochement lent

Au delà de ces pistes de recherche, il est intéressant de s'interroger plus largement sur l'évolution de nos domaines de recherche. Nous reprenons ci-dessous quelques réflexions publiées en préface du numéro spécial TAL et RI de la revue Traitement Automatique des Langues (Claveau et Nie 2016). En 2000, la revue TAL avait déjà consacré un numéro sur les liens entre RI et TAL (Jacquemin 2000). Dans sa préface, Chr. Jacquemin revenait sur la citation de K. Spärck-Jones que nous rappelions dans l'introduction, tout en la modérant. Son constat était alors le suivant : l'apport du TAL en RI reste peu évident en général, mais il peut être bénéfique si les conditions suivantes sont remplies :

1. la tâche de RI doit nécessiter une représentation fine ;
2. les outils du TAL utilisés ne reposent pas sur « des représentations des connaissances riches dont l'adaptation en vraie grandeur est incertaine » ;
3. les outils du TAL n'induisent pas un coût calculatoire élevé.

Vingt ans après, le constat général que nous avons dressé dans (Claveau et Nie 2016), et qui a motivé l'angle de présentation de nos travaux, est toujours identique : la

fertilisation croisée entre ces deux domaines est longtemps restée assez pauvre, même si plusieurs auteurs ont montré le potentiel de l'interaction entre TAL et RI.

Pour autant, notre lecture de la situation diffère de celle proposée alors par Chr. Jacquemin. Tout d'abord, il est frappant de noter que seul le sens *TAL pour la RI* était considéré, la RI étant vue uniquement comme une application, mais pas comme un ensemble de concepts et de techniques pouvant être utiles au TAL. Pourtant, comme nous l'avons rappelé dans la partie I, les apports des techniques de la RI au sein de processus de TAL sont bien réels, que ce soit, par exemple, avec les représentations vectorielles, les pondérations, ou encore avec les procédures et données d'évaluation. À ce titre, il nous semble important d'encourager la diffusion au sein de la communauté TAL des développements récents en RI pour continuer cette fertilisation. Ces développements concernent les points évoqués précédemment sur la représentation des documents, le calcul de similarité, et les méthodologies et données d'évaluation. On peut par exemple citer les représentations des textes utilisées dans certains modèles de RI, plus récentes que les sacs-de-mots, mélangeant modèle de langue et réseaux d'inférences (Metzler et W. Croft 2004 ; Strohman et al. 2005), qui dépassent les limites des sacs de mots en offrant la possibilité de calculer des similarités en tenant compte de la proximité des mots, de phénomènes de synonymie, etc. Et bien entendu, depuis quelques années, il convient aussi de citer les travaux abordant, dans un contexte de RI, ces mêmes points (représentations, similarité) avec des réseaux de neurones (voir ci-dessous).

Les conditions 2 et 3 évoquées par Chr. Jacquemin portent sur la mise en œuvre en condition réelle, c'est-à-dire respectivement sur l'adaptabilité et la scalabilité. Concernant l'adaptabilité, la dépendance des outils du TAL à des systèmes experts ou des ressources créées manuellement a grandement diminué avec le développement des approches par apprentissage (supervisé, semi-supervisé ou non supervisé) et la mise à disposition de ressources généralistes ou spécialisées dans un grand nombre de langues. Cette évolution du TAL, que l'on schématise souvent par le passage des approches expertes ou symboliques aux approches statistiques, rend ce point moins prégnant qu'il a pu être à une certaine époque. Concernant le passage à l'échelle, là aussi l'évolution des ressources de calcul, reposant sur les capacités propres des machines (mémoire, CPU, GPU), et surtout sur le calcul distribué (grappe de calcul, *cloud*), a changé la donne. Il est maintenant possible, et même courant, d'exécuter des processus très lourds sur des grandes masses de textes en des temps compatibles avec les tâches de RI.

En revanche, la question de l'apport du TAL pour aider à représenter l'information

(point 1) reste essentielle pour analyser les succès et les échecs du TAL pour la RI. Les outils et ressources de TAL ne peuvent bénéficier à la RI que s'ils apportent quelque chose vis-à-vis du problème de l'ambiguïté, ou de celui du paraphrasage. Un point crucial est que cette connaissance supplémentaire doit être intégrable dans le système de RI. Cela peut se faire parfois très simplement (par exemple par extension de requête, voir sections 1.1.2 et 1.1.3), mais nécessite parfois une connaissance fine des mécanismes de RI, voire une révision complète de ces mécanismes (modification de la représentation, du calcul de similarité, etc. ; voir section 1.1.2). Dans ce cas, les interactions fructueuses ne peuvent se faire qu'au travers d'une connaissance assez pointue des deux domaines ou d'échanges étroits entre les deux communautés.

6.3 Regard vers le futur : des révolutions à venir ?

Depuis quelques années, l'essor de nouvelles techniques d'apprentissage, notamment l'apprentissage profond (*deep learning*), dessine un nouveau paysage pour l'avenir du TAL et de la RI, et donc de leurs interactions. Ces techniques communes attaquent en effet deux des points de convergence entre TAL et RI : la représentation du texte et le calcul de similarité. Pour le premier point, il est bien sûr question d'apprentissage de représentations, qu'elles soient dites distribuées, continues, spectrales, ou par plongements de mots (*word embeddings*), etc. Tous ces systèmes de représentation, réinventant l'analyse distributionnelle, permettent de dépasser beaucoup des contraintes de la représentation sac de mots : des mots proches sémantiquement seront proches dans l'espace de représentation, les régularités morphologiques (mise au pluriel, par exemple) se traduisent par des régularités géométriques, et certains raisonnements y sont possibles (sous la forme d'analogies, par exemple : `le poulain est au cheval ce que l'agneau est au mouton`). Plusieurs problèmes sont encore ouverts, dont celui, important pour la RI, de savoir comment représenter le contenu d'un texte à partir de la représentation de chacun de ses mots, même si des propositions existent depuis quelques années maintenant (Le et Mikolov 2014, inter alia). Un autre problème restant ouvert est celui de la représentation conjointe des données textuelles et de données structurées issues, par exemple, de bases de connaissances. La thèse en cours de François Torregrossa, que nous encadrons, se place dans ce sujet et y apportera des éléments de réponse nous l'espérons. Nous y étudions notamment l'utilisation d'espaces de plongements autres que euclidiens, comme les espace hyperboliques (Poincaré, Lorentz...) qui offre des propriétés de structuration hiérarchique

tout en gardant le bénéfice des représentations continues (Nickel et Kiela 2017).

Comme nous l'évoquions en section 3.1.2, un moteur de recherche, dont le cœur est de calculer des similarités entre textes, peut être vu comme un classifieur. Il est donc compréhensible que beaucoup de travaux cherchent désormais à apprendre ce classifieur. Le développement conjoint de méthodes d'apprentissage adaptées – *metric learning*, *learn to rank* ou désormais réseaux profonds (Mitra et Craswell 2017 ; Mitra et Craswell 2018) – mais surtout la mise à disposition de données (requêtes et documents associés, logs de moteurs de recherche sur le Web) et la disponibilité de puissance de calcul ont bien entendu été, encore une fois, des facteurs déterminants.

Enfin, il faut noter que cette mutation des pratiques ne concerne pas que le matériau textuel, mais d'une manière plus générale toutes les données non structurées (images, vidéos, son, parole et musique, relevés de capteurs, séries temporelles...), pour lesquelles les mêmes approches d'apprentissage sont employées. Cela interroge la spécificité du texte, de la langue, vis-à-vis d'autres types de données, et donc la spécificité des outils du TAL et de la RI vis-à-vis d'autres outils de fouille de données. Finalement, nous terminons ce manuscrit avec cette constatation : les frontières entre les corpus techniques du TAL et de la RI sont en train de disparaître, non pas parce que ces communautés se rapprochent, mais parce qu'elles se fondent l'une et l'autre dans une communauté plus grande, celle des sciences des données.

Annexe A

Curriculum

A.1 Position et responsabilités actuelles

Chargé de recherche au CNRS IRISA, équipe LinkMedia, depuis 2005

Directeur adjoint du GdR CNRS MaDICS animation de la recherche sur les Big Data et Data science, depuis 2015

Responsable de la revue RIDoWS

- thèmes : recherche d'information, documents numériques, Web sémantique
- revue ISTE OpenScience créée en 2017

A.2 Positions et responsabilités précédentes

Trésorier de l'ARIA

- association savante pour la Recherche d'Information et Applications
- animation de la recherche (organisation de conférences...), 2010-2019

Chercheur sur CDD

- juin 2005 à août 2005, Inserm U729, Paris
- sujet : traitement automatique du lexique biomédical

Chercheur post-doctoral

- février 2004 à mai 2005, OLST, Université de Montréal, CA
- boursier du programme Lavoisier du Ministère des affaires étrangères
- sujet : sémantique lexicale et traitement automatique des langues

Attaché temporaire à l'enseignement et la recherche

— septembre 2003- février 2004, IFSIC, Univ. de Rennes 1

A.3 Encadrements

Encadrements de thèse

— F. Moreau, recherche d'information, soutenue en 2006, encadrement 50 %, désormais MCF Rennes 2

— P. Tirilly, indexation d'image, soutenue en 2010, désormais MCF Lille

— A.-R. Ebadat, découverte de connaissances, soutenue en 2013, désormais employé de la start up MyScript

— A. Ncibi, structuration multimédia, thèse non soutenue, 2011-2015

— G. Jadi, analyse d'opinion, encadrement 50 %, thèse non soutenue, 2014-2017

— C. Maigrot, détection de rumeurs et fausses informations, encadrement 50 %, en cours, 2015-présent

— C. Dalloux, fouille de textes dans les dossiers patients, encadrement 50 %, en cours, 2016-présent

— F. Torregrossa, apprentissage de représentations dans un contexte de *search*, CIFRE avec PagesJaunes, en cours, 2018-présent

Autres encadrements ;

— Anne-Lyse Minard, postdoctorante sur le projet NexGenTV

— Sébastien Le Maguer, postdoctorant sur le projet CominWeb

— Davy Weissenbacher, postdoctorant sur le projet Quaero

— Laurent Ughetto, maître de conférences en conversion thématique

— Emmanuelle Martienne, maître de conférences en conversion thématique

A.4 Projets de recherche majeurs

— Icoda (data-journalism, fake news, fact-checking), projet Inria, participant, 2017-2020

— FigTem (extraction d'informations, domaine médical), projet CNRS-CONFAP (Brésil), porteur, 2016-2019

— BigClin (extraction d'informations, domaine médical), projet du Labex CominLabs, porteur, 2016-2019

— NexGenTV (enrichissement de contenu TV, analyse de sentiment), projet FUI, responsable pour l'IRISA, 2015-2018
— MDK (big data, data science), projet ANR, participant, 2015-2016
— LIMAH (fouille d'opinion, journalisme) projet du Labex CominLabs, participant, 2014-2018
— CominWeb (search), projet du Labex CominLabs, WP leader, 2014, 2017- présent
— Quaero (extraction d'information, RI), projet OSEO, responsable pour Inria des activités Traitement Automatique des Langues, 2008-2013

A.5 Bourses, prix

— Prix :
  > *Best demo award* à ACM Multimedia 2017
  > meilleur article à la conférence CORIA 2016
  > *Best paper award* à la conférence NEM Summit 2008
  > meilleur article à la conférence TALN 2005
  > meilleur article à la conférence TALN 2003
— Bourse de post-doctorat du programme Lavoisier du Ministère des affaires étrangères français (2004-2005)
— Allocation de recherche du Ministère de la recherche français (2000-2003)
— Médaille de la ville de Rennes 2003 pour mes travaux scientifiques

A.6 Vulgarisation, presse

— analyse automatique des données du Grand Débat : extraction de Termes avec TermEx. Cette analyse sert de support à un article du journal Ouest France
— interview sur le TAL dans le contexte du Grand Débat pour AEF info
— plusieurs interviews sur les deepfake et les fake news pour reportage pour Télématin Sciences, France2, 19 Nov 2019, Sciences et avenir hors-série 199, Télérama, Science et vie, Sciences et avenir, 01net N 894, LCI.fr 8/06/2018, Sciences et Avenir 30/05/2018, Le Monde.fr 24/05/2018, VL media, Sciences Ouest n363, Stratégies 06/06/2018 Stratégies 12/10/2018, Le journal du CNRS 27/09/2017
— participation à la conférence de presse du CNRS autour de la loi dite "fake news"
— audition à l'assemblée nationale autour de la loi dite "fake news"

— participation au dossier spécial sur la cybersécurité de l'Univ. Rennes 1
— participation dans un documentaire filmé sur les fake news pour la Cité des Sciences
— co-rédaction d'un chapitre sur les Données Multimédias dans le livre "Big Data à découvert", CNRS Éditions

A.7 Animation de la recherche

— Directeur adjoint du GdR MaDICS (Big Data en sciences) depuis 2015
— Trésorier ARIA (association savante en recherche d'informations) depuis 2010
— Responsable et rédacteur en chef de la revue ISTE RIDoWS
— Membre de nombreux comités de programmes (ECIR, ACL, WI, TALN, CORIA...)
— Membre du comité de rédaction de la revue TAL depuis 2011
— Responsable des séminaires du département MID de l'IRISA 2010-2014
— Participation à l'organisation d'EGC 2014, CORIA 2011 à 2019, école d'été EARIA 2012, 2014, 2016, 2018
— organisation de la conférence TALN-CORIA 2018, Rennes et présidence du comité de programme de TALN
— organisation du symposium MaDICS juin 2019, co-organisation de l'atelier-compétition DeFT 2019, organisation des journées journalisme computationnel 2016 et 2017, organisation de IA et Big Data 2016...

A.8 Enseignement

— Apprentissage sur les séquences, master 2 informatique, Univ Rennes 1
— Recherche d'information, master 2 miage, Univ. Rennes 1
— Apprentissage artificiel, 3e année ingénieur informatique, INSA de Rennes
— Indexation multimédia, 5e année ingénieur, ENSSAT, Lannion
— Fouille de texte, M2 Sciences des Données de Santé, Univ. Rennes 1
— Fouille multimédia, Master Big Data, ENSAI, Rennes

Bibliographie

Acosta, Otavio Costa, Aline Villavicencio et Viviane P. Moreira (2011), « Identification and Treatment of Multiword Expressions Applied to Information Retrieval », in : *Proceedings of the Workshop on Multiword Expressions : From Parsing and Generation to the Real World*, MWE '11, Portland, Oregon : Association for Computational Linguistics, p. 101-109, isbn : 978-1-932432-97-8, url : http://dl.acm.org/citation.cfm?id=2021121.2021141.

Adam, Clémentine, Cécile Fabre et Philippe Muller (2013), « Évaluer et améliorer une ressource distributionnelle : protocole d'annotation de liens sémantiques en contexte », in : *TAL* 54.*1*, p. 71-97.

Amsaleg, Laurent et al. (août 2015), « Estimating Local Intrinsic Dimensionality », in : *21st Conf. on Knowledge Discovery and Data Mining, KDD2015*, Sidney, Australia, url : https://hal.inria.fr/hal-01159217.

Ananiadou, Sophia, Nathalie Friburger et Sophie Rosset, éd. (2013), *Entités nommées*, t. 54-2.

Antol, Stanislaw et al. (2015), « VQA : Visual Question Answering », in : *International Conference on Computer Vision (ICCV)*.

Aronson, Alan R. et François-Michel Lang (2010), « An overview of MetaMap : historical perspective and recent advances », in : *JAMIA* 17.*3*, p. 229-236.

Baroni, M. et A. Lenci (2011), « How we BLESSed distributional semantic evaluation. », in : *Workshop on GEometrical Models of Natural Language Semantics*, p. 1-10.

Beam, Andrew L. et al. (2018), « Clinical Concept Embeddings Learned from Massive Sources of Medical Data », in : *CoRR* abs/1804.01486, arXiv : 1804.01486, url : http://arxiv.org/abs/1804.01486.

Berger, Adam et John Lafferty (1999), « Information retrieval as statistical translation », in : *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '99)*, sous la dir. d'ACM, New York, NY, USA, p. 222-229.

Besançon, Romaric, Martin Rajman et Jean-Cédric Chappelier (1999), « Textual Similarities based on a Distributional Approach », in : *Proceedings of the Tenth Inter-*

national Workshop on Database and Expert Systems Applications (DEXA'99), p. 180-184.

Besse, Philippe C. et al. (2018), « Can everyday AI be ethical. Fairness of Machine Learning Algorithms », in : *CoRR* abs/1810.01729, arXiv : 1810.01729, URL : http://arxiv.org/abs/1810.01729.

Billhardt, Holger, Daniel Borrajo et Victor Maojo (fév. 2002), « A Context Vector Model for Information Retrieval », in : *J. Am. Soc. Inf. Sci. Technol.* 53.*3*, p. 236-249, ISSN : 1532-2882, DOI : 10.1002/asi.10032, URL : http://dx.doi.org/10.1002/asi.10032.

Bingham, Ella et Heikki Mannila (2001), « Random Projection in Dimensionality Reduction : Applications to Image and Text Data », in : *Proc. of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, San Francisco, California : ACM, p. 245-250, ISBN : 1-58113-391-X, DOI : 10.1145/502512.502546, URL : http://doi.acm.org/10.1145/502512.502546.

Bojanowski, Piotr et al. (2016), « Enriching Word Vectors with Subword Information », in : *arXiv preprint arXiv :1607.04606*.

Bollegala, D., Y. Matsuo et M. Ishizuka (2007), « Measuring semantic similarity between words using web search engines », in : *Proc. of he conference WWW'2007*.

Bolukbasi, Tolga et al. (2016), « Man is to Computer Programmer As Woman is to Homemaker ? Debiasing Word Embeddings », in : *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, Barcelona, Spain : Curran Associates Inc., p. 4356-4364, ISBN : 978-1-5108-3881-9, URL : http://dl.acm.org/citation.cfm?id=3157382.3157584.

Boratto, Ludovico et al., éd. (2020), *First International Workshop on Algorithmic Bias in Search and Recommendation (Bias 2020)*, en marge de ECIR 2020, Lisbonne, Portugal.

Boughanem, Mohand et Jacques Savoy, éd. (avr. 2008), *Recherche d'information : états des lieux et perspectives*, français, Hermès Science, URL : http://www.editions-hermes.fr/.

Broda, Bartosz, Maciej Piasecki et Stan Szpakowicz (2009), « Rank-Based Transformation in Measuring Semantic Relatedness », in : *22nd Canadian Conference on Artificial Intelligence*, p. 187-190.

Budanitsky, Alexander et Graeme Hirst (2006), « Evaluating WordNet-based Measures of Lexical Semantic Relatedness », in : *Computational Linguistics* 32.*1*, p. 13-47.

Chapelle, Olivier et Yi Chang (2011), « Yahoo ! Learning to Rank Challenge Overview », in : *Proceedings of the Yahoo ! Learning to Rank Challenge, held at ICML 2010, Haifa, Israel, June 25, 2010*, sous la dir. d'Olivier Chapelle, Yi Chang et Tie-Yan Liu, t. 14, JMLR Proceedings, JMLR.org, p. 1-24, URL : http://www.jmlr.org/proceedings/papers/v14/chapelle11a.html.

Charniak, Eugene (1993), *Statistical Language Learning*, Cambridge, Massachusetts : MIT Press.

Claveau, Vincent (déc. 2003), « Automatic acquisition of semantic lexicons for information retrieval », Theses, Université Rennes 1, URL : https://tel.archives-ouvertes.fr/tel-00524646.

— (2009a), « Translation of Biomedical Terms by Inferring Rewriting Rules », in : *Information Retrieval in Biomedicine : Natural Language Processing for Knowledge Integration*, sous la dir. de Violaine Prince et Mathieu Roche, IGI Global, p. 106-123, DOI : 10.4018/978-1-60566-274-9.ch006, URL : https://hal.inria.fr/hal-00843785.

— (2009b), « Translation of Biomedical Terms by Inferring Rewriting Rules », in : *Information Retrieval in Biomedicine : Natural Language Processing for Knowledge Integration*, sous la dir. de Violaine Prince et Mathieu Roche, IGI - Global.

— (juin 2012), « Vectorisation, Okapi et calcul de similarité pour le TAL : pour oublier enfin le TF-IDF », in : *TALN - Traitement Automatique des Langues Naturelles*, Grenoble, France, URL : https://hal.archives-ouvertes.fr/hal-00760158.

— (août 2013), « IRISA participation to BioNLP-ST 2013 : lazy-learning and information retrieval for information extraction tasks », in : *BioNLP Workshop, colocated with ACL 2013*, Bulgaria, p. 188-196, URL : https://hal.archives-ouvertes.fr/hal-00912308.

— (jan. 2014), « Agrégation de sac-de-sacs-de-mots pour la recherche d'information par modèles vectoriels », in : *14 ème conférence Extraction et Gestion des Connaissances, EGC 2014*, Rennes, France, 6 p. URL : https://hal.archives-ouvertes.fr/hal-01027719.

— (mar. 2016), « Dimensionalité intrinsèque dans les espaces de représentation des termes et des documents », in : *Actes de la conférence CORIA 2016*, sous la dir. d'editor, Toulouse, France.

— (mar. 2018), « Indiscriminateness in representation spaces of terms and documents », in : *ECIR 2018 - 40th European Conference in Information Retrieval*, t. 10772, LNCS,

Grenoble, France : Springer, p. 251-262, DOI : 10.1007/978-3-319-76941-7_19, URL : https://hal.archives-ouvertes.fr/hal-01859568.

Claveau, Vincent et Ewa Kijak (2011), « Morphological Analysis of Biomedical Terminology with Analogy-Based Alignment », in : *Proceedings of the RANLP conference*, Hissar, Bulgarie.

— (oct. 2013), « Analyse morphologique non supervisée en domaine biomédical. Application à la recherche d'information », in : *Traitement Automatique des Langues* 54.*1*, p. 13-45, URL : https://hal.archives-ouvertes.fr/hal-00912301.

— (2015), « Thésaurus distributionnels pour la recherche d'information et vice-versa », in : *Revue des Sciences et Technologies de l'Information - Série Document Numérique* 18.*2-3*, DOI : 10.3166/DN.18.2-3.101-121, URL : https://hal.archives-ouvertes.fr/hal-01226551.

— (déc. 2016a), « Direct vs. indirect evaluation of distributional thesauri », in : *International Conference on Computational Linguistics, COLING*, Osaka, Japan, URL : https://hal.archives-ouvertes.fr/hal-01394739.

— (mai 2016b), « Distributional Thesauri for Information Retrieval and vice versa », in : *Language and Resource Conference, LREC*, Portoroz, Slovenia, URL : https://hal.archives-ouvertes.fr/hal-01394770.

— (avr. 2017), « Strategies to select examples for Active Learning with Conditional Random Fields », in : *CICLing 2017 - 18th International Conference on Computational Linguistics and Intelligent Text Processing*, Budapest, Hungary, p. 1-14, URL : https://hal.archives-ouvertes.fr/hal-01621338.

Claveau, Vincent, Ewa Kijak et Olivier Ferret (août 2014a), « Improving distributional thesauri by exploring the graph of neighbors », in : *Proceedings of the International Conference on Computational Linguistics, COLING*, Dublin, Irlande, URL : https://hal.archives-ouvertes.fr/hal-01027545.

— (août 2014b), « Improving distributional thesauri by exploring the graph of neighbors », in : *International Conference on Computational Linguistics, COLING 2014*, Dublin, Ireland, 12 p. URL : https://hal.archives-ouvertes.fr/hal-01027545.

Claveau, Vincent et Marie-Claude L'Homme (2005a), « Apprentissage par analogie de liens sémantiques entre dérivés morphologiques », in : *Actes de la conférence de Terminologie et Intelligence Artificielle, TIA'05*, Rouen, France.

— (2005b), « Structuring terminology by analogy machine learning », in : *Proceedings of the International conference on Terminology and Knowledge Engineering, TKE'05*, Copenhague, Danemark.
— (2005c), « Structuring Terminology by Analogy-Based Machine Learning », in : *Proc. of the 7th International Conference on Terminology and Knowledge Engineering, TKE'05*, Copenhaguen, Denmark.

Claveau, Vincent et Sébastien Lefèvre (2015), « Topic segmentation of TV-streams by watershed transform and vectorization », in : *Computer Speech and Language* 29.*1*, p. 63-80, doi : 10.1016/j.csl.2014.04.006, url : https://hal.archives-ouvertes.fr/hal-00998259.

Claveau, Vincent et Abir Ncibi (juin 2013), « Découverte de connaissances dans les séquences par CRF non-supervisés », in : *20ème conférence sur le Traitement Automatique des Langues Naturelles, TALN*, t. 1, Sables d'Olonne, France, volume 1, url : https://hal.archives-ouvertes.fr/hal-00912314.

— (avr. 2014), « Knowledge discovery with CRF-based clustering of named entities without a priori classes », in : *Conference on Intelligent Text Processing and Computational Linguistics CICLing*, sous la dir. d'Alexander Gelbukh, t. 8403, LNCS 1-2, Kathmandu, Nepal : Springer, p. 415-428, url : https://hal.archives-ouvertes.fr/hal-01027520.

Claveau, Vincent et Jian-Yun Nie (août 2016), *Recherche d'information et traitement automatique des langues*, t. 56, Traitement Automatique des Langues, TAL 3, ATALA, url : https://hal.archives-ouvertes.fr/hal-01394788.

Claveau, Vincent et Christian Raymond (juin 2017), « IRISA at DeFT2017 : classification systems of increasing complexity », in : *DeFT 2017 - Défi Fouille de texte*, Actes de l'Atelier Défi Fouille de Texte, DeFT, Orléans, France, p. 1-10, url : https://hal.archives-ouvertes.fr/hal-01643993.

Claveau, Vincent, Romain Tavenard et Laurent Amsaleg (mar. 2010), « Vectorisation des processus d'appariement document-requête », in : *7e conférence en recherche d'informations et applications, CORIA'10*, Sousse, Tunisie, p. 313-324.

Dalloux, Clément, Vincent Claveau et Natalia Grabar (sept. 2019), « Speculation and negation detection in french biomedical corpora », in : *RANLP 2019 - Recent Advances in Natural Language Processing*, Varna, Bulgaria, p. 1-10, url : https://hal.archives-ouvertes.fr/hal-02284444.

Deerwester, S. et al. (1990), « Indexing by Latent Semantic Analysis », in : *Journal of the American Society for Information Science*.

Deléger, Louise, Fiammetta Namer et Pierre Zweigenbaum (2008), « Morphosemantic parsing of medical compound words : Transferring a French analyzer to English. », in : *International Journal of Medical Informatics* 78.*Supplement 1*, p. 48-55.

Detyniecki, Marcin (2000), « Mathematical aggregation operators and their application to video querying », thèse de doct., Université de Paris 6.

Domengès, Dominique et Michel Volle (1979), « Analyse factorielle sphérique : une exploration », in : *Annales de l'INSEE* 35, p. 3-83.

Ebadat, Ali-Reza, Vincent Claveau et Pascale Sébillot (mar. 2012a), « Semantic Clustering using Bag-of-Bag-of-Features », in : *CORIA - COnférence en Recherche d'Information et Applications*, Bordeaux, France, p. 229-244, url : https://hal.archives-ouvertes.fr/hal-00753912.

— (2012b), « Semantic Clustering using Bag-of-Bag-of-Features », in : *Actes de la 9e conférence en recherche d'information et applications, CORIA 2012*, Bordeaux, France, p. 229-244, url : http://hal.archives-ouvertes.fr/hal-00753912.

Escoffier, Bernard (1978), « Analyse factorielle et distances répondant au principe d'équivalence distributionnelle », in : *Revue de statistique appliquée* 26.*4*, p. 29-37.

Ferret, Olivier (2013), « Identifying Bad Semantic Neighbors for Improving Distributional Thesauri », in : *51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, Sofia, Bulgaria, p. 561-571.

— (2014), « Typing relations in distributional thesauri », in : *Advances in Language Production, Cognition and the Lexicon*, sous la dir. de N. Gala, R. Rapp et G. Bel, Springer.

Firth, John R. (1957), « Studies in Linguistic Analysis », in : Oxford : Blackwell, chap. A synopsis of linguistic theory 1930-1955, p. 1-32.

Fort, Karen et Vincent Claveau (mai 2012), « Annotating Football Matches : Influence of the Source Medium on Manual Annotation », in : *LREC - Eight International Conference on Language Resources and Evaluation*, Istanbul, Turkey, url : https://hal.archives-ouvertes.fr/hal-00709170.

Gabrilovich, Evgeniy et Shaul Markovitch (2007), « Computing semantic relatedness using wikipedia-based explicit semantic analysis », in : *20th International Joint Conference on Artificial Intelligence (IJCAI 2007)*, p. 6-12.

Gao, J., J.Y. Nie et Ming Zhou (2006), « Statistical query translation models for cross-language information retrieval », in : *ACM Transactions on Asian Information Processing (TALIP)* 5.*4*, p. 296-322.

Gao, Jianfeng, Xiaodong He et Jian-Yun Nie (2010), « Clickthrough-Based Translation Models for Web Search : from Word Models to Phrase Models », in : *Proceedings of the CIKM conference*, sous la dir. d'editor, p. 1139-1148.

Gao, Jianfeng, Jian-Yun Nie et al. (2004), « Dependence Language Model for Information Retrieval », in : *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, Sheffield, United Kingdom : ACM, p. 170-177, isbn : 1-58113-881-4, doi : 10.1145/1008992.1009024, url : http://doi.acm.org/10.1145/1008992.1009024.

Gaussier, Eric (1999), « Unsupervised Learning of Derivational Morphology from Inflectional Corpora », in : *Proceedings of Workshop on Unsupervised Methods in Natural Language Learning, 37th Annual Meeting of the Association for Computational Linguistics, ACL 99*, Maryland, États-Unis, p. 24-30.

Goodfellow, Ian J. et al. (2014), « Generative Adversarial Nets », in : *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, Montreal, Canada : MIT Press, p. 2672-2680, url : http://dl.acm.org/citation.cfm?id=2969033.2969125.

Gosselin, P.H., M. Cord et S. Philipp-Foliguet (oct. 2007), « Kernels on Bags of Fuzzy Regions for Fast Object retrieval », in : *IEEE International Conference on image processing, ICIP 2007*, t. 1, p. 177-180.

Grabar, Natalia, Vincent Claveau et Clément Dalloux (oct. 2018), « CAS : French Corpus with Clinical Cases », in : *LOUHI 2018 - The Ninth International Workshop on Health Text Mining and Information Analysis*, Ninth International Workshop on Health Text Mining and Information Analysis (LOUHI) Proceedings of the Workshop, Bruxelles, France, p. 1-7, url : https://hal.archives-ouvertes.fr/hal-01937096.

Grabar, Natalia, Cyril Grouin et al. (juil. 2019), « Information Retrieval and Information Extraction from Clinical Cases. Presentation of the DEFT 2019 Challenge », in : *Atelier -compétition Défi fouille de texte*, Toulouse, France, url : https://hal.archives-ouvertes.fr/hal-02280852.

Grabar, Natalia et Pierre Zweigenbaum (2002), « Lexically-based terminology structuring : Some inherent limits », in : *Proc. of International Workshop on Computational Terminology, COMPUTERM*, Taipei, Taiwan.

Grau, Brigitte, Anne-Laure Ligozat et Martin Gleize (2015), « Recherche d'information précise : vers des modèles hybrides exploitant des sources d'information structurées et non structurées », in : *Traitement Automatique des Langues* 56.*3*, p. 75-99.

Grefenstette, Gregory (1994), *Explorations in automatic thesaurus discovery*, Kluwer Academic Publishers.

Grefenstette, Gregory, Nasredine Semmar et Faïza Elkateb-Gara (2005), « Modifying a Natural Language Processing System for European Languages to Treat Arabic in Information Processing and Information Retrieval Applications », in : *Computational Approaches to Semitic Languages - Workshop Proceedings*, University of Michigan, p. 31-38.

Grouin, Cyril et al. (2011), « Présentation et résultats du défi fouille de texte DEFT2011 », in : *Acte de l'atelier DeFT, associé à la conférence TALN*.

Haddad, Hatem et Chedi Bechikh Ali (2014), « Performance of Turkish Information Retrieval : Evaluating the Impact of Linguistic Parameters and Compound Nouns », in : *Proceedings of the 15th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 8404*, CICLing 2014, Kathmandu, Nepal : Springer-Verlag New York, Inc., p. 381-391, isbn : 978-3-642-54902-1, doi : 10.1007/978-3-642-54903-8_32, url : http://dx.doi.org/10.1007/978-3-642-54903-8_32.

Hagiwara, Masato, Yasuhiro Ogawa et Katsuhiko Toyama (2006), « Selection of Effective Contextual Information for Automatic Synonym Acquisition », in : *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, Sydney, Australia, p. 353-360.

Hasan, Sadid A. et al. (sept. 2018), « Overview of the ImageCLEF 2018 Medical Domain Visual Question Answering Task », in : *CLEF2018 Working Notes*, CEUR Workshop Proceedings, Avignon, France : CEUR-WS.org <http ://ceur-ws.org>.

Hathout, Nabil (2001), « Analogies morpho-synonimiques. Une méthode d'acquisition automatique de liens morphologiques à partir d'un dictionnaire de synonymes », in : *Actes de la 8^{e} conférence Traitement Automatique du Langage Naturel, TALN'01*, Tours, France.

Haussler, D. (1999), *Convolution kernels on discrete structures*, rapp. tech. UCSC-CRL-99-10, University of California at Santa-Cruz.

Hearst, Marti (1997), « Text-tiling : segmenting text into multi-paragraph subtopic passages », in : *Computational Linguistics* 23.*1*, p. 33-64.

Heineman, George T., Gary Pollice et Stanley Selkow (2008), « Algorithms in a Nutshell », in : Oreilly Media, chap. Chapter 8 :Network Flow Algorithms.

Hochreiter, Sepp et Jürgen Schmidhuber (nov. 1997), « Long Short-Term Memory », in : *Neural Comput.* 9.*8*, p. 1735-1780, issn : 0899-7667, doi : 10.1162/neco.1997.9.8.1735, url : http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Hoffman, Matthew, Francis R. Bach et David M. Blei (2010), « Online Learning for Latent Dirichlet Allocation », in : *Advances in Neural Information Processing Systems 23*, sous la dir. de J.D. Lafferty et al., Curran Associates, Inc., p. 856-864, url : http://papers.nips.cc/paper/3902-online-learning-for-latent-dirichlet-allocation.pdf.

Hofmann, T. (1999), « Probabilistic latent semantic indexing », in : *Proc. of SIGIR*, Berkeley, USA.

Houle, Michael E., Hisashi Kashima et Michael Nett (2012), « Generalized expansion dimension », in : *Proc. of the 12th IEEE International Conference on Data Mining Workshops (ICDMW)*, p. 587-594.

Huang, Eric H. et al. (2012), « Improving word representations via global context and multiple word prototypes », in : *50th Annual Meeting of the Association for Computational Linguistics (ACL'12)*, Jeju Island, Korea, p. 873-882.

Hull, David (1993), « Using Statistical Testing in the Evaluation of Retrieval Experiments », in : *Proceedings of the 16th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'93*, Pittsburgh, États-Unis.

Jacquemin, Christian, éd. (2000), *Traitement automatique des langues pour la recherche d'information*, t. 41-2.

Jiampojamarn, Sittichai, Grzegorz Kondrak et Tarek Sherif (2007), « Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion », in : *Proc. of the conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, New York, USA.

Knight, Kevin et Jonathan Graehl (1998), « Machine Transliteration », in : *Computational Linguistics* 24.*4*, p. 599-612.

Kraaij, W., J.-Y. Nie et M. Simard (2003), « Embedding Web-based statistical translation models in cross-language information retrieval », in : *Computational Linguistics* 29.*3*, p. 381-419.

Kundi, Fazal Masud et al. (2014), *A Review of Text Summarization*, rapp. tech. 4, MAGNT Research Report (ISSN. 1444-8939), p. 309-317.

Kurimo, Mikko, Mathias Creutz et Ville T. Turunen (2009), « Morpho challenge evaluation by information retrieval experiments », in : *Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access*, CLEF'08, Aarhus, Denmark : Springer-Verlag, p. 991-998, url : http://dl.acm.org/citation.cfm?id=1813809.1813961.

Kurimo, Mikko, Sami Virpioja et Ville T. Turunen, éd. (2010), *Proceedings of the MorphoChallenge 2010*, Espoo, Finlande.

Landauer, Thomas K. et Susan T. Dumais (1997), « A solution to Plato's problem : the latent semantic analysis theory of acquisition, induction, and representation of knowledge », in : *Psychological review* 104.*2*, p. 211-240.

Landauer, Thomas et Susan Dumais (1997), « A solution to Plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge », in : *Psychological Review* 104.*2*, p. 211-240.

Langlais, Philippe (2013), « Mapping Source to Target Strings without Alignment by Analogical Learning : A Case Study with Transliteration », in : *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, Sofia, Bulgaria, p. 684-689.

Langlais, Philippe, François Yvon et Pierre Zweigenbaum (avr. 2009), « Improvements in Analogical Learning : Application to Translating multi-Terms of the Medical Domain », in : *12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*, Athens, Greece, p. 487-495.

Le, Quoc V. et Tomas Mikolov (2014), « Distributed Representations of Sentences and Documents », in : *CoRR* abs/1405.4053, url : http://arxiv.org/abs/1405.4053.

Lepage, Yves (2003), « De l'analogie ; rendant compte de la communication en linguistique », Thèse d'habilitation (HDR), Grenoble, France : Université de Grenoble 1.

Levow, Gina-Anne, Douglas W. Oard et Philip Resnik (2005), « Dictionary-based techniques for cross-language information retrieval », in : *Information Processing & Management* 41 (3), p. 523-547.

LIN, Dekang (1998), « Automatic retrieval and clustering of similar words », in : *17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (ACL-COLING'98)*, Montréal, Canada, p. 768-774.

LOVINS, Julie Beth (1968), « Development of a Stemming Algorithm », in : *Mechanical Translation and Computational Linguistics* 1, p. 22-31.

MAIGROT, Cédric (juil. 2019), « Détection de fausses informations dans les réseaux sociaux », Theses, Université Rennes 1.

MAIGROT, Cédric et al. (oct. 2016), « MediaEval 2016 : A multimodal system for the Verifying Multimedia Use task », in : *MediaEval 2016 : "Verfiying Multimedia Use" task*, Hilversum, Netherlands, DOI : 10.1145/1235, URL : https://hal.archives-ouvertes.fr/hal-01394785.

MAISONNASSE, Loïc, Éric GAUSSIER et Jean-Pierre CHEVALLET (2008), « Modélisation de relations dans l'approche modèle de langue en recherche d'information », in : *Actes de la COnférence en Recherche d'Infomations et Applications - CORIA 2008*, Trégastel, France, March 12-14, 2008, p. 305-319, URL : http://asso-aria.org/coria/2008/305.pdf.

MANNING, Christopher D., Prabhakar RAGHAVAN et Hinrich SCHÜTZE (2008), *Introduction to Information Retrieval*, Cambridge University Press.

MARKÓ, Kornél, Stefan SCHULZ et Udo HAN (2005), « Morphosaurus - design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain », in : *Methods of Information in Medicine* 44.*4*.

MARKÓ, Kornél, Stefan SCHULZ, Olena MEDELYAN et al. (2005), « Bootstrapping Dictionaries for Cross-Language Information Retrieval », in : *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*, Salvador, Brésil.

MCCARTHY, D. et R. NAVIGLI (2009), « The English lexical substitution task », in : *Language Resources and Evaluation* 43.*2*, p. 139-159.

MEL'ČUK, Igor (1996), « Lexical Functions : A Tool for the Description of Lexical Relations in a Lexicon », in : *Lexical Functions in Lexicography and Natural Language Processing*, sous la dir. de Leo WANNER, Amsterdam/Philadelphie : Benjamins, p. 37-102.

— (2000), *Cours de morphologie générale, 1993-2000*, t. 1-5, Montréal/Paris : Presses de l'Université de Montréal/CNRS Éditions.

Mel'čuk, Igor (mar. 2006), *Aspects of the Theory of Morphology*, Trends in Linguistics. Studies and Monographs, Mouton de Gruyter, Berlin.

Metzler, D. et W.B. Croft (2004), « Combining the Language Model and Inference Network Approaches to Retrieval », in : *Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval* 40.*5*, p. 735-750.

Mikolov, Tomas, Wen-tau Yih et Geoffrey Zweig (2013), « Linguistic Regularities in Continuous Space Word Representations », in : *2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL HLT 2013)*, Atlanta, Georgia, p. 746-751.

Miller, George A. (1990), « WordNet : An On-Line Lexical Database », in : *International Journal of Lexicography* 3.*4*.

Mitra, Bhaskar et Nick Craswell (2017), « Neural Models for Information Retrieval », in : *CoRR* abs/1705.01509, arXiv : 1705.01509, url : http://arxiv.org/abs/1705.01509.

— (déc. 2018), « An Introduction to Neural Information Retrieval », in : *Foundations and Trends® in Information Retrieval* 13.*1*, p. 1-126, url : https://www.microsoft.com/en-us/research/publication/introduction-neural-information-retrieval/.

Moreau, Fabienne et Vincent Claveau (2006), « Extension de requêtes par relations morphologiques acquises automatiquement », in : *Revue I3 (Information - Interaction - Intelligence)* 6.*2*, p. 31-50.

Moreau, Fabienne, Vincent Claveau et Pascale Sébillot (avr. 2007), « Automatic morphological query expansion using analogy-based machine learning », in : *Proceedings of the European Conference on Information Retrieval, '07*, Rome, Italie.

Moreau, Fabienne et Pascale Sébillot (2005a), *Contributions des techniques du traitement automatique des langues à la recherche d'information*, Rapport de recherche 1690, IRISA.

— (2005b), *Contributions des techniques du traitement automatique des langues à la recherche d'information*, rapp. tech. 1690, IRISA, url : http://www.irisa.fr/bibli/publi/pi/2005/1690/1690.html.

Morin, Emmanuel et Béatrice Daille (2010), « Compositionality and lexical alignment of multi-word terms », in : *Language Resources and Evaluation (LRE)* 44.

MUGGLETON, Stephen et Luc De RAEDT (1994), « Inductive Logic Programming : Theory and Methods », in : *Journal of Logic Programming* 19/20, p. 629-679, URL : citeseer.ist.psu.edu/muggleton94inductive.html.

NÉVÉOL, Aurélie, Sonya E. SHOOSHAN et Vincent CLAVEAU (juin 2008), « Automatic Inference of Indexing Rules for MEDLINE », in : *Proceedings of the ACL 2008 Workshop BioNLP*, Columbus, OH, USA.

NICKEL, Maximillian et Douwe KIELA (2017), « Poincaré Embeddings for Learning Hierarchical Representations », in : *Advances in Neural Information Processing Systems 30*, sous la dir. d'I. GUYON et al., Curran Associates, Inc., p. 6338-6347, URL : http://papers.nips.cc/paper/7213-poincare-embeddings-for-learning-hierarchical-representations.pdf.

NIE, Jian-Yun (2010), *Cross-Language Information Retrieval*, Synthesis Lectures on Human Language Technologies series, Morgan & Claypool.

OCH, Franz Josef et Hermann NEY (2003), « A Systematic Comparison of Various Statistical Alignment Models », in : *Computational Linguistics* 29.*1*, p. 19-51.

PADÓ, Sebastian et Mirella LAPATA (2007), « Dependency-Based Construction of Semantic Space Models », in : *Computational Linguistics* 33.*2*, p. 161-199.

PAROUBEK, Patrick et al., éd. (2012), *Actes du huitième défi fouille de texte, DEFT2012, atelir conjoint à TALN*, Grenoble, France.

PENG, Fuchun et al. (2002), « Investigating the Relationship Between Word Segmentation Performance and Retrieval Performance in Chinese IR », in : *Proceedings of the 19th International Conference on Computational Linguistics, COLING*, Taipei, Taiwan, p. 1-7.

PIRKOLA, Ari et al. (2001), « Dictionary-Based Cross-Language Information Retrieval : Problems, Methods, and Research Findings », in : *Information Retrieval* 4.*3–4*, p. 209-230.

PONTE, J. M. et W. B. CROFT (1998), « A language modeling approach to information retrieval », in : *Proc. of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR '98)*, p. 275-281.

PONTE, Jay M. et W. Bruce CROFT (1998), « A Language Modeling Approach to Information Retrieval », in : *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, Melbourne, Australia, p. 275-281, ISBN : 1-58113-015-5, DOI : 10.1145/290941.291008, URL : http://doi.acm.org/10.1145/290941.291008.

Porter, Martin (1980), « An Algorithm for Suffix Stripping », in : *Program* 14.*3*, p. 130-137.

Pustejovsky, James (1995), *The Generative Lexicon*, Cambridge, MA : MIT Press.

Raymond, Christian et Vincent Claveau (2011), « Participation de l'IRISA à DEFT 2011 : expériences avec des approches d'apprentissage supervisé et non-supervisé », in : *Challenge DeFT (défi fouille de texte)*, France, ?, url : https://hal.archives-ouvertes.fr/hal-00643724.

Řehůřek, Radim et Petr Sojka (mai 2010), « Software Framework for Topic Modelling with Large Corpora », English, in : *Proc. of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, http://is.muni.cz/publication/884893/en, Valletta, Malta : ELRA, p. 45-50.

Robertson, Stephen E., Steve Walker et Micheline Hancock-Beaulieu (1998), « Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive », in : *Proceedings of the 7th Text Retrieval Conference, TREC-7*, p. 199-210.

Ruiz-Casado, M, E. Alfonseca et P. Castells (2005), « Using context-window overlapping in Synonym Discovery and Ontology Extension », in : *Proc. of RANLP*, Borovets, Bulgarie.

Sahami, M. et T.D. Heilman (2006), « A web-based kernel function for measuring the similarity of short text snippets », in : *Proc. of the conference WWW'2006*.

Sahlgren, Magnus (2001), « Vector-Based Semantic Analysis : Representing Word Meanings Based on Random Labels », in : *ESSLLI 2001 Workshop on Semantic Knowledge Acquisition and Categorisation*, Helsinki, Finland.

Salton, G. et M. McGill (1983), *Introduction to modern information retrieval*, McGraw-Hill.

Savoy, Jacques (1993), « Stemming of French Words Based on Grammatical Categories », in : *Journal of the American Society for Information Science (JASIS)* 44.*1*, p. 1-9.

— (2002), *Morphologie et Recherche d'Information*, Rapport technique, Institut interfacultaire d'informatique, Université de Neuchâtel.

Shen, Wei et Jian-Yun Nie (sept. 2015), « Is Concept Mapping Useful for Biomedical Information Retrieval ? », in : *Proceedings of Experimental IR Meets Multilinguality, Multimodality, and Interaction : 6th International Conference of the CLEF Association, CLEF'15*, Toulouse, France : Springer International Publishing, p. 281-

286, ISBN : 978-3-319-24027-5, DOI : 10.1007/978-3-319-24027-5_29, URL : http://dx.doi.org/10.1007/978-3-319-24027-5_29.

SOILLE, P. (2003), *Morphological Image Analysis : Principles and Applications*, Berlin : Springer-Verlag.

SPÄRCK-JONES, Karen (1999), « What is the Role of NLP in Text Retrieval ? », in : *Natural Language Information Retrieval*, sous la dir. de T. STRZALKOWSKI, Kluwer Academic Publishers, p. 1-24.

STEINBERGER, Josef, Mijail A. KABADJOV et Massimo POESIO (2005), « Improving LSA-based summarization with anaphora resolution », in : *In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, p. 1-8.

STROHMAN, T. et al. (2005), *Indri : A language-model based search engine for complex queries (extended version)*, rapp. tech., CIIR.

STRZALKOWSKI, T. et al. (1999), « Evaluating Natural Language Processing Techniques in Information Retrieval », in : *Natural Language Information Retrieval*, sous la dir. de T. STRZALKOWSKI, Kluwer Academic Publishers, p. 113-145.

TIRILLY, Pierre (juil. 2010), « Natural language processing for image indexing », Theses, Université Rennes 1, URL : https://tel.archives-ouvertes.fr/tel-00516422.

TSUJI, Keita, Béatrice DAILLE et Kyo KAGEURA (2002), « Extracting French-Japanese Word Pairs from Bilingual Corpora based on Transliteration Rules », in : *Proc. of the 3rd International Conference on Language Resources and Evaluation, LREC'02*, Las Palmas de Gran Canaria, Spain.

TURNEY, P. et P. PANTEL (2010), « From frequency to meaning : Vector space models of semantics », in : *Journal of Artificial Intelligence Research* 37.*1*, p. 141-188.

TURNEY, P.D. (2001), « Mining the Web for Synonyms : PMIIR versus LSA on TOEFL », in : *Lecture Notes in Computer Science* 2167, p. 491-502.

TURTLE, H. et W.B. CROFT (1991), « Evaluation of an Inference Network-Based Retrieval Model », in : *ACM Transactions on Information System* 9.*3*, p. 187-222.

TUTTLE, Mark et al. (1990), « Using Meta-1 – the 1st Version of the UMLS Metathesaurus », in : *Proc. of the 14th annual Symposium on Computer Applications in Medical Care (SCAMC)*, Washington, USA, p. 131-135.

VAN DE CRUYS, T., T. POIBEAU et A. KORHONEN (2011), « Latent vector weighting for word meaning in context », in : *Proc. of the Conference on Empirical Methods in Natu-*

ral Language Processing, sous la dir. d'Association for Computational Linguistics, p. 1012-1022.

Van de Cruys, Tim (2010), « Mining for Meaning. The Extraction of Lexico-semantic Knowledge from Text », thèse de doct., The Netherlands : University of Groningen.

Vechtomova, Olga et Stephen E. Robertson (2012), « A Domain-Independent Approach to Finding Related Entities », in : *Information Processing and Management* 48.*4*, p. 654-670.

Vicedo, Jose L. et Antonio Ferrandez (2000), « Importance of Pronominal Anaphora resolution in Question Answering systems », in : *In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, ACL*, p. 555-562.

Vincent, L. et P. Soille (1991), « Watersheds in Digital Spaces : An Efficient Algorithm Based on Immersion Simulations », in : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13.*6*, p. 583-598.

Voorhees, Ellen M. (1994), « Query Expansion Using Lexical-semantic Relations », in : *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, Dublin, Ireland : Springer-Verlag New York, Inc., p. 61-69, isbn : 0-387-19889-X, url : http://dl.acm.org/citation.cfm?id=188490.188508.

Vukotić, Vedran, Vincent Claveau et Christian Raymond (juin 2015), « IRISA at DeFT 2015 : Supervised and Unsupervised Methods in Sentiment Analysis », in : *DeFT, Défi Fouille de Texte, joint à la conférence TALN 2015*, Actes de l'atelier DeFT, Défi Fouille de Texte, joint à la conférence TALN 2015, Caen, France, url : https://hal.archives-ouvertes.fr/hal-01226528.

Vukotic, Verdran (2017), « Deep Neural Architectures for Automatic Representation Learning from Multimedia Multimodal Data », 2017ISAR0015, thèse de doct., url : http://www.theses.fr/2017ISAR0015/document.

Ward, Grady (1996), *Moby Thesaurus*, Moby Project.

Yamamoto, Kazuhide et Takeshi Asakura (2010), « Even Unassociated Features Can Improve Lexical Distributional Similarity », in : *Second Workshop on NLP Challenges in the Information Explosion Era (NLPIX 2010)*, Beijing, China, p. 32-39.

Yang, Grace Hui et Sicong Zhang (2017), « Differential Privacy for Information Retrieval », in : *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '17, Amsterdam, The Netherlands : ACM, p. 325-326,

ISBN : 978-1-4503-4490-6, DOI : 10.1145/3121050.3121107, URL : http://doi.acm.org/10.1145/3121050.3121107.

ZARGAYOUNA, Haïfa, Catherine ROUSSEY et Jean-Pierre CHEVALLET (2015), « Recherche d'information sémantique : état des lieux », in : *Traitement Automatique des Langues* 56.*3*, p. 49-73.

ZHITOMIRSKY-GEFFET, Maayan et Ido DAGAN (2009), « Bootstrapping Distributional Feature Vector Quality », in : *Computational Linguistics* 35.*3*, p. 435-461.

ZHONG, Zhi et Hwee Tou NG (2012), « Word Sense Disambiguation Improves Information Retrieval », in : *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, ACL*, ACL '12, Jeju Island, Korea : Association for Computational Linguistics, p. 273-282, URL : http://dl.acm.org/citation.cfm?id=2390524.2390563.

UMR IRISA

UNIVERSITÉ DE RENNES 1

Titre : Du traitement des langues en recherche d'information et vice versa

Mot clés : Traitement automatique des langues, recherche d'information, intelligence artificielle

Résumé : La recherche d'information (RI) et le traitement automatique des langues (TAL) sont deux domaines de recherche de l'informatique partageant en commun leur matériau premier : la langue. Pourtant, ces deux domaines ont longtemps évolué indépendamment, avec peu d'interactions. Au travers d'une sélection de nos travaux passés, nous montrons pourtant tous les bénéfices à croiser les connaissances acquises dans chacun de ces domaines. Précisément, nous avons articulé ce mémoire en deux parties, l'une dédiée aux apports du TAL pour la RI, et l'autre aux apports de la RI pour le TAL. Nous revisitons ainsi plusieurs de nos contributions sur, d'une part, la morphologie, la translittération, la segmentation thématique, l'analyse fine de termes médicaux dans un contexte de RI, et d'autre part, sur l'utilisation des moteurs de recherche comme classifieurs, les tâches de RI comme techniques d'évaluation de techniques de TAL, la sémantique distributionnelle et les plongements de mots par et pour la RI. Nous discutons également de la pertinence de cette dichotomie entre ces deux domaines à l'heure de l'intelligence artificielle, et de la convergence de leur corpus technique (notamment les approches neuronales) ; nous présentons enfin quelques enjeux de recherche à la croisée de ces domaines.

Title: About Natural Language Processing for Information Retrieval and vice versa

Keywords: Natural Language Processing, Information Retrieval, Artificial Intelligence

Abstract: Information retrieval (IR) and Natural Language Processing (ALP) are two areas of computer science research that share their common raw material: language. However, these two fields have long evolved independently, with little interaction. Through a selection of our past contributions, we show however all the benefits of crossing the knowledge acquired in each of these fields. Precisely, we have divided this thesis into two parts, one dedicated to the contributions of NLP to IR tasks, and the other to the contributions of IR techniques to NLP. We thus revisit several of our results on, on the one hand, morphology, transliteration, thematic segmentation, fine analysis of medical terms in an IR context, and, on the other hand, on the use of search engines as classifiers, IR tasks as evaluation benchmarks for NLP techniques, distributional semantics and embedding by and for IR. We also discuss the relevance of this dichotomy between these two domains in the age of artificial intelligence and the convergence of techniques (especially neural approaches) ; finally, we present some research issues at the crossroads of these domains.