Thèse de Doctorat

Adrien BOUGOUIN

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Nantes
sous le label de l'Université de Nantes Angers Le Mans

École doctorale : Sciences et technologies de l'information, et mathématiques

Discipline : Informatique, section CNU 27
Spécialité : Traitement automatique du langage naturel
Unité de recherche : Laboratoire d'informatique de Nantes-Atlantique (LINA)

Soutenue le 27 octobre 2015

Indexation automatique par termes-clés en domaines de spécialité

JURY

Président :	**M. Marc GELGON**, Professeur des universités, Université de Nantes
Rapporteurs :	**M^me^ Brigitte GRAU**, Professeur des universités, ENSIIE
	M. Jacques SAVOY, Professeur des universités, Université de Neuchâtel
Examinateurs :	**M. Marc GELGON**, Professeur des universités, Université de Nantes
	M^me^ Fabienne MOREAU, Maître de conférences, Université de Rennes
Directrice de thèse :	**M^me^ Béatrice DAILLE**, Professeur des universités, Université de Nantes
Co-encadrant de thèse :	**M. Florian BOUDIN**, Maître de conférences, Université de Nantes

« Si à la place de l'éducation massive que nous avons (devons avoir) de nos jours avec un curriculum, dès lors que nous avons tous des ordinateurs, tous connectés à d'importantes bibliothèques où chacun peut poser n'importe quelle question et recevoir une réponse, des références, n'importe quelle chose pouvant l'intéresser (peu importe à quel point cela puisse paraître étrange pour un autre), alors chacun demanderait, et trouverait, et avancerait à son propre rythme, dans la direction qu'il souhaite suivre, quand il le souhaite, alors tout le monde aimerait apprendre. De nos jours, ce que les gens appellent apprendre est imposé et chacun est forcé d'apprendre en classe la même chose que les autres, le même jour et au même rythme. Mais tout le monde est différent. Pour certains c'est trop rapide, pour d'autres trop lent, ou encore inadapté. Donnons à chacun la chance, en plus de l'école (je ne dis pas qu'il faut abolir l'école, mais en supplément de l'école), de suivre leur propre curiosité. »

— Isaac Asimov (1988)

Table des matières

1 Introduction **11**
- 1.1 Contexte . . . 11
- 1.2 Problématique . . . 12
- 1.3 Hypothèses . . . 13
- 1.4 Mise en œuvre . . . 14
- 1.5 Plan de thèse . . . 14

2 Indexation automatique par termes-clés **15**
- 2.1 Introduction . . . 15
- 2.2 Sélection des termes-clés candidats . . . 16
- 2.3 Extraction automatique de termes-clés . . . 18
  - 2.3.1 Approche non supervisée . . . 19
  - 2.3.2 Approche supervisée . . . 27
- 2.4 Assignement automatique de termes-clés . . . 33
- 2.5 Évaluation automatique de l'indexation par termes-clés . . . 34
- 2.6 Conclusion . . . 35

3 Ressources **37**
- 3.1 Introduction . . . 37
- 3.2 Termith . . . 38
- 3.3 DEft . . . 40
- 3.4 Wikinews . . . 44
- 3.5 SemEval . . . 45
- 3.6 DUC . . . 47
- 3.7 Prétraitement des données . . . 47
- 3.8 Conclusion . . . 48

4 Extraction de termes-clés **49**
- 4.1 Introduction . . . 49
- 4.2 Sélection des termes-clés candidats . . . 49
  - 4.2.1 Analyse des propriétés linguistiques des termes-clés . . . 50
  - 4.2.2 Sélection fine des termes-clés candidats . . . 53
  - 4.2.3 Évaluation . . . 55
  - 4.2.4 Bilan . . . 58
- 4.3 Extraction non supervisée de termes-clés . . . 58
  - 4.3.1 TopicRank . . . 59

4.3.2 Évaluation . . . 62
4.3.3 Analyse d'erreurs . . . 69
4.3.4 Bilan . . . 70
4.4 Conclusion . . . 71

5 Indexation par termes-clés en domaines de spécialité 73
5.1 Introduction . . . 73
5.2 Indexation manuelle en domaines de spécialité . . . 74
5.2.1 Principes généraux . . . 74
5.2.2 Ressources . . . 74
5.2.3 Méthodologie . . . 76
5.2.4 Bilan . . . 76
5.3 Indexation automatique en domaines de spécialité . . . 76
5.3.1 TopicCoRank . . . 76
5.3.2 Évaluation . . . 79
5.3.3 Analyse des sorties de TopicCoRank . . . 88
5.3.4 Bilan . . . 90
5.4 Évaluation manuelle en domaines de spécialité . . . 90
5.4.1 Protocole d'évaluation manuelle . . . 90
5.4.2 Évaluation manuelle des méthodes proposées . . . 92
5.4.3 Bilan . . . 94
5.5 Conclusion . . . 95

6 Conclusion et perspectives 97
6.1 Contributions . . . 98
6.2 Perspectives . . . 99

Listes des publications 103
Publication en revue . . . 103
Publication en conférence internationale avec actes . . . 103
Publications en conférence national avec actes . . . 104

Bibliographie 105

Liste des tableaux

2.1 Comparaison des méthodes d'extraction non supervisée de termes-clés de la littérature, lorsque dix termes-clés sont extraits 20
2.2 Matrice de confusion pour l'évaluation des méthodes d'indexation automatique par termes-clés 35

3.1 Détail des revues du corpus de linguistique (Termith) 39
3.2 Détail des revues du corpus de sciences de l'information (Termith) 39
3.3 Détail des revues du corpus d'archéologie (Termith) 39
3.4 Détail des revues du corpus de chimie (Termith) 40
3.5 Corpus Termith 40
3.6 Corpus DEft 42
3.7 Résultats de tests humains sur le corpus DEft 44
3.8 Corpus Wikinews 44
3.9 Accord inter-annotateur κ (Fleiss, 1971) sur le corpus Wikinews 45
3.10 Corpus SemEval 45
3.11 Nombre de termes-clés attribués dans SemEval, en fonction des annotateurs 46
3.12 Corpus DUC 47
3.13 Bilan des corpus 48

4.1 Statistiques des termes-clés de référence des collections DEft, SemEval et DUC 51
4.2 Taux d'adjectifs, par catégorie (relationnel, composé complexe ou qualificatif), au sein des termes-clés de référence 53
4.3 Taux d'adjectifs, par catégorie (relationnel, composé complexe ou qualificatif), au sein des documents 53
4.4 Patrons grammaticaux les plus fréquents parmi les termes-clés français et anglais 54
4.5 Résultats de l'évaluation intrinsèque des méthodes de sélection de termes-clés candidats appliquées aux données Termith 57
4.6 Résultats de l'évaluation intrinsèque des méthodes de sélection de termes-clés candidats appliquées aux collections DEft, SemEval et DUC 57
4.7 Résultats de l'extraction de dix termes-clés avec TF-IDF et KEA sur les données Termith, selon la méthode de sélection des termes-clés candidats utilisée 58

4.8 Résultats de l'extraction de dix termes-clés avec TF-IDF et KEA sur DEft, SemEval et DUC, selon la méthode de sélection des termes-clés candidats utilisée . . . 58
4.9 Résultats de l'extraction de dix termes-clés avec TF-IDF, TextRank, SingleRank et TopicRank sur les données Termith . . . 68
4.10 Résultats de l'extraction de dix termes-clés avec TF-IDF, TextRank, SingleRank et TopicRank sur les collections DEft, Wikinews, SemEval et DUC . . 68
4.11 Résultats de l'extraction de dix termes-clés avec chacune des contributions de TopicRank, appliquées séparément à SingleRank sur les données Termith 69
4.12 Résultats de l'extraction de dix termes-clés avec chacune des contributions de TopicRank, appliquées séparément à SingleRank sur les collections DEft, Wikinews, SemEval et DUC . . . 69
4.13 Résultat de TopicRank sur les données Termith, selon la méthode de sélection des termes-clés candidats utilisée . . . 70
4.14 Résultat de TopicRank sur DEft, SemEval et DUC, selon la méthode de sélection des termes-clés candidats utilisée . . . 70

5.1 Exemple de remplissage de la grille d'indexation de linguistique . . . 75
5.2 Résultat de l'extraction de dix termes-clés avec TF-IDF, TopicRank, KEA++, TopicCoRank$_{extr.}$, TopicCoRank$_{assign.}$ et TopicCoRank appliqués aux collections Termith . . . 82
5.3 Taux moyens d'extraction et d'assignement réalisés par TopicCoRank sur les données Termith . . . 84
5.4 Résultat de l'extraction de dix termes-clés avec TF-IDF, TopicRank, KEA++, TopicCoRank$_{extr.}$, TopicCoRank$_{assign.}$ et TopicCoRank appliqués à DEft, SemEval et DUC . . . 86
5.5 Taux moyens d'extraction et d'assignement réalisés par TopicCoRank sur DEft, SemEval et DUC . . . 86
5.6 Taux de termes-clés avec un score de 0, de 1 ou de 2 pour l'évaluation de la pertinence de TF-IDF et de TopicRank . . . 93
5.7 Performances de TF-IDF et de TopicRank en termes de précision, de rappel et de f1-mesure . . . 93
5.8 Taux de termes-clés de référence avec un score de 0, de 1 ou de 2 pour l'évaluation du silence de TF-IDF et de TopicRank . . . 94
5.9 Taux de termes-clés avec un score de 0, de 1 ou de 2 pour l'évaluation de la pertinence de KEA et de TopicCoRank . . . 94
5.10 Taux de termes-clés de référence avec un score de 0, de 1 ou de 2 pour l'évaluation du silence de KEA et de TopicCoRank . . . 95

Table des figures

2.1 Exemple d'indexation par termes-clés d'une notice bibliographique (résumé) . . . 16
2.2 Chaîne de traitement classique en extraction de termes-clés . . . 19
2.3 Illustration des deux propriétés d'informativité et de grammaticalité induites entre trois modèles de langues (Tomokiyo et Hurst, 2003) . . . 22
2.4 Illustration du fonctionnement de TopicalPageRank (Liu *et al.*, 2010) . . . 26
2.5 Exemple de chaîne lexicale (Ercan et Cicekli, 2007) . . . 31

3.1 Exemple de notices Termith . . . 41
3.2 Exemple de document de DEft . . . 43
3.3 Autre exemple de document de DEft . . . 43
3.4 Exemple de document de Wikinews . . . 44
3.5 Exemple de document de SemEval . . . 46
3.6 Exemple de document de DUC . . . 48

4.1 Exemple d'extraction de termes-clés avec TopicRank, comparé à TF-IDF, TextRank et SingleRank, sur un article journalistique de Wikinews . . . 63
4.2 Résultats de l'extraction de dix termes-clés avec TopicRank, en fonction de la stratégie de regroupement et de la valeur du seuil de similarité ζ . . . 65
4.3 Résultats de l'extraction de dix termes-clés, avec TopicRank, en fonction des différentes stratégies de sélections d'un terme-clé candidats par sujet . 66
4.4 Résultats de l'extraction de dix termes-clés, avec SingleRank, en fonction de la fenêtre de cooccurrences . . . 67

5.1 Illustration du graphe unifié utilisé par TopicCoRank . . . 77
5.2 Exemple d'extraction de termes-clés avec TopicCoRank. . . . 80
5.3 Courbes de rappel-précision de TF-IDF, TopicRank KEA++, TopicCoRank$_{\text{extr.}}$, TopicCoRank$_{\text{assign.}}$ et TopicCoRank appliqués aux données Termith . . . 83
5.4 Performance de TopicCoRank appliqué aux données Termith, lorsque le taux d'assignement varie . . . 84
5.5 Performance de TopicCoRank appliqué aux données Termith, lorsque le paramètre λ varie . . . 85
5.6 Courbes de rappel-précision de TF-IDF, TopicRank, TopicCoRank$_{\text{extr.}}$, TopicCoRank$_{\text{assign.}}$ et TopicCoRank appliqués à DEft, SemEval et DUC . . . 87
5.7 Performance de TopicCoRank, appliqué à DEft, SemEval et DUC, lorsque le taux d'assignement varie . . . 88

5.8 Performance de TopicCoRank, appliqué à DEft, SemEval et DUC, lorsque le paramètre λ varie 89
5.9 Interface d'évaluation manuelle de l'Inist 92

1

Introduction

> « Rechercher des informations est une activité fréquente pour quiconque utilise quotidiennement un ordinateur. Alors qu'Internet est une source abondante d'information de tout genre, trouver les documents pertinents est encore difficile. [...] Les termes-clés aident à organiser et retrouver ces documents d'après leur contenu. »
>
> — Medelyan et Witten (2008)

1.1 Contexte

La société contemporaine dans laquelle nous vivons se situe en pleine ère de l'information. Cette ère succède l'ère moderne, durant laquelle de nombreuses découvertes et avancées scientifiques ont été faites ; durant laquelle des connaissances considérables ont été acquises. Elle est aussi marquée par le début de la mondialisation, qui, sur le plan scientifique, favorise la dissémination et la production de nouvelles connaissances. Jusqu'alors rangées sous la forme de documents papiers dans des bibliothèques, où des documentalistes les indexent et aident ensuite scientifiques et particuliers à y accéder le plus efficacement possible, les connaissances sont devenues trop nombreuses et leur stockage physique inadapté (Rider, 1946). L'ère de l'information débute vers la fin des années 1940 et apporte une solution à ce problème : l'informatisation des données. Cette informatisation présente tout d'abord l'avantage de pouvoir stocker les connaissances sur des supports pérennes, de capacité de plus en plus grande (de quelques mégaoctets à plusieurs gigaoctets) et de taille de plus en plus réduite (du disque dur au DVD). Très vite, la communauté scientifique y voit aussi un moyen pour améliorer la recherche d'information, en indexant les documents qui contiennent les connaissances, en proposant des interfaces pour permettre à un utilisateur de formuler une requête et en cherchant les documents pertinents vis-à-vis de cette requête (Salton *et al.*, 1975).

Produit de cette ère de l'information, le réseau informatique mondial Internet en est aussi devenu l'un des acteurs principaux. En effet, si l'informatisation et l'indexation des données facilite leur recherche, Internet facilite leur accès depuis les bases de données informatisées qui y sont connectées. Médium d'information mondial et accessible de tous[1], il favorise donc la transition depuis les bibliothèques traditionnelles vers des bibliothèques numériques. Ces dernières combinent le savoir faire des documentalistes avec les techniques du Traitement automatique des langues (TAL) et de la Recherche d'information (RI) pour informatiser les données et faciliter leur accès et leur recherche.

Cette thèse s'inscrit dans le cadre du projet ANR Termith (ANR-12-CORD-0029), qui s'intéresse à l'accès à l'information numérique en domaines de spécialité et qui s'articule lui-même autour du travail de l'Institut de l'information scientifique et technique (Inist). Né en 1988 de la fusion du Centre de documentation scientifique et technique (CDST) et du Centre de documentation sciences humaines (CDSH), tout deux fondés en 1970 pendant les débuts de l'informatisation des données, l'Inist possède deux des plus importantes bases de données informatisées d'Europe : PASCAL en sciences exactes et FRANCIS en sciences humaines. Aujourd'hui acteur de la Bibliothèque scientifique numérique (BSN) fondée en 2009 par le ministère de l'enseignement supérieur et de la recherche français, l'une de ses missions est de faciliter l'accès à la recherche mondiale au travers de la production de notices bibliographiques[2] associées à des mots-clés d'indexation, que nous appelons ici termes-clés. Avec la croissance du nombre de productions scientifiques, l'indexation manuelle par termes-clés est de plus en plus difficile. Cette tâche nécessite un travail de maintenance des ressources terminologiques utilisées pour l'indexation (Guinchat et Skouri, 1996) et des effectifs humains conséquents afin de tenir la charge journalière de données à indexer.

Soucieux de faciliter le travail d'indexation par termes-clés de toute sorte de document (résumé d'une notice bibliographique, article scientifique, article journalistique, nouvelle, etc.) et pour toute sorte d'application (indexation, résumé, publicité ciblée, etc.), de nombreux chercheurs s'intéressent à son automatisation. En témoignent le nombre grandissant d'articles scientifiques à ce sujet (Hasan et Ng, 2014) ainsi que l'émergence de campagnes d'évaluation (Kim *et al.*, 2010; Paroubek *et al.*, 2012).

1.2 Problématique

Étant donné un document, l'indexation automatique par termes-clés consiste à trouver les unités textuelles qui décrivent son contenu principal. La difficulté de cette tâche réside dans l'identification des éléments importants vis-à-vis de son contenu, ainsi que leur représentation avec les unités textuelles appropriées. La première difficulté est d'ordre sémantique : il faut réussir à comprendre le document pour en extraire l'essence ; la seconde est d'ordre linguistique et terminologique : il faut déterminer les propriétés linguistiques des termes-clés et connaître le vocabulaire du domaine auquel appartient le document. Par ailleurs, la forme la plus appropriée pour un terme-clé n'est pas nécessairement présente dans le contenu du document, elle peut être implicite.

[1] En 2012, l'Institut national de la statistique et des études économiques (Insee) estimait qu'environ 80 % des français sont connectés à Internet.

[2] Une notice bibliographique contient les informations factuelles d'un document (titre, auteurs, affiliation des auteurs, etc.), ainsi qu'un résumé.

Plutôt que de comprendre le document, les méthodes d'indexation automatique par termes-clés de la littérature se fondent sur des statistiques et des modélisations particulières de celui-ci. Pour ce qui est de l'usage d'unités textuelles appropriées, elles se contentent le plus souvent de celles qui occurrent dans le document. De manière générale, des termes-clés candidats sont sélectionnés dans le document d'après des critères prédéfinis (par exemple, ce doit être des groupes nominaux), ces candidats sont analysés et les termes-clés sont ensuite extraits d'entre eux en fonction du résultat de l'analyse.

L'analyse des termes-clés candidats du document peut être réalisée avec deux approches : supervisée ou non supervisée. En général, l'approche supervisée consiste à analyser les caractéristiques des termes-clés de données manuellement indexées pour apprendre à reconnaître les termes-clés. Elle consiste donc à chercher les candidats qui sont le plus vraisemblablement les termes-clés, tandis que l'approche non supervisée consiste à chercher les candidats les plus importants dans le contenu du document.

Notre objectif est d'améliorer l'indexation par termes-clés en domaines de spécialité. Cette indexation doit être de qualité documentaire, c'est-à-dire qu'elle doit respecter les principes que suivent les documentalistes, ou indexeurs professionnels (Guinchat et Skouri, 1996). Nous travaillons d'abord d'un point de vue généraliste, puis nous nous focalisons sur l'indexation par termes-clés en domaines de spécialité.

1.3 Hypothèses

Dans cette thèse, nous formulons trois hypothèses.

Notre première hypothèse concerne la sélection des termes-clés candidats et leur impact sur la suite du processus d'indexation par termes-clés. Selon nous, l'indexation gagne en efficacité lorsque la qualité de l'ensemble des candidats sélectionnés augmente. Il s'agit là d'une hypothèse triviale : si l'un des composants d'une chaîne de traitement fait des erreurs, alors celles-ci peuvent se répercuter sur les autres composants et dégrader leur performance. Cependant, de nombreux travaux utilisent encore des méthodes de sélection de candidats grossières, ou des filtres linguistiques suffisant pour sélectionner des candidats de la même forme que les termes-clés, mais produisant un nombre important d'erreurs. Ces méthodes présentent l'avantage d'être facile à mettre en œuvre pour des performances d'indexation par termes-clés satisfaisantes. Nous observons toutefois une prise de conscience de l'importance de la sélection des candidats, et des travaux récents montrent que nous gagnerions à affiner sa réalisation (Wang *et al.*, 2014). Pour améliorer la qualité de la sélection, nous pensons qu'il faut s'intéresser à deux propriétés de l'ensemble de candidats sélectionnés : le nombre de termes-clés qui se trouvent parmi les candidats et le nombre total de candidats sélectionnés. Paradoxalement, le premier doit être maximisé, tandis que le second doit être minimisé, car un espace de recherche trop grand augmente la difficulté de l'indexation (Hasan et Ng, 2014). Notre objectif est de trouver des propriétés linguistiques plus fines afin d'obtenir le meilleur compromis entre ces deux conditions.

Notre seconde hypothèse concerne la détection des mots et expressions importants vis-à-vis d'un document. Selon nous ce n'est pas l'importance de ces mots et expressions qui doit être déterminée, mais l'importance de ce qu'ils représentent. Nous parlons de sujet. Les sujets abordés dans un document sont véhiculés par une ou plusieurs unités textuelles. Il faut donc en déterminer l'importance en analysant l'usage de ces unités textuelles. Par ailleurs, si plusieurs unités textuelles sont utilisées pour représenter le même sujet dans un

document il ne faut pas déterminer l'importance de ce sujet indépendamment de chaque unité textuelle, car cela peut engendrer au moins deux types d'erreurs :

- Redondance : plusieurs termes-clés proposés représentent le même sujet ;
- Imprécision : pour chaque unité textuelle, l'importance du sujet est différente car elle ne tient pas compte de toutes les références à ce sujet dans le document.

Notre objectif est de mutualiser l'analyse des unités textuelles qui véhiculent les mêmes sujets afin d'éviter la redondance et mieux capturer l'importance de ces sujets.

Enfin, notre troisième hypothèse concerne l'usage de données indexées manuellement pour l'indexation automatique d'un document du même domaine. Nous pensons qu'il est possible de tirer profit de ces données pour (1) améliorer la précision de l'identification des unités textuelles importantes en situant le document dans son contexte global et (2) assigner des termes-clés du domaine qui sont importants vis-à-vis du document. La première perspective d'amélioration est générique à tout document, tandis que la seconde se fonde plus sur l'indexation par termes-clés telle qu'elle est pratiquée en domaines de spécialité (par les indexeurs professionnels). Utiliser des données déjà indexées doit permettre de proposer des termes-clés conformes au langage documentaire du domaine auquel appartient le document. De plus, cela peut résoudre le problème des termes-clés implicites au document. Notre objectif est de trouver une représentation unifiant celle du document à celle de son domaine, puis de proposer une méthode d'analyse capable d'identifier les unités textuelles importantes vis-à-vis du document et du domaine.

1.4 Mise en œuvre

Les contributions que nous présentons dans cette thèse sont fondées sur les hypothèses que nous venons d'exprimer. Nous proposons trois contributions, appliquées à deux langues : français et anglais. Nous disposons de quatre collections de données pour travailler dans le contexte général : deux en français et deux en anglais. Dans le cadre du projet Termith, l'Inist met à notre disposition quatre autres collections de données en domaines de spécialité. Ces collections sont en français et couvrent la linguistique, les sciences de l'information, l'archéologie et la chimie. En complément, l'Inist met à notre disposition des indexeurs professionnels. Nous tirons profit de leur expertise pour la réalisation de nos travaux, ainsi que pour la mise en place d'une campagne d'évaluation manuelle de nos travaux.

1.5 Plan de thèse

Cette thèse est organisée de la manière suivante. Tout d'abord, le chapitre 2 présente l'état de l'art en indexation automatique par termes-clés, puis le chapitre 3 introduit les données avec lesquelles nous travaillons. Nos contributions sont détaillées dans les chapitres 4 et 5. Le chapitre 4 présente nos travaux fondés sur les deux premières hypothèses et le chapitre 5 sur la troisième hypothèse. Ce dernier ce concentre sur l'indexation en domaine de spécialité. Il présente tout d'abord l'indexation manuelle réalisée par les indexeurs professionnel, puis notre contribution et enfin, une campagne d'évaluation manuelle de nos travaux en domaines de spécialité. Pour terminer, le chapitre 6 dresse le bilan de notre travail et présente quelques perspectives.

2

Indexation automatique par termes-clés

> « Il y a besoin d'outils pouvant créer des termes-clés. Bien que les termes-clés sont très utiles, une infime quantité seulement des documents disponibles sur Internet en contient. »
>
> — Turney (2000)

2.1 Introduction

Les termes-clés, souvent appelés mots-clés[1], sont les unités textuelles (mots ou expressions) qui caractérisent le contenu principal d'un document : les sujets qu'il aborde, ses idées, etc (voir l'exemple figure 2.1). Associés à un document, ils donnent une description précise de son contenu et servent à l'indexer pour la recherche d'information (RI). Nous parlons donc d'indexation par termes-clés. Cette indexation ne doit toutefois pas être confondue avec l'indexation dite « plein texte » au cœur de nombreux systèmes de RI. Celle-ci pondère tous les mots d'un document en fonction de leur importance relative à son contenu, tandis que l'indexation par termes-clés fournit un ensemble restreint de mots ou expressions qui représentent ses sujets importants, explicites ou non (voir la figure 2.1). Dans la suite, lorsque nous parlons d'indexation, nous nous référons à l'indexation par termes-clés.

Dans la littérature, nous distinguons deux catégories d'indexation automatique par termes-clés : l'une libre, l'autre contrôlée. L'indexation libre consiste à extraire d'un document les unités textuelles jugées les plus importantes vis-à-vis de son contenu. Nous parlons d'*extraction automatique de termes-clés*. L'indexation contrôlée fournit les termes-clés en

[1] Un terme-clé est plus communément appelé mot-clé. Cependant, un mot-clé n'étant pas uniquement monolexical, nous utilisons la notion de terme-clé pour lever toute ambiguïté. Lorsque nous parlons de mots-clés, cela ne concerne donc que les monolexicaux.

La cause linguistique

L'objectif est de fournir une définition de base du concept linguistique de la cause en observant son expression. Dans un premier temps, l'A. se demande si un tel concept existe en langue. Puis il part des formes de son expression principale et directe (les verbes et les conjonctions de cause) pour caractériser linguistiquement ce qui fonde une telle notion.

Termes-clés de référence : français ; interprétation sémantique ; conjonction ; expression linguistique ; concept linguistique ; relation syntaxique ; cause.

FIGURE 2.1 – Exemple d'indexation par termes-clés d'une notice bibliographique (résumé). Les termes-clés soulignés sont explicites, c'est-à-dire qu'ils occurrent dans le document, les autres sont implicites.

se fondant sur un vocabulaire contrôlé (une terminologie), sans se restreindre aux unités textuelles présentes dans le document. Nous parlons d'*assignement automatique de termes-clés*.

Dans ce chapitre d'introduction à l'indexation automatique par termes-clés, nous commençons par présenter l'étape de sélection des termes-clés candidats, qui est une étape commune à la plupart des méthodes d'extraction de termes-clés, et qui devient un objet d'étude à part entière (Wang *et al.*, 2014). Ensuite, nous présentons les tâches d'extraction automatique de termes-clés et d'assignement automatique de termes-clés, puis nous terminons par une description du processus d'évaluation des méthodes d'indexation par termes-clés. Les travaux que nous présentons ont été effectués sur l'anglais, à l'exception de quelques travaux sur le chinois (Ding *et al.*, 2011; Zhang, 2008).

2.2 Sélection des termes-clés candidats

La sélection des termes-clés candidats consiste à déterminer quelles sont les unités textuelles qui sont potentiellement des termes-clés, c'est-à-dire les unités textuelles qui ont des particularités similaires à celles des termes-clés définis par des humains, telles que la structure morphosyntaxique nom-adjectif commune à la majorité des termes-clés (par exemple, « interprétation sémantique », « concept linguistique » et « relation syntaxique »). Elle réduit l'espace de recherche et permet ainsi de diminuer le temps de traitement nécessaire pour l'extraction de termes-clés et de supprimer les unités textuelles non pertinentes pouvant affecter négativement ses performances. Pour distinguer les différents candidats sélectionnés, nous définissons deux catégories : les candidats positifs, qui correspondent aux termes-clés assignés par des humains (termes-clés de référence), et les candidats négatifs. Parmi les candidats négatifs, nous distinguons les candidats non importants des candidats erronés, tels que les conjonctions, les déterminants ou les unités textuelles mal segmentées (par exemple, « base du concept » issu du groupe nominal « une définition de base du concept linguistique », lui même composé des groupes nominaux « une définition de base » et « concept linguistique » dans la notice de la figure 2.1).

Il existe plusieurs méthodes de sélection de candidats, de la simple sélection de n-grammes, de *chunks* nominaux ou d'unités textuelles grammaticalement définies, jusqu'à

une méthode plus complexe visant à réduire radicalement le nombre de candidats.

Les n-grammes sont toutes les séquences ordonnées de n mots adjacents (voir l'exemple 1). La sélection des n-grammes est très exhaustive, elle fournit un grand nombre de termes-clés candidats, ce qui maximise la quantité de candidats positifs, la quantité de candidats non importants, mais aussi la quantité de candidats erronés. Pour réduire cette dernière, il est courant de filtrer les n-grammes avec un antidictionnaire regroupant les mots ne pouvant pas être des mots-clés (conjonctions, prépositions, mots d'usage courant, etc.). Si un n-gramme contient un mot de l'antidictionnaire en début ou en fin, alors il n'est pas considéré comme un terme-clé candidat.

Malgré son aspect grossier, la sélection des n-grammes est largement utilisée en extraction de termes-clés (Witten *et al.*, 1999; Hulth, 2003; Medelyan *et al.*, 2009), pour sa simplicité de mise en œuvre.

Exemple 1. $\{1..3\}$-grammes sélectionnés dans le phrase « L'objectif est de fournir une définition de base du concept linguistique de la cause en observant son expression. » :

Uni-gramme	**Bi-gramme**	**Tri-gramme**
« objectif »	« concept linguistique »	« définition de base »
« définition »		« base du concept »
« base »		
« concept »		
« linguistique »		
« cause »		
« expression »		

Les *chunks* nominaux (*NP-chunks*) sont des syntagmes[2] non récursifs (ou minimaux) dont la tête est un nom, accompagné de ses éventuels déterminants et modifieurs usuels (voir l'exemple 2). Ils sont linguistiquement définis et leur sélection, sans considérer les déterminants qui les précèdent, est donc plus fiable que celle des n-grammes pour l'extraction de termes-clés. Hulth (2003) le montre dans ses expériences consacrées à l'apport de connaissances linguistiques pour l'extraction automatique de termes-clés. Cependant, ses propos sont nuancés par un autre de ses constats : tirer profit de la catégorie grammaticale des mots des n-grammes permet d'obtenir de meilleures performances qu'avec les *chunks* nominaux.

Exemple 2. *chunks* nominaux sélectionnés dans le phrase « L'objectif est de fournir une définition de base du concept linguistique de la cause en observant son expression. » :

***Chunk* nominal**	**Candidat sélectionné**
« l'objectif »	« objectif »
« une définition »	« définition »
« base »	« base »
« concept linguistique »	« concept linguistique »
« la cause »	« cause
« expression »	« expression »

[2]Syntagme : unité syntaxique intermédiaire entre le mot et la phrase. Aussi appelé groupe, le syntagme constitue une unité de sens dont chaque constituant conserve sa signification et sa syntaxe propre.

La sélection d'unités textuelles qui forment des séquences grammaticalement définies permet de contrôler avec précision la nature et la grammaticalité des candidats sélectionnés. Pour cela, il faut définir des patrons grammaticaux tels que `/(N|A)+/` (voir l'exemple 3), qui représente les plus longues séquences de noms (`N`) et d'adjectifs (`A`), exprimé avec la syntaxe des expressions rationnelles.

À l'instar des *chunks* nominaux, la sélection des séquences grammaticalement définies est plus fondée linguistiquement que celle des n-grammes. Dans ses travaux, Hulth (2003) sélectionne les candidats à partir des patrons des termes-clés de référence les plus fréquents dans ses données. D'autres chercheurs, tels que Wan et Xiao (2008), se contentent des plus longues séquences de noms (noms propres inclus) et d'adjectifs.

Exemple 3. Séquences `/(N|A)+/` sélectionnés dans le phrase « L'objectif est de fournir une définition de base du concept linguistique de la cause en observant son expression. » :

/(N|A)+/
« objectif »
« définition »
« base »
« concept linguistique »
« cause
« expression »

En plus des trois méthodes de sélection précédentes, Huang *et al.* (2006) proposent un filtrage des candidats sélectionnés à partir des n-grammes, afin de réduire le nombre de candidats redondants (par exemple, « cause » représente « cause linguistique » dans la notice de la figure 2.1, page 16). Tout d'abord, ils suppriment les candidats peu fréquents dans le document, puis filtrent la redondance en les mettant en compétition. Ils construisent des groupes de candidats possédant le même mot, puis un seul candidat par groupe est retenu : celui le plus fréquent. Un candidat peut être en compétition dans différents groupes. Dans ce cas, il doit être le « vainqueur » de chaque groupe pour être retenu.

Le travail de Huang *et al.* (2006), sur la sélection des termes-clés candidats, fait partie d'un travail focalisé sur l'extraction de termes-clés. Leur évaluation ne s'intéresse qu'a cet aspect, l'apport de leur méthode de sélection des termes-clés candidats n'a donc pas été montré.

2.3 Extraction automatique de termes-clés

L'extraction automatique de termes-clés est la tâche la plus utilisée pour l'indexation par termes-clés. Les méthodes d'extraction automatique de termes-clés effectuent soit un ordonnancement par importance des termes-clés candidats vis-à-vis du contenu du document, soit une classification des termes-clés candidats entre les classes « terme-clé » et « non terme-clé ». La figure 2.2 présente la chaîne de traitement de la majorité des méthodes d'extraction de termes-clés. L'ordonnancement est principalement réalisé avec une approche non supervisée et la classification est réalisée avec une approche supervisée, qui requiert des documents d'apprentissage manuellement indexées.

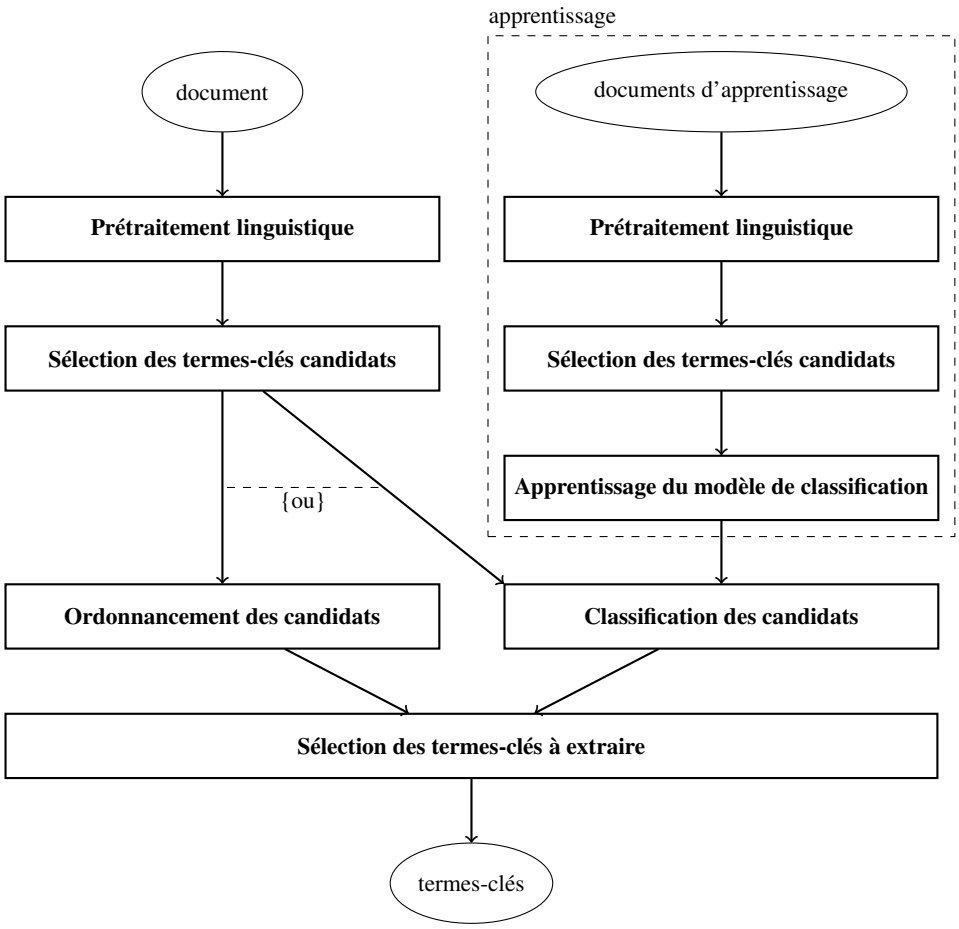


FIGURE 2.2 – Chaîne de traitement classique en extraction de termes-clés

2.3.1 Approche non supervisée

La plupart des méthodes non supervisées d'extraction de termes-clés ordonnent les termes-clés candidats d'après leur importance vis-à-vis du contenu du document (par exemple, l'expression « concept linguistique » est importante vis-à-vis du document de la figure 2.1, page 16), puis extraient les k plus importants en tant que termes-clés. Du fait qu'elles ne requièrent pas de données d'entraînement, elles sont applicables dans toutes les situations et ont la particularité de s'abstraire du domaine des documents qu'elles traitent. Les termes-clés candidats sont analysés avec des règles simples fondées sur des traits statistiques extraits du document ou d'un corpus de référence non indexé.

De nombreuses méthodes sont proposées. Certaines se fondent uniquement sur des statistiques et d'autres les combinent avec des représentations plus complexes du document : des groupes sémantiques et des graphes de cooccurrences de mots.

Nous présentons ces différentes méthodes. Lorsque celles-ci ont été évaluées sur des données disponibles, nous comparons leurs performances aux autres dans le tableau 2.1. Ces dernières sont exprimées en terme de f1-mesure. Cette mesure est exprimée entre 0 et 100 et est d'autant plus élevée si la méthode évaluée extrait un grand nombre de termes-clés corrects (voir la section 2.5, page 34).

Méthode	DUC *(Wan et Xiao, 2008)*	Inspec *(Hulth, 2003)*	NUS *(Nguyen et Kan, 2007)*	ICSI *(Adam* et al.*, 2003)*
TF-IDF*	27,0	36,3	**6,6**	**12,1**
KeyCluster*	14,0	40,6	1,7	3,2
TextRank*	9,7	33,0	3,2	2,7
SingleRank*	25,6	35,3	3,8	4,4
ExpandRank*	26,9	35,3	3,8	4,3
TopicalPageRank	31,2	—	—	—
WordTopic-MultiRank	**34,0**	**48,2**	—	—

TABLE 2.1 – Comparaison des méthodes d'extraction automatique de termes-clés de la littérature, lorsque dix termes-clés sont extraits. Les performances sont exprimées en terme de f1-mesure. DUC est une collection d'articles journalistiques, Inspec est une collection de résumés d'articles scientifiques, NUS est une collection d'articles scientifiques et ICSI est une collections de transcriptions textuelles de réunions. * indique que les résultats ont été reportés par Hasan et Ng (2010).

Méthodes statistiques

Les méthodes statistiques se fondent majoritairement sur le nombre d'occurrences des termes-clés candidats (souvent assimilé à leur fréquence) ou de leur nombre de mots, soit dans le document, soit dans un corpus de référence, ou bien les deux.

TF-IDF (Salton *et al.*, 1975) et Likey (Paukkeri et Honkela, 2010) sont deux méthodes similaires qui comparent le comportement d'une unité textuelle dans le document avec son comportement dans un corpus de référence. Elles font l'hypothèse qu'une unité textuelle a une forte importance vis-à-vis du document si elle y est très fréquente et si elle l'est peu dans le corpus de référence, auquel cas elle est spécifique au document (Spärck Jones, 1972) :

$$\text{TF-IDF}(ut) = \text{TF}(ut) \times \log\left(\frac{N}{\text{DF}(ut)}\right) \tag{2.1}$$

$$\text{Likey}(ut) = \frac{\text{rang}_{\text{document}}(ut)}{\text{rang}_{\text{corpus}}(ut)} \tag{2.2}$$

Dans TF-IDF, TF (*Term Frequency*) représente le nombre d'occurrences d'une unité textuelle *ut* dans le document, DF (*Document Frequency*) représente le nombre de documents du corpus de référence dans lesquels elle occurre et N est le nombre total de documents du corpus de référence. Plus le score TF-IDF d'une unité textuelle est élevé, plus celle-ci est importante vis-à-vis du document. Dans Likey, les rangs d'une unité textuelle dans le document et dans le corpus est obtenu à partir de son nombre d'occurrences dans le document et dans le corpus, respectivement. Plus le rapport entre ces deux rangs est faible, plus l'unité textuelle évaluée est importante dans le document.

La nature linguistique de l'unité textuelle *ut* peut être fixée au mot ou au terme-clé candidat. Si la granularité fixée est le mot, il est courant de déterminer le score d'importance des termes-clés candidat en faisant la somme du score TF-IDF ou Likey des mots qui les composent. Cependant, faire cette somme favorise les plus longues séquences de mots et fait monter dans le classement des candidats redondants qui possèdent un mot important en commun.

Méthode de pondération historique de RI, TF-IDF reste encore aujourd'hui l'une des

méthodes de référence à laquelle il faut se comparer pour montrer la validité d'une nouvelle méthode non supervisée d'extraction de termes-clés. Les résultats du tableau 2.1 (page 20) montrent que TF-IDF est encore compétitive vis-à-vis des méthodes non supervisées récentes.

Okapi (ou BM25) (Robertson *et al.*, 1998) est une mesure alternative à TF-IDF. En RI, celle-ci est préférée à TF-IDF. Bien que l'extraction automatique de termes-clés soit une discipline entre le TAL et la RI, la méthode de pondération Okapi n'a, à notre connaissance, pas été appliquée pour l'extraction de termes-clés. Claveau (2012) décrit Okapi comme un TF-IDF prenant mieux en compte la longueur des documents. Cette dernière est utilisée pour normaliser le TF (TF_{BM25}) :

$$\text{Okapi}(ut) = \text{TF}_{\text{BM25}}(ut) \times \log\left(\frac{N - \text{DF}(ut) + 0,5}{\text{DF}(ut) + 0,5}\right) \tag{2.3}$$

$$\text{TF}_{\text{BM25}}(ut) = \frac{\text{TF}(ut) \times (k_1 + 1)}{\text{TF}(ut) + k_1 \times \left(1 - b + b \times \frac{\text{DL}}{\text{DL}_{\text{moyenne}}}\right)} \tag{2.4}$$

où k_1 est une constante fixée à 2, où b est une constante fixée à $0,75$, où DL (*Document Length*) représente la longueur du document (en nombre de mots) et où $\text{DL}_{moyenne}$ représente la longueur moyenne des documents du corpus de référence.

Le travail de Barker et Cornacchia (2000) est un autre exemple d'utilisation de la fréquence pour extraire les termes-clés. Se reposant sur des fondements plus linguistiques, ils utilisent des groupes nominaux comme termes-clés candidats et tiennent compte à la fois de leur fréquence et de celle de leur tête nominale pour déterminer leur importance.

Selon Barker et Cornacchia (2000), un candidat important est un candidat informatif et fréquent. L'informativité est ici assimilée à sa taille, en nombre de mots : plus il contient de mots, plus il est informatif. Pour éviter les répétitions, jugées inesthétiques, les longs candidats (informatifs) sont parfois abrégés et leur fréquence réelle ne reflète pas leur usage. C'est pourquoi Barker et Cornacchia (2000) proposent d'utiliser la fréquence de la tête des candidats pour décider s'ils doivent être extraits ou non. Leur méthode fonctionne en quatre étapes. Ils extraient tout d'abord les n noms les plus fréquents, ils gardent uniquement les groupes nominaux contenant un de ces noms, puis les ordonnent selon le produit de leur taille et de leur fréquence réelle. Enfin, ils extraient les k groupes nominaux de meilleur rang.

Tomokiyo et Hurst (2003) tentent aussi de vérifier statistiquement deux propriétés que doit respecter un terme-clé candidat pour être extrait :

- informativité : un terme-clé doit capturer au moins une des idées essentielles exprimées dans le document analysé ;
- grammaticalité : un terme-clé doit être bien formé syntaxiquement.

Pour vérifier ces deux propriétés, trois modèles de langue (ML, voir l'équation 2.5) sont utilisés (voir la figure 2.3). Les deux premiers modèles, l'un uni-gramme, $\text{ML}^1_{\text{document}}$, l'un n-gramme, $\text{ML}^N_{\text{document}}$, sont construits à partir du document. Le dernier, un modèle n-gramme,

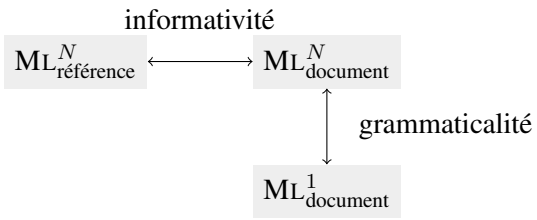


FIGURE 2.3 – Illustration des deux propriétés d'informativité et de grammaticalité induites entre trois modèles de langues (Tomokiyo et Hurst, 2003)

$\mathrm{ML}^N_{\text{référence}}$, est construit à partir d'un corpus de référence, c'est le modèle de référence. Il fournit une vision globale de la distribution des n-grammes dans la langue (français, anglais, etc.). De ce fait, plus la probabilité d'un terme-clé candidat selon le modèle n-gramme du document diverge positivement par rapport à sa probabilité selon le modèle de référence, plus il respecte la propriété d'informativité (voir l'équation 2.6). De manière similaire, plus la probabilité d'un terme-clé candidat selon le modèle n-gramme du document diverge positivement par rapport à sa probabilité selon le modèle uni-gramme du document, plus il respecte la propriété de grammaticalité (voir l'équation 2.7). La divergence est exprimée en terme de coût avec la divergence Kullback-Leibler (voir l'équation 2.8). Les termes-clés candidats sont ordonnés dans l'ordre décroissant de la somme des scores d'informativité et de grammaticalité, puis les k termes-clés candidats de meilleur rang sont extraits comme termes-clés.

$$\mathrm{ML}(\textit{candidat} = m_1\ m_2\ ...\ m_k) = \prod_{i=1}^{k} \mathrm{Prob}(m_i | m_{i-(N-1)} m_{i-((N-1)-1)} ... m_{i-1}) \quad (2.5)$$

$$\text{informativité}(\textit{candidat}) = \mathrm{KL}_{\textit{candidat}}(\mathrm{ML}^N_{\text{document}} \| \mathrm{ML}^N_{\text{référence}}) \quad (2.6)$$

$$\text{grammaticalité}(\textit{candidat}) = \mathrm{KL}_{\textit{candidat}}(\mathrm{ML}^N_{\text{document}} \| \mathrm{ML}^1_{\text{document}}) \quad (2.7)$$

$$\mathrm{KL}_{\textit{candidat}}(\mathrm{ML} \| \mathrm{ML}') = \mathrm{ML}(\textit{candidat}) \log \frac{\mathrm{ML}(\textit{candidat})}{\mathrm{ML}'(\textit{candidat})} \quad (2.8)$$

Tout comme Tomokiyo et Hurst (2003), Ding *et al.* (2011) tentent de définir des propriétés visant à affiner l'extraction de termes-clés. Ils expriment leurs propriétés sous la forme de contraintes dans un système d'optimisation (programmation par les entiers) qui explore l'espace des solutions possibles (toutes les combinaisons de mots à extraire). Les contraintes sont les suivantes :

- taille : les termes-clés extraits ne doivent pas être en nombre supérieur à k ;
- couverture : les termes-clés doivent couvrir le plus possible de sujets abordés dans le document ;
- cohérence : les mots des termes-clés doivent être cohérents entre eux.

La couverture de chaque sujet d'une solution est calculée avec le modèle *Latent Dirichlet Allocation* (LDA) (Blei *et al.*, 2003). LDA est un modèle probabiliste qui permet d'expliquer des ensembles d'observations (ici, des mots) avec des ensembles non observés (ici, des sujets), eux-mêmes définis par des distributions de probabilités calculées à partir de données (ici, des documents). Depuis le modèle LDA, Ding *et al.* (2011) extraient la probabilité conditionnelle des mots des termes-clés d'une solution sachant chaque sujet, ce qui indique quels mots de la solution sont importants pour chaque sujet. L'importance des mots des termes-clés doit excéder un seuil donné pour chaque sujet afin que la contrainte de couverture soit respectée pour la solution. La contrainte de cohérence est calculée entre chaque paire de mots de la solution. Si deux mots cooccurrent, c'est-à-dire apparaissent dans le même contexte dans le document, plus que selon un seuil donné, alors ceux-ci peuvent être présents dans la même solution, sinon la solution n'est pas satisfaisante.

Les deux contraintes réduisent le nombre de solutions acceptables. Il faut ensuite trouver quel ensemble de termes-clés parmi ces solutions est le meilleur. Pour cela, un score d'importance des mots est calculé et l'ensemble de termes-clés pour lequel la somme du score d'importance des mots est la plus élevée est extrait. Ce score est obtenue avec une combinaison linéaire du score TF-IDF du mot, d'un « bonus » s'il occurre dans le titre du document et d'un autre « bonus » s'il occurre dans sa première phrase :

$$\text{importance}(\text{mot}) = \alpha \times \frac{\sum_{d \in D} \text{TF-IDF}_d(\text{mot})}{|D|} + \beta \times \mu_{\text{mot}} + \gamma \times \nu_{\text{mot}} \tag{2.9}$$

$$\mu_{\text{mot}} = \begin{cases} \mu, & \text{si mot} \in T \\ 0, & \text{sinon} \end{cases}$$

$$\nu_{\text{mot}} = \begin{cases} \nu, & \text{si mot} \in P \\ 0, & \text{sinon} \end{cases}$$

où α, β et γ sont les coefficients associés à chaque score ($\alpha + \beta + \gamma = 1$) et où μ et ν sont les « bonus » attribués si le mot occurre dans le titre T ou dans la première phrase P du document, respectivement.

Paramétrée à l'aide de 50 articles journalistiques et évaluée sur 100 autres, la méthode de Ding *et al.* (2011) atteint environ 70 % de précision en moyenne, c'est-à-dire en moyenne quatre termes-clés corrects sur les six demandés, soit une performance très satisfaisante.

Méthodes par groupement

Les méthodes par groupement utilisent des groupes d'unités textuelles partageant une ou plusieurs caractéristiques (similarité lexicale, similarité sémantique, etc.).

Matsuo et Ishizuka (2004) groupent les 30 % de termes-clés candidats les plus fréquents lorsqu'ils cooccurrent (dans la phrase) avec les mêmes autres candidats et avec une fréquence comparable (nous parlons abusivement de lien sémantique), puis extraient les termes-clés en analysant la fréquence de cooccurrences de tous les candidats avec ces groupes. Leur hypothèse est qu'un terme-clé candidat est plus vraisemblablement un terme-clé si sa fréquence de cooccurrence avec les candidats de chaque groupe est plus importante que selon toute probabilité. Dans un premier temps, ils estiment la fréquence de cooccurrence de chaque candidat avec chaque groupe, puis, dans un second temps, ils mesurent le biais

statistique χ^2 entre leur estimation et la fréquence réelle observée (voir l'équation 2.10). Pour estimer la fréquence de cooccurrences d'un candidat avec ceux d'un groupe, ils supposent qu'un terme-clé candidat apparaissant dans de longues phrases a le plus de chance de cooccurrer avec un candidat d'un des groupes. Ainsi, soit n_t le nombre de termes-clés candidats présents dans les phrases où le candidat étudié apparaît et p_g le nombre de candidats présents dans les phrases ou un candidat du groupe g apparaît, alors la fréquence attendue entre le candidat étudié et le groupe g est représenté par le produit $n_t p_g$.

$$\chi^2(\textit{candidat}) = \sum_g \frac{(\text{fréquence}(\textit{candidat}, g) - n_t p_g)^2}{n_t p_g} \tag{2.10}$$

Lors de leurs expériences, les auteurs se sont aperçus que certains candidats peuvent être sémantiquement liées à des candidats fréquents dans un domaine plus général que celui du document. En supposant que ces cas spéciaux soient ceux ayant le plus fort biais statistique, ils suppriment du χ^2 l'argument maximum de la sommation :

$$\chi^{2\prime}(\textit{candidat}) = \chi^2 - \max_g \left\{ \frac{(\text{fréquence}(\textit{candidat}, g) - n_t p_g)^2}{n_t p_g} \right\} \tag{2.11}$$

Les termes-clés extraits sont les k termes-clés candidats ayant le plus fort biais statistique mesuré par $\chi^{2\prime}$.

Dans l'algorithme KeyCluster, Liu *et al.* (2009) utilisent aussi un groupement sémantique, mais dans leur cas, ils ne considèrent que les mots du document (mots d'un antidictionnaire exclus). Le mot le plus central de chaque groupe est sélectionné comme mot de référence et sert à l'extraction des termes-clés : chaque terme-clé candidat contenant au moins un mot de référence est extrait comme terme-clé. Cette méthode présente l'avantage d'offrir une bonne couverture des sujets abordés dans un document, car tous les groupes sémantiques sont représentés par au moins un terme-clé. Cependant, aucune pondération n'est proposée pour ordonner les termes-clés. De plus, Hasan et Ng (2010) ont montré que KeyCluster est en général moins performant que TF-IDF (voir le tableau 2.1, page 20).

Méthodes à base de graphe

Les approches à base de graphe sont actuellement les plus populaires. Utilisés dans de nombreuses applications du TAL (Kozareva *et al.*, 2013), les graphes ont l'avantage de présenter de manière simple et intuitive le document.

Mihalcea et Tarau (2004) proposent TextRank, une méthode d'ordonnancement d'unités textuelles à partir d'un graphe pour le résumé automatique et l'extraction de termes-clés. Pour l'extraction de termes-clés, les nœuds du graphe sont les mots du document et les arêtes qui les connectent représentent leurs relations d'adjacence dans le document, dans une fenêtre de deux mots. Un score d'importance (initialisé à un), est calculé pour chaque mot à partir de l'algorithme itératif PageRank (Brin et Page, 1998). PageRank est un algorithme de marche aléatoire (*random walk*) : un marcheur aléatoire parcours le graphe de mot en mot en se déplaçant vers un mot qui cooccurre avec le mot courant. Le résultat du parcours du marcheur permet de déduire l'importance de chaque mot d'après le principe de la recommandation (du vote) : un mot est d'autant plus important s'il cooccurre avec un

grand nombre de mots (parce qu'il est beaucoup visité par le marcheur) et si les mots avec lesquels il cooccurre sont eux aussi importants (parce qu'il a plus de chance d'être visité par le marcheur). Les mots les plus importants sont considérés comme des mots-clés, ils sont marqués dans le document et les plus longues séquences de mots-clés adjacents sont extraites en tant que termes-clés.

Soit le graphe de cooccurrences de mots non orienté $G = (N, A)$, où les nœuds N représentent les mots du documents, et où les arêtes A les connectent lorsqu'ils cooccurrent dans le document. L'importance de chaque mot n_i est obtenue itérativement selon la formule TextRank suivante :

$$S(n_i) = (1 - \lambda) + \lambda \times \sum_{n_j \in A(n_i)} \frac{S(n_j)}{|A(n_j)|} \tag{2.12}$$

où $A(n_i)$ est l'ensemble des nœuds connectés au nœud n_i et où λ est un facteur d'atténuation. Nombre réel défini entre 0 et 1, ce dernier peut être considéré comme la probabilité pour que le nœud n_i soit important d'après le principe de la recommandation. Brin et Page (1998) suggèrent 0,85 comme valeur par défaut de λ. Selon eux, cette valeur est un bon compromis entre la précision des résultats et la vitesse de convergence de l'algorithme.

Bien qu'intéressant, de par son intuitivité, Hasan et Ng (2010) ont montré que TextRank est moins performant que TF-IDF (voir le tableau 2.1, page 20).

Wan et Xiao (2008) modifient TextRank et proposent SingleRank. Dans un premier temps, leur méthode augmente la précision de l'ordonnancement en utilisant une fenêtre de cooccurrences élargie empiriquement à dix mots et en pondérant les arêtes par le nombre de cooccurrences entre les deux mots qu'elles connectent. La pondération, notée $\text{poids}(n_j, n_i)$, sert à ajuster l'importance du mot n_i acquise à partir de sa recommandation par le mot n_j (voir l'équation 2.13). Dans un second temps, les termes-clés ne sont plus générés à partir des séquences de mots-clés dans le document, mais ordonnés à partir de la somme du score d'importance des mots qui les composent. Comparé à TextRank, dans les expériences de Hasan et Ng (2010) réalisées avec quatre collections de données différentes, SingleRank donne de meilleurs résultats (voir le tableau 2.1). Ils restent cependant plus faibles que ceux de TF-IDF.

$$S(n_i) = (1 - \lambda) + \lambda \times \sum_{n_j \in A(n_i)} \frac{\text{poids}(n_j, n_i) \times S(n_j)}{\sum_{n_k \in A(n_j)} \text{poids}(n_j, n_k)} \tag{2.13}$$

Toujours dans le but d'améliorer l'efficacité de l'ordonnancement proposé par Mihalcea et Tarau (2004), Wan et Xiao (2008) proposent ExpandRank. ExpandRank étend SingleRank en utilisant des documents similaires au document analysé d'après la mesure de similarité vectorielle cosinus. Faisant l'hypothèse que ces documents similaires fournissent des informations supplémentaires relatives aux mots du document et aux relations qu'ils entretiennent, ExpandRank utilise les relations de cooccurrences observées dans les documents similaires pour ajouter et renforcer des arêtes dans le graphe. Dans leurs expériences réalisée avec une collection de 308 articles journalistiques, Wan et Xiao (2008) obtiennent des résultats au-delà de ceux de SingleRank. Ces résultats n'ont cependant jamais pu être reproduit et les expériences de Hasan et Ng (2010) ne montrent globalement pas d'amélioration vis-à-vis de SingleRank (voir le tableau 2.1, page 20).

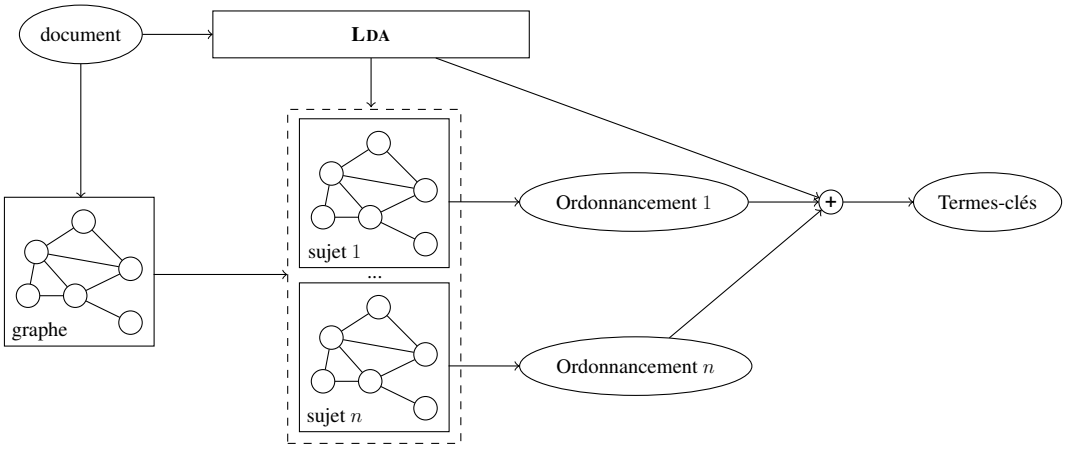


FIGURE 2.4 – Illustration du fonctionnement de TopicalPageRank (Liu *et al.*, 2010)

Liu *et al.* (2010) tentent aussi d'améliorer SingleRank. Ils proposent TopicalPageRank (TPR), une méthode qui cherche cette fois-ci à augmenter la couverture du document par les termes-clés extraits. Pour ce faire, ils détectent les sujets du document et ordonnent les mots en fonction de chaque sujet (voir la figure 2.4). À l'aide du modèle LDA (Blei *et al.*, 2003), ils ajustent chaque ordonnancement avec la probabilité conditionnelle d'un sujet donné sachant chaque mot (voir l'équation 2.14), puis donnent plus d'importance aux candidats dont les mots ont la plus forte importance (le meilleur rang) selon les sujets les plus probables dans le document (voir l'équation 2.15).

$$S(n_i, \textit{sujet}) = (1-\lambda) \times p(\textit{sujet}|n_i) + \lambda \times \sum_{n_j \in A(n_i)} \frac{\text{poids}(n_j, n_i) \times S(n_j)}{\sum_{n_k \in A(n_j)} \text{poids}(n_j, n_k)} \quad (2.14)$$

$$\text{TPR}(\textit{candidat}) = \sum_{\textit{sujet}} \left[p(\textit{sujet}|\textit{document}) \times \sum_{n \in \textit{candidat}} \text{rang}_{\textit{sujet}}(n) \right] \quad (2.15)$$

Contrairement aux précédentes méthodes à base de graphe que nous avons présenté, TopicalPageRank améliore TF-IDF (voir le tableau 2.1, page 20).

Dans la continuité du travail de Liu *et al.* (2010), Zhang *et al.* (2013) proposent WordTopic-MultiRank. WordTopic-MultiRank ajoute les sujets de LDA aux nœuds du graphe de cooccurrences et effectue un seul ordonnancement, qui tient compte de tous les sujets en même temps. Cet ordonnancement est réalisé conjointement entre les mots et les sujets, de sorte que :

- un sujet est d'autant plus important s'il est connecté à un grand nombre de mots importants ;
- un mot est d'autant plus important s'il cooccurre avec un grand nombre de mots importants et s'il est connecté à un grand nombre de sujets importants.

Comme pour SingleRank et TopicalPageRank, les termes-clés candidats sont ensuite ordonnés d'après le score d'importance des mots qu'ils contiennent.

L'ordonnancement conjoint (*co-ranking*) à partir de modèles à base de graphe est une technique qui commence à susciter de l'intérêt en TAL (Wan, 2011; Yan *et al.*, 2012; Liu *et al.*, 2014). Zhang *et al.* (2013) sont les premiers à l'appliquer à l'extraction de termes-clés. Cette approche est intéressante, car elle tient compte à la fois du contexte local du mot (le document) et de son contexte global (les sujets de la collection de données analysée par LDA).

Comparés aux résultats de TopicalPageRank, Zhang *et al.* (2013) montrent que l'ordonnancement conjoint des mots et des sujets est légèrement plus performant que la combinaison de multiples ordonnancements influencés par chaque sujet (voir le tableau 2.1, page 20).

Bilan des méthodes non supervisées

Les méthodes non supervisées d'extraction de termes-clés utilisent des techniques très différentes, mais reposent toutes sur des statistiques simples : la fréquence d'occurrence des mots dans le document, leur fréquence documentaire et la fréquence de cooccurrence entre eux. Les méthodes purement statistiques mises à part, c'est le mode de représentation du document et son analyse qui différencie les méthodes non-supervisées.

Le graphe est le mode de représentation le plus utilisé actuellement. Représentant les relations de cooccurrences entre les mots du document, il est analysé à l'aide d'un algorithme de marche aléatoire qui attribue un score d'importance à chaque nœud (mot). Ce graphe et la manière dont il est analysé sont très intuitifs, mais nous notons quelques défauts. Nous reprochons aux méthodes actuelles de modéliser le document par ses mots et leurs relations, et donc de déterminer l'importance des mots au lieu de celle des termes-clés candidats. Par ailleurs, bien que la notion de sujet ait été introduite, nous nous demandons s'il ne serait pas plus juste de grouper les candidats qui représentent le même sujet. Il est possible que l'ordonnancement gagne en précision en faisant de la sorte.

2.3.2 Approche supervisée

Les méthodes supervisées apprennent principalement à classer les termes-clés en tant que « terme-clé » ou « non terme-clé ». Leur apprentissage se fait à partir d'une collection d'apprentissage (ou d'entraînement) dont les documents sont manuellement indexés par des termes-clés. Les termes-clés candidats sont sélectionnés dans ces documents, ils servent d'exemples lorsqu'il font partie de l'indexation manuelle (de référence), de contre-exemples sinon et certaines de leurs caractéristiques (traits) sont analysées pour apprendre à discriminer « termes-clés » et « non termes-clés ».

Les méthodes proposées emploient des classifieurs. Elles diffèrent selon ces classifieurs et les traits qu'elles utilisent. Nous présentons ces différentes méthodes en les groupant par classifieur et les présentons en soulignant les traits choisis.

Classifieurs probabilistes

Les classifieurs probabilistes utilisent des distributions de probabilités de divers traits. Pour l'extraction des termes-clés d'un document, ces distributions sont combinées pour déterminer le score de vraisemblance, la probabilité, de chaque terme-clé candidat en tant que « terme-clé ». À l'instar des méthodes non supervisées, les méthodes supervisées utilisant un classifieur probabiliste peuvent ordonner les termes-clés candidats classés « termes-clés »

selon leur probabilité afin d'extraire un nombre donné de termes-clés, si nécessaire (pour évaluer les méthodes, entre autre).

KEA (Witten *et al.*, 1999) est la méthode d'extraction de termes-clés la plus populaire. Elle effectue une classification naïve bayesienne pour attribuer le score de vraisemblance de chaque terme-clé candidat. Elle combine les distributions probabilistes de deux traits : la première position du candidat dans le document et son poids TF-IDF. L'intuition de Witten *et al.* (1999) est que les termes-clés ont une certaine importance vis-à-vis du document (leur poids TF-IDF) et qu'ils font leur première apparition dans des zones similaires du document.

KEA est une approche très simple qui considère tous les traits comme indépendants (principe de la classification naïve bayésienne). Sa simplicité et ses bonnes performances ont suscité un grand intérêt et de nombreuses variantes ont été proposées. C'est le cas de la méthode de Frank *et al.* (1999), qui utilise comme trait supplémentaire le nombre de fois qu'un terme-clé candidat est un exemple, c'est-à-dire un terme-clé de l'indexation manuelle d'un document de la collection d'entraînement. Appliquée en domaines de spécialité, cette variante de KEA favorise l'extraction de termes-clés déjà utilisés pour une extraction de termes-clés homogène et améliore significativement les performances de KEA. Par ailleurs, Frank *et al.* (1999) montrent que les performances de KEA se stabilisent à partir de 50 documents d'apprentissage, alors que les performances de leur méthode augmente toujours lorsque le nombre de documents d'apprentissage augmente. Cette méthode est donc intéressante dans un contexte semi-automatique. Si ses sorties sont corrigées manuellement, et si chaque document nouvellement indexé est ajouté pour refaire l'apprentissage, alors sa précision doit toujours augmenter, contrairement à celle de KEA.

Turney (2003) reprend lui aussi KEA. Comme Ding *et al.* (2011), il améliore la cohérence entre les termes-clés candidats extraits (voir la section 2.3.1 page 22). Pour cela, il ajoute une deuxième classification naïve bayesienne après celle de KEA. La première classification sert à ordonner les candidats selon leur vraisemblance et la deuxième attribue un nouveau score de vraisemblance aux candidats, de sorte que les L meilleurs candidats aient un meilleur score de vraisemblance s'ils ont un fort lien sémantique avec un ou plusieurs candidat(s) parmi les K meilleurs ($K < L$). La force du lien sémantique est représenté par deux scores (soit $2 \times K$ traits) : le nombre de pages Web contenant les deux candidats et le nombre de titres de pages Web contenant les deux candidats.

Nguyen et Kan (2007) améliorent KEA pour l'extraction de termes-clés à partir d'articles scientifiques. Faisant l'hypothèse que les termes-clés n'ont pas une répartition homogène dans les sections d'un article scientifique, ils notent les occurrences des termes-clés candidats dans les sections génériques d'un article scientifique (résumé, introduction, motivations, état de l'art et conclusion), puis utilisent le vecteur d'occurrences ainsi construit comme trait supplémentaire. De cette manière, les termes-clés apparaissant dans les sections les plus susceptibles de contenir des termes-clés ont un score de vraisemblance plus élevé.

Caragea *et al.* (2014) utilisent eux aussi un classifieur naïf bayesien. Leur méthode repose sur le même constat que Wan et Xiao (2008) : l'extraction de termes-clés peut bénéficier

des informations extraites dans des documents en liens avec le document à partir duquel les termes-clés doivent être extraits (voir la section 2.3.1 page 24). Travaillant avec des articles scientifiques, Caragea *et al.* (2014) utilisent le réseau de citations des documents afin de déterminer leur influence sur les autres documents et inversement. En plus des traits existant, ils ajoutent un TF-IDF calculés à partir de la fréquence de chaque candidat dans les contextes citationnels, ainsi que deux traits binaires indiquant si (1) le candidat occurre dans une phrase (du document) qui cite un autre document ou si (2) il occurre dans une phrase d'un autre document qui cite le document. Bien que leur méthode obtient de meilleures performances que KEA, les auteurs mettent en évidence un défaut de leur approche. Considérant les contextes citationnels, un document qui vient d'être publié ne peut pas avoir été cité par d'autres articles et leur extraction de termes-clés pour un document tend à s'améliorer dans le temps.

Contrairement à Witten *et al.* (1999), qui utilisent un classifieur naïf bayésien et considèrent que tous les traits sont indépendants, Sujian *et al.* (2003) proposent une méthode utilisant un classifieur d'entropie maximale. Ce classifieur cherche parmi plusieurs distributions (une pour chaque trait) laquelle a la plus forte entropie. La distribution ayant la plus forte entropie est par définition celle qui contient le moins d'informations, ce qui la rend moins arbitraire et donc plus appropriée pour l'extraction automatique de termes-clés. Chaque trait se voit donc attribuer un poids, de sorte que les traits les moins arbitraires ont le plus de poids dans la classification. En plus des traits cités pour les méthodes précédentes, et à l'instar de Nguyen et Kan (2007), ils tirent parti d'autres traits liés à la nature des documents qu'ils traitent. Ainsi, pour des articles journalistiques ils utilisent leur type (information, sport, etc.) et la catégorie d'entité nommée des candidats s'ils en sont une (personne, pays, organisme, etc.).

Plus récemment, le travail de Zhang (2008) montre l'applicabilité d'un CRF (*Conditional Random Field*) à la tâche d'extraction de termes-clés. Ce classifieur a l'intéressante capacité à prédire des séquences de classes, soit à étiqueter tout un document en donnant les classes suivantes pour chaque mot : « terme-clé », « début d'un terme-clé » et « partie d'un terme-clé » (voir l'exemple 4). Zhang (2008) reprend les traits présentés précédemment (TF-IDF, première position, section d'article scientifique, etc.) et y ajoute le contexte de chaque mot. Nous trouvons cette notion de contexte dans les méthodes non supervisées à base de graphe (voir la section 2.3.1 page 24), mais il s'agit de la première fois que nous trouvons celle-ci dans une méthode supervisée. Le fonctionnement particulier du CRF (étiquetage de séquences de mots) se prête plus à l'utilisation du contexte que les autres classifieurs.

Exemple 4. Étiquetage par CRF de la phrase « L'objectif est de fournir une définition de base du concept linguistique de la cause en observant son expression. » de la figure 2.1 (page16) avec les étiquettes `TC_D` (début d'un terme-clé), `TC_P` (partie d'un terme-clé), `N` (nom), `V` (verbe), `PP` (participe présent), `Pro` (pronom), `Det` (déterminant), `Pre` (préposition) :

```
L'/Det objectif/N est/V de/Pre fournir/V une/Det
définition/N de/Pre base/N du/Pre concept/TC_D
linguistique/TC_P de/Pre la/Det cause/N en/Pre
observant/PP son/Pro expression/N ./PONCT
```

Arbres de décision

Les arbres de décision sont des classifieurs dont les branches représentent des tests sur des traits des candidats. Ces tests routent les candidats vers les feuilles de l'arbre représentant leur classe respective (« terme-clé » ou « non terme-clé »).

Dans son article sur l'apprentissage pour l'extraction automatique de termes-clés, Turney (2000) entraîne plusieurs arbres de décision (technique de *random forest*) et réduit la tâche d'extraction de termes-clés à un vote. Les arbres de décision classent indépendamment chaque candidat et les candidats majoritairement classés « terme-clé » sont extraits comme termes-clés. Parmi les nombreux traits qu'utilise Turney (2000), les plus novateurs sont des traits binaires qui visent des catégories grammaticales précises : « contient un nom propre ? », « contient un verbe usuel ? » et « se termine par un adjectif ? ». Contrairement aux autres travaux, celui de Turney (2000) est aussi l'un des seuls à décliner les traits sur deux niveaux de granularité : le candidat (grain expression) et chacun de ses mots (grain mot).

Ercan et Cicekli (2007) utilisent eux aussi des arbres de décision pour extraire les termes-clés. Les termes-clés sont ici restreints aux mots-clés. L'aspect novateur de leur méthode est l'usage de chaînes lexicales pour la définition de nouveaux traits discriminants. Une chaîne lexicale est un graphe de mots liés entre eux hiérarchiquement (voir la figure 2.5). Ercan et Cicekli (2007) tiennent compte des relations hiérarchiques de méronymie[3]/holonymie[4], d'hyponymie[5]/hyperonymie[6] et de synonymie, auxquelles ils donnent un poids (4 pour la méronymie/holonymie, 7 pour l'hyponymie/hyperonymie et 10 pour la synonymie). Chaque mot se voit attribuer quatre traits correspondant à quatre scores obtenus à partir des poids des relations :

1. Score de la chaîne lexicale : somme du poids de toutes les relations de la chaîne lexicale ;
2. Score du mot dans la chaîne lexicale : somme du poids de toutes les relations du mot avec les autres mots de la chaîne lexicale ;
3. Couverture de la chaîne lexicale : différence entre la dernière occurrence, dans le document, d'un mot de la chaîne lexicale avec la première occurrence, dans le document, d'un mot de la chaîne lexicale ;
4. Couverture du mot et de ses voisins dans la chaîne lexicale : identique à la couverture de la chaîne lexicale, mais en tenant compte uniquement du mot et de ses voisins dans la chaîne.

Tirant aussi profit d'arbres de décision, Lopez et Romary (2010) sont les vainqueurs de la campagne d'évaluation SemEval-2010 (Kim *et al.*, 2010). Ils extraient les termes-clés en deux étapes. Tout d'abord, ils ordonnent les termes-clés candidats avec les arbres de

[3]Méronyme : mot dont le signifié est une sous-partie de celui d'un autre mot, son holonyme. Par exemple, « bras » est un méronyme de « corps ».

[4]Holonyme : mot dont le signifié est composé de celui d'un autre mot, son méronyme.

[5]Hyponyme : mot dont le signifié est plus spécifique que celui d'un autre mot, son hyperonyme.

[6]Hyperonyme : mot dont le signifié est plus général que celui d'un autre mot, son hyponyme.

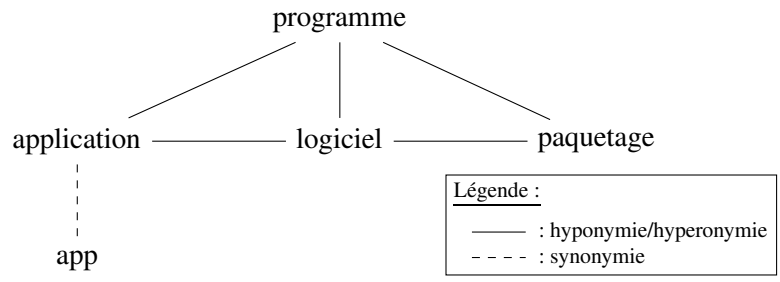


FIGURE 2.5 – Exemple de chaîne lexicale (Ercan et Cicekli, 2007)

décision, puis ils les ré-ordonnent à la manière de Turney (2003) : les candidats bien classés initialement sont d'autant mieux classés qu'ils ont un fort lien sémantique avec d'autres candidats bien classés initialement.

Séparateurs à vastes marges

Les séparateurs à vastes marges (SVM) projettent les exemples et les contre-exemples du corpus d'entraînement sur un plan (ou hyperplan) selon la valeur de leurs traits, puis construisent la droite (l'hyperplan) qui les sépare. Pour classer les termes-clés candidats d'un document, il suffit ensuite de les projeter sur ce même plan et d'utiliser l'hyperplan appris.

Zhang *et al.* (2006) utilisent un SVM pour extraire les termes-clés à partir de ce qu'ils appellent le contexte global et le contexte local des termes-clés candidats. Ils représentent le contexte global d'un candidats par son TF-IDF, sa première position et ses occurrences dans différentes parties du document, tandis qu'ils représentent son contexte local par sa catégorie grammaticale et trois traits encore jamais utilisés auparavant. Les deux premiers traits sont déterminés à partir des dépendances entre les mots. L'un dénote le nombre de fois que le candidat modifie un mot et l'autre dénote le nombre de fois qu'un mot modifie le candidat. Le dernier trait s'appelle le TF-IDF contextuel, il s'agit de la somme du TF-IDF de tous les mots qui cooccurrent avec le candidat. Ce dernier trait est intéressant, il indique si un candidat occurre dans un contexte important vis-à-vis du document.

Jiang *et al.* (2009) extraient les termes-clés à partir d'un type particulier de SVM, baptisé SVM^{rank}. SVM^{rank} construit plusieurs hyperplans qui permettent d'ordonner les termes-clés candidats. Utilisant le score TF-IDF des candidats, leur taille (en nombre de mots), leur première position, leur entropie et d'autres traits, le travail de Jiang *et al.* (2009) montre que le classifieur SVM^{rank} est plus performant qu'un SVM ou qu'un classifieur naïf bayesien utilisant les mêmes traits.

Eichler et Neumann (2010) extraient eux aussi les termes-clés à partir d'un SVM^{rank}. Ils apprennent le SVM^{rank} avec trois valeurs pour le rang. La valeur maximale est attribuée aux exemples d'un document d'apprentissage, la valeur minimale aux contre-exemples du document et une valeur intermédiaire à ses contre-exemples qui sont des exemples d'autres documents d'apprentissage. Cette approche peut être assimilé à celle de Frank *et al.* (1999), qui estime qu'un terme-clé candidat fréquemment utilisé comme terme-clé dans le corpus

d'apprentissage est plus vraisemblablement un terme-clé. Quant aux traits utilisés pour entraîner le SVM^{rank}, le plus notable se réfère à Wikipedia. L'intuition des auteurs est que si un terme-clé candidat fait l'objet d'un article Wikipedia, alors il est plus vraisemblablement un terme-clé.

Perceptrons multicouches

Les perceptrons multicouches sont des classifieurs qui émulent la biologie de l'apprentissage humain. Ce sont des réseaux de neurones répartis sur au moins trois couches. Les neurones de la première couche représentent les traits d'un candidat (un neurone par trait), ceux des couches intermédiaires (couches cachées) propagent des scores obtenus selon la valeur des traits et ceux de la dernière couche donnent un score final pour chaque classe « terme-clé » et « non terme-clé » (un neurone par classe). La classe ayant le plus haut score est celle du terme-clé candidat pour lequel correspond la valeur des traits. Optionnellement, les scores calculés pour chaque classe peuvent être utilisés pour déterminer le degré de confiance du perceptron pour la classe qu'il a attribué (Denker et Lecun, 1991).

Sarkar *et al.* (2010) utilisent un perceptron multicouche. Les traits qu'ils emploient concernent la fréquence du candidat, sa position et sa taille (en nombre de mots) ainsi que celle de ses mots (en nombre de caractères). Cette dernière est rarement utilisée comme trait pour l'extraction supervisée de termes-clés. L'hypothèse de Sarkar *et al.* (2010) se fonde sur la loi de Zipf (1935) : les mots courts étant plus fréquents que les mots longs, alors la taille d'un mot est une indication de sa rareté, donc de sa spécificité vis-à-vis du document.

À la manière des méthodes utilisant un classifieur probabiliste, Sarkar *et al.* (2010) ordonnent les termes-clés candidats afin d'extraire un nombre donné de termes-clés lorsque nécessaire. Pour cela, ils utilisent le degré de confiance attribué à la classification (Denker et Lecun, 1991). En premier sont placés les candidats classés « terme-clé », dans l'ordre décroissant du score de confiance ; en dernier sont placés les candidats classés « non terme-clé », dans l'ordre croissant du score de confiance. Ainsi, s'il y a plus de « terme-clé » que requis, alors ceux ayant la plus haute confiance sont extrait. Inversement, s'il n'y a pas suffisamment de « terme-clé », alors des candidats classés « non terme-clé » avec une confiance faible sont ajoutés.

Algorithmes génétiques

Les algorithmes génétiques sont des algorithmes qui donnent une solution approchée à un problème d'optimisation. Ce type d'algorithme n'effectue pas de classification et n'est pas nécessairement supervisé.

Turney (2000) propose une méthode supervisée, GenEx, dont les paramètres sont valués par un algorithme génétique, appelé *Genitor*. Un algorithme d'extraction de termes-clés, appelé *Extractor*, est appliqué sur le corpus d'apprentissage avec des paramètres initiaux, puis le *Genitor* fait évoluer la valeur de ses paramètres jusqu'à trouver celle qui maximise les performances de l'extraction. L'extraction des termes-clés d'un document se fait ensuite avec l'*Extractor* et ses paramètres configurés par le *Genitor*. Les paramètres appris sont principalement des seuils limitant la taille des candidats, le nombre de mots importants à considérer pour filtrer les candidats, ou encore le nombre de termes-clés à extraire. Ce sont

aussi des facteurs multiplicateurs utilisés notamment pour le calcul de l'importance des mots et des candidats.

Bilan des méthodes supervisées

Les méthodes supervisées reformulent la tâche d'extraction de termes-clés en une tâche de classification des termes-clés candidats en tant que « terme-clé » ou « non terme-clé ». Pour cela, elles utilisent des classifieurs et proposent divers traits pour discriminer les candidats. C'est par les traits qu'elles proposent que les méthodes supervisées rivalisent. Certains traits sont génériques et communs à la majorité des méthodes. C'est le cas de la position de la première occurrence du candidat et de son score TF-IDF. D'autres traits sont spécifiques à certains types de documents. Par exemple, la section dans laquelle occurre un terme-clé est un trait discriminant pour l'extraction de termes-clés à partir d'articles journalistiques.

2.4 Assignement automatique de termes-clés

L'assignement automatique de termes-clés fait l'objet de moins de travaux que l'extraction. Il s'agit aussi d'une tâche plus difficile, car elle doit assigner des entrées d'un vocabulaire contrôlé en tant que termes-clés d'un document indépendamment de leur présence dans celui-ci.

Medelyan et Witten (2006) sont les premiers à proposer une méthode capable de faire de l'assignement de termes-clés. Celle-ci, KEA++, améliore la méthode d'extraction KEA. Pour cela, elle utilise un thésaurus[7] du domaine de spécialité à traiter. Il est mis à profit de deux manières : d'abord pour sélectionner les termes-clés candidats, ensuite pour améliorer la classification.

Medelyan et Witten (2006) décident de réaliser l'assignement en se limitant aux termes-clés qui occurrent dans le document. Ils sélectionnent donc toutes les unités textuelles qui correspondent à une entrée du thésaurus. À l'instar des candidats de KEA, ceux de KEA++ sont ensuite classés en tant que « terme-clé » ou « non terme-clé » par un classifieur naïf bayesien. Ce classifieur est le même que celui de KEA, à l'exception d'un trait supplémentaire : le nombre de relations sémantiques qu'entretient le candidat avec les autres dans le thésaurus. De cette manière, ils déterminent l'importance du candidats dans le domaine.

Évaluée avec des documents du domaine agroalimentaire et le thésaurus Agrovoc[8], la méthode KEA++ double les performances de KEA. Le travail de Medelyan et Witten (2006) montre donc l'efficacité d'une méthode d'assignement de termes-clés, même limitée.

Liu *et al.* (2011) proposent une méthode que nous assimilons à de l'assignement de termes-clés. Ils font l'hypothèse qu'un document et ses termes-clés expriment le même contenu, mais dans deux langues différentes : l'une expressive et l'autre synthétique. Ils reformulent donc la tâche d'indexation par termes-clés en une tâche de traduction du langage naturel vers celui des termes-clés.

Liu *et al.* (2011) apprennent un modèle de traduction IBM *Model-1* (Brown *et al.*, 1993) à l'aide de paires de mots : un mot du langage naturel, l'autre du langage synthétique des

[7] Thésaurus : liste de termes regroupés selon les concepts d'un domaine de connaissance qu'ils représentent.
[8] `http://aims.fao.org/standards/agrovoc`

termes-clés. Les mots du langage naturel sont extraits du document et les mots du langage synthétique sont extrait soit de son titre, soit de son résumé. Le modèle de traduction peut ensuite être appliqué aux termes-clés candidats, pour réaliser de l'extraction de termes-clés à la manière des classifieurs probabilistes présentés dans la section 2.3.2 (page 27), ou être utilisé pour générer une traduction, c'est-à-dire des termes-clés.

Parce qu'elle est capable de générer des termes-clés, cette méthode est intéressante. Cependant, il ne s'agit là que d'un premier pas vers l'assignement de termes-clés, car aucun vocabulaire n'est utilisé pour contrôler la génération.

Alors que l'assignement attribue des termes-clés d'une qualité certifiée grâce au vocabulaire contrôlé, cette tâche n'est encore que très peu étudiée. Seulement deux travaux originaux s'en rapprochent. L'un se fonde sur un vocabulaire contrôlé (un thésaurus), mais ne va pas au-delà du contenu du document ; l'autre génère des termes-clés qui n'occurrent pas nécessairement dans le document, mais ne se fonde pas sur un vocabulaire contrôlé.

2.5 Évaluation automatique de l'indexation par termes-clés

Pour montrer l'apport des nouvelles méthodes d'indexation par termes-clés, celles-ci sont comparées automatiquement aux méthodes existantes dans un processus d'évaluation « à la Cranfield » (Voorhees, 2002). Chaque méthode est appliquée à un ensemble de documents de test (collection de test), les termes-clés qu'elle fournit pour chaque document sont mis en correspondance « exacte » avec les termes-clés attribués manuellement aux documents (jugements de référence)[9], puis évalués selon différents critères. Pour chaque critère, c'est la méthode qui obtient les meilleurs résultats en moyenne qui est jugée la plus efficace.

La mise en correspondance des termes-clés extraits/assignés aux termes-clés de référence sert à distinguer ceux qui sont corrects de ceux qui ne le sont pas. Nous parlons, respectivement, de vrais positifs et de faux positifs (voir le tableau 2.2). De la même manière, toute autre unité textuelle non extraite/assignée par la méthode automatique est appelée faux négatif si elle correspond à un terme-clé de référence, et vrai négatif dans le cas contraire.

D'après le paradigme d'évaluation « à la Cranfield », un vrai positif ne peut être considéré comme tel que s'il est strictement identique à un terme-clé de référence. Cette correspondance « exacte » induit une évaluation pessimiste des méthodes d'indexation automatique par termes-clés, car les variantes des termes-clés de référence sont jugées incorrectes sans distinction avec les autres faux positifs. Pour minimiser ce problème, toutes les évaluations réalisées dans la littérature tiennent compte uniquement du radical des mots des termes-clés, c'est-à-dire leur forme privée de tout suffixe (par exemple, « empir » est le radical de « empirique »). Les différences d'accords en genre et en nombre sont donc autorisées, ainsi que toute autre dérivation suffixale. Cette approche n'est pas parfaite, car elle fait parfois correspondre des mots porteurs de sens différents (par exemple, « empire » et « empirique » possèdent la même racine « empir »). Une approche plus rigoureuse serait d'utiliser les lemmes des mots, c'est-à-dire leur forme conventionnelle (celle que nous retrouvons dans un dictionnaire). Contrairement à la racinisation, qui applique seulement des règles de désuffixage (par exemple, *-es* → - afin d'enlever l'accord du féminin pluriel), la lemmatisation requiert un lexique et doit être appliquée selon le context du mot à

[9] Les termes-clés de référence sont soit les termes-clés des auteurs, soit les termes-clés de lecteurs (personnes lambda ou professionnelles), soit la combinaison des deux.

		Jugement de référence	
		« terme-clé »	« non terme-clé »
Résultat	« terme-clé »	vrai positif (VP)	faux positif (FP)
	« non terme-clé »	faux negatif (FN)	vrai negatif (VN)

TABLE 2.2 – Matrice de confusion pour l'évaluation des méthodes d'indexation automatique par termes-clés

analyser (par exemple, le lemme de « couvant » est soit « couvant », soit « couver » selon son contexte). Pour des raisons pratiques, la lemmatisation n'est donc jamais utilisée pour l'évaluation automatique de termes-clés.

Les critères d'évaluation utilisés pour évaluer et comparer les méthodes d'indexation par termes-clés sont la précision, le rappel et la f1-mesure. La précision capture la capacité d'une méthode à minimiser les erreurs (voir l'équation 2.16). Inversement, le rappel mesure la capacité de la méthode à fournir le plus possible de termes-clés corrects (voir l'équation 2.17). Quant à la f1-mesure, elle évalue le compromis entre précision et rappel, c'est-à-dire la capacité de la méthode à extraire un maximum de termes-clés corrects tout en faisant un minimum d'erreurs (voir l'équation 2.19).

$$\text{précision} = \frac{\#\text{VP}}{\#\text{VP} + \#\text{FP}} \tag{2.16}$$

$$\text{rappel} = \frac{\#\text{VP}}{\#\text{VP} + \#\text{FN}} \tag{2.17}$$

$$\text{f-mesure} = (1 + \beta^2) \times \frac{\text{précision} \times \text{rappel}}{(\beta^2 \times \text{précision}) + \text{rappel}} \tag{2.18}$$

$$\text{f1-mesure} = 2 \times \frac{\text{précision} \times \text{rappel}}{\text{précision} + \text{rappel}} \tag{2.19}$$

2.6 Conclusion

Nous avons présenté la tâche d'indexation par termes-clés, de la sélection des termes-clés candidats aux différentes méthodes d'extraction et d'assignement de termes-clés, en passant par le processus d'évaluation automatique de ces dernières.

La sélection des termes-clés candidats est une étape quasi-systématique de l'extraction automatique de termes-clés. Ne s'agissant cependant pas du cœur de la tâche, la définition de nouvelle méthodes de sélection est négligée au profit de méthodes de sélection simples (sélection de n-grammes, *chunks* nominaux ou séquences de noms et d'adjectifs). Toutefois, l'idée que l'indexation par termes-clés gagnerait en performances si les candidats sélectionnés étaient moins nombreux, c'est-à-dire contiendraient moins d'erreurs, semble faire consensus auprès des chercheurs qui s'y intéressent (Huang *et al.*, 2006; Wang *et al.*, 2014). La méthode qui consiste à définir des patrons grammaticaux, tels que celui pour sélectionner les plus longues séquences de noms et d'adjectifs, pourrait être utilisée pour sélectionner des candidats en s'intéressant plus en profondeur aux classes grammaticales employées dans les termes-clés, ainsi qu'à d'autres propriétés linguistiques de leurs composants.

Le cœur de la tâche d'indexation automatique par termes-clés est réalisé de deux manières différentes : soit les termes-clés sont extraits depuis le contenu du document, soit ils sont assignés en puisant dans un vocabulaire contrôlé. Dans la littérature, l'extraction fait l'objet de plus de travaux que l'assignement. Elle est plus simple à mettre en œuvre, car elle analyse les unités textuelles présentent dans le document, tandis que l'assignement doit pouvoir déterminer si une entrée du vocabulaire contrôlé (n'occurrant pas nécessairement dans le document) est importante vis-à-vis de celui-ci. Par ailleurs, la seule méthode réalisant actuellement l'assignement se limite aux entrées présentent dans le document, tant il peut parfois être difficile d'établir le lien entre celles qui n'y sont pas et le contenu du document.

L'évaluation des méthodes d'indexation par termes-clés est généralement effectuée de manière automatique. Comparé à un jugement de référence unique, le résultat d'une méthode d'indexation par termes-clés est évalué en termes de précision, qui est d'autant plus élevée s'il y a le moins d'erreurs, de rappel, qui est d'autant plus élevée s'il y a beaucoup de termes-clés positifs, et de f1-mesure, qui est le compromis entre précision et rappel. Ce modèle d'évaluation présente l'avantage d'être utilisable dès lors que des données de test sont disponibles. La comparaison à un jugement de référence unique rend cependant l'évaluation pessimiste. Lorsque les conditions le permettent, une évaluation manuelle reste indispensable pour mieux mesurer les forces et les faiblesses d'une méthode d'indexation par termes-clés.

3

Ressources

> « [...] pour comprendre précisément les forces et les faiblesses d'un système d'extraction de termes-clés, il est essentiel de l'évaluer avec différentes collections de données. »
>
> — Hasan et Ng (2010)

3.1 Introduction

En indexation automatique par termes-clés, des collections de données sont nécessaires à l'évaluation et à la comparaison des nouveaux travaux aux précédents. Elles sont généralement réparties en deux ensembles : un ensemble de documents d'entraînement (ou d'apprentissage), dont les termes-clés de référence servent à paramétrer ou à entraîner une méthode, et un ensemble de documents de test, dont les termes-clés de référence servent à évaluer les performances d'une méthode. De nombreuses collections sont accessibles publiquement[1], elles couvrent plusieurs langues (français, anglais, etc.), domaines (météorologie, sciences humaines et sociales, informatique, etc.) et genres de documents (résumés, articles scientifiques, articles journalistiques, etc.). Cette diversité est essentielle à la compréhension des points forts et des points faibles d'une méthode d'indexation par termes-clés, car des facteurs différents peuvent influencer ses performances. Hasan et Ng (2014) en énoncent quatre :

- si un document est long, alors le nombre de termes-clés candidats pour celui-ci est élevé et l'indexation par termes-clés est plus difficile que pour un document court ;
- si le contenu d'un document est structuré (par exemple un article scientifique réparti en sections), alors une méthode tenant compte de cette structure est avantagée ;

[1] Un grand nombre de collections de données est accessible depuis le dépôt GitHub de Su Nam Kim (*snkim*) : `https://github.com/snkim/AutomaticKeyphraseExtraction`.

- si des changements thématiques surviennent dans un document, une méthode qui utilise la position de la première occurrence des candidats peut être désavantagée ;
- si des sujets sans relation sont abordés dans un même document, alors une méthode qui tisse des liens sémantiques entre les termes-clés candidats est pénalisée.

Des variations de performance dues à la mauvaise qualité de certains documents sont aussi constatées. Dans le cas d'articles scientifiques, par exemple, les documents sont fréquemment issus d'un PDF (*Portable Document Format*), puis convertis en texte depuis ses flux de données ou avec des outils d'OCR (*Optical Character Recognition*). Ces procédés ne sont pas parfaits, ils ajoutent parfois des fautes d'orthographe (mauvais caractères et problèmes d'accents) et ils ne traitent pas toujours correctement certains environnements graphiques et textuels tels que les notes de pieds de page, les tableaux, les figures et les équations. Nous identifions deux autres facteurs que Hasan et Ng (2014) n'évoquent pas. Selon les collections de données, la proportion de termes-clés de référence n'apparaissant pas dans les documents varie d'un extrême à l'autre. Dans le cas où la majorité des termes-clés n'apparaissent pas dans les documents, les méthodes d'extraction de termes-clés sont désavantagées. La qualification des annotateurs des termes-clés de référence est le second facteur. Nous observons trois annotateurs différents : les auteurs des articles, des indexeurs professionnels et des lecteurs. Les auteurs cherchent à attirer le lecteur ; les indexeurs professionnels (ou ingénieurs documentalistes) aspirent à une indexation homogène, et conforme au langage du domaine, de tous les documents d'un même domaine grâce à un vocabulaire contrôlé ; les lecteurs sont plus libres que les auteurs et les indexeurs professionnels, ils ne sont influencés ni par les termes-clés « à la mode », ni par un vocabulaire contrôlé.

Dans nos travaux de recherche, nous utilisons cinq collections de données : Termith, DEft, Wikinews, SemEval et DUC. En accord avec le point de vue de Hasan et Ng (2010), celles-ci couvrent deux langues (français et anglais), un large éventail de domaines (sciences humaines et sociales, informatique, météorologie, catastrophes naturelles, etc.) et trois genres de documents (résumés, articles scientifiques et articles journalistiques). De plus, les annotateurs de chaque collection ont des qualifications différentes.

3.2 Termith

Le projet Termith et l'Inist mettent à notre disposition quatre collections de notices bibliographiques en domaines de spécialité. Une notice bibliographique est une entrée d'un catalogue bibliographique. Elle décrit un document avec différentes métadonnées : des informations factuelles (titre, auteurs, affiliation des auteurs, éditeur, résumé, termes-clés des auteurs, etc.) et des informations produites par un indexeur professionnel (résumé — si nécessaire, termes-clés, etc.). Pour produire les termes-clés, l'indexeur professionnel a recours à des ressources et des pratiques documentaire (Guinchat et Skouri, 1996) qui assurent une indexation par termes-clés de qualité : homogène, conforme au langage du domaine du document, spécifique au document, général du point de vue de son domaine, exhaustive et impartiale.

Les tableaux 3.1 à 3.4 représentent la répartition, revue par revue, de chacune des collections. Le projet Termith s'intéresse principalement aux sciences humaines et sociales (SHS). De ce fait, trois de ses collections de notices représentent un domaine de SHS : linguistique, sciences de l'information (sciences de l'info.) et archéologie. La quatrième collection re-

Revue	Quantité de documents	Période de publication
A. sp. Anglais de spécialité	99	2001 – 2012
Bulletin d'études orientales	8	2008 – 2010
Cahiers de grammaire	4	2000
Cahiers de praxématique : (Montpellier)	44	2006 – 2009
Langue française : (Paris. 1969)	39	2011 – 2012
Les Cahiers de l'ILCEA	20	2004 – 2007
Lidil : (Grenoble)	76	2005 – 2011
Linx	146	2001 – 2008
Mots : (Paris. 1980)	112	2002 – 2012
Recherches linguistiques de Vincennes	37	2001 – 2010
Travaux de linguistique : (Gent)	130	2001 – 2005

TABLE 3.1 – Détail des revues du corpus de linguistique (Termith)

Revue	Quantité de documents	Période de publication
Communication : (Montréal)	2	2005
Document numérique	176	2001 – 2012
Documentaliste : (Paris)	385	2007 – 2012
Questions de communication : (Nancy)	135	2007 – 2012
Études de communication	8	2009

TABLE 3.2 – Détail des revues du corpus de sciences de l'information (Termith)

Revue	Quantité de documents	Période de publication
Annales de Bretagne et des pays de l'Ouest	3	2001 – 2008
Archéosciences	24	2007 – 2011
Arts asiatiques : (Paris)	36	2001 – 2006
Bulletin de correspondance hellénique	9	2001 – 2003
Bulletin de l'Ecole française d'Extrême-Orient	4	2001
Bulletin de la Société préhistorique française	163	2007 – 2012
Documents d'archéologie méridionale	51	2001 – 2006
Gallia	22	2001 – 2007
Gallia. Préhistoire	26	2001 – 2007
Journal de la Société des océanistes	2	2001
Paléo : (Les Eyzies de Tayac-Sireuil)	137	2001 – 2009
Paléorient	43	2001 – 2007
Préhistoire anthropologie méditerranéennes	30	2003 – 2005
Revue archéologique de Picardie	10	2001 – 2007
Revue archéologique de l'Est	35	2005 – 2011
Revue archéologique de l'Ouest	39	2006 – 2011
Revue archéologique du centre de la France	13	2001 – 2004
Revue des études grecques : (Paris)	11	2002 – 2012
Revue numismatique : (Paris)	17	2001 – 2005
Syria	7	2001
Syria : (Paris)	18	2004 – 2005
Techniques & culture : (Paris)	16	2003 – 2010

TABLE 3.3 – Détail des revues du corpus d'archéologie (Termith)

Revue	Quantité de documents	Période de publication
Canadian Journal of Chemistry	125	1983 – 1991
Comptes rendus de l'Académie des sciences. Série 2, Mécanique, physique, chimie, sciences de l'univers, sciences de la terre	265	1985 – 1990
Comptes rendus de l'Académie des sciences. Série II, Mécanique, physique, chimie, astronomie	70	1994 – 1997
Comptes rendus de l'académie des sciences. Série IIc, chimie	124	1998 – 2001
Comptes rendus. Chimie	200	2002 – 2012

TABLE 3.4 – Détail des revues du corpus de chimie (Termith)

présente la chimie. Le corpus de linguistique est constitué de 715 notices d'articles français paru entre 2000 et 2012 dans 11 revues ; le corpus des sciences de l'information contient 706 notices d'articles français publiés entre 2001 et 2012 dans cinq revues ; le corpus d'archéologie est composé de 716 notices représentant des articles français paru entre 2001 et 2012 dans 22 revues ; le corpus de chimie est composé de 784 notices d'articles français publiés entre 1983 et 2012 dans cinq revues.

Le tableau 3.5 présente les caractéristiques du corpus Termith. Chaque domaine est divisé en deux sous-ensembles[2] : un ensemble d'apprentissage (appr.) composé de 506 à 582 notices selon le domaine et un ensemble de test contenant 200 notices. S'agissant de résumés, les documents sont courts (voir la figure 3.1), ils ont en moyenne 123,9 mots. Quant aux termes-clés, nous utilisons ceux produits par les indexeurs professionnels. Ils sont concis (un à deux mots principalement) et, du fait de la méthodologie d'indexation, ils ne sont majoritairement pas présents dans les résumés (ils sont « à assigner »).

Corpus	**Documents**				**Termes-clés**			
	Langue	Genre	Quantité	Mots moy.	Annotateur	Quantité moy.	« À assigner »	Mots moy.
Linguistique								
Appr.	Français	Scientifique (résumé)	515	160,5	Professionnel	8,6	60,6 %	1,7
Test	"	"	200	147,0	"	8,9	62,8 %	1,8
Sciences de l'info.								
Appr.	Français	Scientifique (résumé)	506	105,0	Professionnel	7,8	67,9 %	1,8
Test	"	"	200	157,0	"	10,2	66,9 %	1,7
Archéologie								
Appr.	Français	Scientifique (résumé)	518	221,1	Professionnel	16,9	37,0 %	1,3
Test	"	"	200	213,9	"	15,6	37,4 %	1,3
Chimie								
Appr.	Français	Scientifique (résumé)	582	105,7	Professionnel	12,2	75,2 %	2,2
Test	"	"	200	103,9	"	14,6	78,8 %	2,4

TABLE 3.5 – Corpus Termith

3.3 DEft (Paroubek *et al.*, 2012)

DEft est une campagne d'évaluation francophone qui s'intéresse chaque année à un domaine particulier du TAL. Le corpus éponyme que nous utilisons dans nos travaux est la collection de documents construite dans le cadre de l'édition 2012 de DEft, édition portée

[2]Les revues sont réparties équitablement entre chaque ensemble.

La cause linguistique *Linguistique*

L'objectif est de fournir une définition de base du concept linguistique de la cause en observant son expression. Dans un premier temps, l'A. se demande si un tel concept existe en langue. Puis il part des formes de son expression principale et directe (les verbes et les conjonctions de cause) pour caractériser linguistiquement ce qui fonde une telle notion.

Termes-clés : français ; interprétation sémantique ; conjonction ; expression linguistique ; concept linguistique ; relation syntaxique ; cause.

Congrès de l'ABF : les publics des bibliothèques *Sciences de l'info.*

Le cinquante-troisième congrès annuel de l'Association des bibliothécaires de France (ABF) s'est déroulé à Nantes du 8 au 10 juin 2007. Centré sur le thème des publics, il a notamment permis de méditer les résultats de diverses enquêtes auprès des usagers, d'examiner de nouvelles formes de partenariats et d'innovations technologiques permettant aux bibliothèques de conquérir de nouveaux publics, et montré des exemples convaincants d'ouverture et d'"hybridation", conditions du développement et de la fidélisation de ces publics.

Termes-clés : rôle professionnel ; évolution ; bibliothèque ; politique bibliothèque ; étude utilisateur ; besoin de l'utilisateur ; partenariat ; web 2.0 ; centre culturel.

Étude préliminaire de la céramique non tournée micacée du bas Languedoc occidental : typologie, chronologie et aire de diffusion *Archéologie*

L'étude présente une variété de céramique non tournée dont la typologie et l'analyse des décors permettent de l'identifier facilement. La nature de l'argile enrichie de mica donne un aspect pailleté à la pâte sur laquelle le décor effectué selon la méthode du brunissoir apparaît en traits brillant sur fond mat. Cette première approche se fonde sur deux séries issues de fouilles anciennes menées sur les oppidums du Cayla à Mailhac (Aude) et de Mourrel-Ferrat à Olonzac (Hérault). La carte de répartition fait état d'échanges ou de commerce à l'échelon macrorégional rarement mis en évidence pour de la céramique non tournée. S'il est difficile de statuer sur l'origine des décors, il semble que la production s'insère dans une ambiance celtisante. La chronologie de cette production se situe dans le deuxième âge du Fer. La fourchette proposée entre la fin du IV^{e} et la fin du II^{e} s. av. J.-C. reste encore à préciser.

Termes-clés : distribution ; mourrel-ferrat ; olonzac ; le cayla ; mailhac ; micassé ; céramique non-tournée ; celtes ; production ; echange ; commerce ; cartographie ; habitat ; oppidum ; site fortifié ; fouille ancienne ; identification ; décor ; analyse ; répartition ; diffusion ; chronologie ; typologie ; céramique ; étude du matériel ; hérault ; aude ; france ; europe ; la tène ; age du fer.

Réaction entre solvant et espèces intermédiaires apparues lors de l'électroréduction-acylation de la fluorénone et de la fluorénone-anil dans l'acétonitrile *Chimie*

Étude du comportement des différents acylates de fluorénols-9 vis-à-vis des anions CH_2CN (électrogénérés par réduction de l'azobenzène en son dianion dans l'acétonitrile). Réduction de la fluorénone dans l'acétonitrile en présence de chlorures d'acides ou d'anhydrides

Termes-clés : réduction chimique ; acylation ; réaction électrochimique ; acétonitrile ; composé aromatique ; composé tricyclique ; cétone ; cétimine ; effet solvant ; effet milieu ; radical libre organique anionique ; mécanisme réaction ; nitrile ; hydroxynitrile ; composé saturé ; composé aliphatique ; anhydride organique ; fluorénone ; fluorénone,phénylimine ; fluorénol-9,acylate ; fluorènepropiononitrile-9(hydroxy-9) ; bifluorényle-9,9pdiol-9,9p ; fluorèneδ9 :α-acétonitrile ; butyrique acide(chloro-4) chlorure.

FIGURE 3.1 – Exemple de notices Termith. Les termes-clés soulignés peuvent être extraits depuis le contenu des documents.

sur l'extraction de termes-clés, d'une part, et sur l'assignement de termes-clés, d'une autre part. Le corpus DEft est composé de 234 articles français publiés entre 2003 et 2008 dans quatre revues des Sciences Humaines et Sociales : *Anthropologie et Sociétés*, TTR *: traduction, terminologie, rédaction*, META *: Research in Hermeneutics, Phenomenology, and Practical Philosophy* et *Revue des Sciences de l'Éducation*. Il s'agit du corpus utilisé pour la tâche d'extraction de termes-clés de DEft-2012.

Le tableau 3.6 présente les différentes caractéristiques du corpus DEft. Celui-ci est divisé en deux sous-ensembles[3] : un ensemble d'apprentissage composé de 141 articles et un ensemble de test contenant 93 articles. Les documents de DEft étant des articles scientifiques, ils contiennent beaucoup d'informations. Ils ont en moyenne 7 102,9 mots. Les termes-clés de référence fournis avec les documents de DEft sont ceux donnés par les auteurs. Ces derniers en attribuent en moyenne 5,3 par documents. Les termes-clés ne contiennent en moyenne pas plus de 1,7 mots et sont majoritairement présent dans le contenu des documents.

Corpus	**Documents**				**Termes-clés**			
	Langue	Genre	Quantité	Mots moy.	Annotateur	Quantité moy.	« À assigner »	Mots moy.
Appr.	Français	Scientifique	141	7 276,7	Auteur	5,4	18,2 %	1,7
Test	"	"	93	6 839,4	"	5,2	21,1 %	1,6

TABLE 3.6 – Corpus DEft

Le corpus DEft est aussi un corpus au contenu bruité, imparfait. Du fait de la conversion en texte, les caractères spéciaux ne sont pas toujours reconnus et la segmentation en paragraphes a parfois lieu en milieu de phrases. Les figures 3.2 et 3.3 montrent deux exemples d'un document où la segmentation est correcte et d'un document où la segmentation est incorrecte, respectivement.

Paroubek *et al.* (2012) établissent des performances de référence en demandant à des étudiants de master en ingénierie multilingue d'indexer par termes-clés les documents de DEft. Le tableau 3.7 montre les résultats de l'indexation par termes-clés obtenus pour chacun de ces étudiants. Ceux-ci ont jugé la tâche difficile, comme en témoigne les faibles résultats obtenus (f1-mesure moyenne de 21,6 %). L'indexation par termes-clés est subjective et ils éprouvent des difficultés dans le cas ou une expression dans le texte est reformulée. Ils donnent l'exemple de « traduction française et allemande » qui est représentée par le terme-clé « traduction allemande et traduction française ». Ils notent aussi la présence de termes-clés d'un même champ sémantique, tels que « interprète » et « interprétation », et soulignent la contre-intuitivité de ce type de redondance. Les conclusions ne sont pas encourageantes pour l'indexation automatique par termes-clés. Il est possible que les problèmes rencontrés lors des tests humains soient dus à la nature des indexations manuelles. La redondance qui semble contre-intuitive en est un exemple. Nous pouvons en effet supposer qu'un auteur à recours à ce genre de procédé pour être certain d'attirer tout lecteur potentiel, par exemple l'un effectuant une recherche par mot-clé avec « interprète » et l'autre avec « interprétation ».

[3] Les revues sont réparties équitablement entre chaque ensemble.

Considérée comme une « problem solving activity » (Guilford 1975), la créativité, démystifiée, fait partie du quotidien du traducteur. Victimes d'idées préconçues et erronées sur la notion de « fidélité », beaucoup de traducteurs sont insécurisés face à leur créativité. Ils peuvent alors, comme en témoigne un de nos exemples, manquer de courage et jouer la carte de la stratégie du « playing it safe », ou bien, lorsque, comme dans un autre cas, leur statut social et professionnel leur donne une certaine assurance, garder leurs solutions créatives et revendiquer leur « trahison », toutefois sans pour autant essayer de trouver des légitimations à leurs solutions. Légitimations qui restent la plupart du temps au stade de « mécanismes de justification » ponctuels. Une analyse des besoins nous permet de montrer comment ces justifications hétéroclites et éparses peuvent venir s'intégrer dans un édifice théorique cohérent, s'appuyant notamment sur des fondements cognitivistes, susceptible de donner au traducteur le courage de sa créativité.

Pour pouvoir déterminer l.utilité d.un quelconque apport théorique à la pratique du traducteur, il faut commencer par examiner s.il existe un besoin en la matière et quelle en est la nature. Nous le

ferons à l.aide de deux corpus qui se complètent. Le premier est la transcription du débat mené par

un groupe de quatre « semi-professionnels » de l.Institut de traducteurs et interprètes de

[. . .]

Termes-clés : créativité ; didactique de la traduction ; cognitivisme ; analyse conversationnelle ; théorie de la traduction.

FIGURE 3.2 – Exemple de document de DEft. Les termes-clés soulignés sont ceux qui peuvent être extraits.

Bien qu'un grand nombre de travaux ethnographiques novateurs aient été suscités par l'« espace interculturel » que se partagent Australiens autochtones et non autochtones, notamment dans le domaine des arts visuels, les chercheurs ont accordé moins d'attention aux représentations rituelles publiques auxquelles les Aborigènes ont donné un nouvel essor en tant qu'instruments politiques. On a encore moins écrit sur la (re)construction interne de l'identité sociale autochtone et sa projection dans la production de rituels publics sur la scène néocoloniale australienne contemporaine. Tout en effectuant une remise à jour des recherches précédentes sur la question, le présent article montre comment, au cours des dix dernières années, les leaders rituelles aînées d'une petite localité d'Australie centrale ont inauguré une phase entièrement nouvelle de représentations rituelles - une phase qui diffère substantiellement des formes antérieures d'expérience cérémonielle, qui étaient étroitement liées à la négociation et à l'échange des matériaux rituels.

Pour M. Nampijinpa L.

Depuis que l'anthropologie « a découvert » la religion australienne – à partir du milieu du XIXe siècle avec les ouvrages de Spencer et Gillen dont les travaux de terrain ont alimenté les recherches de Durkheim, ethnologue en chambre, et de ses héritiers - on s'est beaucoup intéressé aux manifestations rituelles de la cosmologie aborigène connue sous le nom de « Dreaming », c'est-à-dire « Rêve » ou « Récit du Rêve ». Et bien que la fréquence de ce type de représentations cérémonielles ait diminué chez les Aborigènes, cette diminution quantitative n'affecte en rien les résultats analytiques issus de l'étude des usages contemporains du champ rituel. [. . .]

Termes-clés : dussart ; aborigènes ; femmes ; identité ; rituel ; warlpiri ; australie.

FIGURE 3.3 – Autre exemple de document de DEft. Les termes-clés soulignés sont ceux qui peuvent être extraits.

Mesure	P1	P2	P3	P4	P5	P6	P7
Précision (%)	25,0	20,0	16,7	11,8	29,2	29,2	20,8
Rappel (%)	25,0	20,8	16,7	8,3	29,2	29,2	20,8
F1-mesure (%)	25,0	20,4	16,7	9,8	29,2	29,2	20,8

TABLE 3.7 – Résultats de tests humains (sept personnes — P1..P7) sur le corpus DEft

3.4 Wikinews (Bougouin *et al.*, 2013)

Wikinews est une collection de 100 articles journalistiques en français que nous avons collecté sur le site web d'information collabotatif WikiNews[4] entre les mois de mai et décembre 2012[5].

Le tableau 3.8 donne les détails de ce corpus. Il est constitué de 100 documents de test. S'agissant d'articles journalistiques, ce sont des documents de petite taille (voir la figure 3.4). Les termes-clés de référence sont des termes-clés de lecteurs. Ces derniers ont attribué 9,6 termes-clés par document en moyenne. Parmi ces termes-clés, seuls 7,6 % d'entre eux ne peuvent pas être extraits depuis le contenu des articles.

Corpus	Documents				Termes-clés			
	Langue	Genre	Quantité	Mots moy.	Annotateur	Quantité moy.	« À assigner »	Mots moy.
Test	Français	Journalistique	100	308,5	Lecteur	9,6	7,6 %	1,7

TABLE 3.8 – Corpus Wikinews

Météo du 19 août 2012 : alerte à la canicule sur la Belgique et le Luxembourg

A l'exception de la province de Luxembourg, en alerte jaune, l'ensemble de la Belgique est en vigilance orange à la canicule. Le Luxembourg n'est pas épargné par la vague du chaleur : le nord du pays est en alerte orange, tandis que le sud a était placé en alerte rouge.

En Belgique, la température n'est pas descendue en dessous des 23°C cette nuit, ce qui constitue la deuxième nuit la plus chaude jamais enregistrée dans le royaume. Il se pourrait que ce dimanche soit la journée la plus chaude de l'année. Les températures seront comprises entre 33 et 38°C. Une légère brise de côte pourra faiblement rafraichir l'atmosphère. Des orages de chaleur sont a prévoir dans la soirée et en début de nuit.

Au Luxembourg, le mercure devrait atteindre 32°C ce dimanche sur l'Oesling et jusqu'à 36°C sur le sud du pays, et 31 à 32°C lundi. Une baisse devrait intervenir pour le reste de la semaine. Néanmoins, le record d'août 2003 (37,9°C) ne devrait pas être atteint.

Termes-clés : luxembourg ; alerte ; météo ; belgique ; août 2012 ; chaleur ; température ; chaude ; canicule ; orange ; la plus chaude.

FIGURE 3.4 – Exemple de document de Wikinews. Les termes-clés soulignés sont ceux qui peuvent être extraits.

Le tableau 3.9 montre l'accord inter-annotateur κ (Fleiss, 1971) pour l'indexation ma-

[4] `http://fr.wikinews.org/`

[5] Les documents, et leur indexation par termes-clés, du corpus Wikinews sont disponibles sur le dépôt GitHub suivant : `https://github.com/adrien-bougouin/WikinewsKeyphraseCorpus`

nuelle des termes-clés de Wikinews. Les termes-clés sont attribués librement, c'est-à-dire sans règles ou méthodologie définie, par au moins trois étudiants de master en TAL. L'accord entre les annotateurs est très faible. Il confirme la subjectivité de la tâche d'indexation par termes-clés et le besoin de définir une méthodologie et des règles précises pour associer des termes-clés à un document.

Corpus	κ
Test	-0,1

TABLE 3.9 – Accord inter-annotateur κ (Fleiss, 1971) sur le corpus Wikinews

3.5 SemEval (Kim *et al.*, 2010)

À l'instar de DEft pour le français, SemEval est une campagne d'évaluation pour l'anglais. Le corpus éponyme dont nous disposons est la collection de documents construite pour la tâche 5 de l'édition 2010 de SemEval, tâche consacrée à l'extraction de termes-clés à partir d'articles scientifiques. Le corpus SemEval est constitué de 244 articles scientifiques en anglais issus de la bibliothèque numérique de l'ACM (*Association for Computing Machinery*). Cette bibliothèque regroupe les articles de plusieurs domaines, qu'elle répartie dans différentes catégories. Les documents du corpus SemEval concernent les catégories C2.4 (*Distributed Systems* — Systèmes distribués), H3.3 (*Information Search and Retrieval* — Recherche d'information), I2.11 (*Distributed Artificial Intelligence – Multiagent Systems* — Intelligence artificielle distribuée – Systèmes multi-agents) et J4 (*Social and Behavioral Sciences – Economics* — Sciences sociales et comportementales – Économie) de la classification ACM de 1998.

Le tableau 3.10 présente les caractéristiques de SemEval. La collection est répartie en deux sous-ensembles[6] : un ensemble de 144 documents d'apprentissage et un ensemble de 100 documents de test. À l'instar des documents de DEft, ceux de SemEval sont des articles scientifiques (voir la figure 3.5) de taille importante (en moyenne 5 152,3 mots). Trois jeux d'indexation par termes-clés sont fournis avec SemEval : les termes-clés des auteurs, les termes-clés de lecteurs et l'union des deux jeux précédents. Dans nos travaux, nous utilisons la combinaison des termes-clés des auteurs et des lecteurs. Du fait de cette combinaison, le nombre de termes-clés par document est plus élevé que celui des autres corpus dont nous disposons. Ils contiennent en moyenne 2,1 mots et sont majoritairement extractibles depuis le contenu des articles.

Corpus	**Documents**				**Termes-clés**			
	Langue	Genre	Quantité	Mots moy.	Annotateur	Quantité moy.	« À assigner »	Mots moy.
Appr.	Anglais	Scientifique	144	5 134,6	Auteur / Lecteur	15,4	13,5 %	2,1
Test	"	"	100	5 177,7	"	14,7	22,1 %	2,1

TABLE 3.10 – Corpus SemEval

Le tableau 3.11 montre la quantité de termes-clés attribués par les auteurs, des lecteurs ou la réunion des deux. Ces chiffres montrent la différence de stratégie entre les auteurs

[6]Les quatres catégories ACM sont réparties équitablement entre chaque ensemble.

Deployment Issues of a VoIP Conferencing System in a Virtual Conferencing Environment

ABSTRACT

Real-time services have been supported by and large on circuitswitched networks. Recent trends favour services ported on packet-switched networks. For audio conferencing, we need to consider many issues - scalability, quality of the conference application, floor control and load on the clients/servers - to name a few. In this paper, we describe an audio service framework designed to provide a Virtual Conferencing Environment (VCE). The system is designed to accommodate a large number of end users speaking at the same time and spread across the Internet. The framework is based on Conference Servers [14], which facilitate the audio handling, while we exploit the SIP capabilities for signaling purposes. Client selection is based on a recent quantifier called "Loudness Number" that helps mimic a physical face-to-face conference. We deal with deployment issues of the proposed solution both in terms of scalability and interactivity, while explaining the techniques we use to reduce the traffic. We have implemented a Conference Server (CS) application on a campus-wide network at our Institute.

1. INTRODUCTION

Today's Internet uses the IP protocol suite that was primarily designed for the transport of data and provides best effort data delivery. Delay-constraints and characteristics separate traditional data on the one hand from voice & video applications on the other. Hence, as progressively time-sensitive voice and video applications are deployed on the Internet, the inadequacy of the Internet is exposed. Further, we seek to port telephone services on the Internet. [. . .]

Termes-clés : voip conferencing system ; packet-switched network ; audio service framework ; virtual conferencing environment ; conference server ; loudness number ; partial mixing ; voice activity detectection ; three simultaneous speakers sufficiency ; vad technique ; vce ; voip ; real-time audio ; simultaneous speakers ; sip.

FIGURE 3.5 – Exemple de document de SemEval. Les termes-clés soulignés sont ceux qui peuvent être extraits.

et des lecteurs. Kim *et al.* (2010) expliquent que les auteurs donnent peu de termes-clés comparés aux lecteurs et ils ajoutent que l'intersection des deux indexations de l'ensemble de test ne couvre qu'un tiers des termes-clés des auteurs, soulignant ainsi la différence entre les deux.

Corpus	**Annotateur**		
	Auteur	Lecteur	Combinaison
Appr.	559	1824	2223
Test	387	1217	1482

TABLE 3.11 – Nombre de termes-clés attribués dans SemEval, en fonction des annotateurs

3.6 DUC (Wan et Xiao, 2008)

DUC est une campagne d'évaluation internationale portée sur le résumé automatique. Notre collection de documents DUC est issue du corpus construit dans le cadre de l'édition 2001 de la campagne (Over, 2001). Dans leurs travaux en extraction automatique de termes-clés, Wan et Xiao (2008) utilisent les 308 documents de test du corpus de la campagne et demandent à deux étudiants de les indexer par termes-clés. Il s'agit de 308 articles journalistiques anglais publiés par six média d'information différents : *Associated Press Newswire*, *Foreign Broadcast Information Service*, *Financial Times*, *Los Angeles Times*, *San Jose Mercury News* et *Wall Street Journal*. Ceux-ci couvrent 30 sujets d'actualités (tornades, contrôle des armes à feu, etc.), indiqués explicitement pour chaque documents.

Le tableau 3.12 donne les détails du corpus DUC. Les documents sont des articles journalistiques (voir la figure 3.6). Comme ceux de Wikinews, ces documents sont courts. Ils sont cependant plus élaborés et sont constitués de 900,7 mots en moyenne. Cette différence mise à part, DUC et Wikinews sont très similaires. Leur proportion de termes-clés à assigner est très faible. Ils laissent penser que l'extraction de termes-clés est plus aisée sur les articles journalistiques.

Corpus	**Documents**				**Termes-clés**			
	Langue	Genre	Quantité	Mots moy.	Annotateur	Quantité moy.	« À assigner »	Mots moy.
Test	Anlgais	Journalistique	308	900,7	Lecteur	8,1	3,5 %	2,1

TABLE 3.12 – Corpus DUC

3.7 Prétraitement des données

Afin de les utiliser en entrées des systèmes d'indexation par termes-clés, les documents des collections présentées précédemment subissent des traitements destinés à faciliter leur analyse. Ces traitements permettent de délimiter les phrases dans les documents, de délimiter les mots dans les phrases et d'identifier la catégorie grammaticale de ces derniers.

En français, nous utilisons un segmenteur à base d'expressions rationnelles pour détecter les phrases. Il s'agit du segmenteur `PunktSentenceTokenizer` intégré dans le module python NLTK (*Natural Language ToolKit*) (Bird *et al.*, 2009). Les phrases ainsi détectées sont ensuite segmentées en mots à l'aide de l'outil Bonsai, du *Bonsai* PCFG-LA *parser*[7], dont se sert l'étiqueteur grammatical, MElt (Denis et Sagot, 2009), que nous utilisons.

En anglais, nous utilisons aussi le segmenteur `PunktSentenceTokenizer` pour détecter les phrases des documents. Celles-ci sont ensuite segmentées en mots par le `TreeBankWordTokenizer` intégré dans NLTK. Enfin, les mots sont étiquetés grammaticalement à l'aide du *Stanford* POS *tagger* (Toutanova *et al.*, 2003).

[7] `http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html`

Commodities and Agriculture : Germany sets scientist to work on BSE threat

The German government yesterday announced the launch of a new research project to examine whether the cattle disease bovine spongiform encephalopathy (BSE) can be transmitted to human beings.

The initiative comes as the country is pushing for a European Union ban on British beef imports, arguing that there is still no conclusive evidence that the disease cannot affect humans.

Seven German universities and research institutes will be sponsored by the country's research and technology ministry to examine possible connections between the origins of BSE and two other diseases, Creutzfeldt Jakob disease and Gerstmann Straussler syndrome, which very rarely affect humans. Several German scientists have expressed concern that BSE — popularly known as 'mad cow disease' because of the way it debilitates the brains of cattle — may be transmissible to humans who eat contaminated beef or take medicines made with ingredients from contaminated animals.

"The danger that BSE can be transmitted to humans is minimal or non-existent," said Professor Hans Kretzschmar from Gottingen University. "However, we do not know that it is non-existent. I personally think (British beef) should not be imported."

Contaminated British beef will be discussed at a meeting of EU health ministers on March 30, but a German official said that any decisions about a ban would be made by the union's agriculture ministers, who were likely to argue that existing safeguards were sufficient.

In 1992, the last year for which figures are available, Germany imported 2,092 tonnes of British beef — 2 per cent of all its beef imports from other EU countries — and 13 tonnes of veal.

The research ministry said that more than 100,000 cattle had died as a result of catching BSE in Britain. A further 50 cases of the disease had been recorded in Switzerland and there were two known cases in Germany, one of which affected a cow imported from Britain.

Termes-clés : cattle disease ; bovine spongiform encephalopathy ; bse ; mad cow disease ; British beef imports ; possible connections.

FIGURE 3.6 – Exemple de document de DUC. Les termes-clés soulignés sont ceux qui peuvent être extraits.

Corpus	Documents				Termes-clés			
	Langue	Genre	Quantité	Mots moy.	Annotateur	Quantité moy.	« À assigner »	Mots moy.
Termith								
↳ Linguistique	Français	Scientifique (résumé)	715	156.7	Professionnel	8,7	61,2 %	1,7
↳ Sciences de l'info.	"	"	706	119,7	"	8,5	67,6 %	1,7
↳ Archéologie	"	"	718	219,1	"	16,5	37,1 %	1,3
↳ Chimie	"	"	782	105,2	"	12,8	76,1 %	2,2
DEft	"	Scientifique	234	7 102,9	Auteur	5,3	19,4 %	1,7
Wikinews	"	Journalistique	100	308,5	Lecteur	9,6	7,6 %	1,7
Semeval	Anglais	Scientifique	244	5 152,3	Auteur / Lecteur	15,1	17,0 %	2,1
DUC	"	Journalistique	308	900,7	Lecteur	8,1	3,5 %	2,1

TABLE 3.13 – Bilan des corpus

3.8 Conclusion

Nous avons présenté les ressources que nous utilisons dans nos travaux de recherche. Pour analyser nos travaux, les évaluer et les comparer aux autres, nous disposons de cinq collections de données différentes : Termith, DEft, Wikinews, SemEval et DUC. Elles sont monolingues et uniformes en genre. Nous rappelons leurs propriétés dans le tableau 3.13.

4

Extraction de termes-clés

« [...] il reste encore des progrès à faire sur la tâche. »
— Kim *et al.* (2010)

4.1 Introduction

Dans ce chapitre, nous nous intéressons à la tâche d'extraction automatique de termes-clés. Cette tâche consiste à identifier dans un document ses mots et expressions qui permettent d'en caractériser le contenu. Elle peut être réalisée de manière non supervisée ou supervisée, grâce à la mise en œuvre d'algorithmes d'ordonnancement par importance des mots du document ou grâce à l'entraînement de classificateurs capables de déterminer si une unité textuelle est un terme-clé ou non.

Nous proposons deux contributions à l'extraction automatique de termes-clés. Tout d'abord, nous nous intéressons à l'étape préliminaire de sélection des termes-clés candidats, puis nous nous intéressons à leur ordonnancement par importance. Les méthodes proposées sont non supervisées.

4.2 Sélection des termes-clés candidats

La sélection des termes-clés candidats établit la liste des termes-clés possibles pour un document donné. Bien qu'étudiée en surface, ou de manière ad-hoc à une méthode particulière d'extraction de termes-clés, cette étape est critique. Si un nombre insuffisant de candidats est sélectionné, alors la performance maximale pouvant être atteinte pour l'extraction de termes-clés est faible. Inversement, si un nombre trop important de candidats est sélectionné, alors cela augmente la difficulté de l'extraction (Hasan et Ng, 2014). De nombreux travaux ont montré que les groupes nominaux, souvent approximés par les séquences de noms et d'adjectifs (`/(N|A)+/`), forment de bons termes-clés candidats et sont

très proches des termes-clés de référence (Barker et Cornacchia, 2000; Hulth, 2003; Wan et Xiao, 2008).

Dans notre travail, nous remettons en question la sélection systématique d'un adjectif apposé à un groupe nominal. En nous appuyant sur une analyse linguistique des termes-clés de trois collections de données en français et en anglais, nous proposons une méthode qui juge si un adjectif est utile, c'est-a-dire s'il apporte du sens bénéfique à la caractérisation du contenu du document. S'il est utile, alors il est sélectionné avec le groupe nominal qu'il modifie. Sinon, nous estimons qu'il est superflu et le groupe nominal seul est sélectionné comme terme-clé candidat. Deux évaluations montrent le bien fondé de cette méthode : l'une intrinsèque, l'autre extrinsèque. L'évaluation intrinsèque compare la qualité de l'ensemble de termes-clés candidats sélectionnés par notre méthode à ceux sélectionnés par les méthodes couramment utilisées ; l'évaluation extrinsèque compare l'impact de ces méthodes de sélection sur deux méthodes d'extraction de termes-clés.

4.2.1 Analyse des propriétés linguistiques des termes-clés

Afin de sélectionner plus finement les termes-clés candidats, nous extrayons et analysons des statistiques concernant les termes-clés : leur taille (en nombre de mots) et la catégorie grammaticale des mots qui les composent. Cela nous permet de confirmer les observations faites dans les travaux précédents et d'en inférer de nouvelles, axées sur la catégorie grammaticale des mots des termes-clés.

Notre analyse couvre les deux langues de nos ressources : français et anglais. Elle se porte sur les collections DEft (français), DUC (anglais) et SemEval (anglais). Pour ne pas influencer l'évaluation de notre travail, cette analyse est effectuée sur un sous-ensemble des collections. L'évaluation est réalisée sur les ensembles de test, donc l'analyse est réalisée sur les ensembles normalement destinés à l'apprentissage. DUC n'étant pas réparti en plusieurs sous-ensembles, nous utilisons 100 documents pour l'évaluation et les 208 restant pour l'analyse.

Analyse surfacique

Le tableau 4.1 montre la proportion de termes-clés uni-grammes, bi-grammes et tri-grammes, ainsi que la proportion de termes-clés multi-mots contenant au moins un mot appartenant à l'une des sept catégories grammaticales que nous observons en leur sein[1] : nom commun, nom propre, adjectif, verbe, adverbe, préposition et déterminant. Pour obtenir ces informations, les termes-clés ont été automatiquement segmentés en mots et étiquetés grammaticalement à l'aide des outils utilisés pour prétraiter les collections de données (voir la section 3.7, page 47), puis manuellement corrigés.

Concernant la taille des termes-clés de référence, les uni-grammes, bi-grammes et tri-grammes couvrent plus de 90 % des termes-clés de références. En français, ce sont les uni-grammes qui sont les plus utilisés, suivis par les bi-grammes, tandis qu'en anglais, ce sont les bi-grammes qui sont les plus employés, avec des proportions équivalentes d'uni-grammes et de tri-grammes. Ces premières observations font écho à celles que nous trouvons dans la littérature. Nous en concluons qu'il s'agit de propriétés stables des termes-clés.

[1] Nous nous focalisons sur les expressions (termes-clés multi-mots), car nous avons observé que la quasi totalité des termes-clés composés d'un unique mot sont des noms.

	DEft *(fr)*	**SemEval** *(en)*	**DUC** *(en)*
Taux (en %) de termes-clés :			
Uni-grammes	60,2	20,2	17,1
Bi-grammes	24,5	53,4	60,8
Tri-grammes	8,8	21,3	17,8
Taux (en %) de termes-clés contenant au moins un(e) :			
Nom commun	93,1	98,7	94,5
Nom propre	6,9	4,3	17,1
Adjectif	65,5	50,2	50,0
Verbe	1,0	4,0	1,0
Adverbe	1,3	0,7	1,6
Préposition	31,2	1,5	0,3
Déterminant	20,4	0,0	0,0

TABLE 4.1 – Statistiques des termes-clés de référence des collections DEft, SemEval et DUC

Une approche raisonnable peut donc se restreindre aux $\{1..3\}$-grammes, à l'instar de celle de Witten *et al.* (1999).

Concernant les catégories des mots que contiennent les termes-clés de référence, nous observons que la quasi-totalité des termes-clés contiennent un nom (ce sont majoritairement des groupes nominaux) et que la moitié d'entre eux est modifiée par un adjectif. Les autres catégories de mots, comme le verbe et l'adverbe sont très peu utilisées. L'usage de ces dernières au sein de termes-clés semble être exceptionnel. Les déterminants et prépositions ont un usage presque exclusivement français. En anglais, les modifications nominales (par exemple, *« nature conservation »* – « conservation de la nature ») sont préférées aux formes syntagmatiques (par exemple, *« conservation of nature »* – « conservation de la nature »).

Analyse des adjectifs

Après le nom et le nom propre, c'est l'adjectif qui est le plus utilisé. Nous analysons plus finement sa nature et examinons les trois catégories d'adjectifs suivantes : relationnel, composé complexe et qualificatif.

Un adjectif relationnel est un adjectif dénominal (Bally, 1944). Il est dérivé d'un nom (par exemple, l'adjectif relationnel « culturel » est dérivé du nom « culture ») pour lequel il établit une relation équivalente à celle exprimée par le complément du nom (par exemple, « héritage culturel » équivaut à « héritage de la culture »). Caractéristique du discours du spécialiste (Maniez, 2009), l'adjectif relationnel sert de modificateur dans les titres de catégories, telles que celles de Wikipedia (par exemple « héritage culturel »[2]), qui constituent de bons termes-clés candidats (Medelyan et Witten, 2008; Eichler et Neumann, 2010). Par transitivité, l'adjectif relationnel semble donc être un modificateur qui apporte du sens bénéfique à la caractérisation de tout ou partie du contenu d'un document.

Un adjectif composé complexe est un adjectif constitué de plusieurs mots, souvent délimités graphiquement par un trait d'union (par exemple, « socio-éducatif »). Il complexe

[2] `http://en.wikipedia.org/wiki/Category:Cultural_heritage`

contribue avec précision et concision à la caractérisation du nom qu'il modifie (par exemple, « activité socio-éducative » hyponyme de « activité »). Pour cette raison, nous pensons qu'il est utile pour caractériser tout ou partie du contenu d'un document. De plus, la composition adjectivale est l'un des processus privilégiés pour la formation de néologismes[3] (Boughedaoui, 1997).

Un adjectif qualificatif est un adjectif qui donne une qualification à un nom. Il désigne la qualité ou la manière d'être (par exemple, « grand »). Cette catégorie d'adjectifs est la plus courante. Nous faisons donc l'hypothèse qu'un adjectif appartenant à cette catégorie n'est pas toujours utile à la caractérisation du contenu d'un document.

Pour détecter les adjectifs relationnels, nous utilisons une technique simple, adaptée (ou adaptable) à plusieurs langues et ne requérant pas nécessairement de ressources linguistiques finies.

Dans un premier temps, les adjectifs relationnels sont détectés avec une base de données lexicales. Pour le français, nous utilisons la base WoNeF (Pradet *et al.*, 2013). Pour l'anglais, nous utilisons la base WordNet (Miller, 1995). WoNeF est issue de WordNet, ses entrées ont été obtenues par traduction de WordNet. Pour savoir si un adjectif est relationnel, nous utilisons la propriété `[PERTAINYM]` de WordNet et son équivalent `[DERIVED]` dans WoNeF.

Dans un second temps, les adjectifs relationnels qui ne sont pas présents dans la base de données lexicales sont détectés à l'aide de leur suffixe (Dubois et Dubois-Charlier, 1999). Une liste des suffixes les plus productifs pour les adjectifs relationnels est utilisée pour identifier les adjectifs relationnels potentiels. En français, les suffixes les plus productifs sont *-ain*, *-aire*, *-al*, *-el*, *-esque*, *-estre*, *-eux*, *-ien*, *-ier*, *-if*, *-il*, *-in*, *-ique*, *-ois*, et *-é* (Harastani *et al.*, 2013) ; en anglais, les suffixes utilisés sont *-al*, *-ant*, *-ary*, *-ic*, *-ous* et *-ive* (Grabar et Hamon, 2006).

La détection des adjectifs relationnels, telle que nous la réalisons, n'est pas exacte. En effet, les adjectifs qualificatifs et/ou dénominaux se terminant par un suffixe d'adjectif relationnel sont détectés comme relationnels (par exemple, « principal », « descriptif », et « contemporain »), et les adjectifs à usage tantôt qualificatif, tantôt relationnel selon le contexte (Maniez, 2009) sont toujours détectés comme relationnels (par exemple, « sulfureux » « civil » et « populaire »). Dans la littérature, les approches pour identifier les adjectifs relationnels (dans le cadre de l'extraction terminologique) reposent sur une analyse en corpus (Daille, 2000; Maniez, 2005; Harastani *et al.*, 2013), où il s'agit notamment de trouver des paraphrases avec un complément de nom. Dans le contexte de l'extraction de termes-clés, où de larges corpus ne sont pas toujours disponibles, les paraphrases ne sont pas toutes présentes et de telles approches ne sont pas applicables. De plus, Harastani *et al.* (2013) montrent qu'une approche comme la notre reste une alternative viable.

Pour déterminer si un adjectif est composé, nous regardons s'il possède un trait d'union. Le trait d'union est l'unique marque explicite de composition en français et en anglais, et son usage est le procédé le plus productif. Néanmoins, il existe deux autres procédés que nous ne traitons pas : la séparation avec un espace (par exemple, « vert clair ») et la concaténation sans marque explicite (par exemple, « ethnolinguistique »).

[3] Néologisme : mot nouveau, emprunt récent à une autre langue ou nouvelle emploi d'un mot déjà existant (nouvelle acception).

Le tableau 4.2 donne le taux d'adjectifs, par catégorie, dans les termes-clés de référence. Nous observons que la majorité de ces adjectifs sont relationnels, ce qui conforte notre hypothèse que les adjectifs relationnels sont des modificateurs utiles pour les termes-clés. Ceci est confirmé par le tableau 4.4 qui montre que l'un des patrons grammaticaux les plus productifs de termes-clés représente un nom modifié par un adjectif relationnel. Le cas des adjectifs composés complexes est moins marqué. Ils sont peu employés par rapport aux adjectifs relationnels et aux adjectifs qualificatifs. Ces derniers, quant à eux, ont un emploi non négligeable, en particulier en anglais : le troisième patron grammatical de termes-clés implique l'emploi d'un adjectif qualificatif (voir le tableau 4.4).

		DEft *(fr)*	**SemEval** *(en)*	**DUC** *(en)*
Adjectifs relationnels	(%)	87,1	43,6	53,1
Adjectifs composés complexes	(%)	3,3	16,4	10,6
Adjectifs qualificatifs	(%)	9,6	40,0	36,3

TABLE 4.2 – Taux d'adjectifs, par catégorie (relationnel, composé complexe ou qualificatif), au sein des termes-clés de référence

Le tableau 4.3 montre le taux d'adjectifs, par catégorie, dans les documents. En comparant les taux présentés dans le tableau 4.2 à ceux du tableau 4.3 nous pouvons déduire le degré d'ambiguïté d'un adjectif en tant que modificateur utile dans un terme-clé. Ainsi, ce tableau montre qu'il y a moins d'ambiguïtés quant à l'appartenance d'un adjectif relationnel ou composé à un terme-clé, car ces deux catégories d'adjectifs sont nettement moins utilisées dans les documents que dans les termes-clés de référence. À l'inverse, les adjectifs qualificatifs ont un très fort usage dans les documents et il y a donc plus d'ambiguïté quant à leur nécessité en tant que modificateur dans un terme-clé.

		DEft *(fr)*	**SemEval** *(en)*	**DUC** *(en)*
Adjectifs relationnels	(%)	61,9	30,7	29,9
Adjectifs composés complexes	(%)	0,4	7,9	8,8
Adjectifs qualificatifs	(%)	37,7	61,4	61,3

TABLE 4.3 – Taux d'adjectifs, par catégorie (relationnel, composé complexe ou qualificatif), au sein des documents

4.2.2 Sélection fine des termes-clés candidats

Pour sélectionner les termes-clés candidats, nous proposons une méthode exploitant les propriétés des termes-clés que nous avons observé. Cette méthode commence par présélectionner les termes-clés candidats à l'aide d'un patron grammatical, puis elle filtre les adjectifs qualificatifs jugés inutiles au sein des candidats.

Présélection des termes-clés candidats

L'étape de présélection des candidats utilise un patron grammatical défini sous la forme d'une expression rationnelle. Ce patron est appliqué aux catégories grammaticales des sé-

	Patron	Exemple	%
Français	`Nc Ar`	« concept linguistique »	46,4
	`Nc Sp D Nc`	« besoin de l'utilisateur »	12,5
	`Nc Sp Nc`	« analyse de discours »	8,2
	`Nc A`	« modèle prédictif »	4,3
	`Np Np`	« languedoc roussillon »	3,0
Anglais	`Nc Nc`	*« hurricane expert »* (« expert en ouragans »)	32,5
	`Ar Nc`	*« chinese earthquake »* (« tremblement de terre chinois »)	15,1
	`A Nc`	*« dominant strategy »* (« stratégie dominante »)	9,5
	`Nc Nc Nc`	*« voice activity detection »* (« détection d'activité vocale »)	5,3
	`Ac Nc`	*« packet-switched network »* (« réseau à commutation de paquets »)	4,9

TABLE 4.4 – Patrons grammaticaux les plus fréquents parmi les termes-clés français et anglais. Les classes grammaticales sont exprimées au format Multext (Ide et Véronis, 1994), sauf `Ar` et `Ac` qui représentent, respectivement, un adjectif relationnel et un adjectif composé.

quences de mots adjacents dans le document et sélectionne celles qui le respectent. Il est dit « gourmand », c'est-à-dire qu'il capture les plus longues séquences possibles sans produire de candidats qui se recouvrent dans le texte (comme c'est le cas avec les n-grammes). Il permet ainsi de diminuer le risque d'extraire des termes-clés redondants (Hasan et Ng, 2014).

D'après nos observations, seuls les noms, les adjectifs, les prépositions et les déterminants sont utiles pour sélectionner les termes-clés candidats en permettant une performance maximale quasi-optimale. Dans le cas de l'anglais, les prépositions et les déterminants sont en proportions très faibles et peuvent donc ne pas être considérés lors de la définition du patron grammatical. Dans le cas du français, les prépositions et les déterminants apparaissent au sein de plus de 30 % des termes-clés de référence. Notre étude se portant sur les adjectifs uniquement, nous faisons le choix de ne pas considérer ces deux classes grammaticales lors de la définition du patron de présélection des termes-clés candidats. Pour l'adjectif, nous nous appuyons sur les patrons grammaticaux les plus productifs de termes-clés du tableau 4.4 et décidons de limiter le nombre d'adjectifs à un pour le français et l'anglais[4].

Pour le français, nous définissons le patron `/N+ A?/`, qui accepte une séquence de noms (ou noms propres) se terminant optionnellement par un adjectif. En français, l'adjectif peut être soit antéposé, soit postposé. Le patron que nous avons défini n'accepte que les adjectifs postposés pour deux raisons. La première raison est que les adjectifs relationnels, que nous jugeons les plus utiles au sein des termes-clés, sont toujours postposés. La seconde raison est que l'adjectif antéposé ne fait pas partie des patrons les plus productifs de termes-clés[5] (voir le tableau 4.4).

Pour l'anglais, nous définissons le parton `/A? N+/`, qui accepte une séquence de noms (ou noms propres) modifiée optionnellement par un adjectif antéposé. En anglais, tous les adjectifs sont antéposés. Ce patron ne filtre donc aucun adjectif à cette étape de présélection.

[4] En français et en anglais, 3,3 % et 5,3 % des termes-clés contiennent plus d'un adjectif, respectivement.
[5] En français, seulement 0,7 % des termes-clés commencent par un adjectif.

Algorithme 1 : Sélection fine des termes-clés candidats

Entrée : document
Sortie : candidats

```
1  patron ← Nil
2  Si document.langue = "francais" alors
3  |  patron ← /N+ A?/
4  Sinon
5  |  Si document.langue = "anglais" alors
6  |  |  patron ← /A? N+/
7  candidats ← {}
8  candidats_preliminaires ← preselection(document, patron)
9  Pour chaque cdt ∈ candidats_preliminaires faire
10 |  Si ∃mot ∈ cdt, estAdjectif(mot) ∧ ¬estRelationnel(mot) ∧ ¬estCompose(mot) alors
11 |  |  tete_cdt ← cdt_{\{mot}}
12 |  |  freq_cdt ← document.conter(cdt)
13 |  |  freq_tete_cdt ← document.conter(tete_cdt) − freq_cdt
14 |  |  Si freq_cdt > freq_tete_cdt alors
15 |  |  |  candidats ← candidats ∪ {cdt}
16 |  |  Sinon
17 |  |  |  candidats ← candidats ∪ {tete_cdt}
18 |  Sinon
19 |  |  candidats ← candidats ∪ {cdt}
```

Filtrage des adjectifs superflus

Cette étapte juge, pour chaque terme-clés candidat contenant un adjectif, si l'adjectif est utile au sein du termes-clés candidat et le retire s'il ne l'est pas. Sur la base de notre analyse, les adjectifs relationnels et composés complexes sont systématiquement jugés utiles. Seuls les adjectifs qualificatifs font l'objet d'une prise de décision.

Pour décider si un adjectif qualificatif apporte du sens utile au groupe nominal qu'il modifie dans le terme-clé candidat, nous comparons les usages respectifs du candidat avec et sans l'adjectif. Notre intuition est qu'un adjectif est superflu (inutile) s'il modifie un groupe nominal qui est utilisé de manière autonome (sans l'adjectif) un nombre significatif de fois. Concrètement, si le groupe nominal occurre plus souvent sans l'adjectif qu'avec l'adjectif, alors ce dernier est jugé inutile et est retiré du terme-clé candidat.

L'algorithme 1 résume le fonctionnement de notre méthode de sélection des termes-clés candidats. Les lignes 1 à 8 concernent la présélection des candidats et les lignes 9 à 19 l'identification et le filtrage des adjectifs qualificatifs superflus.

4.2.3 Évaluation

Afin de montrer la validité de notre méthode de sélection de candidats, nous réalisons deux expériences : l'une intrinsèque, où les ensembles de termes-clés candidats de différentes méthodes de sélections sont comparés qualitativement, l'autre extrinsèque, où les différentes méthodes de sélection sont comparées d'après les performances de deux méthodes

d'extraction de termes-clés.

Méthodes de référence

Nous comparons notre méthode de sélection de termes-clés candidats a trois autres méthodes utilisées dans les travaux précédents en extraction automatique de termes-clés :

- sélection des n-grammes ($1 \leqslant n \leqslant 3$) ;
- sélection des plus longues séquences de noms et d'adjectifs : `/(N|A)+/` ;
- sélection des *NP-chunks* :
  - `/Np+|(A? Nc A+)|(A Nc)|Nc+/` en français ;
  - `/Np+|(A+ Nc)|Nc+/` en anglais.

Pour l'évaluation extrinsèque, nous utilisons deux méthodes d'extraction automatique de termes-clés très utilisées : la méthode non supervisée TF-IDF et la méthode supervisée KEA. Bien que des méthodes plus récentes donnent de meilleures performances que TF-IDF et KEA (Kim *et al.*, 2010), ces dernières présentent l'avantage d'être reproductibles, ce qui nous permet de les utiliser avec la même chaîne de prétraitements et avec les termes-clés candidats que nous souhaitons.

Collections de données

Pour évaluer ce travail, nous utilisons les ensembles de test de toutes nos collections de données, sauf Wikinews. Cette dernière collection ne possède pas d'ensemble d'entraînement et ne peut donc pas être utilisée pour la méthode de référence KEA.

Mesures d'évaluation

Pour évaluer la qualité des ensembles de termes-clés candidats sélectionnés par les différentes méthodes de sélection, nous comparons leur nombre moyen de candidats sélectionnés au rappel maximal (R_{max}) qu'ils permettent d'atteindre. Nous estimons que plus un ensemble de termes-clés candidats permet d'atteindre une performance maximale élevée (R_{max}) avec un nombre de candidats réduit, alors plus il est de bonne qualité. Pour cela, nous déterminons la qualité Q d'un ensemble de termes-clés candidats en faisant le rapport entre le rappel maximal et le nombre de candidats :

$$Q = \frac{R_{max}}{\text{Candidats}} \tag{4.1}$$

Plus Q est élevée, meilleure est la qualité de l'ensemble des termes-clés candidats sélectionnés.

Les performances des méthodes d'extraction de termes-clés sont exprimées en termes de précision (P), rappel (R) et f1-mesure (F). En accord avec l'évaluation menée dans les travaux précédents (Kim *et al.*, 2010), les opérations de comparaison entre les termes-clés de référence et les termes-clés extraits sont effectuées à partir de la racine des mots qui les composent. Pour cela, nous utilisons la méthode de Porter (1980)[6].

[6]Initialement proposée pour l'anglais, la méthode de Porter (1980) a été adaptée à d'autres langues (dont le français) dans le cadre du projet Snowball.

Méthode	**Linguistique** *(fr)*			**Sciences de l'information** *(fr)*			**Archéologie** *(fr)*			**Chimie** *(fr)*		
	Candidats	R_{max}	Q	Candidats	R_{max}	Q	Candidats	R_{max}	Q	Candidats	R_{max}	Q
n-grammes	88,2	**35,9**	0,41	94,4	**32,4**	0,43	135,0	**58,0**	0,43	63,5	**20.1**	0,32
/ (N | A) + /	31,5	25,1	0,80	34,5	24,2	0,70	48,0	43,5	0,91	22,4	16,5	0,74
NP-chunks	29,3	24,0	0,82	32,7	24,5	0,75	45.6	43,7	0.96	21,7	16,3	0.75
LR-NP	**28,5**	23,9	**0,84**	**31,8**	24,2	**0,76**	**44,2**	43,5	**0,98**	20,7	16,4	**0,79**

TABLE 4.5 – Résultats de l'évaluation intrinsèque des méthodes de sélection de termes-clés candidats appliquées aux données Termith

Méthode	**DEft** *(fr)*			**SemEval** *(en)*			**DUC** *(en)*		
	Candidats	R_{max}	Q	Candidats	R_{max}	Q	Candidats	R_{max}	Q
n-grammes	2610,4	**74,1**	0,03	1652,3	**71,7**	0,04	478,9	**90,4**	0,19
/ (N | A) + /	810,3	61,1	0,08	518,5	62,0	0,12	147,4	88,3	0,60
NP-chunks	736,5	63,0	**0,09**	478,1	56,3	0,12	141,4	75,6	0,54
LR-NP	**658,2**	60,1	**0,09**	**423,8**	59,0	**0,14**	**135,3**	84,8	**0,63**

TABLE 4.6 – Résultats de l'évaluation intrinsèque des méthodes de sélection de termes-clés candidats appliquées aux collections DEft, SemEval et DUC

Évaluation intrinsèque

L'évaluation intrinsèque a pour objectif d'évaluer la qualité de l'ensemble des termes-clés candidats sélectionnés par les méthodes de référence et de la comparer à celle de l'ensemble de termes-clés sélectionné par notre méthode (LR-NP, pour *Linguistically-Refined Noun Phrases*).

Les tableaux 4.5 et 4.6 présentent les résultats de l'évaluation intrinsèque. Nous y reportons le nombre de candidats sélectionnés par chaque méthode, le rappel maximal pouvant être atteint avec ceux-ci et leur qualité Q.

Globalement, notre méthode sélectionne le moins de candidats sans nécessairement induire le plus faible rappel maximal. C'est, sans surprise, la sélection des n-grammes qui induit le meilleur rappel maximal. Celui-ci est très proche du rappel maximal optimal[7], mais au prix d'un nombre de candidats 4 à 5 fois supérieur à celui des autres méthodes. Par ailleurs, ces derniers se recouvrent mutuellement, donc le risque d'extraire des termes-clés redondants est plus grand et la difficulté de la tâche d'extraction est donc plus élevée (Hasan et Ng, 2014). Comme le montre la valeur de Q, notre méthode est de meilleure qualité que les autres : LR-NP $>$ / (N | A) + / $>$ *NP-chunks* $>$ n-grammes.

Évaluation extrinsèque

L'évaluation extrinsèque a pour objectif d'évaluer l'efficacité de notre méthode de sélection de termes-clés en situation réelle d'extraction de termes-clés et de la comparer à celle des méthodes de référence. Il s'agit aussi de valider notre hypothèse qu'une sélection de candidats de meilleure qualité induit de meilleures performances lors de l'extraction.

Les tableaux 4.7 et 4.8 présentent les résultats de l'évaluation extrinsèque. Nous y reportons la performance des méthodes TF-IDF et KEA, en termes de précision, rappel et f1-mesure. Globalement, la performance des méthodes d'extraction de termes-clés est corrélée

[7]En extraction automatique de termes-clés, le rappel maximal optimal correspond au pourcentage de termes-clés qui occurrent dans les documents.

Méthode	**Linguistique** *(fr)*						**Sciences de l'information** *(fr)*						**archéologie** *(fr)*						**Chimie** *(fr)*					
	TF-IDF			KEA			TF-IDF			KEA			TF-IDF			KEA			TF-IDF			KEA		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
n-grammes	12,2	14,7	13,1	13,5	16,1	14,4	11,6	12,2	11,5	12,4	13,1	12,3	23,3	16,2	18,6	23,2	16,1	18,6	11,0	8,9	9,4	11,4	9,4	9,9
/ (N|A) +/	13,2	15,5	14,0	13,8	16,3	14,7	13,3	13,8	13,2	12,7	13,1	12,5	27,9	19,0	22,1	29,9	20,6	23,9	15,0	11,7	12,6	15,0	12,1	12,8
NP-chunks	**13,3**	**15,8**	**14,2**	14,0	16,5	14,9	**13,7**	**14,3**	**13,5**	**13,1**	**13,7**	**13,0**	**28,4**	**19,4**	**22,5**	29,9	20,7	23,9	15,3	12,0	12,9	15,0	12,0	12,7
LR-NP	**13,3**	**15,8**	**14,2**	**14,1**	**16,6**	**15,0**	13,5	14,2	13,4	12,7	13,2	12,5	28,2	19,2	22,3	**30,3**	**20,8**	**24,1**	**15,8**	**12,3**	**13,2**	**15,3**	**12,1**	**12,9**

TABLE 4.7 – Résultats de l'extraction de dix termes-clés avec TF-IDF et KEA sur les données Termith, selon la méthode de sélection des termes-clés candidats utilisée

Méthode	**DEft** *(fr)*						**SemEval** *(en)*						**DUC** *(en)*					
	TF-IDF			KEA			TF-IDF			KEA			TF-IDF			KEA		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
n-grammes	6,9	12,9	8,9	**15,5**	**29,1**	**20,0**	9,7	6,5	7,7	19,7	13,9	16,2	15,7	20,9	17,7	12,5	17,3	14,3
/ (N|A) +/	10,3	19,1	13,2	14,3	26,7	18,4	13,2	8,9	10,5	20,9	14,6	17,1	**24,5**	32,1	27,3	**14,7**	20,2	16,8
NP-chunks	10,0	18,6	12,8	14,5	27,2	18,7	13,3	9,0	10,6	20,6	14,5	16,9	21,4	28,5	24,1	13,0	19,0	15,7
LR-NP	**10,4**	**19,1**	**13,3**	14,8	28,0	19,2	**13,6**	**9,3**	**10,9**	**21,6**	**15,2**	**17,7**	**24,5**	**32,3**	**27,4**	**14,7**	**20,4**	**16,9**

TABLE 4.8 – Résultats de l'extraction de dix termes-clés avec TF-IDF et KEA sur DEft, SemEval et DUC, selon la méthode de sélection des termes-clés candidats utilisée

à la qualité de l'ensemble des termes-clés candidats sélectionnés. Les candidats sélectionnés avec notre méthode induisent les meilleures performances dans la quasi-totalité des cas de figure étudiés.

4.2.4 Bilan

Nous avons proposé une méthode de sélection des termes-clés candidats d'un document. Développée à l'issue d'une analyse des propriétés linguistiques des termes-clés de référence de trois collections de données, cette méthode présélectionne des termes-clés candidats composés uniquement de noms et d'au plus un adjectif, puis détermine si chaque adjectif apporte du sens selon sa catégorie (relationnel, composé complexe ou qualificatif) et son usage dans le document. Vis-à-vis des méthodes de sélection de termes-clés candidats les plus utilisées, celle que nous proposons présente l'avantage de sélectionner moins de candidats sans réduire significativement le nombre de termes-clés corrects qui s'y trouvent. Les méthodes d'extraction de termes-clés sont aussi plus performantes avec les candidats qu'elle sélectionne. La qualité de l'ensemble de candidats proposés est donc meilleure.

Nous nous sommes principalement intéressés aux adjectifs relationnels, qui constituent une sous-partie des adjectifs dénominaux. À l'avenir, il serait intéressant d'élargir notre étude à tous les adjectifs dérivés : dénominaux et déverbaux. De même, nous devrions élargir notre étude aux prépositions et déterminants pour le français.

4.3 Extraction non supervisée de termes-clés

L'extraction non supervisée de termes-clés consiste, le plus souvent, à ordonner les termes-clés candidats, ou leurs mots, selon leur importance au sein du document. Actuellement, l'approche la plus étudiée pour cela est l'approche à base de graphe. Celle-ci représente le document avec un graphe de cooccurrences de mots et les ordonne par importance avec un algorithme qui simule le concept de recommandation (ou de vote). Les termes-clés sont ensuite générés à partir des séquences de mots les plus importants (mots-clés), ou extraits

à partir des termes-clés candidats ordonnés selon la somme du score d'importance de leurs mots.

Dans notre travail, nous remettons en question l'ordonnancement des mots, plutôt que des termes-clés candidats. Nous émettons aussi l'hypothèse que ce n'est pas l'importance des candidats qui doit être déterminée, mais celle de ce qu'ils représentent. Nous parlons de sujets. Par ailleurs, certains candidats peuvent représenter le même sujet. Nous proposons une nouvelle méthode à base de graphe, TopicRank, qui se fonde sur cette hypothèse.

4.3.1 TopicRank

TopicRank est une méthode à base de graphe pour extraire des termes-clés représentant chacun un sujet important dans le document. Elle repose sur quatre étapes : identification des sujets, construction d'un graphe de sujets, ordonnancement des sujets et sélection du terme-clé candidat le plus représentatif de chaque sujet.

Identification des sujets

Un sujet représente un concept (ou une idée), véhiculé par une ou plusieurs unités textuelles du document. Dans TopicRank, les unités textuelles qui véhiculent les sujets sont les termes-clés candidats. L'identification des sujets consiste donc à les grouper lorsqu'ils sont supposés appartenir au même sujet. Afin de proposer une méthode générique n'utilisant pas de données supplémentaires, nous appliquons un groupement « naïf » des candidats, fondé sur les mots qu'ils partagent : leur similarité lexicale.

Deux candidats c_1 et c_2 sont considérés comme des ensembles de mots (sacs de mots) et leur degré de similarité est calculé à l'aide de la mesure de Jaccard (voir l'équation 4.2), de sorte qu'ils soient très similaires s'ils partagent un grand nombre de mots. À l'instar des systèmes d'évaluation automatique, nous ajoutons plus de souplesse à la mesure de similarité en tronquant les mots avec la méthode de racinisation de Porter (1980).

$$\text{sim}(c_1, c_2) = \frac{|\text{Porter}(c_1) \ \cap \ \text{Porter}(c_2)|}{|\text{Porter}(c_1) \ \cup \ \text{Porter}(c_2)|} \tag{4.2}$$

Cette mesure est « naïve », car l'ordre des mots, leur ambiguïté et leur synonymie ne sont pas pris en compte. À cela s'ajoute aussi des erreurs introduites par l'usage de la méthode de Porter (1980) (par exemple les mots « empire » et « empirique » partagent le même radical « empir »).

Le groupement des termes-clés candidats en sujets est effectué avec l'algorithme de groupement hiérarchique agglomératif (*Hierarchical Agglomerative Clustering* – HAC). L'algorithme 2 décrit le fonctionnement classique du groupement HAC. Initialement, chaque candidat représente un groupe et, jusqu'à l'obtention d'un nombre prédéfini de groupes (nb_groupes), ceux qui ont la plus forte similarité (groupe_sim) sont unis pour ne plus former qu'un seul groupe. Afin de ne pas fixer le nombre de sujets (groupes) à identifier comme condition d'arrêt de l'algorithme, nous définissons un seuil de similarité ζ devant être dépassé ou égalé afin de pouvoir unifier deux groupes. La similarité groupe_sim entre deux groupes est déterminée à partir de la similarité de Jaccard calculée entre tous les candidats

de chaque groupe. Il existe trois stratégies pour la calculer :

$$\text{groupe\_sim}_{\text{simple}}(g_1, g_2) = \max_{c_1 \in g_1, c_2 \in g_2} \text{sim}(c_1, c_2) \tag{4.3}$$

$$\text{groupe\_sim}_{\text{complète}}(g_1, g_2) = \min_{c_1 \in g_1, c_2 \in g_2} \text{sim}(c_1, c_2) \tag{4.4}$$

$$\text{groupe\_sim}_{\text{moyenne}}(g_1, g_2) = \frac{\sum_{c_1 \in g_1} \sum_{c_2 \in g_2} \text{sim}(c_1, c_2)}{|g_1| \times |g_2|} \tag{4.5}$$

La stratégie simple favorise le rappel : les groupes contiennent théoriquement le plus grand nombre de candidats véhiculant effectivement le même sujet, mais aussi un nombre potentiellement élevé d'intrus. À l'inverse, la stratégie complète favorise la précision : les groupes contiennent théoriquement moins d'intrus que la stratégie simple, mais ils ne sont pas exhaustifs. La stratégie moyenne est le compromis entre les deux première : les groupes contiennent potentiellement moins d'intrus que ceux obtenus avec les stratégie simple et sont théoriquement plus exhaustifs que ceux obtenus avec la stratégie complète.

Algorithme 2 : HAC

Entrée : candidats $= \{c_1, \ldots, c_n\}$
Entrée : nb_groupes
Sortie : groupes

1 groupes $\leftarrow \{\{c_1\}, \ldots, \{c_n\}\}$
2 **Tant que** |groupes| > nb_groupes **faire**
3 $\quad \langle g_i, g_j \rangle \leftarrow \arg\max_{g_i, g_j \in \text{groupes}, g_i \neq g_j} \text{groupe\_sim}(g_i, g_j)$
4 $\quad \text{groupes} \leftarrow \text{groupes}_{\setminus \{g_i, g_j\}} \cup \{g_i \cup g_j\}$

Construction du graphe

Soit le graphe complet et non orienté $G = (N, A)$, composé d'un ensemble de nœuds N et d'arêtes[8] A. Les nœuds du graphe représentent les sujets du document et les arêtes qui les connectent représentent la force de leur lien sémantique dans le document. Contrairement aux travaux précédent, nous ne souhaitons pas utiliser de fenêtre de cooccurrences et n'exprimons donc pas la force du lien sémantique entre deux sujets par le nombre de co-occurrences entre leurs candidats respectifs. Pour préserver l'intuition derrière l'usage du nombre de cooccurrences, nous interconnectons tous les nœuds et exprimons la force de leur lien sémantique en fonction de la distance (en nombre de mots) qui les sépare dans le document :

$$\text{poids}(n_i, n_j) = \sum_{c_i \in n_i} \sum_{c_j \in n_j} \text{dist}(c_i, c_j) \tag{4.6}$$

$$\text{dist}(c_i, c_j) = \sum_{p_i \in \text{pos}(c_i)} \sum_{p_j \in \text{pos}(c_j)} \frac{1}{|p_i - p_j|} \tag{4.7}$$

où $\text{poids}(n_i, n_j)$ est le poids de l'arête entre les sujets n_i et n_j, et où $\text{dist}(c_i, c_j)$ représente la force sémantique entre les candidats c_i et c_j, calculée à partir de toutes leurs positions respectives, $\text{pos}(c_i)$ et $\text{pos}(c_j)$, dans le document.

[8] $A = \{(n_1, n_2) \mid \forall n_1, n_2 \in N, n_1 \neq n_2\}$, car G est un graphe complet.

Ordonnancement des sujets

L'ordonnancement des sujets doit établir un ordre d'importance des sujets dans le document. Pour cela, nous appliquons l'algorithme d'ordonnancement de SingleRank (Wan et Xiao, 2008) à notre graphe de sujets. Cet algorithme se fonde sur le principe de recommandation (ou du vote). Un sujet est d'autant plus important s'il est fortement connecté avec un grand nombre de sujets et si les sujets avec lesquels il est fortement connecté sont importants :

$$S(n_i) = (1-\lambda) + \lambda \times \sum_{n_j \in A(n_i)} \frac{\text{poids}(n_j, n_i) \times S(n_j)}{\sum_{n_k \in A(n_j)} \text{poids}(n_i, n_j)} \tag{4.8}$$

où $A(n_i)$ est l'ensemble des sujets[9] connectés au sujet n_i et où λ est un facteur d'atténuation. Défini entre 0 et 1, ce dernier peut être considéré comme la probabilité pour que le sujet n_i soit utilisé par recommandation. Nous suivons Brin et Page (1998) et fixons λ à 0,85.

Sélection des termes-clés

La sélection des termes-clés est la dernière étape de TopicRank. Elle consiste à chercher les termes-clés candidats qui représentent le mieux les sujets importants. Dans le but de ne pas extraire de termes-clés redondants, un seul candidat est sélectionné par sujet. Ainsi, pour k sujets, k termes-clés non redondants couvrant théoriquement k sujets sont extraits.

La difficulté de cette étape réside dans la capacité à trouver parmi plusieurs termes-clés candidats d'un même sujet celui qui le représente le mieux. Nous proposons trois stratégies de sélection pouvant répondre à ce problème :

- position : en supposant qu'un sujet est tout d'abord introduit par sa forme la plus appropriée, le terme-clé candidat sélectionné pour un sujet est celui qui apparaît en premier dans le document ;
- fréquence : en supposant que la forme la plus représentative d'un sujet est sa forme la plus fréquente, le terme-clé candidat sélectionné pour un sujet est celui qui est le plus fréquent dans le document ;
- centroïde : le terme-clé candidat sélectionné pour un sujet est celui qui est le plus similaire aux autres candidats du sujet.

La stratégie position est liée au trait « première position » utilisé en extraction supervisée de termes-clés. Ce trait ayant montré son efficacité et sa fiabilité (Lim *et al.*, 2012), la stratégie position est un très bon candidat à la sélection du terme-clé de chaque sujet. Dans le chapitre 3 (page 37), nous évoquons le fait que les documents dans lesquels il y a des changements thématiques invalident l'usage de la première position. Cela n'est pas vrai dans notre situation, car la stratégie position est appliquée à des groupes représentant un seul sujet.

La stratégie fréquence est aussi un bon candidat à la sélection du terme-clé de chaque sujet. Intuitivement, l'unité textuelle la plus utilisée d'un sujet est plus probablement sa forme préférée. Cependant, nous pensons que cette stratégie est moins généralisable à tout type

[9] $A(n_i) = \{n_j \mid \forall n_j \in N, n_j \neq n_i\}$, car G est un graphe complet.

de document et style discursif que la stratégie position. Certains auteurs, par exemple, préfèrent utiliser des unités textuelles concises et moins précise que la forme préférée qui rend la lecture moins agréable lorsqu'elle est répétée. Dans l'exemple de notice de linguistique Termith présentée dans la figure 3.1 (page 41), nous pouvons citer le termes-clés « concept linguistique », qui est simplifié par « concept ».

La stratégie centroïde, contrairement aux deux autres, n'est pas fondée sur l'usage des termes-clés candidats dans le document. Elle se concentre uniquement sur les mots qu'ils partagent pour déterminer le candidat qui est le cœur de tous les autres. Les termes-clés qu'elle sélectionne sont donc ni trop spécifiques, ni trop généraux. Certains termes-clés pouvant être très spécifiques (par exemple, « concept linguistique » au lieu de « concept » pour le document de linguistique de la figure 3.1, page 41) et d'autres plus généraux (par exemple, « cause » au lieu de « cause linguistique » pour le document de linguistique de la figure 3.1, page 41), cette stratégie semble être la moins adaptée des trois.

Exemple

La figure 4.1 (page 63) donne un exemple d'extraction de termes-clés avec TopicRank à partir de l'articles journalistiques de la collection DUC présenté dans la figure 3.6 (page 48). Dans cet exemple, nous observons un groupement correct de toutes les variantes de « alertes », mais aussi un groupement erroné de « août 2003 » avec « août 2012 ». Dans ce dernier cas, TopicRank est tout de même capable d'extraire « août 2012 » grâce à la sélection du candidat apparaissant en premier. Globalement, l'extraction des termes-clés est correcte et huit des dix termes-clés extraits sont corrects. Comparée à TF-IDF, TextRank et SingleRank, TopicRank est la méthode la plus performante pour ce document.

4.3.2 Évaluation

Pour valider notre approche, nous réalisons deux série d'expériences. Une première série pour déterminer les paramètres de TopicRank et une seconde série pour le comparer aux travaux précédents et analyser l'impact de chacune de nos contributions.

Méthodes de référence

Dans nos expériences, nous comparons TopicRank à trois autres méthodes non supervisées d'extraction automatique de termes-clés. Nous choisissons la méthode TF-IDF et les deux méthodes à base de graphe TextRank et SingleRank[10].

Dans le but de comparer TopicRank à SingleRank avec ses performances observées dans la littérature (Hasan et Ng, 2010), nos expériences sont d'abord réalisées lorsque les termes-clés candidats sont sélectionnés avec le patron grammatical `/(N|A)+/`. Après cette comparaison nous étudions le comportement de TopicRank selon la méthode de sélection des candidats utilisée, comme nous l'avons fait dans la section 4.2 (page 49).

Mesures d'évaluation

Les performances des méthodes d'extraction de termes-clés sont exprimées en termes de précision (P), rappel (R) et f1-mesure (F). En accord avec l'évaluation menée dans les tra-

[10]Toutes les méthodes sont implémentées par nos soins et intégrées à la même chaîne de traitement.

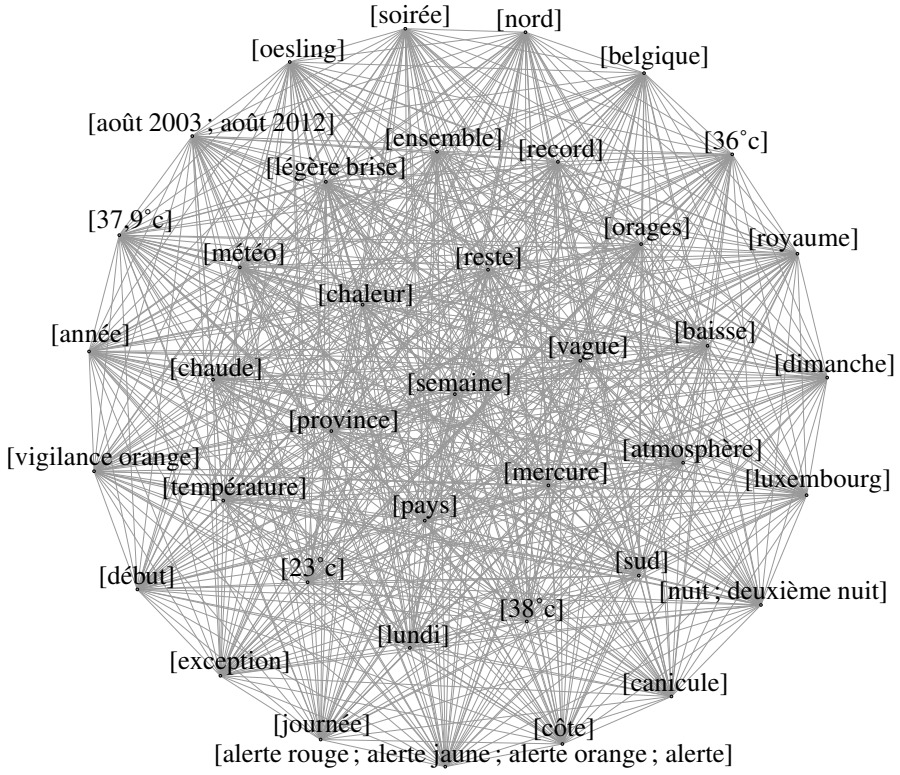
Météo du 19 août 2012 : alerte à la canicule sur la Belgique et le Luxembourg

A l'exception de la province de Luxembourg, en alerte jaune, l'ensemble de la Belgique est en vigilance orange à la canicule. Le Luxembourg n'est pas épargné par la vague du chaleur : le nord du pays est en alerte orange, tandis que le sud a était placé en alerte rouge.

En Belgique, la température n'est pas descendue en dessous des 23°C cette nuit, ce qui constitue la deuxième nuit la plus chaude jamais enregistrée dans le royaume. Il se pourrait que ce dimanche soit la journée la plus chaude de l'année. Les températures seront comprises entre 33 et 38°C. Une légère brise de côte pourra faiblement rafraichir l'atmosphère. Des orages de chaleur sont a prévoir dans la soirée et en début de nuit.

Au Luxembourg, le mercure devrait atteindre 32°C ce dimanche sur l'Oesling et jusqu'à 36°C sur le sud du pays, et 31 à 32°C lundi. Une baisse devrait intervenir pour le reste de la semaine. Néanmoins, le record d'août 2003 (37,9°C) ne devrait pas être atteint.

Termes-clés de référence : luxembourg ; alerte ; météo ; belgique ; août 2012 ; chaleur ; température ; chaude ; canicule ; orange ; la plus chaude.



Sortie de TopicRank : luxembourg ; alerte ; nuit ; belgique ; août 2012 ; chaleur ; température ; chaude ; canicule ; dimanche.

Sortie de TF-IDF : luxembourg ; belgique ; alerte ; canicule ; chaude ; nuit ; chaleur ; sud ; dimanche ; deuxième nuit.
Sortie de TextRank : août 2012 ; août 2003 ; alerte orange ; vigilance orange ; deuxième nuit ; légère brise.
Sortie de SingleRank : alerte orange ; alerte jaune ; alerte rouge ; alerte ; deuxième nuit ; août 2012 ; août 2003 ; vigilance orange ; légère brise ; luxembourg.

FIGURE 4.1 – Exemple d'extraction de termes-clés avec TopicRank, comparé à TF-IDF, TextRank et SingleRank, sur un article journalistique de Wikinews. Les termes-clés soulignés sont les termes-clés correctement extraits.

vaux précédents, nous considérons correcte l'extraction d'une variante flexionnelle d'un terme-clé de référence (Kim *et al.*, 2010), les opérations de comparaison entre les termes-clés de référence et les termes-clés extraits sont donc effectuées à partir de la racine des mots qui les composent. Nous utilisons la méthode de Porter (1980).

Analyse empirique de TopicRank

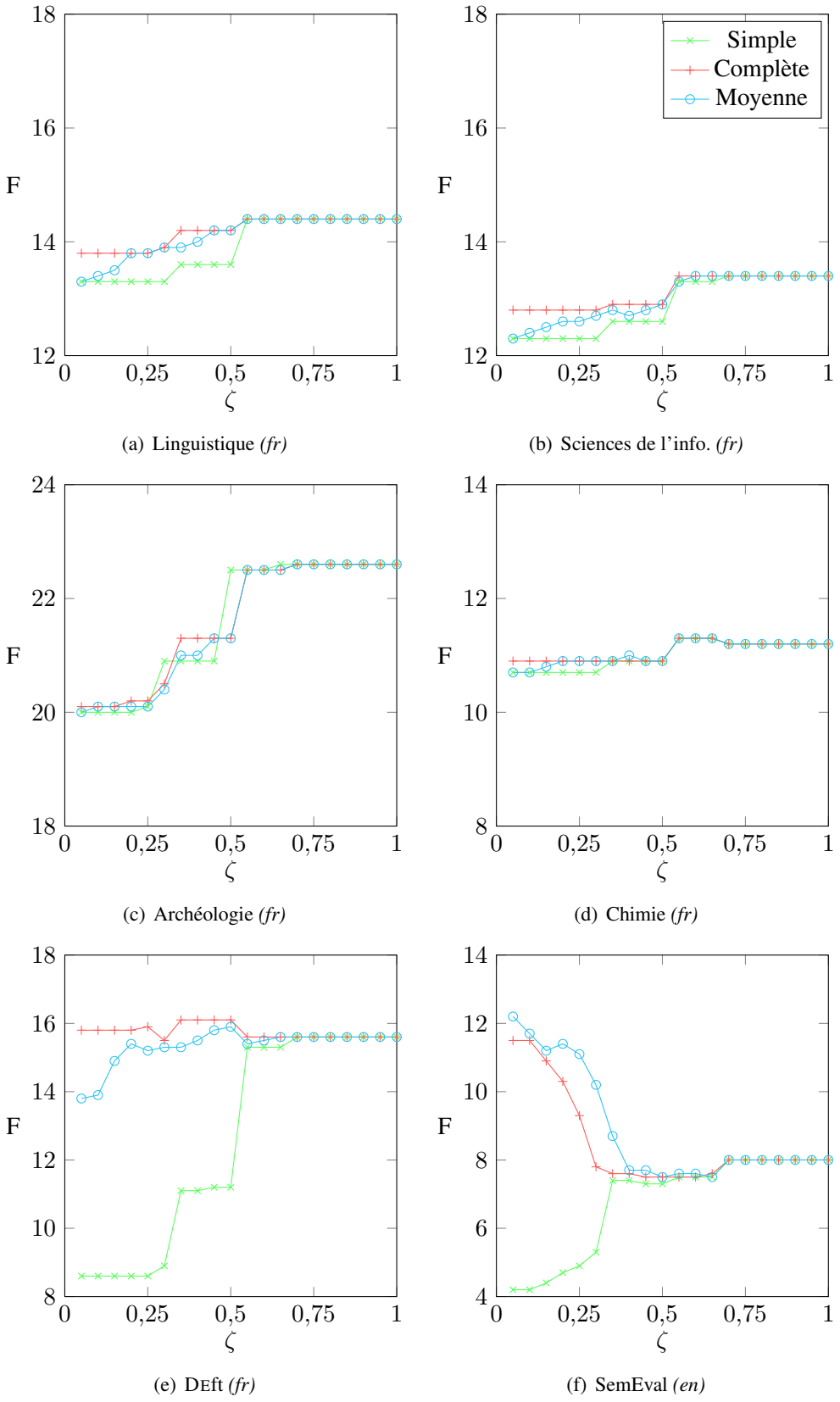
Dans cette section, nous effectuons une première série d'expériences afin de déterminer la configuration la plus générique de TopicRank. En utilisant les ensembles d'entraînement des collections Termith, de DEft, de SemEval et de DUC, nous réalisons deux expériences dans lesquelles nous faisons varier le seuil de similarité (ζ) avec la stratégie de groupement (simple, complète et moyenne), puis la stratégie de sélection du terme-clé candidat le plus représentatif de chaque sujet.

La figure 4.2 présente les résultats de TopicRank lorsque nous faisons varier le seuil ζ avec un pas de 0,05 pour chaque stratégie de groupement[11]. Avec les collections Termith, nous observons des comportements et des performances similaires quelque soit la valeur du seuil ζ et la stratégie de groupement utilisée. La petite taille des documents fait que très peu de termes-clés candidats sont groupés et les performances évoluent peu jusqu'à stabilisation lorsque ζ vaut 0,55. Avec DEft et SemEval, nous observons que chaque stratégie de groupement a un comportement qui lui est propre jusqu'à un point de convergence lorsque ζ vaut 0,70. Ce point de convergence correspond à la situation où les sujets créés sont les mêmes quelque soit la stratégie. Avec la stratégie simple, les résultats s'améliorent lorsque ζ augmente. En effet, elle prend en compte la similarité maximale entre les candidats de deux groupes, donc elle à tendance à trop grouper (à créer des groupes représentant plusieurs sujets) lorsque ζ est faible et à mieux grouper lorsque ζ augmente. La stratégie complète ayant le fonctionnement contraire, ses résultats se dégradent au fur et à mesure que ζ augmente. Enfin, la stratégie moyenne agit en compromis. Pour DEft, son comportement est le même que celui de la stratégie simple, mais ses résultats sont très supérieurs jusqu'au point de convergence. Pour SemEval, son comportement est le même que celui de la stratégie complète, mais ses résultats sont supérieurs jusqu'au point de convergence.

Dans la suite de nos expériences, nous ne reportons pas les résultats avec la meilleure configuration pour chaque collection. À la place, nous proposons la configuration suivante par défaut : le terme-clé de chaque sujet est sélectionné avec la stratégie moyenne et le seuil ζ est fixé à 0,25, c'est-à-dire que deux termes-clés candidats sont groupés s'ils ont au moins $1/4$ des mots en commun.

La figure 4.3 présente les résultats obtenus avec TopicRank et les différentes stratégies de sélection du terme-clé de chaque sujet. Dans la majorité des cas, la stratégie fréquence donne les meilleures performances, suivie par la stratégie position, qui donne des résultats compétitifs. Néanmoins, la performance de la stratégie fréquence obtenue sur SemEval montrent que cette dernière n'est pas stable. À l'échelle d'articles de plusieurs pages, où anaphores et autres figures rhétoriques sont nombreuses, sélectionner le candidat le plus fréquent n'est pas pertinent. Les résultats montrent donc que la stratégie la plus fiable pour sélectionner le

[11] La stratégie de sélection du terme-clé le plus représentatif par sujet utilisée dans cette expérience est la stratégie position.



(a) Linguistique *(fr)*

(b) Sciences de l'info. *(fr)*

(c) Archéologie *(fr)*

(d) Chimie *(fr)*

(e) DEft *(fr)*

(f) SemEval *(en)*

FIGURE 4.2 – Résultats de l'extraction de dix termes-clés avec TopicRank, en fonction de la stratégie de regroupement et de la valeur du seuil de similarité ζ

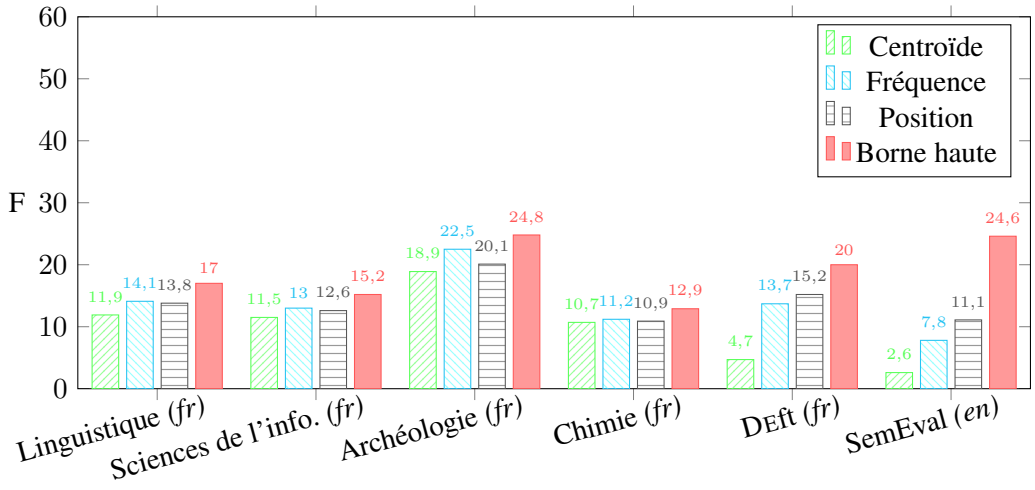


FIGURE 4.3 – Résultats de l'extraction de dix termes-clés, avec TopicRank, en fonction des différentes stratégies de sélections d'un terme-clé candidats par sujet

terme-clé de chaque sujet est la stratégie position. Par ailleurs, la borne haute obtenue par un oracle sélectionnant toujours un candidat positif, lorsque c'est possible, montre que cette stratégie donne des performances quasi-optimales.

Dans la suite de nos expériences, le terme-clé de chaque sujet est donc sélectionné avec la stratégie position.

Paramétrage empirique de SingleRank

Contrairement aux autres méthodes de référence, SingleRank possède un paramètre qui est définit arbitrairement : la fenêtre de cooccurrences fixée à dix par Wan et Xiao (2008). De même que pour TopicRank, nous utilisons les ensembles d'entrainement des collections Termith, de DEft et de SemEval pour déterminer qu'elle est la valeur optimale de la fenêtre de cooccurrences pour SingleRank[12].

La figure 4.4 présente les résultats de SingleRank lorsque nous faisons varier la fenêtre de cooccurrences de deux à vingt mots, avec un pas de un. Globalement, nous observons une stabilité des performances de SingleRank lorsque la fenêtre dépasse cinq mots. Les résultats montrent que la valeur de la fenêtre fixée à dix par Wan et Xiao (2008) est effectivement l'une des meilleures valeurs. Dans les expériences suivantes, nous utilisons donc la valeur recommandée par Wan et Xiao (2008).

Comparaison de TopicRank avec l'existant

Les tableaux 4.9 et 4.10 montrent les performances de TopicRank comparées à celles des trois méthodes de référence. De manière générale, les performances des méthodes d'extraction de termes-clés sont basses. Il est avéré que les documents de grande taille, tels que

[12]Nous ne répétons pas cette expérience pour TextRank, car le critère d'adjacence (fenêtre de valeur 2) est un critère fort dans la méthode TextRank.

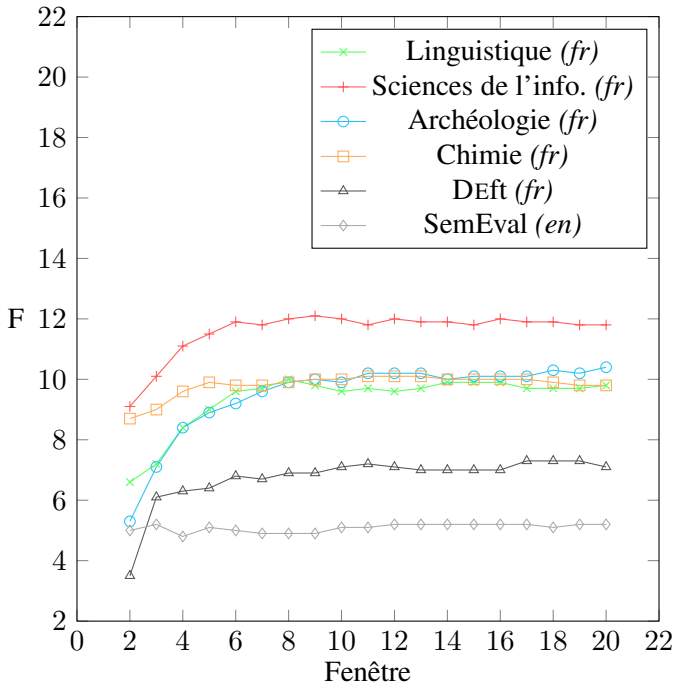


FIGURE 4.4 – Résultats de l'extraction de dix termes-clés, avec SingleRank, en fonction de la fenêtre de cooccurrences

ceux de SemEval et de DEft, sont plus difficiles à traiter que les autres documents. Hasan et Ng (2014) explique qu'un grand nombre de termes-clés candidats sont sélectionnés dans ces documents (ils sont en moyenne 647 pour SemEval et 915 pour DEft), l'espace de recherche est plus grand et la difficulté de l'extraction de termes-clés est donc plus élevée. Le cas des données Termith est encore plus particulier. En effet, elles sont constituées de documents courts et les méthodes d'extraction de termes-clés devraient donc obtenir de meilleures performances, mais 37 à 76 % de leurs termes-clés n'occurrent pas dans les documents et ne peuvent donc pas être extraits. Globalement, TopicRank est plus performant que les méthodes de référence à base de graphe et confirme donc que le groupement des candidats permet de rassembler des informations pour améliorer la précision de l'ordonnancement. Comparée à la méthode TF-IDF, TopicRank donne aussi de meilleurs résultats, pour les collections DEft, Wikinews et SemEval. Cette supériorité vis-à-vis de TF-IDF est importante à noter, car cette méthode obtient de bons résultats en tirant parti de statistiques extraites de documents supplémentaires, alors que TopicRank utilise le document seul.

De nouveau, les performances de TopicRank sur les collections Termith ne sont pas en adéquation avec celles obtenues sur les autres collections de données. TopicRank est bien plus performant que les autres méthodes à base de graphe, mais il ne fait pas mieux que TF-IDF. Notre hypothèse est que la nature très spécifique des données Termith (domaines de spécialité) permet à TF-IDF de mieux détecter les termes-clés candidats spécifiques au document grâce aux statistiques recueillies.

Dans le but de confirmer la pertinence de tous les apports de TopicRank, nous réalisons une expérience supplémentaire dans laquelle nous appliquons individuellement à SingleRank

Méthode	**Linguistique** *(fr)*			**Sciences de l'info.** *(fr)*			**Archéologie** *(fr)*			**Chimie** *(fr)*		
	P	R	F	P	R	F	P	R	F	P	R	F
TF-IDF	**13,0**	**15,4**	**13,9**	**13,4**	**14,0**	**13,2**	**28,1**	**19,1**	**22,2**	**14,1**	**11,1**	**11,9**
TextRank	7,1	6,1	6,4	5,8	4,3	4,8	10,2	5,3	6,8	9,4	5,3	6,5
SingleRank	9.0	10,6	9,6	9,5	10,0	9,4	12,7	8,9	10,2	13,0	10,4	11,0
TopicRank	11,2	13,1	11,9	12,1	12,8	12,1	27,5	18,7	21,8	13,8	11,1	11,8
Borne haute	**14,5**	**17,0**	**15,4**	**15,0**	**15,6**	**14,9**	**32,5**	**22,2**	**25,8**	**15,8**	**12,5**	**13,3**

TABLE 4.9 – Résultats de l'extraction de dix termes-clés avec TF-IDF, TextRank, SingleRank et TopicRank sur les données Termith

Méthode	**DEft** *(fr)*			**Wikinews** *(fr)*			**SemEval** *(en)*			**DUC** *(en)*		
	P	R	F	P	R	F	P	R	F	P	R	F
TF-IDF	10,3	19,1	13,2	33,9	35,9	34,3	13,2	8,9	10,5	**23,8**	**30,7**	**26,4**
TextRank	4,9	7,1	5,7	9,3	8,3	8,6	7,9	4,5	5,6	4,9	5,4	5,0
SingleRank	4,5	9,0	5,9	19,4	20,7	19,7	4,6	3,2	3,7	22,3	28,4	24,6
TopicRank	**11,7**	**21,7**	**15,1**†	**35,0**	**37,5**	**35,6**†	**14,9**	**10,3**	**12,1**†	18,3	23,8	20,4
Borne haute	**14,5**	**27,0**	**18,7**	**41,8**	**44,1**	**42,2**	**30,0**	**20,7**	**24,3**	**30,5**	**38,7**	**33,7**

TABLE 4.10 – Résultats de l'extraction de dix termes-clés avec TF-IDF, TextRank, SingleRank et TopicRank sur les collections DEft, Wikinews, SemEval et DUC. † indique une amélioration significative de TopicRank vis-à-vis de TextRank et SingleRank, à 0,001 pour le t-test de Student.

toutes les modifications successives permettant d'obtenir la méthode TopicRank depuis la méthode SingleRank : l'usage d'un graphe complet (+ complet), la projection des termes-clés candidats dans le graphe (+ candidats) et la projection des sujets dans le graphe (+ sujets). Les résultats de ces trois variantes de SingleRank sont présentés dans les tableaux 4.11 et 4.12. Globalement, l'usage des termes-clés candidats et des sujets induit une amélioration des performances de SingleRank, avec une amélioration plus importante en utilisant les sujets. Cela confirme la pertinence d'ordonner directement les candidats, plutôt que les mots, ainsi que la pertinence de grouper les candidats représentant le même sujet pour mutualiser les relations qu'ils entretiennent avec les candidats représentant d'autres sujets. Dans le cas des collections Termith, nous observons que le groupement des candidats est moins efficace que l'utilisation des candidats seuls. Toutefois, la combinaison du groupement avec le graphe complet et la nouvelle pondération des arêtes pallie ce défaut. L'usage d'un graphe complet, quant à lui, n'améliore pas significativement les résultats de SingleRank. Ceux-ci sont équivalents à ceux obtenus en construisant un graphe de cooccurrences, mais nous pensons que l'usage du graphe complet est à privilégier afin d'éviter d'avoir à fixer le paramètre de la fenêtre de cooccurrences.

Sélection des candidats pour TopicRank

Nous reprenons ici les expériences réalisées dans la section 4.2 (page 49) à propos de la sélection des termes-clés candidats. Les tableaux 4.13 et 4.14 montrent les performances obtenues par TopicRank utilisé avec les quatre méthodes de sélection de termes-clés candidats : n-grammes, `/(N|A)+/`, *NP-chunks* et LR-NP. Globalement, la méthode LR-NP

Méthode	**Linguistique** *(fr)*			**Sciences de l'info.** *(fr)*			**Archéologie** *(fr)*			**Chimie** *(fr)*		
	P	R	F	P	R	F	P	R	F	P	R	F
SingleRank	9.0	10,6	9,6	9,5	10,0	9,4	12,7	8,9	10,2	13,0	10,4	11,0
+ complet	10,0	11,9	10,7	9,9	10,2	9,8	13,5	9,5	11,0	13,0	10,7	11,2
+ candidats	10,8	12,7	11,5	11,1	11,6	11,0	25,7	17,4	20,3	**14,2**	**11,1**	**11,9**
+ sujets	10,6	12,5	11,3	10,9	11,5	10,8	26,5	18,0	20,9	13,5	10,7	11,5
TopicRank	**11,2**	**13,1**	**11,9**	**12,1**	**12,8**	**12,1**	**27,5**	**18,7**	**21,8**	13,8	11,1	11,8

TABLE 4.11 – Résultats de l'extraction de dix termes-clés avec chacune des contributions de TopicRank, appliquées séparément à SingleRank sur les données Termith

Méthode	**DEft** *(fr)*			**Wikinews** *(fr)*			**SemEval** *(en)*			**DUC** *(en)*		
	P	R	F	P	R	F	P	R	F	P	R	F
SingleRank	4,5	9,0	5,9	19,4	20,7	19,7	4,6	3,2	3,7	**22,3**	**28,4**	**24,6**
+ complet	4,4	9,0	5,8	20,0	21,4	20,3	5,5	3,8	4,4	22,2	28,1	24,5
+ candidats	10,3	19,2	13,2†	28,5	30,0	28,8†	9,4	6,8	7,8†	10,4	13,5	11,6
+ sujets	11,1	20,4	14,2†	30,7	32,6	31,1†	14,2	9,9	11,6†	18,9	24,2	21,0
TopicRank	**11,7**	**21,7**	**15,1**†	**35,0**	**37,5**	**35,6**†	**14,9**	**10,3**	**12,1**†	18,3	23,8	20,4

TABLE 4.12 – Résultats de l'extraction de dix termes-clés avec chacune des contributions de TopicRank, appliquées séparément à SingleRank sur les collections DEft, Wikinews, SemEval et DUC. † indique une amélioration significative vis-à-vis de SingleRank, à 0,001 pour le t-test de Student.

est, ici aussi, la méthode qui induit les meilleures performances. Son apport comparé à la méthode `/(N|A)+/` est tout de même plus modéré que sur TF-IDF et KEA. Cela montre que des méthodes au mode de fonctionnement différent ne réagissent pas de la même façon selon les candidats. TopicRank est peu sensible aux légères variations dans la qualité des candidats : les méthodes LR-NP et `/(N|A)+/` dont la qualité est très proche (voir les tableaux 4.5 et 4.6, page 57) peuvent donc être appliquées sans distinction à TopicRank.

4.3.3 Analyse d'erreurs

Dans cette section, nous analysons les erreurs de TopicRank. La première source d'erreurs est le mauvais groupement de certains candidats en sujets. La seconde source d'erreurs concerne la spécialisation des termes-clés extraits.

Les erreurs liées au groupement en sujets sont dues à la présence, dans le même groupe, de candidats ne véhiculant pas le même sujet, auquel cas la stratégie de sélection du terme-clé du sujet peu échouer. La principale cause de cela est la simplicité de notre mesure de similarité. En effet, elle ne tient compte ni du sens des candidats selon leur contexte, ni de leur sémantique latente. Par ailleurs, elle n'est pas adaptée à toutes les tailles de candidats. Par exemple, si deux candidats sont constitués de deux mots dont un en commun, alors ils sont groupés. Concrètement, nous observons le groupement de « représentation structurale » avec « représentation culturelle », parce qu'ils partagent le même nom, ou encore le groupement de « force économique » avec « délabrement économique », parce qu'ils partagent le même adjectif.

Les erreurs liées à la spécialisation des termes-clés extraits concerne à la fois les pro-

Méthode	**Linguistique** *(fr)*			**Sciences de l'info.** *(fr)*			**Archéologie** *(fr)*			**Chimie** *(fr)*		
	P	R	F	P	R	F	P	R	F	P	R	F
n-grammes	7,4	8,5	7,8	7,8	8,4	7,8	12,0	8,2	9,5	7,1	6,0	6,1
/ (N | A) +/	11,2	13,1	11,9	12,1	12,8	12,1	27,5	18,7	21,8	13,8	11,1	11,8
NP-chunks	11,4	13,3	12,1	**12,5**	**13,2**	**12,5**	28,5	19,3	22,5	14,1	11,3	12,0
LR-NP	**11,8**	**13,8**	**12,5**	12,2	12,8	12,2	**29,9**	**20,3**	**23,7**	**14,6**	**11,5**	**12,3**

TABLE 4.13 – Résultat de TopicRank sur les données Termith, selon la méthode de sélection des termes-clés candidats utilisée

Méthode	**DEft** *(fr)*			**Wikinews** *(fr)*			**SemEval** *(en)*			**DUC** *(en)*		
	P	R	F	P	R	F	P	R	F	P	R	F
n-grammes	8,2	15,0	10,5	22,7	24,8	23,3	13,2	9,2	10,7	9,5	13,3	10,9
/ (N | A) +/	**11,7**	**21,7**	**15,1**	**35,0**	**37,5**	**35,6**	14,9	10,3	12,1	**18,4**	**23,8**	**20,4**
NP-chunks	11,6	21,6	14,9	33,7	35,9	34,2	15,7	10,6	12,7	16,1	21,1	18,0
LR-NP	11,6	21,5	14,9	33,9	36,0	34,3	**16,6**	**11,5**	**13,5**	17,9	23,7	20,1

TABLE 4.14 – Résultat de TopicRank sur DEft, SemEval et DUC, selon la méthode de sélection des termes-clés candidats utilisée

blèmes de sous- et sur-spécialisation. Le problème de sous-spécialisation survient lorsque le terme-clé extrait est moins précis que le terme-clé de référence. Nous pouvons citer, par exemple, « papillons » qui est extrait à la place de « papillons mutants » dans l'article Wikinews *Fukushima fait muter les papillons*[13]. Le problème de sur-spécialisation survient lorsque le terme-clé extrait est plus précis que le terme-clé de référence. Nous pouvons citer, par exemple, « député Antoni Pastor » qui est extrait à la place de « Antoni Pastor » dans l'article Wikinews *Îles Baléares : le Parti populaire exclut le député Antoni Pastor pour avoir défendu la langue catalane*[14]. La présence simultanée de ces deux problèmes les rend difficiles à résoudre. Pour beaucoup, il s'agit là d'un problème d'évaluation (Zesch et Gurevych, 2009).

4.3.4 Bilan

Nous avons présenté TopicRank, une méthode non supervisée qui groupe les termes-clés candidats en sujets, détermine quels sont les sujets les plus importants, puis extrait le terme-clé candidat qui représente le mieux chacun d'eux. Cette nouvelle méthode offre plusieurs avantages vis-à-vis des précédentes méthodes à base de graphe. Le groupement des termes-clés potentiels en sujets distincts permet de rassembler des informations relatives au même sujet et le choix d'un seul terme-clé pour représenter un sujet important permet d'extraire un ensemble de termes-clés non redondants (pour k termes-clés extraits, exactement k sujets sont couverts).

TopicRank a quelques limitations. Premièrement, le groupement que nous proposons est « naïf » et il serait intéressant d'expérimenter d'autres méthodes de groupement en sujets. Lorsque les données disponibles le permettent, nous pourrions par exemple suivre Liu *et al.* (2010); Zhang *et al.* (2013) en utilisant LDA. Le choix du termes-clés d'un sujet peut aussi

[13] http://fr.wikinews.org/w/index.php?oldid=432477
[14] http://fr.wikinews.org/w/index.php?oldid=479948

être amélioré. Une solution intéressante serait d'utiliser une méthode de titrage automatique de sujets (Lau *et al.*, 2011). Étant donner les candidats d'un sujet, une telle méthode peut proposer celui qui le représente le mieux, voir une unité textuelle qui n'est pas présente dans le document.

4.4 Conclusion

Nous avons présenté deux contributions à l'extraction automatique de termes-clés. Dans un premier temps, nous avons analysé les propriétés linguistiques des termes-clés de référence de trois de nos collections de données, puis nous avons exploité cette analyse pour sélectionner les termes-clés candidats plus finement, en portant une attention particulière à leurs adjectifs. Dans un second temps, nous avons proposé une nouvelle méthode à base de graphe pour l'ordonnancement par importance des sujets d'un document et l'extraction d'un terme-clé représentatif de chacun des sujets les plus importants.

5

Indexation par termes-clés en domaines de spécialité

> « La multiplication des bases de données et l'information devenue « marché » (donc rentable) ont entraîné d'autres corps de métier à s'intéresser à la pratique de l'indexation. Mais ce sont les bibliothécaires et documentalistes qui en ont défini les méthodes, les usages et les outils. »
>
> — Guinchat et Skouri (1996)

5.1 Introduction

Dans ce chapitre, nous nous intéressons à l'indexation par termes-clés en domaines de spécialité. Dans la littérature, l'indexation par termes-clés se divise en deux catégories : l'extraction de termes-clés, qui fournit des termes-clés apparaissant dans le contenu du document, et l'assignement de termes-clés, qui fournit des termes-clés appartenant à un vocabulaire contrôlé et n'apparaissant pas nécessairement dans le document. Alors que dans la littérature, l'indexation par termes-clés est principalement réalisée au seul moyen de l'extraction de termes-clés, nous montrons que l'assignement de termes-clés joue un rôle important en domaines de spécialité.

Nous commençons par décrire le comportement des indexeurs professionnels qui maintiennent les bases des données bibliographiques de l'Inist (Institut de l'information scientifique et technique), puis nous en proposons une automatisation. Les indexeurs professionnels assignent à chaque document des termes-clés du domaine (d'un vocabulaire contrôlé), et extraient des termes-clés spécifiques au document (hors du vocabulaire contrôlé), voir des concepts nouveaux dans le domaine. Pour reproduire ce comportement, nous étendons nos travaux sur TopicRank en intégrant dans le graphe de sujets les entrées du vocabulaire du domaine.

Enfin, nous présentons les premiers résultats d'une campagne d'évaluation manuelle de nos travaux en domaines de spécialité. Pour cette campagne, nous proposons un protocole et des métriques permettant d'évaluer deux aspects : la pertinence des termes-clés extraits/assignés et la quantité d'information importante capturée par les termes-clés.

5.2 Indexation manuelle en domaines de spécialité

En nous fondant sur les propos recueillis auprès des indexeurs professionnels de l'Inist, qui maintient une partie des bases de données bibliographiques de la BSN (Bibliothèque scientifique numérique), nous présentons la méthodologie d'indexation manuelle en domaines de spécialité.

5.2.1 Principes généraux

L'indexation manuelle de l'Inist est régie par cinq principes généraux :

1. Conformité : les termes-clés doivent être conformes au contenu du document et au langage documentaire utilisé dans son domaine ;
2. Exhaustivité : les termes-clés doivent représenter tous les aspects importants du document, même lorsque ceux-ci sont implicites ;
3. Homogénéité : les termes-clés des documents d'un même domaine doivent être cohérents et identiques lorsqu'ils représentent le même concept ;
4. Spécificité : les termes-clés doivent décrire le contenu d'un document au niveau le plus spécifique et peuvent parfois être accompagnés de termes-clés plus génériques afin de restituer le contenu du document dans son domaine ;
5. Impartialité : les termes-clés ne doivent pas être représentatifs d'un jugement apporté par l'indexeur.

Ces principes généraux de l'indexation manuelle par termes-clés remettent en cause la séparation entre extraction et assignement de termes-clés dans le contexte de l'indexation automatique en domaines de spécialité. En effet, une indexation par termes-clés à un niveau professionnel doit respecter le langage de spécialité employé dans le domaine des documents indexés (tâche d'assignement), mais elle doit aussi être exhaustive et donc fournir des termes-clés très spécifiques, voir de nouveaux concepts (tâche d'extraction).

5.2.2 Ressources

L'indexation par termes-clés réalisée par les indexeurs professionnels de l'Inist s'appuie sur plusieurs ressources. Ces dernières sont représentatives d'une expertise de terrain dans chaque domaine de spécialité (grille d'indexation), d'une expertise terminologique (vocabulaire contrôlé) et d'une expertise documentaire (règles de préindexation). Elles assurent des conditions de travail propices au respect des cinq principes généraux énoncés précédemment.

Grille d'indexation

De nos jours, la grille d'indexation est un guide transmis de manière informelle aux indexeurs. Elle indique les notions à indexer selon le domaine de spécialité et peut se traduire par un formulaire à compléter pour chaque document (voir le tableau 5.1). C'est un canevas donné à titre indicatif aux indexeurs. Ces derniers sont les seuls juges pour décider si elle est adaptée ou non pour indexer un document.

Champ	Exemple
Théorie linguistique	concept linguistique
Objet d'étude	français ; conjonction ; expression linguistique ; cause
Niveau de description	interprétation sémantique ; relation syntaxique

TABLE 5.1 – Exemple de remplissage de la grille d'indexation de linguistique, pour la notice de linguistique présentée dans la figure 3.1 (page 41)

En définissant les notions à indexer, la grille d'indexation contribue fortement à l'homogénéité de l'indexation : les documents d'un même domaine sont en partie indexés d'après les mêmes notions. Elle contribue aussi à l'exhaustivité : même les notions implicites doivent faire partie de l'indexation.

Vocabulaire contrôlé

Le vocabulaire contrôlé est une liste de termes-clés possibles dans un domaine de spécialité donné. Cette liste est plus ou moins structurée en fonction des domaines[1]. Les termes-clés sont mis en relations s'il sont associés à un même concept (par exemple, « nom composé » et « substantif composé » en linguistique) ou si l'un est l'hyperonyme de l'autre, c'est-à-dire plus générique (par exemple, « allemand » par rapport à « haut-allemand » et « bas-allemand »).

En définissant le langage documentaire à utiliser pour indexer les documents du même domaine, le vocabulaire contrôlé contribue à la conformité et à l'homogénéité de l'indexation. Il n'assure cependant pas l'exhaustivité et doit être mis à jour régulièrement, soit par une veille terminologique, soit au fur et à mesure des indexations manuelles, pour intégrer les nouveaux concepts.

Règles de préindexation

Les règles de préindexation sont des règles qui définissent les termes-clés (du vocabulaire contrôlé ou non) à assigner en fonction soit (1) d'une unité textuelle qui occurre dans le document, soit (2) d'un terme-clé assigné au document. À l'instar du vocabulaire contrôlé, les règles de préindexation nécessitent un gros effort de maintenance manuelle.

Couplées avec le vocabulaire contrôlé, les règles de préindexation permettent d'assurer la conformité et l'homogénéité de l'indexation. Elles contribuent aussi à l'exhaustivité, en permettant l'assignement d'aspects implicites dans le document (1), et à la spécificité, en restituant le contenu du document dans son domaine grâce à des termes-clés génériques (2).

[1] Lorsqu'elles sont structurées, les listes respectent la spécification des thésaurus utilisés par la méthode KEA++ (voir la section 2.4, page 33).

5.2.3 Méthodologie

Nous distinguons cinq phases lors de l'indexation manuelle par termes-clés :

1. Choix des ressources à utiliser (grille d'indexation, vocabulaire contrôlé et règles de préindexation) ;
2. Utilisation d'un système automatisé de proposition de termes-clés à partir des règles de préindexation ;
3. Assignement de termes-clés respectant le langage documentaire (dans le vocabulaire contrôlé) ;
4. Assignement de termes-clés génériques afin de replacer les termes-clés trop spécifiques dans leur domaine ;
5. Extraction des termes-clés ne respectant pas le langage documentaire mais utiles pour décrire le contenu le plus important du document.

L'indexation Inist peut être qualifiée de semi-automatique. En effet, la deuxième phase est automatisée et systématiquement validée par l'indexeur, de sorte à réduire le temps d'indexation et à minimiser d'éventuels oublis. Cette phase montre la prise de conscience, dans les organismes gestionnaires de bases de données bibliographiques, que l'indexation est une tâche difficile et coûteuse à entreprendre manuellement.

5.2.4 Bilan

L'indexation manuelle en domaines de spécialité que nous avons présenté suit des principes d'indexation que nous retrouvons soit en extraction, soit en assignement. Alors qu'en indexation automatique par termes-clés, les méthodes d'extraction sont plus étudiées que les méthodes d'assignement, nous avons vu que l'indexation réalisée par des indexeurs professionnels donne la priorité à l'assignement, et que l'extraction doit uniquement servir à la compléter. Actuellement, aucune méthode n'effectue à la fois extraction et assignement de termes-clés.

5.3 Indexation automatique en domaines de spécialité

L'indexation automatique par termes-clés est définie comme la tâche qui consiste à extraire des termes-clés du contenu d'un document, ou à en assigner à partir d'un vocabulaire contrôlé. Alors que dans la littérature, l'indexation automatique par termes-clés est presque toujours réduite à la seule extraction de termes-clés, nous avons vu en domaines de spécialité que les deux catégories d'indexation (extraction et assignement) jouent chacune un rôle qui lui est propre.

Pour réaliser une indexation automatique par termes-clés, nous proposons la méthode TopicCoRank. TopicCoRank est, à notre connaissance la seule méthode capable de réaliser conjointement extraction et assignement de termes-clés.

5.3.1 TopicCoRank

TopicCoRank est une méthode supervisée à base de graphe qui réalise simultanément extraction et assignement de termes-clés. Issue de TopicRank, elle en modifie les étapes sui-

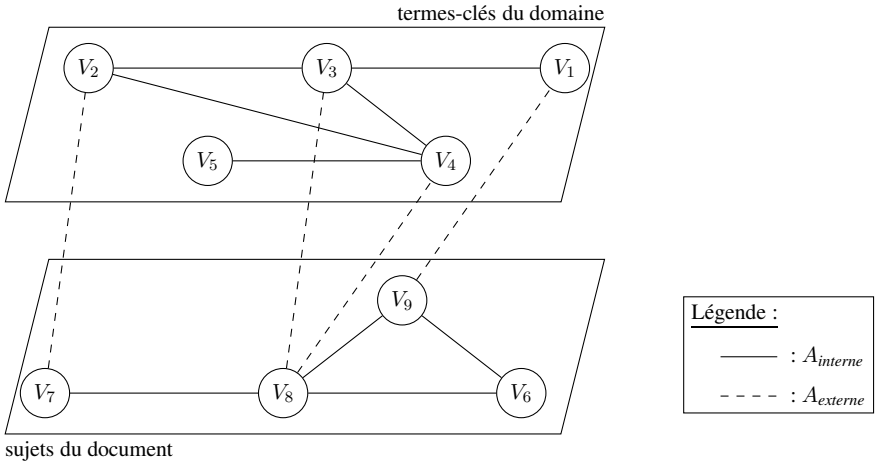


FIGURE 5.1 – Illustration du graphe unifié utilisé par TopicCoRank

vantes : construction du graphe, ordonnancement et sélection des termes-clés. La construction du graphe étend le graphe de sujet initial de TopicRank en l'unifiant à un graphe des termes-clés de référence du domaine ; l'ordonnancement est désormais conjoint pour les sujets du document et les termes-clés du domaine ; la sélection des termes-clés ajoute la possibilité de puiser dans le graphe du domaine afin de réaliser de l'assignement.

Construction du graphe

Afin de réaliser simultanément extraction et assignement de termes-clés, TopicCoRank unifie deux graphes représentant le document (graphe de sujets) et les termes-clés de référence de son domaine (graphe du domaine). Ce dernier graphe est construit à partir des termes-clés de référence de documents d'entraînement. Comme Chaimongkol et Aizawa (2013) l'ont fait avant nous pour l'extraction de termes techniques, nous faisons l'hypothèse que les termes-clés de référence des documents d'entraînement constituent la terminologie du domaine et nous les utilisons comme substituts au vocabulaire contrôlé. Contrairement aux termes-clés candidats sélectionnés dans le document, les termes-clés de référence ne sont pas redondants et ne sont donc pas groupés. Cette hypothèse est forte lorsque les données d'entraînement sont issues d'une indexation établie dans un contexte professionnel. Elle l'est moins pour les autres données (voir l'exemple figure 5.2).

Soit le graphe unifié non orienté $G = (N, A = A_{interne} \cup A_{externe})$. N dénote indifféremment les sujets et les termes-clés du domaine. A regroupe les arêtes $A_{interne}$, qui connectent deux sujets ou deux termes-clés du domaine, et les arêtes $A_{externe}$, qui connectent un sujet à un terme-clé de référence (voir la figure 5.1). Le graphe de sujets et le graphe du domaine sont unifiés grâce aux arêtes $A_{externe}$. En considérant le domaine comme une carte conceptuelle, l'objectif des arêtes $A_{externe}$ est de connecter le document à son domaine par l'intermédiaire des concepts qu'ils partagent. Une arête $A_{externe}$ est donc créée entre un sujet et un terme-clé du domaine si ce dernier appartient au sujet, c'est-à-dire correspond à l'un de ses termes-clés candidats.

Pour permettre un ordonnancement conjoint des sujets et des termes-clés du domaine, le schéma de connexion entre deux sujets et entre deux termes-clés du domaine (arêtes $A_{interne}$)

doit être homogène. En effet, si les conditions de connexion et si la pondération des arêtes ne sont pas équivalentes et du même ordre de grandeur, alors l'impact du domaine sur le document, et du document sur le domaine, sera marginal. Pour obtenir un graphe unifié homogène, TopicCoRank connecte deux sujets ou deux termes-clés du domaine n_i et n_j lorsqu'ils apparaissent dans le même contexte et pondère leur arête par le nombre de fois que cela se produit (poids(n_i, n_j)). Lorsqu'il s'agit des sujets, le contexte est une phrase du document ; lorsqu'il s'agit des termes-clés du domaine, le contexte est l'ensemble des termes-clés de référence d'un document d'entraînement. Les contextes étant utilisés pour la création du graphe, le graphe de sujets n'est plus complet comme celui de TopicRank. Ici, nous utilisons la phrase comme alternative à la fenêtre de cooccurrence.

Ordonnancement conjoint des sujets et des termes-clés du domaine

L'ordonnancement conjoint des sujets et des termes-clés du domaine établit leur ordre d'importance vis-à-vis du contenu du document et du domaine. Pour cela, un score d'importance est attribué simultanément aux sujets et aux termes-clés du domaine. Nous reprenons le principe de la recommandation de TopicRank et l'adaptons au problème d'ordonnancement conjoint. Les premières hypothèses de recommandation sont donc les mêmes que celle de TopicRank :

- un sujet est d'autant plus important s'il est fortement connecté à un grand nombre de sujets et si les sujets avec lesquels il est fortement connecté sont importants ;
- un terme-clé du domaine est d'autant plus important s'il est fortement connecté à un grand nombre de termes-clés du domaine et si les termes-clés du domaine avec lesquels il est connecté sont importants.

Ces hypothèses de recommandation, que nous qualifions d'internes, permettent d'établir l'importance des sujets les uns par rapport aux autres et l'importance des termes-clés du domaine les uns par rapport aux autres. Cependant, elles ne permettent pas de tirer profit des relations entre sujets et termes-clés du domaine. Par ailleurs, l'importance des termes-clés du domaine ne dépend pas du document. Nous ajoutons donc deux nouvelles hypothèses de recommandation, que nous qualifions d'externes :

- un sujet est d'autant plus important s'il est représenté par (connecté à) des termes-clés du domaine importants ;
- un terme-clé du domaine est d'autant plus important vis-à-vis du contenu du document s'il véhicule (est connecté à) l'un de ses sujets importants.

Sujets et termes-clés du domaine sont ainsi évalués d'après leur usage dans le document et leur importance dans le domaine. L'ordonnancement des uns joue un rôle sur celui des autres et permet ainsi d'effectuer conjointement extraction et assignement.

L'équation 5.1 montre le calcul de l'importance d'un sujet ou d'un terme-clé du domaine à partir de sa recommandation interne $R_{interne}$ et de sa recommandation externe $R_{externe}$:

$$S(n_i) = (1-\lambda)\ R_{externe}(n_i) + \lambda\ R_{interne}(n_i) \tag{5.1}$$

$$R_{interne}(n_i) = \sum_{n_j \in A_{interne}(n_i)} \frac{\text{poids}(n_j, n_i) \times S(n_j)}{\sum_{n_k \in A_{interne}(n_j)} \text{poids}(n_j, n_k)} \tag{5.2}$$

$$R_{externe}(n_i) = \sum_{n_j \in A_{externe}(n_i)} \frac{S(n_j)}{|A_{externe}(n_j)|} \tag{5.3}$$

où $A_{interne}(n_i)$ représente l'ensemble de tous les nœuds connectés au nœud n_i par une arête $A_{interne}$, où $A_{externe}(n_i)$ représente l'ensemble de tous les nœuds connectés au nœud n_i par une arête $A_{externe}$ et où le facteur λ permet désormais de définir la recommandation la plus influente entre $R_{interne}$ et $R_{externe}$. Par défaut, nous définissons $\lambda = 0,5$.

Sélection des termes-clés

TopicCoRank utilise l'ordre d'importance établit avec le score S des sujets et termes-clés du domaine pour déterminer les termes-clés du document. Les k nœuds du graphe unifié ayant obtenu les meilleurs scores sont retenus, qu'ils soient des sujets ou des termes-clés du domaine.

Lorsqu'un terme-clé du domaine doit être assigné, une étape de vérification supplémentaire est effectuée pour s'assurer que son importance calculée relève aussi bien du domaine que du document. En effet, il est possible que le graphe du domaine soit constitué de composantes connexes, soit de sous-graphes dont les nœuds ne sont connectés qu'entre eux. Dans ce cas, il se peut qu'un terme-clé du domaine d'un sous-graphe ne soit connecté, ni directement, ni indirectement (par l'intermédiaire d'un autre nœud), à un sujet du document. Son importance est donc déterminée uniquement à partir du domaine et il n'est donc pas pertinent de l'assigner au document.

Lorsqu'un nœud retenu représente un sujet, c'est la même stratégie que celle de TopicRank qui est appliquée. Pour un sujet donné, le terme-clé extrait est son terme-clé candidat qui apparaît en premier dans le document.

Exemple

La figure 5.2 donne un exemple d'extraction et d'assignement de termes-clés avec TopicCoRank à partir de la notice d'archéologie présentée dans la figure 3.1 (page 41). Dans cet exemple, nous observons une meilleure indexation par termes-clés qu'avec TopicRank. Tout d'abord, nous voyons que le graphe du domaine permet l'assignement du termes-clés générique « France ». Ensuite, nous voyons que les relations de « diffusion », « analyse » et « répartition » dans le graphe du domaine permettent de mieux les ordonner.

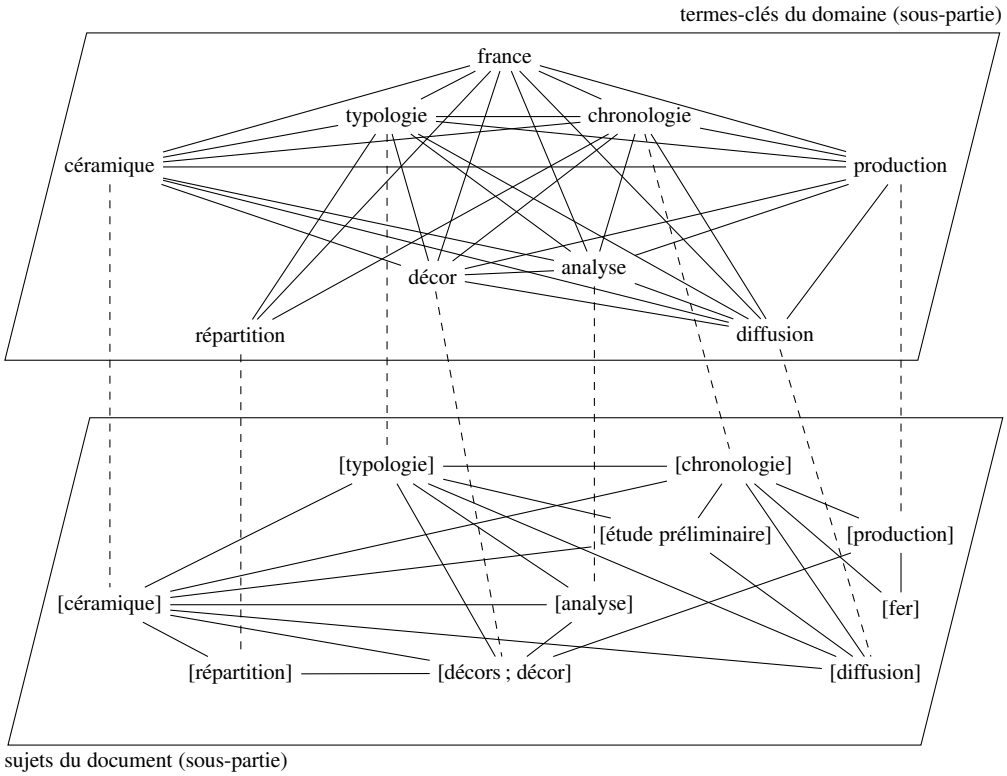
5.3.2 Évaluation

Pour valider notre approche, nous réalisons deux séries d'expériences. Dans un premier temps, nous comparons TopicCoRank à plusieurs méthodes de référence et analysons son

Étude préliminaire de la céramique non tournée micacée du bas Languedoc occidental : typologie, chronologie et aire de diffusion

L'étude présente une variété de céramique non tournée dont la typologie et l'analyse des décors permettent de l'identifier facilement. La nature de l'argile enrichie de mica donne un aspect pailleté à la pâte sur laquelle le décor effectué selon la méthode du brunissoir apparaît en traits brillant sur fond mat. Cette première approche se fonde sur deux séries issues de fouilles anciennes menées sur les oppidums du Cayla à Mailhac (Aude) et de Mourrel-Ferrat à Olonzac (Hérault). La carte de répartition fait état d'échanges ou de commerce à l'échelon macrorégional rarement mis en évidence pour de la céramique non tournée. S'il est difficile de statuer sur l'origine des décors, il semble que la production s'insère dans une ambiance celtisante. La chronologie de cette production se situe dans le deuxième âge du Fer. La fourchette proposée entre la fin du IV^{e} et la fin du II^{e} s. av. J.-C. reste encore à préciser.

Termes-clés de référence : distribution ; mourrel-ferrat ; olonzac ; le cayla ; mailhac ; micassé ; céramique non-tournée ; celtes ; production ; echange ; commerce ; cartographie ; habitat ; oppidum ; site fortifié ; fouille ancienne ; identification ; décor ; analyse ; répartition ; diffusion ; chronologie ; typologie ; céramique ; etude du matériel ; hérault ; aude ; france ; europe ; la tène ; age du fer.



Sortie de TopicCoRank : céramique ; décors ; typologie ; chronologie ; production ; étude préliminaire ; diffusion ; analyse ; france ; répartition.

Sortie de TopicRank : décors ; céramique ; chronologie ; typologie ; production ; fin ; étude préliminaire ; fer ; deuxième âge ; aire.

FIGURE 5.2 – Exemple d'extraction de termes-clés avec TopicCoRank sur le résumé de la notice d'archéologie présentée dans la figure 3.1 de la section 3.2 (page 3.1). Les termes-clés soulignés sont les termes-clés correctement extraits.

comportement en domaines de spécialité. Dans un second temps, nous étudions l'application de TopicCoRank dans le cas général, afin de vérifier si nos hypothèses fortement liées à l'indexation manuelle en domaines de spécialité peuvent se généraliser.

Méthodes de référence

Dans nos expériences, nous comparons TopicCoRank à TF-IDF, TopicRank et KEA++. Pour cette dernière, nous utilisons les thésaurus décrivant les vocabulaires contrôlés de l'Inist en linguistique, sciences de l'information, archéologie et chimie. Pour les ressources autres que Termith, nous ne disposons pas de vocabulaires contrôlés adéquats et n'appliquons donc pas KEA++.

Afin de mesurer l'efficacité de l'ordonnancement conjoint, nous comparons aussi TopicCoRank à deux variantes. La première, TopicCoRank$_{extr.}$, ne réalise que l'extraction de termes-clés ; la seconde, TopicCoRank$_{assign.}$, n'effectue que l'assignement.

Pour toutes les méthodes réalisant de l'extraction, les termes-clés sont issus des candidats sélectionnés avec la méthode que nous présentons dans la section 4.2 (page 49).

Collections de données

Nous utilisons les collections Termith pour l'évaluation en domaines de spécialité et les collections DEft, SemEval et DUC pour l'évaluation dans le cas général[2].

Pour DUC, qui n'est pas divisé en deux ensembles d'entraînement et de test, nous tirons partie des 30 sujets d'actualité répertoriés en construisant un graphe « de domaine » unique pour chaque document à partir des autres documents du même sujet d'actualité.

Pour SemEval, nous construisons quatre graphes de domaine à partir des documents d'entraînement des quatre catégories ACM (C2.4, H3.3, I2.11 et J4) et utilisons l'un ou l'autre de ces graphes selon la catégorie du document de test.

Mesures d'évaluation

Les performances des méthodes d'extraction de termes-clés sont exprimées en termes de précision (P), rappel (R) et f1-mesure (F). En accord avec l'évaluation menée dans les travaux précédents, les opérations de comparaison entre les termes-clés de référence et les termes-clés extraits sont effectuées à partir de la racine des mots qui les composent. Pour cela, nous utilisons la méthode de Porter (1980).

Nous représentons aussi les résultats sous la forme de courbes de rappel–précision. Celles-ci permettent d'observer si une méthode domine les autres pour les critères de rappel et de précision. En optimisation multi-critères, nous parlons de front de Pareto optimal, c'est à dire de la méthode pour laquelle aucune autre méthode n'obtient de meilleures performances. Pour générer ces courbes, nous calculons la précision et le rappel lorsque le nombre de termes-clés extraits/assignés varie de un jusqu'au plus grand nombre commun de termes-clés pouvant être extraits/assignés[3].

[2] Constituées de documents scientifiques, les ressources DEft et SemEval peuvent aussi être considérées comme des données en domaines de spécialité. Cependant, elles regroupent des documents de sous-disciplines très éloignées et leurs termes-clés n'ont pas été attribués avec la rigueur documentaire.

[3] Si, parmi tous les documents de test, le nombre minimum de termes-clés extraits/assignés pour un document est de 73, alors la précision et le rappel sont calculés pour un jusqu'à 73 termes-clés en moyenne pour tous les documents.

Méthode	**Linguistique** *(fr)*			**Sciences de l'info.** *(fr)*			**Archéologie** *(fr)*			**Chimie** *(fr)*		
	P	R	F	P	R	F	P	R	F	P	R	F
TF-IDF	13,3	15,8	14,2	13,5	14,2	13,4	28,2	19,2	22,3	15,8	12,3	13,2
TopicRank	11,8	13,8	12,5	12,2	12,8	12,2	29,9	20,3	23,7	14,6	11,5	12,3
KEA++	11,6	13,0	12,1	9,5	10,2	9,6	23,5	16,2	18,8	11,4	8,5	9,2
TopicCoRank$_{extr.}$	14,3	16,5	15,1	15,4	15,9	15,2‡	36,7	24,6	28,8†	15,8	12,1	13,1
TopicCoRank$_{assign.}$	**24,5**	**28,3**	**25,8**	**19,7**	**19,8**	**19,2**‡	**47,8**	**32,3**	**37,7**†	**20,0**	**14,8**	**16,3**†
TopicCoRank	18,8	21,9	19,9	17,3	17,7	17,0‡	38,3	25,7	30,1†	17,2	13,4	14,4‡

TABLE 5.2 – Résultat de l'extraction de dix termes-clés avec TF-IDF, TopicRank, KEA++, TopicCoRank$_{extr.}$, TopicCoRank$_{assign.}$ et TopicCoRank appliqués aux collections Termith. † et ‡ indiquent une amélioration significative vis-à-vis des méthodes de référence, à 0,001 et 0,05 pour le t-test de Student, respectivement.

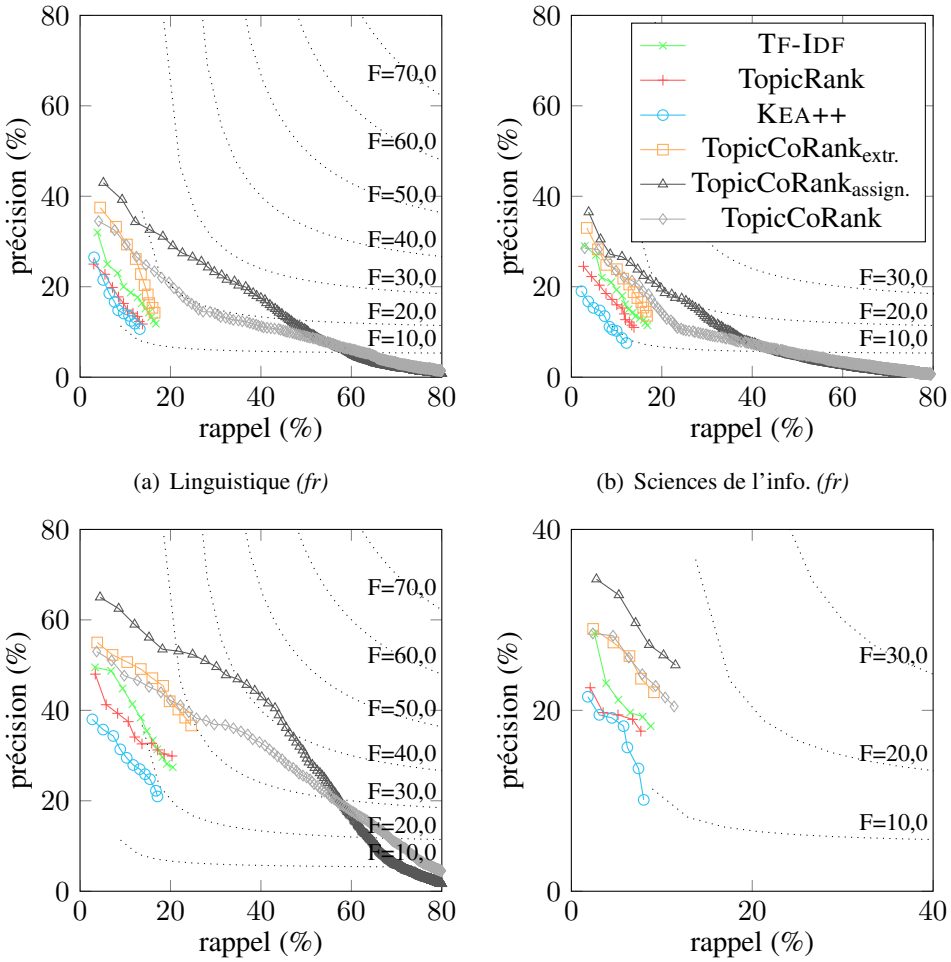
Évaluation de TopicCoRank en domaines de spécialité

Nous réalisons ici une série d'expériences destinées à comparer TopicCoRank à l'existant, puis à observer son comportement selon différentes configurations.

Le tableau 5.2 montre les performances de TopicCoRank en domaines de spécialité (linguistique, sciences de l'information, archéologie, chimie) comparées à celles des méthodes de référence. De manière générale, les résultats montrent le bien fondé de TopicCoRank : la variante TopicCoRank$_{assign.}$ réalise les meilleures performances, suivie par TopicCoRank et TopicCoRank$_{extr.}$. Les faibles performances de KEA++ sont surprenantes, d'autant plus que la seule autre méthode d'assignement, TopicCoRank$_{assign.}$, est celle qui réalise les meilleures. Contrairement à TopicCoRank$_{assign.}$, KEA++ se limite aux entrées du thésaurus qui occurrent dans le document, alors que la majorité des termes-clés des collections Termith n'apparaissent pas dans les documents. De plus les thésaurus de l'Inist ne sont pas aussi riches que ceux utilisés par Medelyan et Witten (2006) dans leurs expériences : moins de relations y sont définies entre les concepts. TopicCoRank et ses variantes sont significativement meilleurs que les méthodes de référence. Comparées à celles de TopicRank, les performances de TopicCoRank$_{extr.}$ montrent que le domaine apporte des informations permettant d'ordonner plus précisément les sujets du document. Le fait que TopicCoRank$_{assign.}$ obtienne les meilleures performances montre aussi que les termes-clés du domaine sont ordonnés efficacement d'après le contenu du document (ses sujets). La prédominance de termes-clés à assigner dans les données Termith est la principale raison pour laquelle la variante TopicCoRank$_{assign.}$ est plus performante que TopicCoRank.

La figure 5.3 permet de comparer le comportement respectif des méthodes de référence, de TopicCoRank et de ses variantes. Elle montre que TopicCoRank et ses variantes dominent les méthodes de référence (front de Pareto) selon les critères de précision et de rappel. Parmi elles, nous observons aussi que la variante TopicCoRank$_{assign.}$ domine la variante TopicCoRank$_{extr.}$, mais que TopicCoRank n'est, ni dominante, ni dominé par elles. Bien que l'amélioration significative de TopicRank par TopicCoRank et ses variantes montrent l'apport de l'ordonnancement conjoint entre sujets du document et termes-clés du domaine, la réalisation simultanée de l'extraction et de l'assignement reste difficile.

Afin d'observer la place que prend l'assignement dans TopicCoRank, et pour comprendre pourquoi sa variante TopicCoRank$_{assign.}$ est plus performante, nous nous intéressons maintenant aux taux de termes-clés extraits et assignés par TopicCoRank, présentés dans le tableau 5.3. Nous observons que l'extraction est légèrement prédominante face à l'as-



(a) Linguistique *(fr)*

(b) Sciences de l'info. *(fr)*

(c) Archeologie *(fr)*

(d) Chimie *(fr)*

FIGURE 5.3 – Courbes de rappel-précision de TF-IDF, TopicRank KEA++, $\text{TopicCoRank}_{\text{extr.}}$, $\text{TopicCoRank}_{\text{assign.}}$ et TopicCoRank appliqués aux données Termith

	Extraction (%)	Assignement (%)
Linguistique *(fr)*	61,7	38,3
Sciences de l'info. *(fr)*	66,4	33,6
Archéologie *(fr)*	69,1	30,9
Chimie *(fr)*	68,4	31,6

TABLE 5.3 – Taux moyens d'extraction et d'assignement réalisés par TopicCoRank sur les données Termith

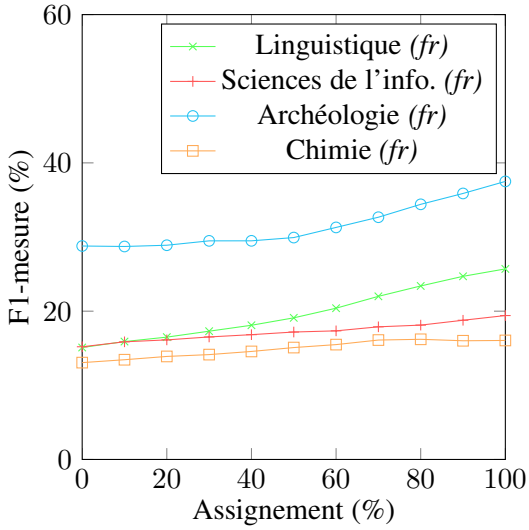


FIGURE 5.4 – Performance de TopicCoRank appliqué aux données Termith, lorsque le taux d'assignement varie

signement. Les deux catégories d'indexation par termes-clés sont effectivement réalisées conjointement, mais l'ordonnancement donne plus d'importance aux sujets du document qu'aux termes-clés de référence du domaine. En domaines de spécialité où l'assignement est préféré, cela peut être résolu en travaillant sur un affinage des schémas de connexion des nœuds de chaque graphe et d'unification de ceux-ci.

Au delà du fait que TopicCoRank$_{\text{assign.}}$ obtient de meilleures performances que TopicCoRank et TopicCoRank$_{\text{extr.}}$, nous faisons une expérience dans laquelle nous forçons le taux d'assignement afin de déterminer si l'ordonnancement des termes-clés du domaine est efficace. Un ordonnancement efficace des termes-clés du domaine doit induire une courbe de performance cumulative quand nous faisons croître le taux d'assignement[4]. La figure 5.4 montre la performance de TopicCoRank lorsque le taux d'assignement varie de 0 % à 100 % avec un pas de 10 %. À chaque augmentation du taux d'assignement, la performance de TopicCoRank augmente. L'ordonnancement des termes-clés du domaine fait donc émerger efficacement ceux les plus importants vis-à-vis du document.

Enfin, nous réalisons une dernière expérience dans laquelle nous faisons varier la valeur du paramètre λ. Plus sa valeur est élevée, plus l'influence de la recommandation interne est

[4]Dans cette situation, cela signifie que la performance obtenue avec TopicCoRank$_{\text{assign.}}$ est la performance maximale avec TopicCoRank

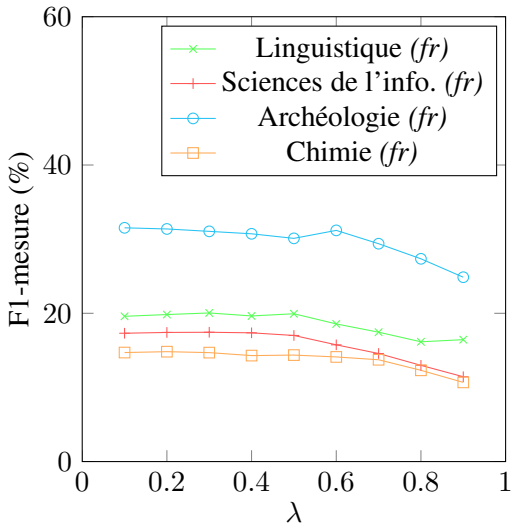


FIGURE 5.5 – Performance de TopicCoRank appliqué aux données Termith, lorsque le paramètre λ varie

forte. La figure 5.5 montre le comportement de TopicCoRank lorsque nous faisons varier sa valeur de 0 à 1 avec un pas de 0,1. En accord avec notre hypothèse que sujets et termes-clés du domaine doivent se recommander les uns les autres, les résultats montrent que les performances de TopicCoRank se dégradent au delà de $\lambda = 0,5$, valeur quasi-optimale.

Évaluation de TopicCoRank hors domaines de spécialité

Nous réalisons ici la même série d'évaluations précédentes, mais hors domaines de spécialité. L'objectif est de déterminer si TopicCoRank et ses hypothèses, fortement liées à notre étude de l'indexation manuelle en domaines de spécialité, se généralisent à tout type de documents.

Le tableau 5.4 montre les performances de TopicCoRank hors domaines de spécialité (DEft, SemEval et DUC) comparées à celles des méthodes de référence. Les résultats montrent de plus faibles performances qu'en domaines de spécialité. TopicCoRank échoue à améliorer TopicRank sur DEft, l'améliore légèrement sur SemEval et l'améliore significativement sur DUC, pour lequel nous utilisons des graphes de domaine très centrés sur le sujet d'actualité de chaque document de test. Contrairement à ce que nous observons en domaines de spécialité, c'est TopicCoRank qui est majoritairement le plus performant et c'est sa variante $\text{TopicCoRank}_{\text{assign.}}$ qui est la moins performante. L'explication tient à la nature des termes-clés de référence de DEft et SemEval. Ceux-ci n'ont pas été produit à l'aide d'un vocabulaire contrôlé et les cinq principes sur lesquels nous fondons nos hypothèses ne sont pas respectés. La contrainte de conformité n'étant pas respectée, la nécessité de l'assignement n'est pas garantie ; la contrainte d'homogénéité n'étant pas respectée non plus, le graphe de domaine que nous construisons contient des termes-clés de référence redondants. La redondance des termes-clés de référence requiert une étape de normalisation. Celle-ci peut s'effectuer à l'aide de notre méthode de groupement en sujets.

La figure 5.6 permet de comparer le comportement respectif des méthodes de référence,

Méthode	DEft *(fr)*			SemEval *(en)*			DUC *(en)*		
	P	R	F	P	R	F	P	R	F
TF-IDF	10,4	19,1	13,3	13,6	9,3	10,9	24,9	32,1	27,7
TopicRank	**11,9**	**21,5**	**14,9**	16,6	11,5	13,5	17,9	23,7	20,1
~~KEA++~~									
$\text{TopicCoRank}_{\text{extr.}}$	10,1	19,1	13,0	17,4	12,3	14,3	24,6	35,5	27,2
$\text{TopicCoRank}_{\text{assign.}}$	6,8	12,8	8,7	11,8	8,4	9,7	25,8	33,1	28,6
TopicCoRank	8,7	16,2	11,2	**17,6**	**12,5**	**14,5**	**28,2**	**36,3**	**31,3**†

TABLE 5.4 – Résultat de l'extraction de dix termes-clés avec TF-IDF, TopicRank, KEA++, $\text{TopicCoRank}_{extr.}$, $\text{TopicCoRank}_{assign.}$ et TopicCoRank appliqués à DEft, SemEval et DUC. † indique une amélioration significative vis-à-vis des méthodes de référence, à 0,001 pour le t-test de Student.

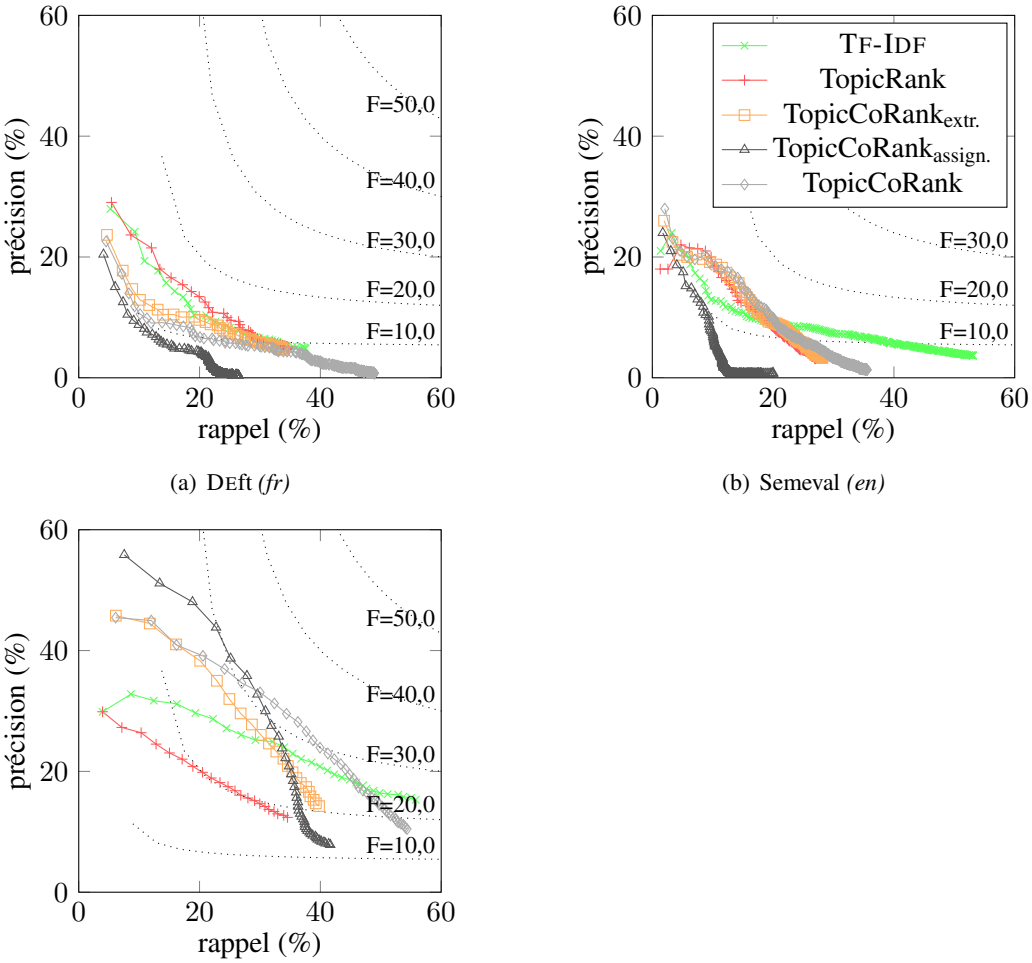
de TopicCoRank et de ses variantes. Contrairement aux domaines de spécialité, où TopicCoRank et ses variantes dominent les méthodes de référence, nous observons ici qu'aucune méthode dominante ne se dégage et, sauf sur DUC, il est difficile de statuer sur l'apport de TopicCoRank à TopicRank. Par ailleurs, le rappel maximal atteint par TopicCoRank n'excède, dans la plupart des cas, pas celui de TF-IDF. TopicCoRank étant capable d'assigner des termes-clés qui n'occurrent pas dans le document, son rappel maximal devrait être plus grand. Il existe deux raisons à ce problème. La première est aussi valable pour TopicRank. Comme les termes-clés candidats sont groupés en sujets et qu'un seul d'entre eux est extrait par sujet, si un terme-clé erroné est extrait d'un sujet contenant un terme-clé correct, alors le rappel maximal observable est plus faible que celui de TF-IDF, qui peut extraire tous les candidats lorsque nous le lui demandons. La seconde raison nous intéresse particulièrement, car elle n'est pas observable en domaines de spécialité. En effet, le problème est aussi due à l'inconsistance des données d'entraînement pour représenter un « domaine » de manière homogène et conforme à son vocabulaire. Les termes-clés employés dans les documents d'entraînement ne sont pas les mêmes que ceux employés pour les documents de test et l'assignement ne fonctionne donc pas.

Comme pour l'évaluation en domaines de spécialité, le tableau 5.5 reporte les taux d'extraction et d'assignement réalisés par TopicCoRank sur DEft, SemEval et DUC. Ceux-ci montrent aussi que les deux catégories d'indexation sont réalisées conjointement, cette fois-ci sans prédominance de l'une face à l'autre.

	Extraction (%)	Assignement (%)
DEft *(fr)*	48,4	51,6
SemEval *(en)*	61,4	38,6
DUC *(en)*	46,9	53,1

TABLE 5.5 – Taux moyens d'extraction et d'assignement réalisés par TopicCoRank sur DEft, SemEval et DUC

La figure 5.7 montrent les performances de TopicCoRank lorsque nous forçons le taux d'assignement, de 0 % à 100 % avec un pas de 10 %. Contrairement à la courbe de performances en domaines de spécialité, celle de TopicCoRank hors domaines de spécialité est



(a) DEft *(fr)*

(b) Semeval *(en)*

(c) DUC *(en)*

FIGURE 5.6 – Courbes de rappel-précision de TF-IDF, TopicRank, $\text{TopicCoRank}_{\text{extr.}}$, $\text{TopicCoRank}_{\text{assign.}}$ et TopicCoRank appliqués à DEft, SemEval et DUC

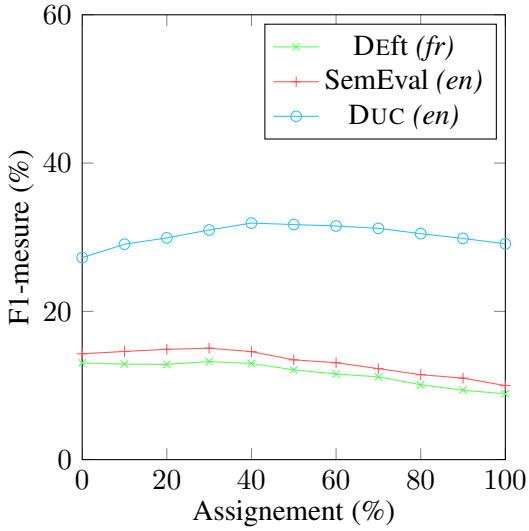


FIGURE 5.7 – Performance de TopicCoRank, appliqué à DEft, SemEval et DUC, lorsque le taux d'assignement varie

croissante puis décroissante pour Duc et est principalement décroissante pour DEft et SemEval. Cela signifie que l'ordonnancement des termes-clés du domaine est moins efficace hors domaines de spécialité. Il est toutefois intéressant de noter que, sur DUC, le taux d'assignement effectué par TopicCoRank sans que nous ne le forcions est proche de sa valeur optimal.

La figure 5.8 montre le comportement de TopicCoRank lorsque nous faisons varier la valeur de λ de 0 à 1 avec un pas de 0,1. Sur DUC, comme en domaines des spécialité, l'indexation par termes-clés est meilleure lorsque l'ordonnancement est fortement influencé par la recommandation externe que lorsqu'il ne l'est pas. Sur DEft et SemEval, c'est l'inverse. Comme observé précédemment, les données d'entraînement sont telles que le graphe du domaine est moins utile pour ces données.

5.3.3 Analyse des sorties de TopicCoRank

Dans cette section, nous analysons les termes-clés corrects (vrais positifs) et incorrects (faux positifs) issus du graphe du domaine des collections Termith.

Analyse des vrais positifs

Parmi les termes-clés assignés, ceux qui sont corrects sont en grande partie présents dans le contenu du document. Ils sont directement connectés aux sujets du document et leur importance respective évolue de manière similaire. Il est fréquent qu'un terme-clé candidat d'un sujet soit extrait en même temps qu'un terme-clé de référence du domaine connecté à ce sujet. Dans cette situation, nous distinguons deux cas de figure : un seul terme-clé est produit, car le terme-clé extrait et celui assigné sont les mêmes ; deux termes-clés corrects complémentaires sont produits, car le terme-clé extrait et celui assigné ne sont pas les mêmes.

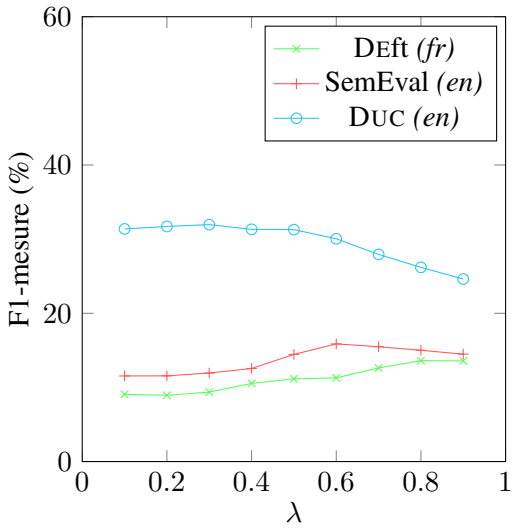


FIGURE 5.8 – Performance de TopicCoRank, appliqué à DEft, SemEval et DUC, lorsque le paramètre λ varie

Il est plus difficile pour les termes-clés de référence du domaine connectés indirectement aux sujets du document d'émerger. Nous observons toutefois l'émergence de quelques termes-clés absents du documents (par exemple, « analyse du discours » en linguistique et les noms de composés « composé aliphatique », « composé benzénique », etc. en chimie), et de termes-clés génériques (par exemple « français » en linguistique, ou encore « Europe » en archéologie). Nous observons dans les données Termith des cadres d'études très récurrents (par exemple, la langue française, en linguistique, ou des fouilles de sites européens, en archéologie) et, de ce fait, certains termes-clés génériques sont très fréquemment utilisés (« français » apparaît dans 48,9 % des documents de linguistique et « Europe » apparaît dans 52,5 % des documents d'archéologie).

Analyse des faux positifs

Les termes-clés génériques évoqués dans l'analyse des vrais positifs sont aussi sources d'erreurs. En effet, comme ils sont associés à un nombre conséquent de documents d'entraînement, ils sont connectés à beaucoup d'autres termes-clés du domaine et gagnent donc de l'importance quelque soit le document. Pour l'exemple du terme-clé « français » en linguistique, nous observons des documents qui traitent de l'arabe mais, parce que les termes techniques employés sont les mêmes (« syntaxe », « sémantique », etc.), « français » est assigné.

Enfin, nous observons quelques problèmes liés à la présence de termes-clés de référence redondants, c'est-à-dire des synonymes. C'est le cas, par exemple, du terme-clé « rite funéraire », qui fait parti du vocabulaire contrôlé d'archéologie, et qui est parfois remplacé par le terme-clé « pratique funéraire », qui ne fait pas partie du vocabulaire contrôlé d'archéologie. TopicCoRank échoue parfois parce qu'il a assigné l'un alors que c'était l'autre qu'il fallait trouver.

5.3.4 Bilan

Nous avons présenté TopicCoRank, une extension de la méthode TopicRank. Proposée dans le but de simuler le comportement d'un indexeur professionnel, cette extension apporte à TopicRank la capacité à assigner des termes-clés. Pour ce faire, TopicCoRank utilise les termes-clés de référence des documents d'entraînement comme vocabulaire contrôlé, crée un graphe du domaine à partir de ces termes-clés, puis unifie ce graphe au graphe de sujets de TopicRank. Termes-clés du domaines et sujets du document sont ensuite ordonnés conjointement pour extraire des termes-clés à partir des sujets et en assigner à partir des termes-clés du domaine.

Les résultats de l'évaluation de TopicCoRank en domaines de spécialité montrent une amélioration significative vis-à-vis de l'état de l'art. Ceux hors domaines de spécialité sont moins probants. Les hypothèses que nous faisons en proposant TopicCoRank sont fondées sur une étude de l'indexation manuelle en domaines de spécialité qui ne sont pas directement généralisables à tout type de documents.

5.4 Évaluation manuelle en domaines de spécialité

L'évaluation manuelle des méthodes d'indexation par termes-clés consiste à faire valider par un ou plusieurs évaluateurs humains les termes-clés proposés par une méthodes automatique. Parce qu'elle est coûteuse, cette évaluation est quasi-systématiquement remplacée par une évaluation automatique, d'après le paradigme de l'évaluation « à la Crandield », comme nous l'avons fait dans nos travaux. Néanmoins, ce paradigme d'évaluation n'est pas adapté à la tâche d'indexation par termes-clés, car il ne considère qu'une seule et unique réponse exacte, c'est-à-dire un seul et unique ensemble correct de termes-clés, alors que certaines variantes d'ensembles de termes-clés sont acceptables (par exemple, un ensemble avec « rite funéraire » et une variante avec « pratique funéraire »). Ayant à notre disposition des indexeurs professionnels de l'Inist, nous réalisons donc une campagne d'évaluation manuelle.

Dans la suite, nous présentons le protocole d'évaluation et les métriques que nous avons proposé, puis nous analysons les premiers résultats de la campagne d'évaluation manuelle de TopicRank, TopicCoRank et de deux méthodes de référence en domaine de spécialité avec la collection de linguistique Termith.

5.4.1 Protocole d'évaluation manuelle

L'évaluation manuelle concerne dix termes-clés extraits/assignés par chaque méthode d'indexation par termes-clés. Le protocole que nous proposons permet d'évaluer deux aspects de l'indexation automatique par termes-clés :

1. Pertinence (validité) : chaque terme-clé fourni par la méthode d'indexation automatique par termes-clés est-il important pour la compréhension du contenu principal du document ?
2. Silence : quel est le degré d'importance des informations perdues entre les termes-clés de référence et les termes-clés fournis par la méthode d'indexation automatique par termes-clés ?

L'évaluation de la pertinence s'intéresse au même aspect que l'évaluation automatique : le nombre de termes-clés corrects doit être maximisé et le nombre d'erreurs minimisé pour obtenir la meilleure performance. L'évaluation du silence s'intéresse à un aspect qui lui est propre. Elle a une dimension plus sémantique : les termes-clés corrects dont l'information est la plus capitale pour la compréhension du contenu principal du document doivent être priorisés pour obtenir la meilleure performance.

Afin de minimiser les problèmes d'ambiguïté et de subjectivité de certains cas de figure, la pertinence et le silence sont évalués sur une échelle à trois valeurs : une valeur représentant le succès, une autre l'échec et une dernière le cas intermédiaire.

Évaluation de la pertinence

Pour évaluer la pertinence d'un terme-clé fourni par une méthode d'indexation par termes-clés, l'évaluateur doit lui attribuer un score sur une échelle de 0 à 2. Ce score distingue les termes-clés incorrects (0), les termes-clés corrects (2) et les variantes de ces derniers (1).

Pour permettre une étude précise de cette évaluation, les indexeurs professionnels doivent indiquer la forme préférée des termes-clés auxquels ils donnent un score de 1 (variantes). Une variante peut faire référence à deux catégories de formes préférées, qui induisent deux raisonnements :

- variante d'un terme-clé déjà fourni (score de 2) ⇒ la méthode d'indexation par termes-clés fourni des termes-clés redondants (voir la figure 5.9) ;
- variante d'un terme-clé non fourni mais présent dans le texte ⇒ la méthode d'indexation par termes-clés identifie correctement les sujets importants du document, mais peine à trouver la forme la plus appropriée pour les représenter.

Lorsque la forme préférée n'est pas présente dans le document, nous estimons que la méthode d'indexation a fourni un terme-clé correct, auquel cas il se voit attribuer le score de 2 (voir la figure 5.9). Les formes variantes résultant d'un accord en nombre (pluriel) obtiennent aussi un score de 2, lorsque la forme normalisée (singulier) ne se trouve pas parmi les termes-clés fournis.

Évaluation du silence

Pour évaluer le silence, l'évaluateur doit attribuer à chaque terme-clé de référence un score indiquant le degré d'importance de l'information qu'il véhicule et qui n'est pas capturée par les termes-clés fournis par une méthode d'indexation par termes-clés. Sur une échelle de 0 à 2, ce score permet d'indiquer s'il n'y pas de perte d'information (0), si l'information perdue est capitale (2) ou si elle est secondaire (1). Lorsqu'un terme-clé de référence obtient un score de 0, cela signifie soit qu'il fait partie des termes-clés fournis par la méthode d'indexation par termes-clés, soit que l'indexeur juge qu'il ne devrait pas être un terme-clé de référence, c'est-à-dire une erreur parmi les termes-clés de référence.

Une perte d'information est jugée secondaire (score de 1) dans deux cas de figure :

- terme-clé de référence secondaire : le terme-clé de référence n'apporte pas l'information la plus importante ;
- terme-clé de référence générique : le terme-clé de référence n'est pas suffisamment spécifique au contenu du document, il a un usage classificatoire.

FIGURE 5.9 – Interface d'évaluation manuelle de l'Inist

Afin de minimiser les pertes d'informations dues à des termes-clés de référence qui ne sont pas présents dans le document, les évaluateurs leur attribuent un score de 1.

5.4.2 Évaluation manuelle des méthodes proposées

Dans cette section, nous analysons l'évaluation manuelle de TopicRank, TopicCoRank, d'une méthode de référence non supervisée, TF-IDF, et d'une méthode de référence supervisée, KEA[5]. L'évaluation est effectuée par un indexeur professionnel, sur la collection de linguistique Termith.

Dans un premier temps, nous analysons l'évaluation manuelle de TopicRank et la comparons à celle de TF-IDF, puis, dans un second temps, nous analysons l'évaluation manuelle de TopicCoRank et la comparons à celle de KEA. Tous les termes-clés sont obtenus à partir des documents prétraités, comme nous l'avons présenté dans la section 3.7 (page 47). Les candidats sont sélectionnés à l'aide du patron grammatical `/(N|A)+/`.

Évaluation de TopicRank

Le tableau 5.6 montre les scores de pertinence moyens de TopicRank et TF-IDF. Pour le score de 1, qui indique qu'un terme-clé est un forme variante, nous distinguons le cas où la variante est redondante du cas où elle ne l'est pas. Nous observons que TopicRank est meilleur que TF-IDF. TopicRank fournit plus de termes-clés pertinents que TF-IDF, mais fait aussi plus d'erreurs. Les termes-clés ayant un score de 1 donnent un explication intéressante à cette contradiction : TF-IDF à une forte tendance à extraire des termes-clés redondants, c'est-à-dire des termes-clés variantes de termes-clés déjà extraits, alors que TopicRank remplit son objectif de ne pas extraire de termes-clés redondants, avec seulement

[5]La raison pour laquelle les évaluations manuelles sont réalisées sur KEA et non pas KEA++ est temporelle. L'évaluation manuelle étant coûteuse, nous n'avons pas pu la réitérer avec KEA++.

0,9 % de redondance.

Méthode	0	1		2
		redondant	non redondant	
TF-IDF	**53,8 %**	6,8 %	4,2 %	35,3 %
TopicRank	56,3 %	**0,9 %**	**5,7 %**	**37,1 %**

TABLE 5.6 – Taux de termes-clés avec un score de 0, de 1 ou de 2 pour l'évaluation de la pertinence de TF-IDF et de TopicRank

Les résultats présentés dans le tableau 5.6 sont contraires à ceux de l'évaluation automatique, puisque c'est ici TopicRank qui est meilleur que TF-IDF sur les documents de linguistique. Afin de mieux observer la différence entre l'évaluation manuelle et l'évaluation automatique, nous calculons la précision (P), le rappel (R) et la f1-mesure (F) résultantes de l'évaluation manuelle et les comparons aux résultats automatiques que nous avons montré dans le tableau 4.9 (page 68). Pour calculer ces performances, les termes-clés ayant un score de 2 sont considérés corrects, de même que ceux ayant un score de 1 non redondants. Les résultats de l'évaluation manuelle comparés à ceux de l'évaluation automatique sont présentés dans le tableau 5.7.

La difficulté d'évaluer automatiquement la tâche d'indexation par termes-clés se confirme. Avec l'évaluation manuelle, les conclusions ne sont pas les mêmes, puisque de manière automatique TopicRank est moins performant que TF-IDF alors qu'il est plus performant selon l'évaluation manuelle. Nous observons aussi un écart conséquent entre les performances évaluées manuellement et celles évaluées automatiquement. Le gain d'environ 20 points atteste le pessimisme de l'évaluation automatique.

Méthode	Évaluation manuelle			Évaluation automatique		
	P	R	F	P	R	F
TF-IDF	39,5	29,7	33,5	**13,0**	**15,4**	**13,9**
TopicRank	**42,8**	**32,2**	**36,2**	11,2	13,1	11,9

TABLE 5.7 – Performances de TF-IDF et de TopicRank en termes de précision (P), de rappel (R) et de f1-mesure (F)

Enfin, le tableau 5.8 montre les scores de silence attribués en moyenne par méthode. D'après la description donnée pour chacun des scores, la méthode qui capture le plus d'information est celle qui maximise le nombre de termes-clés de référence ayant un score de silence 0 et qui minimise ceux ayant un score de 1 et de 2. Nous observons donc que TopicRank couvre mieux le contenu principal des documents que TF-IDF. Parce que TopicRank groupe les termes-clés candidats en sujets et n'extrait qu'un seul terme-clé par sujet, il y a moins de redondance parmi les termes-clés qu'il extrait (voir le tableau 5.6) et le nombre de sujets couverts est donc meilleur.

Méthode	0	1	2
TF-IDF	31,4 %	48,5 %	20,1 %
TopicRank	**35,0 %**	**48,3 %**	**16,8 %**

TABLE 5.8 – Taux de termes-clés de référence avec un score de 0, de 1 ou de 2 pour l'évaluation du silence de TF-IDF et de TopicRank

Évaluation de TopicCoRank

Les résultats que nous montrons dans cette section sont obtenus avec seulement 25 % de l'ensemble de test de la collection linguistique Termith[6]. Les différentes revues qui composent la collection sont réparties de manière homogène dans ces 25 %.

Le tableau 5.9 montre les scores de pertinence moyens de TopicCoRank et KEA. Nous observons que TopicCoRank est meilleur que KEA. TopicCoRank fournit plus de termes-clés pertinents que KEA, mais fait aussi plus d'erreurs. À l'instar de TopicRank, TopicCoRank est moins redondant que la méthode de référence. Il est tout de même plus redondant que TopicRank. Cela est du au fait qu'il peut assigner un terme-clé et extraire une variante pouvant être jugée importante, car plus spécifique, ou redondante.

Méthode	0	1		2
		redondant	non redondant	
KEA	**45,2 %**	10,2 %	**8,0 %**	36,6 %
TopicCoRank	49,8 %	**4,4 %**	6,2 %	**39,6 %**

TABLE 5.9 – Taux de termes-clés avec un score de 0, de 1 ou de 2 pour l'évaluation de la pertinence de KEA et de TopicCoRank

Comparé à TopicRank, TopicCoRank est effectivement plus performant. Il trouve plus de termes-clés corrects et fait moins d'erreurs. Cependant, comme nous l'avons dit ci-dessus, il est plus redondant. Comparée à TF-IDF et KEA, cette redondance est tout de même plus faible.

Enfin, le tableau 5.10 montre les scores de silence attribués en moyenne par méthode. Les conclusions qu'ils induisent ne sont pas en faveur de TopicCoRank. En effet, même s'il ne capture pas autant de termes-clés que TopicCoRank, KEA capture ceux qui sont sémantiquement les plus indispensables. Bien que les deux méthodes sont supervisées, KEA est la seule des deux à apprendre à détecter les termes-clés à partir des traits des termes-clés du domaine. Si l'ordonnancement conjoint à base de graphe est plus efficace pour identifier les termes-clés, l'apprentissage permet une meilleure précision quant à l'identification de ceux les plus indispensables.

5.4.3 Bilan

Nous avons réalisé une campagne d'évaluation manuelle de nos travaux en domaine de spécialité, avec la collection de linguistique Termith. Pour cette campagne, nous avons pro-

[6]En raison de cette incomplétude de l'évaluation manuelle, nous ne comparons pas l'évaluation manuelle à l'évaluation automatique comme nous l'avons fait pour TopicRank et TF-IDF.

Méthode	0	1	2
KEA	**37,3 %**	46,0 %	**16,7** %
TopicCoRank	35,9 %	**45,3 %**	18,8 %

TABLE 5.10 – Taux de termes-clés de référence avec un score de 0, de 1 ou de 2 pour l'évaluation du silence de KEA et de TopicCoRank

posé un protocole d'évaluation et des métriques permettant de capturer deux aspects de l'indexation par termes-clés : la pertinence des termes-clés extraits/assignés et leur silence, c'est-à-dire la quantité d'information importante qu'ils ne capturent pas. Contrairement au premier aspect, qui est similaire à ce qu'évalue un système automatique, le dernier aspect permet d'évaluer les termes-clés d'un point de vu sémantique, jamais considéré auparavant.

Les résultats montrent que, contrairement à ce que montrait l'évaluation automatique, TopicRank effectue une indexation par termes-clés de meilleure qualité que celle de TF-IDF. TopicRank extrait peu de termes-clés redondants et couvre mieux le document, en partie grâce à son groupement des termes-clés candidats en sujets. Entre TopicCoRank, TopicRank, TF-IDF et KEA, c'est TopicCoRank qui trouve le plus de termes-clés corrects. L'évaluation du silence montre tout de même que c'est KEA qui identifie ceux jugés les plus indispensables par les indexeurs professionnels. Les deux aspects (pertinence et silence) de l'évaluation manuelle et leurs conclusions paradoxales soulèvent une nouvelle perspective pour l'évaluation automatique. En effet, il serait intéressant d'ordonner par importance les termes-clés de référence et d'en tenir compte pour identifier les méthodes qui capturent les termes-clés les plus indispensables.

5.5 Conclusion

Nous nous sommes intéressé à l'indexation automatique par termes-clés en domaines de spécialité. Nous avons tout d'abord présenté l'indexation manuelle réalisée par des indexeurs professionnels dans ce contexte, nous avons ensuite proposé une nouvelle méthode automatique se rapprochant le plus possible de cette indexation, puis nous avons présenté les premiers résultats d'une campagne d'évaluation manuelle que nous avons réalisé avec des indexeurs professionnels.

Contrairement aux méthodes d'indexation automatique, l'indexation manuelle n'est pas divisée entre extraction et assignement. L'indexation manuelle en domaines de spécialité préfère l'assignement, car cela permet une indexation homogène des documents d'un même domaine et une conformité vis-à-vis du vocabulaire spécialisé du domaine. Elle a aussi besoin de l'extraction, afin de fournir des termes-clés très spécifiques au document, ainsi que pour y identifier d'éventuels nouveaux concepts.

Pour remédier à la fracture entre extraction et assignement en indexation automatique par termes-clés, nous proposons TopicCoRank. Conçu sur la base de TopicRank, TopicCoRank utilise les données d'entraînement pour représenter le domaine avec un graphe unifié à celui des sujets. Cette unification permet d'améliorer l'ordonnancement des sujets, en tenant compte de leurs relations avec le domaine, et d'assigner des termes-clés, en puisant dans le domaine. À notre connaissance, TopicCoRank est la seule méthode qui réalise conjointement extraction et assignement.

Pour valider les deux méthodes TopicRank et TopicCoRank, nous avons réalisé une campagne d'évaluation manuelle en domaine de spécialité. Le protocole d'évaluation que nous avons proposé permet d'évaluer chaque méthode selon le degré de pertinence des termes-clés qu'elle propose et selon le degré d'information qui lui échappe. Les résultats de l'évaluation manuelle de TopicRank son plus encourageant que ceux de l'évaluation automatique. Ils montrent que TopicRank est en réalité plus performant que TF-IDF, en partie parce qu'il couvre mieux les sujets du document grâce à son groupement en sujets des termes-clés candidats. Ils montrent aussi que TopicCoRank est la méthode la plus performante comparée à TopicRank, TF-IDF et KEA. Cependant, c'est KEA qui trouve les termes-clés les plus indispensables. Au delà de cela, cette campagne a montré les limites de l'évaluation automatique, qui suit un paradigme trop strict pour la tâche d'indexation par termes-clés. Parce que l'évaluation manuelle est trop coûteuse pour être systématiquement mise en œuvre, il est donc important de s'intéresser de plus près aux méthodes d'évaluation automatique. Les ressources de notre campagne, annotées étape par étape, seront donc rendues disponibles gratuitement à toute la communauté scientifique. Cette disponibilité permettra d'évaluer de nouvelles méthodes d'évaluation, en vérifiant leur corrélation avec l'évaluation manuelle des indexeurs professionnels.

6

Conclusion et perspectives

« [...] la tâche est loin d'être résolue [...] »
— Hasan et Ng (2014)

Dans cette thèse, nous nous sommes intéressé à la tâche d'indexation par termes-clés en domaines de spécialité. Étant donné un document textuel, cette tâche consiste à lui attribuer les unités textuelles qui décrivent son contenu. Ces unités textuelles, les termes-clés, permettent de le résumer, de le catégoriser et, surtout, de l'indexer pour la recherche d'information. Les termes-clés peuvent être attribués par les auteurs, des lecteurs ou des indexeurs professionnels, mais seuls ces derniers réalisent un travail impartial, homogène et conforme à une indexation de qualité en domaines de spécialité. Notre objectif est de proposer une alternative automatique aux indexeurs professionnels pour une indexation par termes-clés de documents numériques en domaines de spécialité. Tout d'abord, nous avons fait le choix de traiter ce problème dans sa généralité, puis nous nous sommes ensuite concentré sur les documents en domaines de spécialité et sur leur indexation particulière effectuée par des indexeurs professionnels.

Dans la littérature, de nombreuses méthodes sont proposées pour l'indexation automatique par termes-clés. Elles sont réparties en deux catégories : l'extraction et l'assignement de termes-clés. La première extrait les termes-clés depuis le contenu du document. Il s'agit de la catégorie d'indexation par termes-clés la plus étudiée. Elle est plus simple à mettre en œuvre, car elle traite les unités textuelles du document. Cependant, ces unités textuelles ne sont pas toujours sous une forme appropriée. La seconde assigne les termes-clés depuis un vocabulaire contrôlé représentatif du langage documentaire. Elle assure donc une indexation de meilleure qualité. Cependant, elle est plus difficile à mettre en œuvre, car les entrées du vocabulaire contrôlé ne sont pas nécessairement présentes dans le document. Par ailleurs, elle n'est pas capable d'identifier des termes-clés très spécifiques au document s'ils ne sont pas dans le vocabulaire contrôlé, ni même des nouveaux concepts. Pour traiter le problème d'indexation par termes-clés dans sa généralité, nous avons travaillé sur l'extraction de termes-clés. Pour l'indexation en domaines de spécialité, nous avons proposé une

extension de ce travail afin d'y intégrer la capacité à réaliser l'assignement.

6.1 Contributions

Nos travaux ont fait l'objet de trois contributions : deux contributions pour l'extraction de termes-clés et une contribution pour l'indexation par termes-clés en domaines de spécialité. Nous avons aussi proposé un protocole d'évaluation pour une campagne d'évaluation manuelle de nos travaux.

Notre première contribution à l'extraction de termes-clés concerne l'étape préliminaire de sélection des termes-clés candidats. Elle consiste à identifier les unités textuelles du document susceptibles d'être des termes-clés. Dans la littérature, cette étape est très souvent réalisée à l'aide de règles simples qui ont tendance à sélectionner beaucoup de candidats. Or, nous émettons l'hypothèse que l'indexation gagne en efficacité lorsque la sélection des candidats fournie un ensemble de petite taille contenant le plus possible de termes-clés corrects. En nous fondant sur une analyse des propriétés linguistiques des termes-clés de référence, nous avons d'abord proposé une méthode qui limite le nombre de candidats sélectionnés en ciblant les séquences de noms modifiés, ou non, par un adjectif utile (non superflu). Un adjectif utile se distingue par sa catégorie, relationnel ou composé complexe, et son usage fréquent dans le document ; un adjectif superflu est un adjectif qualificatif qui modifie n'importe quel nom. Nous avons ensuite vérifié notre hypothèse en évaluant la qualité de l'ensemble de candidats sélectionnés par notre méthode, ainsi que son impact sur deux méthodes d'extraction de termes-clés : TF-IDF (Salton *et al.*, 1975) et KEA (Witten *et al.*, 1999). Les résultats ont montré que notre méthode est capable de réduire le nombre de candidats sélectionnés sans éliminer un nombre significatif de termes-clés corrects. Ils montrent aussi qu'elle a une meilleure influence sur la performance des méthodes d'extractions employées.

Notre seconde contribution à l'extraction de termes-clés s'intéresse aux méthodes qui ordonnent les termes-clés candidats par importance, puis extraient les k plus importants en tant que termes-clés. Selon nous, ce ne sont pas les termes-clés candidats qui doivent être ordonnés par importance, mais ce qu'ils représentent : leur sujet. De plus, si plusieurs unités textuelles représentent le même sujet, alors elles doivent être considérées comme une entité unique. Nous avons donc proposé, TopicRank, une méthode à base de graphe qui commence par grouper les termes-clés candidats en sujets, représente le document avec un graphe des sujets, ordonne ces sujets à l'aide d'un algorithme de marche aléatoire dans le graphe, puis extrait un, et un seul, terme-clé candidat pour chacun des k meilleurs sujets. Nos expériences ont montré une amélioration de l'extraction de termes-clés avec TopicRank, comparée à celle réalisée avec les autres méthodes à base de graphe, TextRank (Mihalcea et Tarau, 2004) et SingleRank (Wan et Xiao, 2008). Au travers d'évaluations manuelles, nous avons aussi pu montrer que TopicRank extrait des termes-clés non redondants, lui permettant de mieux couvrir les sujets du document que les autres méthodes.

Notre troisième contribution s'intéresse à l'indexation par termes-clés en domaines de spécialité telle qu'elle est effectuée par un indexeur professionnel. Ces indexeurs mêlent extraction et assignement, avec une préférence pour l'assignement. L'assignement permet d'obtenir une indexation homogène de tous les documents d'un même domaine, une indexation conforme au vocabulaire de ce domaine et une généralisation du contenu de chaque document afin de le situer dans son domaine. L'extraction, quant à elle, permet d'améliorer

l'exhaustivité de l'indexation en ajoutant des termes-clés très spécifiques au document, voir même de nouveaux concepts. Nous faisons donc l'hypothèse qu'extraction et assignement doivent être réalisés conjointement grâce à une contextualisation du contenu du document dans son domaine. Cette contextualisation doit (1) permettre de déterminer l'importance des sujets en tenant aussi compte de la place qu'ils occupent dans le domaine et (2) permettre l'assignement en déterminant les termes-clés du domaines importants vis-à-vis du document. Pour cela, nous avons étendu notre seconde contribution en ajoutant un graphe du domaine, représenté par son vocabulaire contrôlé (ses termes-clés). Notre nouvelle méthode, TopicCoRank, représente le domaine du document avec un graphe des termes-clés de référence attribués à des documents du même domaine, le connecte au graphe de sujets et ordonne conjointement sujets et termes-clés du domaine. Les termes-clés obtenus à partir du graphe de sujets sont extraits et les termes-clés obtenus à partir du graphe du domaine sont assignés. En domaine de spécialité, TopicCoRank obtient des résultats supérieurs à l'état de l'art. À notre connaissance, il s'agit de la première méthode capable de réaliser simultanément extraction et assignement.

Enfin, nous avons participé à la mise en place d'une campagne d'évaluation manuelle en domaines de spécialité. Nous avons proposé, en collaboration avec les indexeurs professionnels de l'Inist, un protocole permettant d'évaluer deux aspects de l'indexation : la pertinence (validité) et le silence (perte d'information). Le premier aspect est celui qui est aussi évalué par l'évaluation automatique. Les résultats obtenus lors de la campagne ont tout de même montré l'importance de réaliser une évaluation manuelle. Celle-ci est moins pessimiste (plus juste) et nous observons un gain de plus de 20 points de f1-mesure par rapport à l'évaluation automatique. Le second aspect est nouveau pour l'évaluation de méthodes d'indexation par termes-clés. Il est purement sémantique et permet de déterminer si, au delà de fournir un grand nombre de termes-clés corrects, la méthode capture les informations les plus importantes du document. Toutes les ressources de cette campagne d'évaluation seront rendues disponibles. Elles sont annotées avec chaque étape (indexation automatique par termes-clés et évaluation manuelle) afin de permettre l'étude de nouvelle méthodes d'évaluation automatique, notamment vérifier leur corrélation avec le jugement humain.

6.2 Perspectives

Nos contributions ont montré des améliorations en matière de sélection de termes-clés, d'extraction de termes-clés et d'indexation par termes-clés en domaines de spécialité. Elles ont toutefois des limites et il reste encore plusieurs perspectives de travail.

Nous identifions trois limites de notre travail s'intéressant à la sélection des termes-clés candidats pour l'extraction de termes-clés. Premièrement, notre étude linguistique des termes-clés s'est limitée aux adjectifs composés complexes et aux adjectifs relationnels, alors qu'il existe d'autres catégories d'adjectifs. Les adjectifs relationnels, qui sont des dénominaux, ont montré leur utilité au sein des termes-clés. Nous envisageons donc d'étudier n'importe quel adjectif dénominal, en nous affranchissant de ses autres propriétés, ainsi que les adjectifs déverbaux. Ensuite, nous avons mis de côté les prépositions (et les déterminants) en français, alors qu'ils sont constitutifs d'environ $1/3$ des termes-clés. L'attachement prépositionnel au nom est ambigu (Colonna et Pynte, 2002) et les prépositions (et les déterminants) étant d'usage fréquent dans tout discours (prépositions et déterminants font partie des mots les plus fréquents du français), cette ambiguïté doit absolument être résolue pour

éviter d'ajouter un nombre conséquent d'erreurs dans l'ensemble des termes-clés candidats.

Notre travail sur TopicRank possède aussi quelques limitations. Tout d'abord, le groupement en sujets que nous avons proposé est naïf. Il ne tient pas compte du lien de synonymie des mots, ni même de leur sens dans leurs contextes (problème d'ambiguïté). Nous avons proposé ce groupement car il permet à TopicRank d'être applicable dans toutes les situations, sans nécessiter de ressources particulières. Lorsque les données mises à dispositions le permettent, il serait intéressant d'envisager d'autres méthodes de groupement. Nous pourrions analyser la sémantique latente au sein d'une collection donnée, comme l'ont fait Liu *et al.* (2010), Ding *et al.* (2011) et Zhang *et al.* (2013) avec LDA (Blei *et al.*, 2003). Nous pourrions aussi nous appuyer sur des ressources lexicales, ou des méthodes automatiques, pour détecter les candidats qui sont synonymes et les grouper. Une autre limite de notre travail est le choix non optimal du terme-clé candidat à extraire pour un sujet. Bien que la stratégie que nous employons donne des résultats satisfaisants, nous avons établi que la stratégie optimale permettrait d'atteindre des performances deux fois supérieures dans certains cas. Une première solution serait d'utiliser une collection d'entraînement pour apprendre à reconnaître le terme-clé au sein d'un sujet. Certains traits utilisés par les méthodes supervisés pourraient servir (position de la première occurrence, nombre de mots, etc.), ainsi que de nouveaux traits liés à la relation qu'entretient le candidat avec les autres candidats du sujet (degré de similarité avec tous les autres, catégorie grammaticale des mots en communs avec les autres, etc.). Une autre solution serait d'appliquer des méthodes de titrage automatique, telles que celle de Lau *et al.* (2011). L'avantage de cette méthode est de permettre de générer des unités textuelles à partir d'un sujet. Dans notre cas, elle permet donc de proposer des termes-clés qui n'occurrent pas nécessairement dans le document. Correctement paramétrée avec un vocabulaire contrôlé, une méthode de titrage automatique générative peut donc être une alternative à TopicCoRank pour réaliser extraction et assignement.

TopicCoRank, qui permet de réaliser simultanément extraction et assignement de termes-clés possède aussi quelques limites. Premièrement, bien que l'ordonnancement conjoint des sujets du document et des termes-clés du domaine améliore l'ordonnancement, nous avons observé que TopicCoRank est plus performant en domaines de spécialité lorsqu'il ne réalise que l'assignement. Cela signifie que les termes-clés du domaine sont correctement identifiés parmi tous les termes-clés de celui-ci, mais que leur importance est trop faible comparée à celle qui est attribuée aux sujets du document. Ce problème peut être résolu de deux manières. Il existe peut être un schéma de pondération des arêtes et d'unification des deux graphes plus performant que celui que nous proposons. Pour unifier les deux graphes, nous pourrions par exemple nous intéresser au contenu des documents de référence. Ainsi, un terme-clé du domaine pourrait être connecté à un sujet lorsqu'il est le terme-clé d'un document dans lequel le sujet apparaît. Une autre manière de résoudre ce problème serait aussi de faire varier l'influence de la recommandation interne (paramètre λ) différemment pour les sujets et les termes-clés de référence. En augmentant plus fortement l'impact de la recommandation issue des sujets (recommandation externe) avec une faible valeur pour λ, l'importance des termes-clés du domaine devrait augmenter, leur permettant de rivaliser avec les sujets dans le classement par importance. Le paramétrage de λ peut aussi résoudre un autre problème de TopicCoRank. En effet, nous avons aussi observé que TopicCoRank fonctionne moins bien hors domaines de spécialité, il est difficilement généralisable. Utiliser un paramètre λ différent pour calculer l'importance des sujets et l'importance des termes-clés peut aussi résoudre ce problème. Nous pourrions le paramétrer empiriquement

avec les données d'entraînement, mais nous aimerions aussi chercher à prédire sa valeur. En effet, nous avons expliqué que le problème de généralisation de TopicCoRank est dû au fait que les données d'entraînement des collections que nous avons utilisé hors domaines de spécialité ne sont pas adaptées. Il serait donc intéressant de voir si nous pouvons évaluer avec quel degré les données sont adaptées et de voir s'il existe une corrélation entre celui-ci et le paramétrage de λ. Enfin, les données en domaines de spécialité que nous utilisons ne contiennent que des notices. Il serait intéressant de voir les différences de performance en utilisant les articles complets que représentent ces notices.

L'ensemble des travaux de cette thèse permet d'améliorer l'indexation automatique par termes-clés. Nos travaux en extraction de termes-clés sont applicables dans presque tous les scénarios d'utilisation. Aucune ressource externe n'est nécessaire à l'exception de celles requises par les outils de prétraitement (segmentation et étiquetage grammatical). Ceux en extraction et assignement simultanés de termes-clés sont applicables en domaines de spécialité, dès lors que des données d'entraînement sont disponibles. Dans le cadre du projet Termith, les résultats de nos travaux seront exploités pour faciliter la recherche d'information avec les outils que propose l'Inist. TopicCoRank identifie plus de termes-clés corrects que les autres méthodes, ce qui facilitera le travail des indexeurs professionnels. Il est aussi envisagé de s'en servir pour réaliser de la veille terminologique, c'est-à-dire mettre à jour des bases terminologiques de domaines de spécialité avec les termes-clés validés par un indexeur professionnel.

Liste des publications

Publication en revue

TopicRank : ordonnancement de sujets pour l'extraction automatique de termes-clés
Adrien Bougouin et Florian Boudin
(Bougouin et Boudin, 2014)

Résumé : Les termes-clés sont les mots ou les expressions polylexicales qui représentent le contenu principal d'un document. Ils sont utiles pour diverses applications telles que l'indexation automatique ou le résumé automatique, mais ne sont cependant pas disponibles pour la plupart des documents. La quantité de ces documents étant de plus en plus importante, l'extraction manuelle des termes-clés n'est pas envisageable et la tâche d'extraction automatique de termes-clés suscite alors l'intérêt des chercheurs. Dans cet article nous présentons *TopicRank*, une méthode non supervisée à base de graphe pour l'extraction de termes-clés. Cette méthode groupe les termes-clés candidats en sujets, ordonne les sujets et extrait de chacun des meilleurs sujets le terme-clé candidat qui le représente le mieux. Les expériences réalisées montrent une amélioration significative vis-à-vis de l'état de l'art des méthodes à base de graphe pour l'extraction non supervisée de termes-clés.

Publié dans la revue Traitement Automatique des Langues (TAL 55-1).

Publication en conférence internationale avec actes

TopicRank : Graph-Based Topic Ranking for Keyphrase Extraction
Adrien Bougouin, Florian Boudin et Béatrice Daille
(Bougouin *et al.*, 2013)

Abstract : *Keyphrase extraction is the task of identifying single or multi-word expressions that represent the main topics of a document. In this paper we present TopicRank, a graph-based keyphrase extraction method that relies on a topical representation of the document. Candidate keyphrases are clustered into topics and used as vertices in a complete graph. A graph-based ranking model is applied to assign a significance score to each topic. Keyphrases are then generated by selecting a candidate from each of the top-ranked topics. We conducted experiments on four evaluation datasets of different languages and domains. Results show that TopicRank significantly outperforms state-of-the-art methods on three datasets.*

Publié dans les actes de la conférence *International Joint Conference on Natural Language Processing (*IJCNLP*)*.

Publications en conférence national avec actes

Influence des domaines de spécialité dans l'extraction de termes-clés
Adrien Bougouin, Florian Boudin et Béatrice Daille
(Bougouin *et al.*, 2014)

Résumé : Les termes-clés sont les mots ou les expressions polylexicales qui représentent le contenu principal d'un document. Ils sont utiles pour diverses applications, telles que l'indexation automatique ou le résumé automatique, mais ne sont pas toujours disponibles. De ce fait, nous nous intéressons à l'extraction automatique de termes-clés et, plus particulièrement, à la difficulté de cette tâche lors du traitement de documents appartenant à certaines disciplines scientifiques. Au moyen de cinq corpus représentant cinq disciplines différentes (archéologie, linguistique, sciences de l'information, psychologie et chimie), nous déduisons une échelle de difficulté disciplinaire et analysons les facteurs qui influent sur cette difficulté.

Publié dans les actes de la conférence Traitement Automatique du Langage Naturel (TALN).

État de l'art des méthodes d'extraction automatique de termes-clés
Adrien Bougouin, Florian Boudin et Béatrice Daille
(Bougouin, 2013)

Résumé : Cet article présente les principales méthodes d'extraction automatique de termes-clés. La tâche d'extraction automatique de termes-clés consiste à analyser un document pour en extraire les expressions (phrasèmes) les plus représentatives de celui-ci. Les méthodes d'extraction automatique de termes-clés sont réparties en deux catégories : les méthodes supervisées et les méthodes non supervisées. Les méthodes supervisées réduisent la tâche d'extraction de termes-clés à une tâche de classification binaire (tous les phrasèmes sont classés parmi les termes-clés ou les non termes-clés). Cette classification est possible grâce à une phase préliminaire d'apprentissage, phase qui n'est pas requise par les méthodes non-supervisées. Ces dernières utilisent des caractéristiques (traits) extraites du document analysé (et parfois d'une collection de documents de références) pour vérifier des propriétés permettant d'identifier ses termes-clés.

Publié dans les actes de la conférence Rencontre de Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL).

Bibliographie

Janin ADAM, Don BARON, Jane EDWARDS, Dan ELLIS, David GELBART, Nelson MORGAN, Barbara PESKIN, Thilo PFAU, Elizabeth SHRIBERG, Andreas STOLCKE et Chuck WOOTERS : The ICSI Meeting Corpus. *In Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, volume 1, pages I–364–I–367 vol.1, April 2003. 20

Charles BALLY : Linguistique générale et linguistique française. *Paris : Ernest Leroux*, 1944. 51

Ken BARKER et Nadia CORNACCHIA : Using Noun Phrase Heads to Extract Document Keyphrases. *In Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence : Advances in Artificial Intelligence*, pages 40–52, London, UK, 2000. Springer-Verlag. ISBN 3-540-67557-4. 21, 50

Steven BIRD, Ewan KLEIN et Edward LOPER : *Natural Language Processing with Python*. O'Reilly Media, 2009. 47

David M. BLEI, Andrew Y. NG et Michael I. JORDAN : Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. ISSN 1532-4435. 23, 26, 100

Mourad BOUGHEDAOUI : Contribution à l'amélioration de la compréhension et de la traduction des adjectifs composés en classe de langue de spécialité. *Anglais et français de spécialité (ASp)*, 15–18:525–541, 1997. 52

Adrien BOUGOUIN : État de l'art des méthodes d'extraction automatique de termes-clés. *In Actes de la 15e Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL'2013)*, pages 96–109, Sables d'Olonne, France, 2013. 104

Adrien BOUGOUIN et Florian BOUDIN : TopicRank : ordonnancement de sujets pour l'extraction automatique de termes-clés. *TAL*, 55(1):45–69, 2014. 103

Adrien BOUGOUIN, Florian BOUDIN et Béatrice DAILLE : Topicrank : Graph-Based Topic Ranking for Keyphrase Extraction. *In Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 543–551, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing. 44, 103

Adrien BOUGOUIN, Florian BOUDIN et Béatrice DAILLE : Influence des domaines de spécialité dans l'extraction de termes-clés. *In Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles*, pages 13–24, Marseille, France, July 2014. Association pour le Traitement Automatique des Langues. 104

Sergey BRIN et Lawrence PAGE : The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1):107–117, 1998. 24, 25, 61

Peter F. BROWN, Vincent J. Della PIETRA, Stephen A. Della PIETRA et Robert L. MERCER : The Mathematics of Statistical Machine Translation : Parameter Estimation. *Computational Linguistics*, 19(2):263–311, 1993. 33

Cornelia CARAGEA, Florin Adrian BULGAROV, Andreea GODEA et Sujatha Das GOLLAPALLI : Citation-Enhanced Keyphrase Extraction from Research Papers : A Supervised Approach. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1435–1446, Doha, Qatar, October 2014. Association for Computational Linguistics. 28, 29

Panot CHAIMONGKOL et Akiko AIZAWA : Utilizing LDA Clustering for Technical Term Extraction. *In Proceedings of the 19th Annual Meeting of the Association for Natural Language Processing (ANLP)*, pages 686–689, Nagoya, Japan, March 2013. Association for Natural Language Processing. 77

Vincent CLAVEAU : Vectorisation, Okapi et calcul de similarité pour le TAL : pour oublier enfin le TF-IDF (Vectorization, Okapi and Computing Similarity for NLP : Say Goodbye to TF-IDF) [in French]. *In Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, Volume 2 : TALN*, pages 85–98, Grenoble, France, June 2012. ATALA/AFCP. 21

Saveria COLONNA et Joël PYNTE : La levée des ambiguïtés syntaxiques : apport des recherches inter-langues. *L'Année Psychologique*, 102(1):151–187, 2002. 99

Béatrice DAILLE : Morphological Rule Induction for Terminology Acquistion. *In Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00, pages 215–221, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics. ISBN 1-55860-717-X. 52

Pascal DENIS et Benoît SAGOT : Coupling an Annotated Corpus and a Morphosyntactic Lexicon for State-of-the-Art POS Tagging with Less Human Effort. *In Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC)*, pages 110–119, Hong Kong, December 2009. City University of Hong Kong. 47

John DENKER et Yann LECUN : Transforming Neural-Net Output Levels to Probability Distributions. *In Advances in Neural Information Processing Systems 3*, pages 853–859. Morgan Kaufmann, 1991. 32

Zhuoye DING, Qi ZHANG et Xuanjing HUANG : Keyphrase Extraction from Online News Using Binary Integer Programming. *In Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 165–173, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing. 16, 22, 23, 28, 100

Jean DUBOIS et Françoise DUBOIS-CHARLIER : *La dérivation suffixale en français*. Nathan, 1999. 52

Kathrin EICHLER et Günter NEUMANN : DFKI KeyWE : Ranking Keyphrases Extracted from Scientific Articles. *In Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval)*, pages 150–153, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. 31, 51

Gonenc ERCAN et Ilyas CICEKLI : Using Lexical Chains for Keyword Extraction. *Information Processing and Management*, 43(6):1705–1714, nov 2007. ISSN 0306-4573. 9, 30, 31

Joseph L FLEISS : Measuring Nominal Scale Agreement among many Raters. *Psychological bulletin*, 76(5):378, 1971. 7, 44, 45

Eibe FRANK, Gordon W. PAYNTER, Ian H. WITTEN, Carl GUTWIN et Craig G. Nevill MANNING : Domain-Specific Keyphrase Extraction. *In Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*, pages 668–673, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. 28, 31

Natalia GRABAR et Thierry HAMON : Terminology Structuring Through the Derivational Morphology. *In Proceedings of the 5th International Conference on Advances in Natural Language Processing*, pages 652–663. Springer-Verlag, 2006. 52

Claire GUINCHAT et Yolande SKOURI : *Guide pratique des techniques documentaires*. Edicef, 1996. 12, 13, 38, 73

Rima HARASTANI, Béatrice DAILLE et Emmanuel MORIN : Identification, alignement, et traductions des adjectifs relationnels en corpus comparables. *In Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, pages 313–326, Les Sables d'Olonne, France, 2013. 52

Kazi Saidul HASAN et Vincent NG : Conundrums in Unsupervised Keyphrase Extraction : Making Sense of the State-of-the-Art. *In Proceedings of the 23rd International Conference on Computational Linguistics : Posters (COLING)*, pages 365–373, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. 20, 24, 25, 37, 38, 62

Kazi Saidul HASAN et Vincent NG : Automatic Keyphrase Extraction : A Survey of the State of the Art. *In Proceedings of the Association for Computational Linguistics (ACL)*, Baltimore, Maryland, June 2014. Association for Computational Linguistics. 12, 13, 37, 38, 49, 54, 57, 67, 97

Chong HUANG, Yonghong TIAN, Zhi ZHOU, Ling C.X. et Tiejun HUANG : Keyphrase Extraction Using Semantic Networks Structure Analysis. *In Data Mining, 2006. ICDM '06. Sixth International Conference on*, pages 275–284, December 2006. 18, 35

Anette HULTH : Improved Automatic Keyword Extraction Given More Linguistic Knowledge. *In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 216–223, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. 17, 18, 20, 50

Nancy IDE et Jean VÉRONIS : MULTEXT : Multilingual Text Tools and Corpora. *In Proceedings of the 15th conference on Computational Linguistics-Volume 1*, pages 588–592. Association for Computational Linguistics, 1994. 54

Xin JIANG, Yunhua HU et Hang LI : A Ranking Approach to Keyphrase Extraction. *In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 756–757, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-483-6. 31

Su Nam KIM, Olena MEDELYAN, Min-Yen KAN et Timothy BALDWIN : SemEval-2010 task 5 : Automatic Keyphrase Extraction from Scientific Articles. *In Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval)*, pages 21–26, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. 12, 30, 45, 46, 49, 56, 64

Zornitsa KOZAREVA, Irina MATVEEVA, Gabor MELLI et Vivi NASTASE, éditeurs. *Proceedings of TextGraphs-8 Graph-Based Methods for Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, October 2013. 24

Jey Han LAU, Karl GRIESER, David NEWMAN et Timothy BALDWIN : Automatic Labelling of Topic Models. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, pages 1536–1545, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. 71, 100

Vicky Min-How LIM, Siew Fan WONG et Tong Ming LIM : Examining the Value of Attribute Scores for Author-Supplied Keyphrases in Automatic Keyphrase Extraction. *International Scholarly and Scientific Research & Innovation*, 6(12):455–460, 2012. 61

Kang LIU, Liheng XU et Jun ZHAO : Extracting Opinion Targets and Opinion Words from Online Reviews with Graph Co-ranking. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 314–324, Baltimore, Maryland, June 2014. Association for Computational Linguistics. 27

Zhiyuan LIU, Xinxiong CHEN, Yabin ZHENG et Maosong SUN : Automatic Keyphrase Extraction by Bridging Vocabulary Gap. *In Proceedings of the 15th Conference on Computational Natural Language Learning*, pages 135–144, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-92-3. 33

Zhiyuan LIU, Wenyi HUANG, Yabin ZHENG et Maosong SUN : Automatic Keyphrase Extraction Via Topic Decomposition. *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing(EMNLP)*, pages 366–376, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. 9, 26, 70, 100

Zhiyuan LIU, Peng LI, Yabin ZHENG et Maosong SUN : Clustering to Find Exemplar Terms for Keyphrase Extraction. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 1 (EMNLP)*, pages 257–266, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. 24

Patrice LOPEZ et Laurent ROMARY : HUMB : Automatic Key Term Extraction from Scientific Articles in GROBID. *In Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 248–251, Uppsala, Sweden, July 2010. Association for Computational Linguistics. 30

François MANIEZ : Identification automatique des adjectifs relationnels : approche fondée sur l'utilisation d'un corpus en langue de spécialité. *De la mesure dans les termes : hommage à Philippe Thoiron*, page 134, 2005. 52

François MANIEZ : L'adjectif dénominal en langue de spécialité : étude du domaine de la médecine. *Revue française de linguistique appliquée*, 14(2):117–130, 2009. ISSN 1386-1204. 51, 52

Yutaka MATSUO et Mitsuru ISHIZUKA : Keyword Extraction from a Single Document Using Word Co-occurrence Statistical Information. *International Journal on Artificial Intelligence Tools*, 13(01):157–169, 2004. 23

Olena MEDELYAN, Eibe FRANK et Ian H. WITTEN : Human-competitive tagging using automatic keyphrase extraction. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1318–1327, Singapore, August 2009. Association for Computational Linguistics. 17

Olena MEDELYAN et Ian H WITTEN : Thesaurus Based Automatic Keyphrase Indexing. *In Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 296–297. ACM, 2006. 33, 82

Olena MEDELYAN et Ian H. WITTEN : Domain-Independent Automatic Keyphrase Indexing with Small Training Sets. *Journal of the American Society for Information Science and Technology*, 59(7):1026–1040, may 2008. ISSN 1532-2882. 11, 51

Rada MIHALCEA et Paul TARAU : TextRank : Bringing Order Into Texts. *In* Dekang LIN et Dekai WU, éditeurs : *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics. 24, 25, 98

George A. MILLER : WordNet : a Lexical Database for English. *Communications of the Association for Computational Linguistics*, 38(11):39–41, 1995. 52

Thuy Dung NGUYEN et Min-Yen KAN : Keyphrase Extraction in Scientific Publications. *In Proceedings of the 10th International Conference on Asian Digital Libraries : Looking Back 10 Years and Forging New Frontiers*, pages 317–326, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 3-540-77093-3, 978-3-540-77093-0. 20, 28, 29

Paul OVER : Introduction to DUC-2001 : an Intrinsic Evaluation of Generic News Text Summarization Systems. *In Proceedings of DUC 2001 Document Understanding Conference*, 2001. 47

Patrick PAROUBEK, Pierre ZWEIGENBAUM, Dominic FOREST et Cyril GROUIN : Indexation libre et contrôlée d'articles scientifiques. Présentation et résultats du défi fouille de textes DEFT2012 (Controlled and Free Indexing of Scientific Papers. Presentation and Results of the DEFT2012 Text-Mining Challenge) [in French]. *In Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, Workshop DEFT 2012 : DÉfi Fouille de Textes (DEFT 2012 Workshop : Text Mining Challenge)*, pages 1–13, Grenoble, France, June 2012. ATALA/AFCP. 12, 40, 42

Mari-Sanna PAUKKERI et Timo HONKELA : Likey : Unsupervised Language-Independent Keyphrase Extraction. *In Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval)*, pages 162–165, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. 20

Martin F PORTER : An Algorithm for Suffix Stripping. *Program : Electronic Library and Information Systems*, 14(3):130–137, 1980. 56, 59, 64, 81

Quentin PRADET, Jeanne Baguenier DESORMEAUX, Gaël de CHALENDAR et Laurence DANLOS : WoNeF : amélioration, extension et évaluation d'une traduction française automatique de WordNet. *In Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, pages 76–89, Les Sables d'Olonne, France, 2013. 52

Fremont RIDER : *The great dilemma of world organization*. The Hadham press, 1946. 11

Stephen E. ROBERTSON, Walker STEVE et Hancock-Beaulieu MICHELINE : Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive Track. *In Proceedings of the Text REtrieval Conference (TREC)*, pages 199–210, 1998. 21

Gerard SALTON, Andrew WONG et Chungshu YANG : A Vector Space Model for Automatic Indexing. *Communication ACM*, 18(11):613–620, November 1975. ISSN 0001-0782. 11, 20, 98

Kamal SARKAR, Mita NASIPURI et Suranjan GHOSE : A New Approach to Keyphrase Extraction Using Neural Networks. *International Journal of Computer Science Issues Publicity Board 2010*, 2010. 32

Karen Spärck JONES : A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 28(1):11–21, 1972. 20

Li SUJIAN, Wang HOUFENG, Yu SHIWEN et Xin CHENGSHENG : News-Oriented Keyword Indexing with Maximum Entropy Principle. *In Proceedings of the 17th Pacific Asia Conference*. COLIPS Publications, 2003. 29

Takashi TOMOKIYO et Matthew HURST : A Language Model Approach to Keyphrase Extraction. *In Proceedings of the ACL 2003 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment - Volume 18*, pages 33–40, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. 9, 21, 22

Kristina TOUTANOVA, Dan KLEIN, Christopher D. MANNING et Yoram SINGER : Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technology - Volume 1 (NAACL)*, pages 173–180, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. 47

Peter D. TURNEY : Learning Algorithms for Keyphrase Extraction. *Information Retrieval*, 2(4):303–336, may 2000. ISSN 1386-4564. 15, 30, 32

Peter D. TURNEY : Coherent Keyphrase Extraction via Web Mining. *In Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pages 434–439, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc. 28, 31

Ellen M. VOORHEES : The Philosophy of Information Retrieval Evaluation. *In Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, pages 355–370, London, UK, 2002. Springer-Verlag. ISBN 3-540-44042-9. 34

Xiaojun WAN : Using Bilingual Information for Cross-Language Document Summarization. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1*, pages 1546–1555. Association for Computational Linguistics, 2011. 27

Xiaojun WAN et Jianguo XIAO : Single Document Keyphrase Extraction Using Neighborhood Knowledge. *In Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, pages 855–860. AAAI Press, 2008. ISBN 978-1-57735-368-3. 18, 20, 25, 28, 47, 50, 61, 66, 98

Rui WANG, Wei LIU et Chris MCDONALD : How Preprocessing Affects Unsupervised Keyphrase Extraction. *In* Alexander GELBUKH, éditeur : *Computational Linguistics and Intelligent Text Processing*, volume 8403 de *Lecture Notes in Computer Science*, pages 163–176. Springer Berlin Heidelberg, 2014. ISBN 978-3-642-54905-2. 13, 16, 35

Ian H. WITTEN, Gordon W. PAYNTER, Eibe FRANK, Carl GUTWIN et Craig G. Nevill MANNING : KEA : Practical Automatic Keyphrase Extraction. *In Proceedings of the 4th ACM Conference on Digital Libraries*, pages 254–255, New York, NY, USA, 1999. ACM. ISBN 1-58113-145-3. 17, 28, 29, 51, 98

Rui YAN, Mirella LAPATA et Xiaoming LI : Tweet Recommendation with Graph Co-Ranking. *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 516–525, Jeju Island, Korea, July 2012. Association for Computational Linguistics. 27

Torsten ZESCH et Iryna GUREVYCH : Approximate Matching for Evaluating Keyphrase Extraction. *In Proceedings of the International Conference RANLP-2009*, pages 484–489, Borovets, Bulgaria, September 2009. Association for Computational Linguistics. 70

Chengzhi ZHANG : Automatic keyword extraction from documents using conditional random fields. *Journal of Computational Information Systems*, 4(3):1169–1180, 2008. 16, 29

Fan ZHANG, Lian'en HUANG et Bo PENG : WordTopic-MultiRank : A New Method for Automatic Keyphrase Extraction. *In Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 10–18, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing. 26, 27, 70, 100

Kuo ZHANG, Hui XU, Jie TANG et Juanzi LI : Keyword Extraction Using Support Vector Machine. *In Proceedings of the 7th International Conference on Advances in Web-Age Information Management*, pages 85–96, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-35225-2, 978-3-540-35225-9. 31

George K. ZIPF : The Psycho-Biology of Language. 1935. 32

l'université nantes angers le mans
COMMUNAUTÉ D'UNIVERSITÉS ET ÉTABLISSEMENTS

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Adrien BOUGOUIN

Indexation automatique par termes-clés en domaines de spécialité

Automatic Domain-Specific Keyphrase Annotation

Résumé

Les termes-clés, ou mots-clés, sont des mots ou des expressions qui représentent le contenu d'un document. Ils en donnent une représentation synthétique et permettent de l'indexer pour la recherche d'information. Cette thèse s'intéresse à l'indexation automatique par termes-clés de documents en domaines de spécialité. La tâche est difficile à réaliser et les méthodes actuelles peinent encore à atteindre des résultats satisfaisants. Notre démarche s'organise en deux temps. Dans un premier temps, nous nous intéressons à l'indexation par termes-clés en général. Nous proposons une méthode pour sélectionner des termes-clés candidats dans un document en nous focalisant sur la catégorie des adjectifs qu'ils peuvent contenir, puis proposons une méthode pour les ordonner par importance. Cette dernière, TopicRank, se situe en aval de la sélection des candidats. C'est une méthode à base de graphe qui groupe les termes-clés candidats véhiculant le même sujet, projettent les sujets dans un graphe et extrait un terme-clé par sujet. Nos expériences montrent que TopicRank est significativement meilleur que les précédentes méthodes à base de graphe. Dans un second temps, nous adaptons notre travail à l'indexation par termes-clés en domaines de spécialité. Nous étudions la méthodologie d'indexation manuelle de documentalistes et la simulons à l'aide de TopicCoRank. TopicCoRank ajoute à TopicRank un graphe qui représente le domaine de spécialité du document. Grâce à ce second graphe, TopicCoRank possède la rare capacité à fournir des termes-clés qui n'apparaissent pas dans les documents. Appliqué à quatre domaines de spécialité, TopicCoRank améliore significativement TopicRank.

Mots clés

Indexation automatique, terme-clé, mot-clé, domaine de spécialité, méthode à base de graphe, recherche d'information, traitement automatique des langues.

Abstract

Keyphrases are words or multi-word expressions that represent the content of a document. Keyphrases give a synoptic view of a document and help to index it for information retrieval. This Ph.D thesis focuses on domain-specific automatic keyphrase annotation. Automatic keyphrase annotation is still a difficult task, and current systems do not achieve satisfactory results. Our work is divided in two steps. First, we propose a keyphrase candidate selection method that focuses on the categories of adjectives relevant within keyphrases and propose a method to rank them according to their importance within the document. This method, TopicRank, is a graph-based method that clusters keyphrase candidates into topics, ranks the topics and extracts one keyphrase per important topic. Our experiments show that TopicRank significantly outperforms other graph-based methods for automatic keyphrase annotation. Second, we focus on domain-specific documents and adapt our previous work. We study the best practice of manual keyphrase annotation by professional indexers and mimic it with a new method, TopicCoRank. TopicCoRank adds a new graph representing the specific domain to the topic graph of TopicRank. Leveraging this second graph, TopicCoRank possesses the rare ability to provide keyphrases that do not occur within documents. Applied on four corpora of four specific domains, TopicCoRank significantly outperforms TopicRank.

Key Words

Document indexing, keyphrase, keyword, specific domain, graph-based method, micro summary, information retrieval, natural language processing.