Accès à l'information dans les grandes collections textuelles en langue arabe

Abdelkader El Mahdaouy

Communauté UNIVERSITÉ Grenoble Alpes

Université Sidi Mohamed Ben Abdellah
Faculté des Sciences Dhar Mahraz

جامعة سيدي محمد بن عبدالله
كلية العلوم ظهر المهراز

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

préparée dans le cadre d'une cotutelle entre la *Communauté Université Grenoble Alpes* et l'Université Sidi Mohamed Ben Abdellah Fès

Spécialité : **Informatique**

Arrêté ministériel : 25 mai 2016

Présentée par

Abdelkader El Mahdaouy

Thèse dirigée par **Eric Gaussier** et **Saïd Ouatik El ALaoui**

préparée au sein des **Laboratoires d'Informatique de Grenoble** et **Informatique et Modélisation de la Faculté des Sciences Dhar El Mahraz Fès**

dans **l'École Doctorale Mathematiques, Sciences et Technologies de l'Information, Informatique** et **le Centre d'Études Doctorales « Sciences et Technologies »**

Accès à l'information dans les grandes collections textuelles en langue arabe

Thèse soutenue publiquement le **16/12/2017**,
devant le jury composé de :

Mohammed Ouçamah Cherkaoui Malki
Professeur, Faculté des Sciences Dhar El Mahraz -Fès, Président
Mohand Boughanem
Professeur, Université Toulouse 3 - CNRS-IRIT, Rapporteur
Pierre Zweigenbaum
Directeur de recherche, Université Paris-Saclay LIMSI-CNRS, Rapporteur
Brahim Ouhbi
Professeur, Ecole Nationale Supérieure d'Art et Métiers- Meknès, Examinateur
Eric Gaussier
Professeur, Université Grenoble Alpes -Grenoble, Directeur de thèse
Saïd Ouatik El Alaoui
Professeur, Faculté des Sciences Dhar El Mahraz -Fès, Directeur de thèse

THÈSE en Cotutelle

présentée pour obtenir le grade de **DOCTEUR**

UNIVERSITÉ SIDI MOHAMMED BEN ABDALLAH

FACULTÉ DES SCIENCES DHAR EL MAHRAZ -FÈS

Centre d'études doctorales "Sciences et Technologies"

Formation Doctorale : Sciences et Technologies de l'Information et de la Communication

Spécialité : Informatique

Laboratoire : Laboratoire Informatique et Modélisation (LIM)

&

UNIVERSITÉ GRENOBLE ALPES

Ecole Doctorale : Mathématiques, des Sciences et Technologies de l'Information et de l'Informatique MSTII (ED 217)

Spécialité : Informatique

Laboratoire : Laboratoire d'Informatique de Grenoble (LIG)

Par

Abdelkader El Mahdaouy

Accès à l'information dans les grandes collections textuelles en langue arabe

Soutenue le 16/12/2017 devant le jury composé de :

Pr. Mohammed Ouçamah Cherkaoui Malki	*Faculté des Sciences Dhar El Mahraz -Fès*	*Rapporteur, Président*
Pr. Mohand Boughanem	*Université Toulouse 3 - CNRS-IRIT -Toulouse*	*Rapporteur*
Pr. Pierre Zweigenbaum	*Université Paris-Saclay LIMSI-CNRS -Paris*	*Rapporteur*
Pr. Brahim Ouhbi	*Ecole Nationale Supérieure d'Art et Métiers- Meknès*	*Examinateur*
Pr. Saïd Ouatik El Alaoui	*Faculté des Sciences Dhar El Mahraz -Fès*	*Directeur de thèse*
Pr. Eric Gaussier	*Université Grenoble Alpes -Grenoble*	*Directeur de thèse*

Année Universitaire : 2017-2018

Remerciements

En rédigeant cette recherche, je suis convaincu que je n'aurais jamais pu la réaliser sans l'aide de nombreuses personnes que je tiens à remercier :

Tout d'abord mes remerciements s'adressent à mes directeurs de thèse, Monsieur **Saïd El Alaoui Ouatik** Professeur à l'Université Sidi Mohamed Ben Abdellah de Fès, et Monsieur **Eric Gaussier** Professeur à l'Université Grenoble Alpes et Directeur du Laboratoire d'Informatique Grenoble (LIG) : Vous avez bien voulu me confier ce travail riche d'intérêt et me guider à chaque étape de sa réalisation, vous m'avez toujours réservé le meilleur accueil malgré vos obligations professionnelles. Vos encouragements inlassables, votre amabilité, et votre gentillesse méritent toute admiration. Je saisis cette occasion pour vous exprimer ma profonde gratitude tout en vous témoignant mon respect.

Mes remerciements vont également aux membres de mon jury, pour avoir accepté avec une grande amabilité de siéger parmi ce jury. Je suis très honoré et très reconnaissant à Monsieur **Mohammed Ouçamah Cherkaoui Malki** Professeur à l'Université Sidi Mohamed Ben Abdellah de Fès et président du jury ; à Monsieur **Brahim Ouhbi** Professeur à l'Ecole Nationale Supérieure d'Art et Métiers de Meknès ; à Monsieur **Pierre Zweigenbaum** Professeur à l'Université Paris-Saclay, pour l'immense intérêt qu'il a porté à mon travail et pour ses remarques judicieuses et ses conseils précieux. Et particulièrement, je suis redevable à Monsieur **Mohand Boughanem** Professeur à l'Université Toulouse 3, pour l'intérêt qui il a porté à mon rapport en acceptant de l'examiner en détail, en ajoutant des commentaires scientifiques et en fournissant des suggestions ce qui a permis de l'améliorer et de l'enrichir.

Un profond respect à tout le cadre administratif et professoral de la Faculté des Sciences Dhar El Mehraz, et un remerciement particulier à tous mes enseignants.

Un grand merci à Madame **Zilora Zouaoui** Gestionnaire de l'école doctorale MSTII à l'Université de Grenoble Alpes pour son aide pércieuse et son dévouement constant.

Je tiens aussi à remercier tous ceux qui m'ont aidé à rendre ce travail possible, que ce soit par des idées ou par des encouragements.

Enfin je tiens à exprimer toute ma gratitude et toute ma reconnaissance à ceux qui ont plus particulièrement assuré le soutien affectif de ce travail doctoral : mes parents ainsi que mes frères.

Résumé : Face à la quantité d'information textuelle disponible sur le web en langue arabe, le développement des Systèmes de Recherche d'Information (SRI) efficaces est devenu incontournable pour retrouver l'information pertinente. La plupart des SRIs actuels de la langue arabe reposent sur la représentation par sac de mots et l'indexation des documents et des requêtes est effectuée souvent par des mots bruts ou des racines. Ce qui conduit à plusieurs problèmes tels que l'ambigüité et la disparité des termes, etc.

Dans ce travail de thèse, nous nous sommes intéressés à apporter des solutions aux problèmes d'ambigüité et de disparité des termes pour l'amélioration de la représentation des documents et le processus de l'appariement des documents et des requêtes. Nous apportons quatre contributions au niveau de processus de représentation, d'indexation et de recherche d'information en langue arabe. La première contribution consiste à représenter les documents à la fois par des termes simples et des termes complexes. Cela est justifié par le fait que les termes simples seuls et isolés de leur contexte sont ambigus et moins précis pour représenter le contenu des documents. Ainsi, nous avons proposé une méthode hybride pour l'extraction de termes complexes en langue arabe, en combinant des propriétés linguistiques et des modèles statistiques. Le filtre linguistique repose à la fois sur l'étiquetage morphosyntaxique et la prise en compte des variations pour sélectionner les termes candidats. Pour sectionner les termes candidats pertinents, nous avons introduit une mesure d'association permettant de combiner l'information contextuelle avec les degrés de spécificité et d'unité. La deuxième contribution consiste à explorer et évaluer les systèmes de recherche d'informations permettant de tenir compte de l'ensemble des éléments d'indexation (termes simples et complexes). Par conséquent, nous étudions plusieurs extensions des modèles existants de RI pour l'intégration des termes complexes. En outre, nous explorons une panoplie de modèles de proximité. Pour la prise en compte des dépendances de termes dans les modèles de RI, nous introduisons une condition caractérisant de tels modèle et leur validation théorique. La troisième contribution permet de pallier le problème de disparité des termes en proposant une méthode pour intégrer la similarité entre les termes dans les modèles de RI en s'appuyant sur les représentations distribuées des mots (RDMs). L'idée sous-jacente consiste à permettre aux termes similaires à ceux de la requête de contribuer aux scores des documents. Les extensions des modèles de RI proposées dans le cadre de cette méthode sont validées en utilisant les contraintes heuristiques d'appariement sémantique. La dernière contribution concerne l'amélioration des modèles de rétro-pertinence (Pseudo Relevance Feedback PRF). Étant basée également sur les RDM, notre méthode permet d'intégrer la similarité entre les termes d'expansions et ceux de la requête dans les modèles standards PRF.

La validation expérimentale de l'ensemble des contributions apportées dans la cadre de cette thèse est effectuée en utilisant la collection standard TREC 2002/2001 de la langue arabe.

Mots clés : Recherche d'Information ; Traitement Automatique de la Langue Arabe ; Dépendance de Termes ; Termes Complexes ; Proximité de Termes ; Disparité des

mots ; Représentations Distribuées des Mots ; Modèle probabilistes de RI ; Racinisation ; Racinisation Légère.

Abstract : Given the amount of Arabic textual information available on the web, developing effective Information Retrieval Systems (IRS) has become essential to retrieve relevant information. Most of the current Arabic SRIs are based on the bag-of-words representation, where documents are indexed using surface words, roots or stems. Two main drawbacks of the latter representation are the ambiguity of Single Word Terms (SWTs) and term mismatch.

The aim of this work is to deal with SWTs ambiguity and term mismatch. Accordingly, we propose four contributions to improve Arabic content representation, indexing, and retrieval. The first contribution consists of representing Arabic documents using Multi-Word Terms (MWTs). The latter is motivated by the fact that MWTs are more precise representational units and less ambiguous than isolated SWTs. Hence, we propose a hybrid method to extract Arabic MWTs, which combines linguistic and statistical filtering of MWT candidates. The linguistic filter uses POS tagging to identify MWTs candidates that fit a set of syntactic patterns and handles the problem of MWTs variation. Then, the statistical filter rank MWT candidate using our proposed association measure that combines contextual information and both termhood and unithood measures. In the second contribution, we explore and evaluate several IR models for ranking documents using both SWTs and MWTs. Additionally, we investigate a wide range of proximity-based IR models for Arabic IR. Then, we introduce a formal condition that IR models should satisfy to deal adequately with term dependencies. The third contribution consists of a method based on Distributed Representation of Word vectors, namely Word Embedding (WE), for Arabic IR. It relies on incorporating WE semantic similarities into existing probabilistic IR models in order to deal with term mismatch. The aim is to allow distinct, but semantically similar terms to contribute to documents scores. The last contribution is a method to incorporate WE similarity into Pseud-Relevance Feedback PRF for Arabic Information Retrieval. The main idea is to select expansion terms using their distribution in the set of top pseudo-relevant documents along with their similarity to the original query terms.

The experimental validation of all the proposed contributions is performed using standard Arabic TREC 2002/2001 collection.

Keywords : Information Retrieval ; Arabic Natural Language Processing ; Term Dependencies, Multi-Word Terms ; Term Proximity ; Distributed Representation of Word vectors, Word Embedding ; Probablistic IR models ; Stemming.

Table des matières

Liste des tableaux

Table des figures

Notations

Notation	Description
x_w^q	Nombre d'occurrences du mot w dans la requête q
x_w^C	Nombre d'occurrences du mot w dans la collection C
x_w^d	Nombre d'occurrences de w dans le document d
t_w^d	Version normalisée de x_w^d
x_p^q	Nombre d'occurrences de de termes complexes p dans la requête q
x_p^C	Nombre d'occurrences de de termes complexes p dans la collection C
x_p^d	Nombre d'occurrences de de termes complexes p dans le document d
t_p^d	Version normalisée de x_p^d
l_d	Longueur de document d
l_{avg}	Longueur moyenne des documents de la collection
N	Nombre de documents de la collection
N_w	Nombre de documents contenant w
$\|C\|$	Nombre de termes dans la collection C
$RSV(q,d)$	Score de d par rapport à q
	Notations des modèles PRFs
F	Ensemble de documents d'expansion ou de feedback
k	Nombre de documents d'expansion
n	Nombre de termes d'expansion
$F(w)$	Poids ou distribution de w dans l'ensemble F
$TF(w)$	Nombre d'occurrences de w dans F

Introduction générale

Contexte

Le développement du web et des supports de stockage ont permis d'archiver de vastes quantités d'informations ce qui a conduit rapidement à une explosion d'information. Cette quantité d'information serait non exploitable si l'information ne peut pas être analysée et retrouvée afin que chaque utilisateur puisse trouver l'information pertinente correspondant à ses besoins. Ces constats ont donné une naissance naturelle au domaine de la Recherche d'Information (*Information Retrieval*). En effet, la RI s'intéresse au développement des modèles, des techniques et des outils permettant l'accès à l'information pertinente. L'information recherchée peut se trouver dans des documents, et pour y retrouver, l'utilisateur a besoin de passer par un intermédiaire communément appelé Système de Recherche d'Information (SRI). Celui-ci permet de mesurer la correspondance entre un ensemble de documents et le besoin en information de l'utilisateur, souvent, exprimé par une requête en langage naturel. Les documents retrouvés sont classés, en vue de les retourner à l'utilisateur, par leur degré de correspondance par rapport à la requête.

Le processus de RI se compose de deux phases principales autour desquelles se situent la plupart des travaux de recherche de la communauté de RI :

- **Indexation :** la phase de représentation du contenu des documents. Le but principal de cette phase est de construire une meilleure représentation du contenu informatif véhiculé par les documents, et de faciliter leur traitement.
- **Interrogation :** cette phase concerne d'une part l'interprétation et la représentation de la requête et, d'autre part, l'utilisation d'un modèle de RI pour la mise en correspondance ou bien l'appariement des représentations construites pour les documents et la requête de l'utilisateur.

La construction de la représentation du contenu des documents et des requêtes est un processus essentiel en RI. En revanche, la construction de telle représentation à partir des textes, non structurés et exprimés en langage naturel, est une tâche très complexe, et elle nécessite une bonne compréhension de leurs contenus. Cependant, le traitement automatique du langage naturel pose des défis majeurs liés principalement à la nature des langues souvent, implicites et ambigus [Lefévre 2000]. C'est pourquoi, le couplage des techniques de Traitement Automatique de la Langue (TAL) et de RI devient particulièrement intéressant [Moreau & Sébillot 2005].

Motivations et problématiques

La part grandissante des collections textuelles en langue arabe, sur internet par exemple, rend nécessaire le développement d'outils permettant d'accéder au mieux à ces informations.

En effet, le nombre d'internautes arabes a augmenté entre 2000 et 2017 de plus de 7000% [1]. Face à la masse documentaire en langue arabe qui ne cesse d'augmenter, l'utilisateur a besoin des SRI efficaces lui permettant d'accéder aux documents pertinents et répondant à ses besoins.

Dans cette langue, le traitement morphologique devient particulièrement important pour l'accès à l'information, parce qu'un SRI doit déterminer une forme appropriée d'index à partir des mots contenus dans les documents du corpus (phase d'indexation). Pour cela, la plupart des travaux de recherche ont focalisé sur le traitement de la morphologie de l'arabe et représentent les documents et les requêtes par un ensemble de mots-clés (représentation par sac-de-mots) sous l'hypothèse d'indépendance de termes[Darwish & Magdy 2014, Ben Guirat *et al.* 2016]. Cependant, les termes simples isolés sont à l'origine des problèmes d'ambigüité et de disparité des termes, en particulier en langue arabe due principalement à l'absence de voyellation. Ce qui augmente considérablement le nombre de documents non-pertinents retrouvés.

Pour remédier aux limitations de la représentation par sac-de-mots, plusieurs travaux de recherche ont été proposés pour aller au-delà de cette représentation. Parmi ces travaux de recherche, les modèles basés sur l'intégration des dépendances de termes. L'intégration de ces dépendances en RI peut être effectuée en utilisant deux approches. La première approche consiste à extraire des combinaisons des unités lexicales (phrases, collocations, termes complexes, etc.) permettant de mieux représenter le contenu sémantique véhiculé par les documents et les requêtes[Croft *et al.* 1991, Nie & Dufort 2002, Hammache *et al.* 2014]. Ces combinaisons des unités lexicales sont moins ambigües et moins polysémiques que les termes simples isolés. L'idée sous-jacente consiste à traiter le texte en considérant les rapports syntagmatiques qu'entretiennent leurs unités lexicales. Cependant, la deuxième approche repose sur l'utilisation de la proximité entre les termes dans les documents [Metzler & Croft 2005, Peng *et al.* 2007, He *et al.* 2011, Sordoni *et al.* 2013]. L'idée sous-jacente est que plus les occurrences des termes de la requête sont apparues proches (à proximité) dans un document, plus ce document est considéré pertinent pour la requête.

L'intégration des dépendances de termes, que ce soit par l'extraction et l'indexation des termes complexes ou l'utilisation des modèles de proximité, peut améliorer la performance de RI. C'est dans ce cadre que s'inscrivent nos deux premières contributions.

L'intégration des dépendances de termes dans les modèles de RI permet de limiter l'ambigüité des termes simples et de construire une représentation plus précise du contenu des documents et des requêtes. Cependant, il ne permet pas de pallier le problème de disparité des termes (*term mismatch*) lors de la phase d'interrogation (appariement des documents et des requêtes). Ce problème est lié au fait qu'un document pertinent ne partage pas forcément les mêmes termes avec la requête. En effet, les auteurs des documents et les utilisateurs peuvent se référer aux mêmes concepts en utilisant des termes lexicalement différents. Souvent, les utilisateurs n'ont pas de connaissances sur les paramètres fournis par le SRI et sur la collection de documents sur laquelle ils effectuent leur recherche, comme

1. http ://www.internetworldstats.com/stats7.htm

le Web par exemple. De plus, il n'est pas évident pour les utilisateurs de bien préciser leur besoin en information à travers des requêtes courtes. Par conséquent, une part importante de documents pertinents n'est pas retrouvée par le SRI. Compte tenu de la morphologie riche et complexe de l'arabe, le problème de disparité des termes devient particulièrement crucial. Les techniques de racinisation légère produisent des *stems* différents pour les variantes morphologiques du même terme (pluriel irrégulier, agglutination des conjonctions, variantes orthographiques, etc.).

Pour pallier le problème de disparité des termes, plusieurs travaux de recherche ont été réalisés, à savoir la prise en compte des similarités entre les termes et la reformulation de la requête [Li & Xu 2014]. En effet, de considérables efforts ont été déployés pour la prise en compte de la similarité entre les termes dans le processus d'appariement des documents et des requêtes à travers l'extension des modèles de base de RI [Fang & Zhai 2006, Karimzadehgan & Zhai 2010, Ganguly *et al.* 2015, Zuccon *et al.* 2015]. L'idée sous-jacente consiste à permettre aux termes des documents qui sont sémantiquement proches à ceux de la requête de contribuer aux scores de ces documents. De plus, les techniques de reformulation de requêtes consistent à construire une nouvelle requête à partir de la requête initiale pour mieux spécifier le besoin en information de l'utilisateur [Croft *et al.* 2011, Manning *et al.* 2008, Clinchant & Gaussier 2013]. L'amélioration de l'expression du besoin d'utilisateur est effectuée par l'ajout de nouveaux termes (synonymes, termes sémantiquement proches, etc.), la réévaluation des poids des termes de la requête initiale, ou l'extraction des sous-requêtes à partir des requêtes longues.

Remédier au problème de disparité des termes, que ce soit par l'intégration de similarité entre les termes dans les modèles de RI ou la reformulation de requêtes, permet d'améliorer le processus d'appariement des documents et des requêtes. C'est dans ce cadre que se situent nos deux dernières contributions.

Contributions

Cette thèse s'inscrit dans le domaine de l'accès à l'information textuelle en langue arabe et, plus précisément, la recherche d'information. Deux problématiques majeures sont soulevées et étudiées : l'ambigüité des termes simples et leur disparité. Dans le premier volet, en plus de l'identification de la meilleure technique de racinisation/racinisation légère des termes, notre défi était l'amélioration des techniques d'extraction des termes complexes. Ces deux tâches constituent une étape essentielle dans la représentation et l'indexation des documents. De plus, nous étudions les apports des dépendances explicites basées sur l'extraction et l'indexation des termes complexes et des dépendances implicites basées sur la proximité des termes pour aller au-delà de la représentation par sac-de-mots. Dans le deuxième volet, nous abordons le problème de disparité des termes. Dans ce sens, nous nous intéressons particulièrement à l'exploitation des Représentations Distribuées des Mots (RDMs) pour la prise en compte de la similarité entre les termes dans les modèles de RI et les modèles de rétro-pertinence (PRF). Les principales contributions que nous

avons apportées dans cette thèse sont :

1. Nous avons proposé une méthode hybride combinant le filtrage linguistique et statistique pour l'extraction des termes complexes. Le filtre linguistique procède par l'identification des termes candidats en utilisant des patrons syntaxiques à partir d'un corpus étiqueté. Ce filtre traite également les variantes graphiques, flexionnelles, morpho-syntaxiques et syntaxiques des termes candidats identifiés. Pour le filtrage statique, nous avons introduit une mesure d'association qui consiste à combiner le degré de spécificité [Frantzi *et al.* 2000], le degré d'unité [Dunning 1993] et l'information contextuelle [Frantzi *et al.* 2000].
2. Nous avons étudié le problème d'indexation et de recherche d'information à base des termes complexes extraits en utilisant notre méthode hybride. En outre, nous avons exploré une panoplie de modèles de proximité pour la RI en langue arabe. Les deux approches d'intégration de dépendances des termes sont évaluées en utilisant différentes approches de racinisation/racinisation légère. De plus, nous avons introduit une condition pour caractériser les modèles de RI pour la prise en compte des dépendances de termes. Par ailleurs, nous avons comparé l'approche basée sur les termes complexes et l'approche basée proximité pour l'intégration de dépendances dans la RI en langue arabe.
3. Pour la prise en compte de l'aspect sémantique dans le processus d'appariement, nous avons introduit une méthode pour intégrer la similarité entre les termes dans les modèles de RI en utilisant les représentations distribuées des mots [Mikolov *et al.* 2013, Pennington *et al.* 2014]. Nous avons également validé les extensions des modèles de RI proposés dans le cadre de cette méthode en utilisant les contraintes heuristiques d'appariement sémantique [Fang & Zhai 2006]. Ces extensions sont comparées avec l'approche d'indexation sémantique basée sur l'ontologie ArabicWordNet [Abderrahim *et al.* 2016] et trois extensions basées RDMs du modèle de langue de RI[Vulić & Moens 2015, Ganguly *et al.* 2015, Zuccon *et al.* 2015].
4. Dans le cadre de la reformulation de requête, nous avons proposé une méthode pour intégrer la similarité entre les termes dans les modèles PRF en utilisant les représentations distribuées des mots [Mikolov *et al.* 2013, Pennington *et al.* 2014]. L'idée de base consiste à intégrer la similarité entre les termes candidats d'expansion et la requête initiale pour améliorer le processus de sélection des termes d'expansion. La pondération des termes d'expansion s'effectue en combinant leurs poids dans l'ensemble des documents d'expansion et leurs similarités à la requête initiale.

Organisation du mémoire

Ce mémoire est structuré comme suit :

Le **Chapitre 1** présente un état de l'art sur la recherche d'information et le traitement automatique de la langue arabe. Dans un premier temps, ce chapitre présente les concepts

de base de la RI, les techniques d'indexation et les modèles classiques de RI. Il décrit aussi les approches et les modèles de RI pour aller au-delà de la représentation par sac-de-mots. Ensuite, une présentation des techniques de reformulation de la requête et les méthodes d'évaluation des SRI est effectuée. Enfin, ce chapitre présente la langue arabe, ses particularités et les différentes techniques de TAL liées à cette langue .

Le **Chapitre 2** présente notre première contribution pour l'extraction des termes complexes de la langue arabe. Dans les premières sections, nous passons en revue des méthodes d'extraction des termes complexes. Ce chapitre se termine par une description de la méthode d'évaluation et la discussion des résultats obtenus.

Le **Chapitre 3** décrit notre deuxième contribution pour l'intégration de dépendance de termes dans la RI en langue arabe. Dans un premier temps, ce chapitre présente nos motivations. Puis, il décrit la méthode d'indexation des termes complexes. Ensuite, il présente les extensions des modèles de RI pour l'intégration des termes complexes et la proximité entre les termes. Enfin, ce chapitre se termine par la description des évaluations effectuées et la discussion des résultats obtenus.

Dans le **Chapitre 4**, nous présentons nos deux dernières contributions qui se rapportent au problème de disparité des termes. Nous commençons par nos motivations pour l'exploitation des RDMs. Puis, nous passons en revue des travaux reliés et la description des modèles utilisés pour l'apprentissage des RDMs. Ensuite, nous présentons notre méthode basée RDMs pour l'extension des modèles de RI et leur validation théorique. En outre, nous introduisons notre méthode PRF basée RDMs pour l'extension des modèles PRF. Enfin, ce chapitre se termine par la description des évaluations effectuées et la discussion des résultats obtenus.

Conclusion générale récapitule les contributions apportées dans le cadre de ce mémoire et ouvre la porte sur leurs perspectives.

Chapitre 1

Recherche d'Information et Traitement Automatique de la Langue Arabe

Sommaire

1.1 Introduction

Historiquement, la naissance du domaine de Recherche d'Information (RI) remonte à la fin des années 1940, après l'invention de l'ordinateur. Le concept "Recherche d'Information (*Information Retrieval*) fut introduit par Calvin N. Mooers, en 1948, afin de désigner les techniques et les algorithmes permettant l'accès à l'information pertinente. Ce domaine de recherche concerne la représentation, le stockage, l'organisation et l'accès aux sources d'information [Salton & McGill 1986]. Contrairement à la recherche de données structurées dans les bases de données, la recherche d'information traite l'information non structurée [Salton 1989]. Les premiers systèmes de recherche d'information (SRI), établis pour automatiser la RI, ont été développés pour mieux accéder aux archives des bibliothèques et des publications scientifiques. Puis, ils sont rapidement déployés à d'autres formes de

contenus, mais en restant réservés à des professionnels et à des secteurs gouvernementaux [Manning *et al.* 2008]. L'avènement de l'Internet et plus particulièrement du Web ont conduit à révéler la RI aux internautes, notamment par le biais des moteurs de recherche. La prolifération exponentielle de l'information numérique disponible a rendu indispensable des moyens et des outils de recherche performants et automatiques, permettant de mieux accéder et de retrouver l'information pertinente.

De nos jours, la RI n'est néanmoins plus réduite à cette recherche documentaire et se rapproche de l'accès à l'information en général. La RI est un domaine de recherche interdisciplinaire qui tire profit essentiellement de la science de l'information, de l'informatique, et du traitement automatique de la langue (TAL). En effet, le vocable recherche d'information regroupe divers tâches, notamment la recherche documentaire, le filtrage et l'extraction d'information, les systèmes de questions-réponses, la recherche d'information sur le Web, etc. Le but principal de la RI est de retrouver un ensemble documents pertinents en réponse aux besoins informationnels de l'utilisateur, exprimés par des requêtes en langage naturel. Contrairement aux langages formels qui sont conçus et optimisés dans l'optique de manipulations algorithmiques, le langage naturel pose des difficultés majeures à son traitement automatique, liées principalement à sa nature implicite, redondante et ambigüe [Lefévre 2000]. Une solution assez naturelle souvent envisagée pour faire face à ces difficultés est d'intégrer au sein des SRI des techniques de TAL pour une meilleure compréhension du contenu des documents et des requêtes et, par conséquent, améliorer le processus d'indexation des documents et construire une représentation plus riche de leur contenu [Rau & Jacobs 1989, Lewis *et al.* 1989, Gaussier & Stéfanini 2003, Moreau & Sébillot 2005]. Ce couplage de RI et TAL est réalisé afin d'améliorer la pertinence du processus d'appariement de contenu des documents et des besoins informationnels des utilisateurs [Gaussier *et al.* 2000, Tannier 2006].

Le reste de ce chapitre est organisé comme suite : dans le premier volet (Section 1.2), nous passons en revue de différentes techniques et modèles de RI. En particulier, nous présentons les concepts de base de la RI dans la Section 1.2.1. Nous introduisons les différentes techniques d'indexation et les modèles standards de RI. Puis, nous introduisons les techniques et les modèles utilisés pour aller au-delà de l'hypothèse d'indépendance de termes dans la RI. Nous présentons aussi les techniques utilisées pour l'expansion de requêtes. Nous terminons le premier volet de ce chapitre par les méthodes d'évaluation des SRI. Le deuxième volet de ce chapitre a pour but de présenter la langue arabe et son traitement automatique (Section 1.3). Dans un premier temps, nous introduisons la langue arabe. Ensuite, nous passons en revue de leurs particularités, notamment de niveaux morphologique et syntaxique. Enfin, nous présentons les différentes techniques de TAL arabe dans la Section 1.3.3, en particulier les techniques de racinisation et ceux d'étiquetage morphosyntaxique.

1.2 Recherche d'Information

D'une manière générale, la recherche d'information (RI) concerne la représentation, le stockage, l'organisation et l'accès à l'information [Salton & McGill 1986]. Face à l'explosion de l'information numérique et les avancées de la société de l'information, les utilisateurs ont besoin des systèmes et des outils efficaces pour accéder à l'information pertinente. L'information recherchée peut se trouver dans des documents, et pour y récupérer, l'utilisateur a besoin de passer par un intermédiaire communément appelé système de recherche d'information (SRI) qui mesure la correspondance entre un ensemble de documents avec la requête.

1.2.1 Concepts de base de la recherche d'information

1.2.1.1 Le document et la collection de documents

Les documents jouent un rôle central en RI car ils sont les sources ou bien les porteurs de l'information. Un document peut être défini comme étant le support physique de l'information, qui peut avoir différentes formes (texte, image, vidéo, etc.). Dans cette thèse nous nous intéressons aux documents textuels en arabe. Ces documents sont indexés par le SRI en vue de les retrouver pour répondre à des besoins informationnels des utilisateurs. L'ensemble des documents interrogés ou manipulés par le SRI lors de l'exécution d'une requête est communément appelé, collection de documents.

1.2.1.2 La requête

La requête est la spécification du besoin en information de l'utilisateur, exprimée par des mots-clés ou des phrases en langage naturel. L'expression du besoin informationnel de l'utilisateur est une étape cruciale en recherche d'information et peut avoir des effets négatifs sur la pertinence des documents retournés. En effet, les requêtes peuvent ne pas exprimer le besoin d'information de l'utilisateur de façons assez précises pour le SRI. Cela est dû, d'une part, au fait que l'utilisateur n'a pas forcément une idée précise à propos de la collection de documents sur laquelle il effectue sa recherche. D'autre part, l'utilisateur ignore les paramètres fournis par le système de recherche pour mieux exprimer sa requête. Pour remédier partiellement à ce problème, une reformulation de la requête est souvent envisagée pour améliorer l'expression du besoin en information de l'utilisateur.

1.2.1.3 Notion de pertinence

D'une manière idéale, un SRI doit retrouver tous les documents pertinents, et en même temps de récupérer aussi peu de documents non pertinents que possible [Rijsbergen 1979]. La pertinence est ainsi une notion fondamentale en RI, car toutes les évaluations s'articulent autour de cette notion [Borlund 2003]. La définition de cette notion est complexe, car elle peut intervenir aux différentes étapes de la RI [Cooper 1971, Saracevic 1975,

Schamber *et al.* 1990] et fait intervenir plusieurs notions [Mizzaro 1997]. Généralement, cette notion est définie par le degré de correspondance entre un document et une requête ou une mesure d'informativité du document à la requête. Cette complexité vient principalement du fait que les utilisateurs des SRI ont des besoins informationnels variés et qu'ils ont des critères très différents pour juger la pertinence d'un document. Différents utilisateurs peuvent ainsi avoir des opinions différentes sur la pertinence de certains documents pour une même requête, voire un même utilisateur peut juger différemment un document en cas d'évolution des connaissances au sujet. Donc, d'un point de vue utilisateur, la pertinence se traduit par les jugements de pertinence de l'utilisateur par rapport aux documents retournés par le SRI en réponse à une requête [Mizzaro 1997]. Cependant, d'un point de vue système, la pertinence se traduit par un score attribué par le SRI qui représente le degré d'appariement entre les représentations de contenu de la requête et du document [Saracevic 1996]. Contrairement à la pertinence utilisateur, la pertinence système est objective, elle permet de mesurer la probabilité de pertinence d'un document de la collection par rapport à la requête.

1.2.1.4 Le processus de la RI

Le but d'un système SRI est de retrouver, parmi une collection de documents préalablement indexés, les documents qui correspondent au besoin d'information de l'utilisateur exprimé sous forme de requête. Pour cela, un SRI est caractérisé par trois composants [Gaussier & Yvon 2012] :

1. un module d'indexation des requêtes
2. un module d'indexation des documents
3. un module d'appariement des documents et des requêtes

Les deux modules d'indexation consistent à analyser les documents et les requêtes afin d'établir une représentation plus riche de leur contenu, dans l'optique d'améliorer la pertinence du processus d'appariement et, par conséquent, la performance de SRI. Le module d'appariement, qui se base sur un formalisme précis défini par un modèle de RI, consiste à mettre en correspondance les documents et les requêtes et à calculer le degré d'appariement de leurs représentations internes. Ce degré d'appariement est appelé aussi le score de pertinence ou encore score de similarité d'un document par rapport à la requête . Les documents qui correspondent le mieux à la requête, ou documents jugés pertinents par le SRI, sont alors retournés à l'utilisateur, dans une liste ordonnée par ordre décroissant de leur score de pertinence. Afin d'améliorer la qualité des résultats de la recherche, le système peut être doté d'un mécanisme de reformulation de la requête.

1.2.2 Indexation

L'indexation est le processus de représentation du contenu des documents. Le but principal de ce processus est de construire une meilleure représentation du contenu informatif véhiculé par les documents et faciliter leur manipulation algorithmique, pour

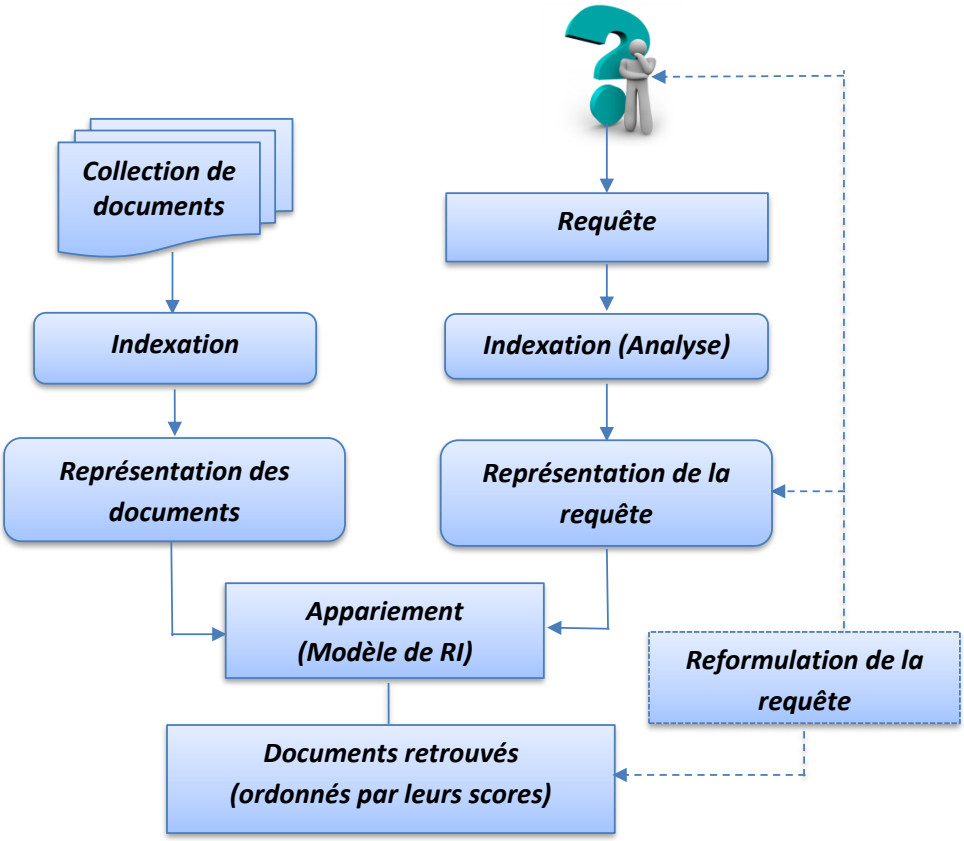


FIGURE 1.1: Processus générale d'un système de recherche d'information

répondre d'une manière efficace et rapide aux requêtes des utilisateurs [Sparck Jones 1974, Salton & McGill 1986, Lewis *et al.* 1989, Manning *et al.* 2008]. Généralement, l'indexation consiste à associer à chaque document de la collection un ensemble de descripteurs (souvent des mots-clés ou encore la représentation de type sac de mots), qui représentent mieux son contenu sémantique. Cette étape est primordiale, car elle spécifie la structure du contenu des documents, l'extraction et la sélection des descripteurs représentatifs. Ces descripteurs peuvent être des mots simples, des groupes de mots (phrases, termes complexes, collocations) ou des concepts, extraits d'une façon manuelle, semi-automatique ou automatique [Salton 1969, Salton 1971a].

Dans ce travail de thèse, nous nous intéressons à l'indexation automatique, où les processus d'analyse de contenu textuel, des documents et des requêtes, et de sélection des descripteurs ou des termes représentatifs sont entièrement automatisés. Dans un tel contexte, les techniques de TAL permettent d'extraire à partir du texte des informations plus riches que des mots simples, issues d'une segmentation de textes. Ces informations de type morphologique, syntaxique ou sémantique peuvent être utilisées pour améliorer la représentation des

contenus des documents [Gaussier & Stéfanini 2003, Moreau & Sébillot 2005]. Le choix des descripteurs représentatifs dépend des niveaux d'analyse de TAL du contenu des documents et des requêtes. Généralement, un SRI peut reposer sur un ou plusieurs niveaux d'analyse de TAL afin d'assurer une indexation pertinente du contenu textuel [Strzalkowski 1999, Yvon 2010]. Pour identifier et caractériser les termes importants des documents, l'indexation peut reposer sur une pondération des descripteurs représentatifs (cas des modèles vectoriels), sous l'hypothèse que les descripteurs importants doivent avoir des poids supérieurs.

Dans les sous-sections suivantes, nous détaillons les niveaux d'analyse de TAL utilisés pour extraire les descripteurs représentatifs du contenu textuel ainsi leur pondération.

1.2.2.1 Analyse lexicale

Cette étape consiste à reconnaitre les unités linguistiques dotées de sens et leurs propriétés morphologiques à partir des chaines de caractères (*tokens*) identifiées par la segmentation en mots (*Tokenisation*) [Strzalkowski 1999, Yvon 2010].

Segmentation de textes

La segmentation des documents textes en mots consiste à reconnaitre et regrouper les chaines de caractères pour former des unités lexicales. Cette étape repose sur une liste de signes qui désignent les séparateurs entre les unités lexicales.

Élimination des mots vides

Les mots vides (articles, propositions, conjonctions, etc.) sont des mots fréquents qui apparaissent dans tous les documents de la collection et ne portent pas de sens. Pour les éliminer, soit on utilise un anti-dictionnaire (*stop word-list*) ou éliminer les mots qui dépassent un certain seuil dans la phase l'indexation.

Normalisation

La normalisation d'un texte consiste à apporter des modifications légères à quelques unités lexicales identifiées lors de la phase segmentation. Pour les langues latines (français et anglais par exemple), cette étape repose principalement sur le traitement des majuscules, la reconnaissance des sigles, des abréviations et des acronymes.

Analyse morphologique

La morphologie est un domaine de la linguistique qui s'intéresse à la structure des mots. D'une manière générale, elle concerne l'étude de différentes combinaisons des morphèmes ou tout simplement les plus petites unités de sens qui constituent les mots [Porter 1980, Harman 1991, Gaussier & Stéfanini 2003]. L'analyse morphologique consiste à reconnaitre les différentes variations des mots. L'exploitation des connaissances morphologiques au

sein des applications de TAL, particulièrement en RI, est de permettre d'apparier les différentes variations d'un même mot apparaissant dans les documents et les requêtes. Pour traiter ces variations lors de l'indexation, deux techniques sont largement utilisées : la lemmatisation et la racinisation. La lemmatisation consiste à transformer les formes fléchies (variantes morphologiques) d'un mot à leur forme canonique, appelé communément lemme [1] (i.e. l'entrée lexicale d'un dictionnaire). Cette technique repose principalement sur l'utilisation des ressources linguistiques et nécessite souvent l'identification de la fonction grammaticale du mot fléchi (contexte). Cependant, la racinisation consiste à transformer les formes fléchies d'un mot à leur racine en supprimant les affixes (préfixes et suffixes). À l'opposé des lemmes qui sont des mots réels de la langue, les racines peuvent ne pas être des mots réels [2]. Plusieurs travaux de recherche ont montré que la racinisation et la lemmatisation améliorent significativement la performance de la RI [Porter 1980, Harman 1991, Gaussier *et al.* 2000, Abu El-Khair 2007].

1.2.2.2 Analyse syntaxique

L'analyse syntaxique est un domaine de la linguistique qui s'intéresse à la structure des phrases, des séquences consécutives qui les forment (appelées communément "syntagmes") et les fonctions grammaticales de leurs éléments (termes simples). L'utilisation de ces connaissances syntaxiques en RI consiste à identifier principalement les phrases, les syntagmes nominaux, les termes complexes, etc. Ils sont utilisés comme des candidats descripteurs aux termes simples [Fagan 1987, Salton & Buckley 1988, Mitra *et al.* 1997]. Ces candidats descripteurs sont moins ambigus que les termes simples qui les composent, exploités pour représenter le contenu des documents et des requêtes et, par conséquent, améliorer la performance de RI (précision des résultats retournés) [Jacquemin *et al.* 1997, Gaussier *et al.* 2000, Haddad 2002]. D'une manière générale, les descripteurs syntaxiques sont identifiés en utilisant des patrons syntaxiques ou les grammaires à base de règles de réécriture. Toutefois cette étape d'identification seule est insuffisante, ces descripteurs sont sujets de plusieurs types de variations (principalement les variantes morpho-syntaxiques et syntaxiques) et nécessite une phase de normalisation [Daille 1994]. Cette normalisation qui est à peu près similaire à celle de l'analyse morphologique consiste à traiter les variantes d'un même syntagme ou terme complexe. De plus, une phase de filtrage statistique est souvent envisagée pour ne considérer que les descripteurs pertinents [Frantzi *et al.* 2000, Jacquemin *et al.* 1997, Daille *et al.* 1994, Kageura & Umino 1996].

Dans le deuxième chapitre, nous allons passer en revue de différentes approches d'extraction des termes complexes qui visent à exploiter des connaissances morphologiques, syntaxiques et statistiques permettant d'enrichir la représentation du contenu textuel pour améliorer la performance des applications de TAL.

1. Le mot "rechercher", verbe à l'infinitif ni accordé ni conjugué est un lemme. Il possède différentes formes fléchies qui correspondent à ses formes conjuguées : "il recherche", "nous recherchons", "vous avez recherché", etc.

2. La racine du mot "rechercher" est "cherch" qui ne correspond pas à un mot réel

1.2.2.3 Analyse sémantique

L'analyse sémantique s'intéresse à l'analyse et l'identification des sens des mots et des phrases. D'une manière générale, cette analyse permet d'identifier les sens/concepts représentés par les mots et les phrases d'un document en s'appuyant sur des ressources sémantiques (graphes conceptuels, réseaux sémantiques) [Baziz 2005, Hliaoutakis *et al.* 2006, Fernández *et al.* 2011, Dragoni *et al.* 2012]. L'analyse sémantique permet construire une représentation plus riche, en gardant les relations sémantiques entre les concepts extraits à partir des textes, que celles obtenues par l'analyse morphologique et syntaxique. En effet, l'indexation sémantique permet de pailler le problème d'ambigüité des termes (lexicale et syntaxique) et leur disparité [Krovetz & Croft 1992, Krovetz 1997, Boubekeur 2008].

- **Ambigüité lexicale :** Ce type d'ambigüité concerne les mots lexicalement identiques qui ont des sens différents selon leur contexte. Ce type d'ambigüité est lié à l'homonymie et la polysémie. L'homonymie est une relation entre plusieurs formes linguistiques ayant la même forme (graphique et/ou phonique) mais leurs sens sont totalement différents. La polysémie concerne les mots ou les formes linguistiques ayant plusieurs sens. Contrairement à l'homonymie qui caractérise des mots radicalement distincts dont la forme est accidentellement la même, la polysémie caractérise la capacité des mots à prendre des sens différents selon leurs contextes.
- **Ambigüité syntaxique :** Ce type d'ambigüité se rapporte à des formes linguistiques (mots ou phrases) ayant plusieurs sens et/ou plusieurs analyses grammaticales.
- **Disparité de termes (*term mismatch*) :** Ce problème est lié au fait que des termes différents peuvent représenter le même sens (synonymie). En RI, les auteurs des documents et les utilisateurs peuvent référer aux mêmes sens en utilisant des termes lexicalement différents. Ceci implique que les documents pertinents peuvent ne pas utiliser les mêmes termes que ceux de la requête et, par conséquent, ne sont pas retournés par le SRI.

1.2.2.4 Pondération des descripteurs d'indexation

La pondération des termes ou des descripteurs d'indexation consiste à leur associer des poids en fonction de leur pouvoir de représentativité ou d'informativité. L'objectif principal de cette pondération consiste à accorder des poids forts aux descripteurs représentatifs. Cette pondération repose sur des mesures statistiques des descripteurs caractérisant leur degré d'apparition dans le document (mesure locale) et/ou dans la collection de documents (mesure globale). D'une manière générale, les mesures d'importance locale (importance dans le document) consistent à calculer la fréquence du descripteur d'index (ou tout simplement du terme) dans le document. L'idée sous-jacente est que plus un terme est fréquent dans un document, plus il est important dans la représentation du contenu de ce document. Les mesures d'importance globale sont calculées en fonction de la fréquence du descripteur dans la collection, de telle sorte que les descripteurs qui apparaissent dans peu de documents de la collection sont plus représentatifs du contenu de ces documents que

ceux qui apparaissent dans tous les documents de la collection.

La pondération $TF \cdot IDF$ consiste à calculer le produit de la fréquence du terme dans le document (*Term Frequency*) et sa fréquence inverse du document (*Inverse Document Frequency*) [Sparck Jones 1972], donné par Équation 1.1 :

$$TF \cdot IDF(w_d) = x_w^d \cdot log(\frac{N}{N_w}) \tag{1.1}$$

où N est le nombre de documents de la collection et N_w est le nombre de documents contenant w.

Cette mesure est une approximation de la représentativité d'un terme dans un document. Ainsi, une valeur de $TF \cdot IDF$ élevée signifie que ce terme est important dans le document et apparait peu dans les autres documents de la collection. La pondération $TF \cdot IDF$ est calculée indépendamment de la longueur des documents, tandis qu'il est important de considérer ce dernier facteur pour pénaliser les documents plus longs. L'idée sous-jacente est que la fréquence d'un terme est plus élevée dans les documents longs. D'autres pondérations ont été proposées pour pallier le problème de la longueur des documents en introduisant des facteurs de normalisation de leur longueur [Robertson & Walker 1997, Singhal *et al.* 1996].

1.2.3 Modèles de recherche d'information

L'appariement des documents et des requêtes s'appuie sur la notion de modèle de RI qui fait la correspondance entre les représentations de leurs contenus (construites dans l'étape d'indexation). Les modèles de RI sont des formalismes théoriques utilisés pour estimer le degré de pertinence ou le score d'un document par rapport à une requête. Ce formalisme est noté par une fonction d'appariement $RSV(d, q)$ (*Retrieval Status Value*), où d et q représentent respectivement un document de la collection et une requête de l'utilisateur. Pour cela, plusieurs familles de modèles ont été proposées : modèles booléens, modèles vectoriels et modèles probabilistes [Baeza-Yates *et al.* 1999, Dominich 2001].

1.2.3.1 Modèle booléen

Le modèle booléen est l'un des premiers modèles de RI, il repose sur la théorie des ensembles et l'algèbre de Boole [Salton 1971a]. Un document est présenté sous forme d'un ensemble de termes non pondérés, ou encore un vecteur booléen dont les composantes sont désignées par des variables logiques qui caractérisent la présence ou l'absence d'un terme. Les requêtes des utilisateurs sont représentées par des expressions logiques contenant des termes reliés par des opérateurs logiques : ET ($\wedge$) , OU ($\vee$) et SAUF ($\neg$). Dans ce modèle, l'appariement a pour objectif de retourner l'ensemble des documents qui impliquent au sens logique l'expression logique de la requête. D'une manière générale, la fonction d'appariement est donnée par Équation 1.2 :

$$RSV(d, q) = \begin{cases} 1 & \text{si} \quad d \Rightarrow q \\ 0 & \text{sinon} \end{cases} \tag{1.2}$$

où d, w_1^q et w_2^q sont respectivement un document et deux termes de la requête, cette équation peut être détaillée comme suite (Équation 1.3) :

$$\begin{cases} RSV(d, w_1^q) = 1 & \text{si} \quad w_1^q \in d \ ; \ 0 \ \text{sinon.} \\ RSV(d, w_1^q \wedge w_2^q) = 1 & \text{si} \quad RSV(d, w_1^q) = 1 \ \textbf{et} \ RSV(d, w_2^q) = 1 \quad ; \ 0 \ \text{sinon.} \\ RSV(d, w_1^q \vee w_2^q) = 1 & \text{si} \quad RSV(d, w_1^q) = 1 \ \textbf{ou} \ RSV(d, w_2^q) = 1 \quad ; \ 0 \ \text{sinon.} \\ RSV(d, \neg w_1^q) = 1 & \text{si} \quad RSV(d, w_1^q) = 0 \ ; 0 \ \text{sinon.} \end{cases} \tag{1.3}$$

Malgré sa facilité d'implantation dans un SRI et son pouvoir expressif, le modèle booléen de base présente plusieurs inconvénients :

- l'appariement est binaire : les documents retournés ne sont pas ordonnés selon leur score pertinence ;
- les termes d'un document sont traités de la même façon, ne sont pas pondérés selon leur importance ou informativité ;
- il n'est pas toujours facile pour tous les utilisateurs à formuler une requête combinant plusieurs opérateurs logiques, en particulier pour les requêtes complexes ;
- il ne supporte pas la reformulation automatique des requêtes par retour de pertinence ;

Des extensions du modèle booléen de base ont été proposées pour remédier aux inconvénients du modèle de base et améliorer la performance de RI [Waller & Kraft 1979, Kraft & Buell 1983, Salton *et al.* 1983].

1.2.3.2 Modèle vectoriel

Le premier modèle vectoriel a été introduit par Salton [1971a] dans le cadre du système SMART. Cette famille de modèles consiste à représenter les documents et les requêtes par des vecteurs dans le même espace vectoriel dont la dimension est le nombre de termes descripteurs extraits (taille du vocabulaire d'indexation) dans la phase d'indexation. Les composantes de ces vecteurs représentent les poids d'importance des termes correspondants dans le document ou dans la requête. D'une manière générale, la pondération des termes des documents est obtenue dans l'étape d'indexation en utilisant les mesures d'importance locale et globale (Section 1.2.2.4). Formellement, une requête q et un document d sont représentés respectivement par $\overrightarrow{q} =< w_1^q, w_2^q, ..., w_M^q >$ et $\overrightarrow{d} =< w_1^d, w_2^d, ..., w_M^d >$, où M est la taille du vocabulaire d'indexation et w_i^q et w_i^d sont les poids du terme w_i dans la requête q et le document d respectivement. L'idée sous-jacente est que les documents pertinents sont ceux qui partagent les mêmes termes avec la requête et leurs vecteurs sont plus proches du vecteur de la requête. Le score de pertinence est calculé en utilisant des mesures de similarité entre le vecteur de la requête et les vecteurs des documents de la collection. Plusieurs mesures de similarité ont été proposées pour estimer le degré de pertinence ($RSV(d, q)$) d'un document par rapport à une requête :

- Le produit scalaire (Équation 1.4) :

$$RSV(d,q) = \sum_{i}^{M} w_i^q \cdot w_i^d \tag{1.4}$$

- La mesure cosinus introduit un facteur de normalisation pour ne pas favoriser les documents plus longs (Équation 1.5) :

$$RSV(d,q) = \frac{\sum_i^M w_i^q \cdot w_i^d}{\sqrt{\sum_i^M {w_i^q}^2 \cdot \sum_i^M {w_i^d}^2}} \tag{1.5}$$

- Coefficient de Dice (Équation 1.6) :

$$RSV(d,q) = \frac{2 \cdot \sum_i^M w_i^q \cdot w_i^d}{\sum_i^M {w_i^q}^2 + \sum_i^M {w_i^d}^2} \tag{1.6}$$

- Mesure de Jaccard (Équation 1.7) :

$$RSV(d,q) = \frac{\sum_i^M w_i^q \cdot w_i^d}{\sum_i^M {w_i^q}^2 + \sum_i^M {w_i^d}^2 - \sum_i^M w_i^q \cdot w_i^d} \tag{1.7}$$

où M est le nombre de composantes des vecteurs de document et de la requête.

Le modèle vectoriel de base ne prend pas en considération les dépendances entre les termes, l'appariement s'effectue sous l'hypothèse d'indépendance (représentation en sac de mots). Pour remédier aux limitations du modèle de base plusieurs extensions ont été proposées [Wong & Raghavan 1984].

1.2.3.3 Modèles probabilistes

Le premier modèle probabiliste a été introduit au début des années 1960, il consiste à retourner une liste de documents ordonnés par la probabilité de leur pertinence par rapport à une requête [Maron & Kuhns 1960]. Les familles des modèles probabilistes sont fondées sur la théorie des probabilités pour modéliser la notion de pertinence et la prise en considération des facteurs de l'incertitude dans l'expression des besoins et l'imprécision de la représentation d'information. Ces modèles consistent à estimer les probabilités d'observer des évènements liés aux documents et aux requêtes, la probabilité qu'un document d soit pertinent $P(R|d,q)$ ou non-pertinent $P(\overline{R}|d,q)$ vis-à-vis une requête q [Robertson 1977]. D'une manière générale, les modèles probabilistes utilisent principalement les fréquences des termes dans le document (*term frequency*) et dans la collection (*document frequency*), ils se différencient selon les évènements observés et les distributions de probabilités utilisées [Roelleke 2013, Fang & Zhai 2014].

Les modèles probabilistes peuvent être classifiés en quatre familles de modèles : modèles probabilistes de pertinence [Robertson *et al.* 1994], modèles de langue [Ponte & Croft 1998], modèles de déviation à l'aléatoire [Amati & Van Rijsbergen 2002] et les modèles d'information [Clinchant & Gaussier 2010]. Dans ce qui suit, nous allons présenter les principaux modèles probabilistes de RI.

Modèle BM25

Le modèle BM25 est l'un des modèles de RI les plus utilisés en RI, introduit par Robertson & Walker [1996] en se basant sur les hypothèses du modèle BIR (binary Independence Model) [Robertson & Jones 1976]. Le modèle BM25 est développé sous l'hypothèse que les fréquences des termes sont distribuées selon un mélange de deux distributions de Poisson, où une distribution représente l'ensemble élite (distribution des fréquences des mots dans les documents pertinents) et l'autre représente l'ensemble non-pertinente. La fonction d'appariement finale consiste à combiner et normaliser les poids du terme dans la requête, le document et la collection (Équation 1.8) :

$$RSV(q,d) = \sum_{w \in q \cap d} \overbrace{\frac{(k_3+1)\cdot x_w^q}{k_3 + x_w^q}}^{A(w,q)} \cdot \underbrace{\frac{(k_1+1)\cdot x_w^d}{K + x_w^d} \cdot \log \frac{N - N_w + 0.5}{N_w + 0.5}}_{B(w,d,C)} \tag{1.8}$$

où $K = k_1 \cdot ((1-b) + b \cdot \frac{d_l}{l_{avg}})$ est le paramètre de normalisation de la fréquence du terme w dans le document, k_1 est un paramètre utilisé pour équilibrer l'échèle de la fréquence du w dans le document , b est le paramètre de normalisation de la longueur du document et k_3 est le paramètre de normalisation de la fréquence du terme dans la requête.

Modèle de langue

Un modèle de langue peut être défini comme une distribution de probabilité sur ensemble des mots. Le modèle de langue de RI, introduit par Pont & Croft [1998], consiste à ordonner les documents par leur probabilité de générer la requête. Formellement, pour une requête $q = w_1, w_2, ..., w_n$ et un document d, la probabilité de générer la requête par le document d est $P(q|d)$. La fonction d'appariement générale est donnée par Équation 1.11 :

$$RSV_{LM}(q,d) = P(q|d) = \prod_{w \in |C|} P(w|d)^{x_w^q} \tag{1.9}$$

L'inconvénient principal de ce modèle est que l'absence d'un terme de la requête dans un document implique un score nul même si le document sous-jacent est pertinent (cas de disparité des termes et utilisation de synonymes). Pour éviter le problème des probabilités nulles, plusieurs méthodes de lissage ont été proposées [Zhai & Lafferty 2001a]. Ces méthodes consistent à associer à la collection un modèle de langue supplémentaire.

— *Méthode de lissage de Jelinek-Mercer* :

$$\begin{aligned} P(w|d) &= \alpha \cdot P(w|d) + (1-\alpha) \cdot P(w|C) \\ &= \alpha \cdot \frac{x_w^d}{l_d} + (1-\alpha) \cdot \frac{x_w^C}{|C|} \end{aligned} \tag{1.10}$$

où α est le paramètre de lissage de Jelinek-Mercer, l_d la longueur de document d, et $|C|$ le nombre de termes dans la collection.

— *Méthode de lissage de Dirichlet* :

$$\begin{aligned} RSV(q,d) &= \log P(q|d) \\ &= \sum_{w \in q} \underbrace{x_w^q}_{A(w,q)} \underbrace{\log[\frac{x_w^d + \mu \frac{x_w^C}{|C|}}{l_d + \mu}]}_{B(w,d,C)} \end{aligned} \tag{1.11}$$

où μ est le paramètre de lissage de Dirichlet.

Modèles de déviation à l'aléatoire (DFR)

Les modèles DFR sont introduits par Amati & Rijsbergen [2002] en se basant sur l'hypothèse de deux distributions de Poisson. Ces modèles consistent à estimer la quantité d'information apportée par un terme vis-à-vis un document de la collection en utilisant l'information de Shannon : $Inf_1 = -\log P(X_w = x_w^d|\lambda_w)$, où λ_w est un paramètre quantifiant l'importance du terme w dans la collection. L'idée sous-jacente est que plus la distribution d'un terme dans le document diverge de sa distribution dans la collection, traduit par une grande quantité de $P(X_w = x_w^d|\lambda_w)$, plus ce terme est important pour représenter le contenu du document. Pour remédier aux limitations de l'utilisation de l'information de Shannon pour quantifier l'importance du terme dans le document, les modèles DFR utilisent deux principes de normalisation :

- *Premier principe de normalisation* consiste à normaliser la quantité d'information Inf_1 par une nouvelle distribution de probabilité $Prob_2$ pour corriger le risque d'accepter un terme en tant que descripteur informatif de document (surestimer l'importance du terme).

$$(1 - Prob_2) \cdot Inf_1 \tag{1.12}$$

- *Deuxième principe de normalisation* est utilisé pour normaliser les fréquences des termes pour prendre en considération la variation des longueurs des documents, en utilisant l'une des formules suivantes :

$$t_w^d = c \cdot x_w^d \cdot \frac{l_{avg}}{l_d} \tag{1.13}$$

$$t_w^d = x_w^d \cdot \log(1 + c \cdot \frac{l_{avg}}{l_d}) \tag{1.14}$$

où c est le paramètre de normalisation

La fonction d'appariement globale est donnée par (Équation 1.15) :

$$RSV(d,q) = \sum_{w \in q \cap d} x_w^q \cdot (1 - Prob2(t_w^d)) \cdot Inf_1(t_w^d) \tag{1.15}$$

Plusieurs instances du modèle DFR global ont été introduites par Amati & Rijsbergen [2002], nous nous intéressons particulièrement au modèle PL2, donné par Équation 1.16 :

$$RSV(d,q) = \sum_{w \in q \cap d} \underbrace{\frac{x_w^q}{x_{w_{max}}^q}}_{A(w,q)} \cdot \underbrace{\frac{1}{t_w^d + 1}(t_w^d \cdot log_2(\frac{t_w^d}{\lambda_w}) + (\lambda_w - t_w^d) \cdot \log_2(e) + 0.5 \cdot \log_2(2\pi \cdot t_w^d))}_{B(w,d,C)} \tag{1.16}$$

où $\lambda_w = \frac{N_w}{N}$ est le paramètre dépendant de la collection du terme w et c est le paramètre paramètre de normalisation de la fréquence de w par rapport à la longueur de document.

Modèles d'information

La famille de modèles d'information est récemment introduite par Clinchant & Gaussier [2010]. Cette famille peut être considérée comme une sous-famille des modèles DFR. Les modèles d'information consistent aussi a mesurer la déviation du comportement du terme dans le document de son comportement moyen dans la collection en utilisant l'information de Shannon ($\log P(X_w \geq t_w^d | \lambda_w)$). Cependant, elle permet de pailler les limitations des modèles DFR, principalement l'utilisation de la distribution discrète de Poisson avec des fréquences continues, en utilisant des distributions de probabilités continues. La fonction d'appariement globale consiste à estimer l'information moyenne apportée par un document d à la requête q, donnée par Équation 1.17 :

$$RSV(q,d) = \sum_{w \in q \cap d} -x_w^q \log P(X_w \geq t_w^d | \lambda_w) \tag{1.17}$$

où $t_w^d = x_w^d \cdot \log(1 + c \cdot \frac{l_{avg}}{l_d})$ est la fréquence normalisée du terme w dans le document d, $\lambda_w = \frac{N_w}{N}$ est le paramètre de la collection de w et c est le paramètre normalisation de la fréquence de d par rapport à la longueur du document.

Deux distributions en rafale ont été proposées pour estimer l'information moyenne apportée par un document par rapport à une requête [Clinchant & Gaussier 2010] :

1. Le modèle log-logistique (LGD) consiste à fixer le paramètre β de la distribution sous-jacente à 1 :

$$RSV_{LGD}(q,d) = \sum_{w \in q \cap d} \underbrace{\frac{x_w^q}{l_q}}_{A(w,q)} \underbrace{(-\log(\frac{g_w}{g_w + t_w^d}))}_{B(w,d,C)} \tag{1.18}$$

2. La loi SPL (*Smoothed Power Law*) :

$$RSV_{SPL}(q,d) = \sum_{w \in q \cap d} \underbrace{\frac{x_w^q}{l_q}}_{A(w,q)} \underbrace{(-\log(\frac{g_w^{\frac{t_w^d}{g_w + t_w^d}} - g_w}{1 - g_w}))}_{B(w,d,C)} \tag{1.19}$$

1.2.4 Au-delà des termes simples

Les modèles traditionnels de RI, présentés dans la section précédente, sont basés sur l'hypothèse d'indépendance de termes (représentation en sac de mots ou termes simples). Ils effectuent un appariement exact des termes simples partagés par la requête et les documents de la collection pour estimer les scores de ces documents par rapport à cette requête. Malgré que ces modèles fonctionnent assez bien dans la pratique, l'appariement exact des termes seuls est insuffisant, voire imprécis pour assurer une meilleure performance des SRIs. Cela dû principalement à l'ambigüité et la disparité des termes. Afin de surmonter ces limitations, plusieurs méthodes ont été proposées pour aller au-delà des termes simples en RI en utilisant des connaissances sémantiques, syntaxiques et/ou statistiques. Ces méthodes peuvent être classées en deux catégories : les approches utilisant des bases de connaissances [Baziz 2005, Hliaoutakis *et al.* 2006, Fernández *et al.* 2011, Dragoni *et al.* 2012] et les approches basées corpus [Hofmann 1999, Wei & Croft 2006, Ganguly *et al.* 2015, Zuccon *et al.* 2015].

Les approches utilisant des bases de connaissances exploitent des ressources existantes telles que les ontologies, les taxonomies et les réseaux sémantiques pour calculer une similarité sémantique entre les représentations des documents et des requêtes. Ces représentations sont créées à l'aide d'une analyse sémantique utilisée lors de l'indexation. Les ressources sémantiques sont généralement créées et maintenues de façon manuelle ou semi-automatique. Par conséquent, peu de domaines ont des ressources sémantiques explicites appropriées. Les approches basées sur corpus s'appuient sur de grands corpus textuels pour en déduire une représentation plus riche en utilisant des informations telles que des phrases, des syntagmes nominaux, des informations de cooccurrence de mots et/ou leurs contextes. Le principal avantage de ces approches réside dans la facilité d'adaptation à tout domaine où un corpus suffisamment grand est disponible.

En adoptant l'approche basée sur corpus, de nombreux chercheurs ont proposé d'aller au-delà d'hypothèse d'indépendance de termes et remédier au problème de disparité des termes en RI en utilisant diverses méthodes : la reformulation de la requête, les modèles de dépendances de RI, les modèles de traduction et les modèles de thèmes (*topic models*) [Li & Xu 2014]. La reformulation de requêtes [Croft *et al.* 2011] consiste à transformer une requête en une autre qui peut mieux représenter le besoin d'information de l'utilisateur et, par conséquent, améliorer le processus d'appariement des documents et des requêtes. Les modèles de dépendances [Nie & Dufort 2002, Metzler & Croft 2005, Hammache *et al.* 2014] consistent à intégrer des dépendances de termes telles que les termes complexes, les collocations et les dépendances de proximité en RI. Les modèles de traduction de RI [Karimzadehgan & Zhai 2010, Berger & Lafferty 1999] effectuent un appariement sémantique par estimation des probabilités de traduction entre les termes de la requête et ceux des documents afin de permettre aux termes distincts, mais sémantiquement similaires de contribuer aux scores de pertinence. Les modèles de thèmes reposent sur la cooccurrence des mots au niveau des documents pour modéliser les associations entre les termes. Les modèles de thème non-probabilistes [Hofmann 1999], dans lesquels les documents sont représentés dans un espace latent de dimension réduite, sont obtenus en

utilisant des méthodes de décomposition de la matrice matrices-documents. Les modèles de thème probabilistes [Wei & Croft 2006] représentent des associations de termes en supposant que chaque thème est une distribution probabiliste sur un ensemble de termes du vocabulaire et que chaque document de la collection est défini comme une distribution probabiliste sur l'ensemble de thèmes.

Dans cette thèse nous nous intéressons particulièrement à l'intégration et l'exploitation des dépendances entre les termes dans la RI.

1.2.4.1 RI à base des dépendances entre les termes

L'intégration de dépendance entre les termes dans la RI peut être effectuée en utilisant deux approches. La première approche consiste à extraire des unités des termes additionnels permettant de mieux représenter le contenu sémantique véhiculé par les documents et les requêtes. Cependant, la deuxième approche repose sur l'indexation positionnelle des documents de la collection et l'utilisation de proximité entre les termes dans les documents.

Dépendances explicites

L'exploitation des dépendances de termes telle que les syntagmes, les n-grammes et les termes complexes n'est pas récente dans la RI. Fagan [1987] a étudié l'apport des termes complexes en RI. Les termes étaient extraits à base d'une méthode de filtrage statistique des termes et une autre syntaxique. L'indexation des termes repose principalement sur deux étapes : l'identification et la normalisation. La méthode d'extraction de termes à base de filtrage statistique identifie les termes en utilisant leurs propriétés statistiques telles que la cooccurrence, la proximité des constituants, et la fréquence documentaire (*document frequency*). En revanche, la méthode d'extraction syntaxique repose sur un analyseur syntaxique pour l'identification des termes. La phase de normalisation des termes consiste à traiter l'ordre de leurs constituants et leurs variations. L'évaluation est effectuée en utilisant plusieurs collections de RI (CACM, CRAN, MED et CISI). Les résultats ont montré que l'utilisation des termes complexes améliore la performance de la RI. Cependant, la différence en termes de performance en RI entre les termes issus d'une analyse statistique ou syntaxique n'est pas statistiquement significative. Croft et al. [1991] ont introduit une méthode pour la construction des requêtes structurées à partir des requêtes exprimées en langage naturel. Pour l'appariement, ils ont utilisé un modèle probabiliste à base des réseaux d'inférence. Les résultats, obtenus en utilisant la collection CACM, ont montré que l'utilisation des phrases donne une meilleure performance par rapport aux termes simples. De plus, les phrases extraites automatiquement donnent une performance comparable avec ceux qui sont sélectionnés manuellement. Dans un autre travail, Mitra et al. [1997] ont conclu que l'intégration des phrases statistiques ou syntaxiques en RI n'a pas d'effet majeur sur la performance lorsqu'un bon modèle de RI est utilisé. De plus, Haddad [2003] a proposé une méthode pour combiner les syntagmes nominaux avec les règles d'association pour l'indexation la recherche des documents en langue française. Les résultats ont montré

que l'exploitation de ces paramètres donne une meilleure performance par rapport aux termes simples.

D'autres méthodes ont été introduites pour l'intégration des termes complexes en RI. Ces termes sont extraits en utilisant trois approches : statistiques [Kageura & Umino 1996], linguistique [Jacquemin 1997] ou hybride [Daille 1994]. Jacquemin et al. [1997] ont proposé un module pour l'indexation et l'expansion des termes complexes en utilisant des connaissances morphologiques et syntaxiques pour la RI. Nie & Dufort [2002] ont étendu le système SMART pour la prise en compte des termes complexes pour la RI et la RI translinguistique. Pour l'extraction des termes complexes, ils ont utilisé une méthode hybride (filtrage linguistique à base de patrons syntaxiques et filtrage statistique à base du nombre d'occurrences). Les résultats des évaluations ont montré que l'intégration de ces termes améliore la performance pour la RI et la RI translinguistique. Dans un autre travail, Boulaknadel et al. [2008b] ont proposé une méthode pour l'indexation des termes complexes pour la langue arabe. L'extraction de ces termes est effectuée en utilisant une méthode hybride inspirée du système ACABIT [Daille 1994]. L'évaluation est réalisée en utilisant un petit corpus du domaine d'environnement, constitué de 1052 documents contenant 54.705 mots différents, en utilisant le modèle BM25. Les résultats ont montré que l'utilisation des termes complexes améliore la performance de RI en langue arabe par rapport à l'utilisation des termes simples seuls. SanJuan & Ibekwe-SanJuan [2010] ont proposé une méthode pour l'indexation des termes complexes pour la RI précise. L'intégration de ces termes a été évaluée pour l'expansion automatique, l'expansion interactive et la combinaison des deux méthodes d'expansion des requêtes. Les résultats expérimentaux, obtenus sur trois collections de tests, ont montré que l'exploitation des termes complexes pour les différentes méthodes d'expansion des requêtes donne une meilleure performance par rapport aux termes simples. Zhang et al. [2011] ont étudié les différentes représentations des textes pour la RI et la classification de textes (anglais et chinois) en utilisant la représentation en sac de mots TF*IDF, par la méthode LSI (Latent Semantic Indexing) et par des termes complexes. Les résultats ont montré que l'indexation des termes complexes donne une meilleure performance par rapport au modèle TF*IDF pour la RI en langue anglaise. Dans un travail plus récent, Hammache et al. [2014] ont proposé un modèle de langue mixte pour l'intégration des termes complexes dans la RI. Ils ont utilisé l'outil text-NSP [3] pour l'extraction des termes complexes. Les évaluations sont effectuées en utilisant trois collections TREC (AP88, WSJ90–92, et WT10ng). Ils ont obtenu des résultats statistiquement significatifs par rapport à l'utilisation des termes simples.

1.2.4.2 Dépendances implicites : modèles de proximité

Les modèles de proximité consistent à étendre le modèle de base de RI (modèle unigramme) pour la prise en considération des dépendances des termes en utilisant des opérateurs de proximités entre les positions de ces termes dans les documents. L'idée sous-jacente est que plus les occurrences des termes de la requête apparaissent proches (à

3. http ://search.cpan.org/dist/Text-NSP/

proximité) dans un document, plus ce document est considéré pertinent pour la requête. Le score de proximité est, donc, non nul lorsque les termes de la requête se trouvent à proximité dans les documents de la collection.

Plusieurs modèles de proximité ont été introduits pour intégrer les dépendances dans les familles de modèles de RI. Pour la famille de modèle de déviation à l'aléatoire, Peng et al. [2007] ont proposé un modèle pour combiner les poids (scores) des termes simples et les paires de termes des requêtes. Dans le contexte du modèle BM25, plusieurs extensions ont été proposées pour aller au-delà de l'hypothèse d'indépendance de termes [Zhao *et al.* 2011, He *et al.* 2011, Zhu *et al.* 2012]. Pour la famille de modèles de langue, Meltzer et al. [2007] ont proposé une extension du modèle de base en utilisant le formalise de champ aléatoire de Markov, appelé MRF (*Markov Random Field*). Dans un autre travail, Lv & Zhai [2009] ont introduit le modèle de langue positionnel, noté PLM (*Positional Languge Model*), pour combiner et unifier les opérateurs de proximité et de passage. De plus, Shi & Nie [2010] ont proposé un modèle pour intégrer différents types de dépendances et les pondérer selon leurs utilités. Tous les modèles cités précédemment, sauf le modèle PLM, consistent à combiner les scores des termes simples et ceux des dépendances. Ce qui introduit une renormalisation des poids des termes simples en tant que constituants des dépendances. Pour éviter ce problème de renormalisation des dépendances, Sordoni et al. [2013] ont adopté le formalisme probabiliste de la mécanique quantique pour modéliser les dépendances. En effet, ils ont introduit un modèle de langue généralisé, appelé QLM (*Quantum Language Model*), où les dépendances sont considérées comme étant des évènements de superposition des évènements de leurs constituants.

1.2.5 Reformulation de la requête

Les utilisateurs d'un SRI spécifient leur besoin informationnel par une requête constituée d'un ensemble mots-clés en langage naturel. Pour qu'un SRI retrouve l'information pertinente, il nécessite une meilleure formulation et spécification du besoin informationnel. Cependant, le langage naturel pose des difficultés majeures liées principalement à son ambigüité et le problème de disparité des termes (*term mismatch*). De plus, les utilisateurs standards n'utilisent pas forcément les paramètres fournis par le SRI pour bien cibler les documents pertinents. Les techniques de reformulation de la requête consistent à construire une nouvelle requête à partir d'une requête initiale pour mieux représenter le besoin informationnel de l'utilisateur et, par conséquent, améliorer la performance du processus d'appariement. Cette reformulation est effectuée par l'ajout de nouveaux termes et/ou la suppression des termes inutiles, la réévaluation des poids des termes de la requête initiale, ou l'extraction des sous-requêtes à partir des requêtes longues. Dans la littérature, les techniques de reformulation de requêtes sont classées selon divers critères [Baeza-Yates *et al.* 1999, Manning *et al.* 2008, Carpineto & Romano 2012] :

- le type des ressources utilisé pour l'expansion des requêtes ;
- la méthode de sélection et de pondération des termes d'expansion ;
- l'intervention de l'utilisateur ;

Selon Manning et al. [2008], les techniques de reformulation des requêtes sont classées en deux catégories : les méthodes globales et les méthodes locales. Les méthodes globales, appelées méthodes d'expansion de requêtes, consistent à construire une nouvelle requête indépendamment des résultats retournés par la requête initiale en utilisant des ressources externes. Ces ressources peuvent être construites à partir d'une collection de documents (méthodes de classification des termes, relation de cooccurrences, etc.) [Lesk 1969, Peat & Willett 1991, Jing & Croft 1994] ou des ressources sémantiques comme Wordnet [Voorhees 1994, Liu *et al.* 2004a, Fang 2008, Zhang *et al.* 2009]. Les méthodes locales, ou méthodes de retour de pertinence, consistent à construire une nouvelle requête à partir des documents pertinents retourner par la requête initiale [Salton 1971b, Salton 1997, Lavrenko & Croft 2001, Lv & Zhai 2009a].

Dans cette thèse, nous nous intéressons particulièrement aux modèles de rétro-pertinence (*Pseudo-Relevance Feedback (PRF)*) pour l'expansion des requêtes.

1.2.5.1 Modèles de rétro-pertinence

Les modèles de retro-pertinence (Pseudo Relevance Feedback PRF) adoptent l'approche locale et consistent à modifier la requête, de manière automatique, à partir des documents mieux classés. Ils se basent sur l'hypothèse que l'ensemble des documents mieux classés, issus d'une première recherche (résultats de la requête initiale), contient souvent les documents pertinents. Le processus d'expansion de la requête est effectué en quatre étapes :

- Sélectionner les tops k documents mieux classés qui sont retrouvés par la requête initiale ;
- Sélectionner les n meilleurs termes à partir des documents mieux classés ;
- Pondérer l'ensemble des termes sélectionnés et les ajouter à la requête ;
- Retrouver les documents en utilisant la requête modifiée ;

Modèle de divergence Kullback-Liebler (KLD)

Le modèle KLD a été proposé par Carpineto et al. [2001] pour l'expansion automatique de la requête. Les poids des termes candidats d'expansion sont obtenus en mesurant la distance de Kullback-Liebler entre la distribution de ces termes dans l'ensemble F des documents retenu pour l'expansion et leurs distributions dans la collection C. Le score d'un terme candidat est donné par l'équation suivante :

$$F(w) = score_{KLD}(w) = P(w|F) \cdot \log(\frac{P(w|F)}{P(w|C)}) \tag{1.20}$$

où $P(w|F) = \frac{TF(w)}{\sum_{d \in F} l_d}$ est le poids de w dans l'ensemble F des documents de *feedback* et $P(w|C) = \frac{x_w^C}{|C|}$ est le poids de w dans la collection C.

Modèles DFR et modèles d'information

La famille de modèles DFR (*Divergence From Randomness*) et la famille de modèles d'information modifient la requête pour la prise en compte des termes candidats d'expansion en utilisant la formule suivante :

$$q'_w = \frac{q_w}{\max_w q_w} + \beta \cdot \frac{Info(w)}{\max_w Info(w)} \tag{1.21}$$

où le paramètre β contrôle le poids des termes candidat d'expansion de l'ensemble F et les termes de la requête initiale. q'_w désigne le nouveau poids de w dans la requête q. $Info(w)$ est le contenu informatif du terme w dans l'ensemble F.

Modèles Bo : Dans les modèles Bo [Amati 2003] de la famille DFR, le contenu informatif des termes d'expansion, est calculé à base de leurs nombres d'occurrences dans l'ensemble F, défini par :

$$F(w) = Info(w) = \log_2(1 + g_w) + TF(w) \cdot \log_2(\frac{1 + g_w}{g_w}) \tag{1.22}$$

où

$$g_w = \begin{cases} \frac{N_w}{N} & \text{pour le modèle Bo1} \\ P(w|C) \cdot \sum_{d \in F} l_d & \text{pour le modèle Bo2} \end{cases}$$

Modèles d'information : Pour renforcer la prise en compte de la contrainte DF et la contrainte sur la longueur des documents, Clinchant & Gaussier [2013] ont proposé deux instances de modèle d'information pour la rétro-pertinence. Dans ces modèles, le contenu informatif d'un terme d'expansion est calculé à base de la moyenne de son contenu informatif dans l'ensemble F, donné par :

$$F(w) = Info(w) = \frac{1}{|F|} \sum_{d \in F} -\log P(X_w \geq t_w^d | \lambda_w) \tag{1.23}$$

où la quantité d'information $-\log P(X_w \geq t_w^d|\lambda_w)$ est calculé en utilisant le modèle LGD ou SPL, présentées dans la Section 1.2.3.3.

1.2.6 Evaluation d'un SRI

L'évaluation des SRIs est un domaine de recherche qui joue un rôle central dans leur conception et développement, a émergé dans les années 1950 pour comparer les différentes techniques d'indexation automatique et de recherche d'information [Cleverdon 1962]. Le paradigme d'évaluation de Cranfield a été le premier cadre expérimental d'évaluation dans

le domaine de RI [Cleverdon 1962, Cleverdon *et al.* 1966], où une collection de documents contenant 225 requêtes avec leurs jugements de pertinence, a été construite à partir de 1398 résumés des articles d'un journal de l'aérodynamique. Avec cette collection à disposition, il était possible de mener les premières expériences d'évaluation des résultats retournés par un SRI. Ce paradigme d'évaluation a inspiré les chercheurs de domaine de RI [Harman 1992, Peters & Braschler 2001, Gövert & Kazai , Harman & Voorhees 2006], principalement par les possibilités de reproduction des résultats et des expériences, pour la mise à disposition des compagnes d'évaluation telles que TREC [4], INEX [5], CLEF [6], FIRE [7] et NTCIR [8]. D'une manière générale, les collections d'évaluation construites par ces initiatives sont constituées de trois éléments principaux :

- la collection de documents ;
- un ensemble de requêtes représentant le besoin informationnel des utilisateurs ;
- les jugements de pertinence qui sont constitués d'une liste de documents pertinents, voire non-pertinents pour chaque requête ;

Ainsi, le SRI est utilisé pour retrouver les documents correspondant aux requêtes données par la collection d'évaluation, ces documents retrouvés sont comparés avec la liste de documents pertinents pour chaque requête en utilisant les jugements de pertinence. La performance de SRI est donc évaluée en utilisant des mesures d'évaluation qui reposent principalement sur le nombre de documents pertinent, voire non-pertinents retrouvés pour chaque requête.

1.2.6.1 Collections de test TREC 2001/2002

Pour évaluer la performance des SRIs pour la langue arabe, la collection TREC 2001/2002 a été créée par l'initiative TREC. La collection de documents [9] est constituée de 383372 documents contenant 76 millions mots, collectés des articles arabes (Arabic Newswire) de l'Agence France-Presse (AFP), publiés entre mai 1994 et décembre 2000. La collection standard TREC 2001 contient 25 requêtes, où les jugements de pertinence sont constitués des meilleurs 70 documents retournés par 30 évaluations réalisées par 10 équipes de recherche. Le nombre moyen de documents pertinents par requête est 164. Cette collection est largement critiquée [Gey & Oard 2001, Voorhees 2001], d'un coté les requêtes ont de longs titres. D'un autre côté, la plupart des documents pertinents sont retrouvés par le système d'une seule équipe. Ce dernier problème a été évité dans la création de la collection TREC 2002, constituée de 50 nouvelles requêtes avec leurs jugements de pertinence, en fixant un seuil de 6% pour la contribution à l'annotation des documents pertinents pour chaque équipe.

4. http ://trec.nist.gov/
5. http ://inex.mmci.uni-saarland.de/
6. http ://www.clef-initiative.eu/
7. http ://www.isical.ac.in/ clia/
8. http ://research.nii.ac.jp/ntcir/index-en.html
9. https ://catalog.ldc.upenn.edu/LDC2001T55

1.2.6.2 Mesures d'évaluation

Les mesures d'évaluation consistent à mesurer la capacité qu'un SRI retourne tous les documents pertinents correspondant au besoin informationnel de l'utilisateur (requêtes). Nous présentons dans ce qui suit les mesures d'évaluation les plus utilisées :

Précision : cette mesure consiste à calculer le pourcentage des documents pertinents retrouvés parmi tous les documents retournés par le SRI. L'idée sous-jacente est qu'un SRI est très précis si presque tous les documents retrouvés sont pertinents.

$$Précision = \frac{\text{Nombre de documents pertinents retrouvés}}{\text{Nombre de documents retrouvés}} \tag{1.24}$$

Rappel : cette mesure consiste à calculer le pourcentage des documents pertinents retrouvés parmi tous les documents pertinents de la collection.

$$Précision = \frac{\text{Nombre de documents pertinents retrouvés}}{\text{Nombre total de documents pertinents de la collection}} \tag{1.25}$$

La précision et le rappel sont deux mesures dépendantes, quand l'une augmente l'autre diminue. En effet, un SRI qui retourne la totalité des documents pertinents de la collection aura un rappel élevé et une faible précision. Tandis qu'un SRI qui restitue uniquement les documents pertinents aura une précision élevée et un rappel très faible. Donc un bon SRI aura de bons taux de précision et de rappel en même temps.

F1-mesure : consiste à combiner les deux mesures rappel et précision en calculant leur moyenne harmonique [Rijsbergen 1979].

$$F1_{mesure} = 2 \cdot \frac{\text{précision } \cdot \text{ rappel}}{\text{précision } + \text{ rappel}} \tag{1.26}$$

Précision@k : est le rapport entre le nombre de documents pertinents parmi les k premiers renvoyés (ordonnés selon leur score de pertinence). L'idée sous-jacente est que les utilisateurs ne peuvent pas examiner tous les documents retournés vis-à-vis leurs requêtes. Cette mesure est appelée la précision à k, notée $P@k$.

$$P@k = \sum_{i=1}^{k} \frac{R(d_i)}{k} \tag{1.27}$$

où $R(d_i) \in \{0, 1\}$ désigne la pertinence du document d_i. Les mesures d'évaluation ci-dessus ne prennent pas en considération l'ordre de pertinence des documents.

Précision moyenne MAP (*Mean Average Precision*) cette mesure consiste d'abord à calculer la précision moyenne AP (*Aveage Precision*) pour chaque requête au niveau des k premiers documents ordonnés par leur score, ainsi pour chaque document pertinent au rang i on calcule sa précision $P@i$ (le nombre de documents pertinents retrouvés au rang de document i). Ce qui permet de prendre en considération l'ordre des documents pertinents.

$$AP_q = \sum_{i \in \{R(d_i)=1\}} \frac{P@i}{|R_q|} \tag{1.28}$$

où $|R_q|$ est le nombre total de documents pertinents pour la requête q. Cette mesure est souvent calculée pour les 1000 premiers documents retournés par le SRI.

Finalement, la MAP consiste à calculer la moyenne des précisions moyennes des requêtes données par la collection :

$$MAP = \frac{1}{|q|} \sum_{q_i} AP_{q_i} \tag{1.29}$$

1.3 Traitement automatique de la langue arabe

Le Traitement Automatique de la Langue (TAL) concerne l'étude de l'ensemble des méthodes et techniques permettant de modéliser, analyser, interpréter et reproduire le langage humain. Cette discipline se trouve à la frontière de la linguistique, l'informatique et l'intelligence artificielle. Ce domaine de recherche à connu son grand jour au début des années 1950 avec la proposition du fameux test de Turing [1950] et la mise au point des premiers systèmes de traduction automatique. Le point fort du TAL dans le traitement de contenu textuel, par exemple, réside dans les possibilités qu'il offre pour le traiter d'un point de vue linguistique. En effet, le TAL permet d'extraire à partir des textes des informations plus riches de nature morphologique, syntaxique et sémantique. Ces informations sont exploitées pour construire une meilleure représentation du contenu facilitant leur manipulation automatique.

Dans les sous-sections suivantes, nous présentons les particularités de la langue arabe et nous passons en revue des techniques de TAL arabe.

1.3.1 La langue arabe

L'arabe est la langue sémitique la plus répandue. Elle partage plusieurs propriétés avec les autres langues sémitiques qui se rapportent à la morphologie, le vocabulaire, l'ordre libre des mots de la phrase, l'utilisation des voyelles courtes et longues, etc. Elle est parlée dans plus de 22 pays du Maroc jusqu'à l'Irak, par plus de 300 millions de personnes. L'arabe est aussi la langue officielle de l'Érythrée, le Tchad et la Somalie.

Aujourd'hui, l'arabe est devenue l'une des langues les plus utilisées sur Internet, où le nombre d'utilisateurs de cette langue est près de 184,6 millions utilisateurs ce qui constitue

4,6% du nombre total des utilisateurs d'Internet dans le monde[10]. La progression du contenu en arabe sur Internet est dû principalement à la disponibilité de cette langue sur les réseaux sociaux telle que *Twitter* et *Facebook* qui ont servi de relais dans les mouvements sociaux et politiques du printemps arabe.

Malgré que la langue arabe a un statut officiel dans plusieurs pays et caractérisé par un grand nombre de locuteurs, le terme arabe est générique et rassemble plusieurs variétés. En effet, cette langue présente une véritable situation diglossique où des variétés écrites et orales répondant à un spectre très diversifié sont utilisées au sein d'une communauté linguistique. Ces variétés peuvent être classées en trois catégories : l'Arabe Classique, l'Arabe Standard Moderne (ASM) et l'Arabe dialectale. La version de l'arabe à laquelle nous nous intéressons dans cette thèse est l'arabe standard moderne (ASM).

- **Arabe Classique :** c'est variante associée à la religion et à la littérature arabe classique.
- **Arabe Standard Moderne (ASM) :** est la langue officielle du monde arabe, basée sur la syntaxe, la morphologie et la phonologie de l'arabe classique. C'est la variété qui est normalisée et étudiée à l'école.
- **Arabe dialectale :** les dialectes sont les vraies formes de la langue maternelle, utilisée dans la vie quotidienne. On peut regrouper les dialectes arabes en cinq groupes : dialecte Égyptien, dialecte Maghrébin, dialecte du Golfe, dialecte Levantin et dialecte Irakien.

L'arabe présente un vrai défi aux applications de TAL dû principalement à ses propriétés morphologiques et syntaxiques [Fehri 1993, Attia 2008]. Nous détaillons dans les sections suivantes les particularités de l'arabe, ainsi les méthodes proposées pour son traitement automatique.

1.3.2 Particularités de la langue Arabe

1.3.2.1 Alphabet de l'arabe

La langue arabe est une langue sémitique qui s'écrit d'une manière cursive de droite à gauche et dont l'alphabet comporte 28 consonnes. Ainsi, l'alphabet comporte six voyelles standards, dont trois voyelles longues qui sont notées dans l'écriture arabe "ا" (*alef*), "و" (*waw*) et "ي" (*yeh*) et trois voyelles courtes (signes diacritiques dont l'emploi est facultatif) " ◌َ " (*fatha*), " ◌ِ " (*kasra*) et " ◌ُ " (*damma*). En plus des voyelles courtes, il existe d'autres diacritiques de syllabation " ◌ّ ", " ◌ْ " et de *tanwïn* (*nounation*) qui sont utilisées pour désigner les mots indéfinis " ◌ً ", " ◌ٍ ", " ◌ٌ ".

La forme des lettres de l'alphabet change en fonction de leur position dans le mot. Ce qui étend l'alphabet arabe à des représentations différentes [Tayli & Al-Salamah 1990]. Le Tableau 1.1 représente les différentes lettres d'alphabet selon leurs positions dans le mot. De plus, l'écriture arabe est monocamérale (absence de la notion de majuscule et minuscule dans l'écriture) ce qui pose un vrai problème pour la reconnaissance des entités

10. http ://www.internetworldstats.com/stats7.htm

nommées. Les mots peuvent être allongés par l'insertion du caractère *kashida*, par exemple le mot "قال" */qaala/ (a dit)* peut prendre plusieurs formes : "قـــال", "قـــــــــــال", etc.

Tableau 1.1: Les différentes formes des lettres de l'alphabet arabe selon leurs positions. I, D, M, F désignent respectivement lettre isolée, lettre au début, lettre au milieu, et lettre à la fin des mots. La colonne IPA représente leurs représentations phonétiques. (Source [Nwesri 2008], page 13)

I	D	M	F	IPA	I	D	M	F	IPA	I	D	M	F	IPA
ء	–	–	–	/ʔ/	ر	ر	ـر	ـر	/r/	ف	فـ	ـفـ	ـف	/f/
ا	ا	ـا	ـا	/aa/	ز	ز	ـز	ـز	/z/	ق	قـ	ـقـ	ـق	/q/
ب	بـ	ـبـ	ـب	/b/	س	سـ	ـسـ	ـس	/s/	ك	كـ	ـكـ	ـك	/k/
ت	تـ	ـتـ	ـت	/t̪/	ش	شـ	ـشـ	ـش	/ʃ/	ل	لـ	ـلـ	ـل	/l/
ث	ثـ	ـثـ	ـث	/θ/	ص	صـ	ـصـ	ـص	/sˤ/	م	مـ	ـمـ	ـم	/m/
ج	جـ	ـجـ	ـج	/ʒ/	ض	ضـ	ـضـ	ـض	/d̪ˤ/	ن	نـ	ـنـ	ـن	/n/
ح	حـ	ـحـ	ـح	/ħ/	ط	طـ	ـطـ	ـط	/t̪ˤ/	ه	هـ	ـهـ	ـه	/h/
خ	خـ	ـخـ	ـخ	/x/	ظ	ظـ	ـظـ	ـظ	/ð̪ˤ/	و	و	ـو	ـو	/w/
د	د	ـد	ـد	/d̪/	ع	عـ	ـعـ	ـع	/ʕ/	ي	يـ	ـيـ	ـي	/j/
ذ	ذ	ـذ	ـذ	/ð̪/	غ	غـ	ـغـ	ـغ	/ɣ/	ى	–	–	ـى	/aa/
ة	–	–	ـة	/t̪/	–	–	–	–		–	–	–	–	–

1.3.2.2 Absence de voyellation

L'absence quasi systématique de voyelles courtes (signes diacritiques) augmente principalement l'ambigüité morphologique et syntaxique. Cette ambigüité est expliquée par le fait que la majorité des mots arabes acceptent plusieurs voyellation potentielles, ce qui nécessite l'analyse de contexte pour en choisir la meilleure. Ce qui amplifie considérablement la complexité du TAL arabe.

Les voyelles de la structure interne du mot (lexèmes ou forme de base du mot sans leur dernière lettre), ne changent pas en fonction de leurs positions dans la phrase, permettent de déterminer le sens du mot et sa catégorie grammaticale. Tandis que les autres voyelles (casuelles) servent à identifier le rôle syntaxique du mot dans la phrase et se changent en fonction de la position dans la phrase [Diab *et al.* 2007]. Selon Debili & Achour [1998] , dans un texte non-voyellé, 74% acceptent plus d'une voyellation lexicale et 89,9% des noms qui le constituent, acceptent plus d'une voyellation casuelle. Pour illustrer cette complexité, nous avons utilisé l'analyseur morphosyntaxique AlKhalil [Boudchiche *et al.* 2016] afin d'analyser le mot كتب */ktb/*. Le système nous renvoie 17 analyses possibles, nous citons :

- كَتَبَ /kataba "il a écrit"

- كُتِبَ /kutiba “il a été écrit”
- كُتُبٌ /kutub “livres”
- كَتُوبٌ /katob “un écrit”
- كَتَّبَ /kattaba “il a fait écrire”
- كُتِّبَ /kuttiba “faire écrire - forme factitive”
- كَتِّبُ /kattibo “fais écrire”
- ...

1.3.2.3 Absence de ponctuation régulière

Le texte arabe est également caractérisé par l'utilisation irrégulière des signes de ponctuation. Ces signes de ponctuation ont été introduits récemment dans le système d'écriture arabe. Cependant, ils ne sont pas assez essentiels à la compréhension du sens et la segmentation des phrases, et même dans le cas où ils sont notés, comme les autres langues latines telles que le français et l'anglais [Belguith *et al.* 2005, Attia 2008]. Le déplacement entre les idées s'effectue en utilisant des particules comme les conjonctions de subordination et de coordination au lieu de signes de ponctuation. Ces particules s'attachent à la forme fléchie des mots (agglutination des morphèmes), et nécessitent une analyse morphologique rigoureuse pour les identifier.

1.3.2.4 La morphologie de l'arabe

La morphologie s'intéresse à l'étude de la formation des mots à travers l'analyse de flexion, de dérivation et de composition. Il s'agit donc de déterminer les unités minimales de sens des mots, appelés morphèmes (forme de base du mot ou *stem*, affixes et clitiques). L'arabe est caractérisé par sa morphologie riche et complexe [Kouloughli 1994, Watson 2007], constitue un défi majeur aux applications de TAL [Habash 2009, Attia 2012, Zitouni 2014].

Les mots de la langue arabe sont classés en trois catégories de mots : les verbes, les noms et les particules (adverbes, conjonctions, prépositions).

- **Les verbes :** sont des entités qui expriment un sens dépendant de temps, sont dérivés à partir d'un ensemble de racines de trois ou quatre consonnes en utilisant des schèmes. Les verbes de l'arabe ont un nombre limité de schèmes (12 schèmes).
- **Les noms :** sont classés en trois catégories : les noms qui sont dérivés à partir de racine verbale, les noms primitifs (nom propre et communs) et les nombres.
- **Les particules :** sont les prépositions et les conjonctions qui sert à lier des verbes, des noms et phrases.

Cette section a été inspirée du livre de Nizar Habash [Habash 2009]. De plus, la plupart des exemples donnés dans cette section sont pris du même livre.

Problème d'agglutination

L'un des problèmes principaux de l'analyse morphologique de l'arabe est celui de l'agglutination des morphèmes. En effet, les affixes et les clitiques collent à la forme de base du mot. Ce qui augmente la complexité la segmentation des unités lexicales, où un mot peut se traduire en une phrase entière en anglais ou français. Par exemple la segmentation du mot **وسيكتبونها** *wasayaktubuwnahA* (Et ils vont l'écrire) est $wa + sa + y + aktub + uwna + hA$. D'une manière générale un mot en langue arabe est constitué de sa forme de base $BASE$ (*stem* dérivé de racine en utilisant un schème), autour de laquelle s'attachent des préfixes PRF, des suffixes SUF, des proclitiques $PROC$ et des enclitiques ENC.

$$\begin{array}{ccccccccc} PROC & + & PRE & + & [BASE] & + & SUF & + & ENC \\ \overbrace{wa + sa} & + & \overbrace{y} & + & \overbrace{aktub} & + & \overbrace{uwna} & + & \overbrace{hA} \end{array}$$

Habash [2009] a défini l'ordre et les nivaux dans lesquels les clitiques peuvent s'attacher à la forme de base du mot :

$$[\ QST\ +\ [\ CNJ\ +\ [\ PRT\ +\ [\ DET\ +\ BASE\ +\ PRO\ ]\ ]\ ]\ ]$$

- QST conjonction d'interrogation **أ** ;
- CNJ conjonctions de coordination et de subordination : **و** /wa (et) et **ف** /fa (et, alors). Ces conjonctions s'attachent à n'importe quel mot (verbes et noms).
- PRT représente la classe de particules : la marque de future **س** /sa (s'attache aux verbes seulement), les prépositions **ب** /bi (avec) et **ك** /ka (comme) qui s'attachent aux noms et aux adjectifs, et la préposition **ل** /li (pour) qui s'attache aux verbes et aux noms.
- DET est l'article définit **ال** qui s'attache seulement aux noms et aux adjectifs.
- PRO un membre de la classe des clitiques pronominaux, représente les pronoms qui s'attachent à la fin de la forme de base du mot que ce soit un nom ou un verbe.

Morphologie dérivationnelle

Les mots (les noms, les verbes conjugués, etc.) de la langue arabe sont dérivés à partir de racines en appliquant des schèmes "**أوزان**" (patrons morphologiques). La racine "**الجذر**" est un morphème abstrait et irréductible, formée de 2 à 4 consonnes discontinues, relie les mots apparentés morphologiquement [Habash 2009]. Le *stem* "**الجذع**" correspond à la forme non fléchie du mot, obtenu par le croisement d'une racine et d'un schème [Beesley 1998, Dichy & Farghaly]. L'application de schème transforme la racine par l'ajout des préfixes, des infixes (consonne, doublement de consonne ou transformation de voyelle) et des suffixes. Par exemple, le *stem* "**مكتوب**" *ma**kt**uw**b*** est dérivé de la racine "**ك ت ب**" ***k-t-b*** en appliquant le schème *maC1C2uwC3*. La notion de racine de la langue arabe est parfois confondue par les chercheurs avec celle de *stem* pour diverses raisons. L'une de ces raisons est que la notion de racine en anglais et dans les autres langues européennes, ayant

une morphologie concaténative, est plus proche à celle de *stem* de l'arabe. De plus, plusieurs chercheurs réfèrent à la racine en utilisant des consonnes continues, et vu l'omission de voyelles courtes celle-ci peut être confondue avec le *stem* dérivé de la racine en utilisant le schème $C1aC2aC3a$. Par exemple, si la racine "ك ت ب" ***k-t-b*** est référée par "كتب", elle est confondue avec le stem *kataba* (il a écrit) ce qui augmente la confusion entre la racine et le *stem*. En revanche, le lemme correspond à la forme canonique du mot utilisée comme entrée dans les dictionnaires. Dans des cas particuliers, le *stem* peut coïncider avec le lemme [Habash 2009]. La Figure 1.2 illustre un exemple du système de dérivation de l'arabe en utilisant la racine "ك ت ب" ***k-t-b***.

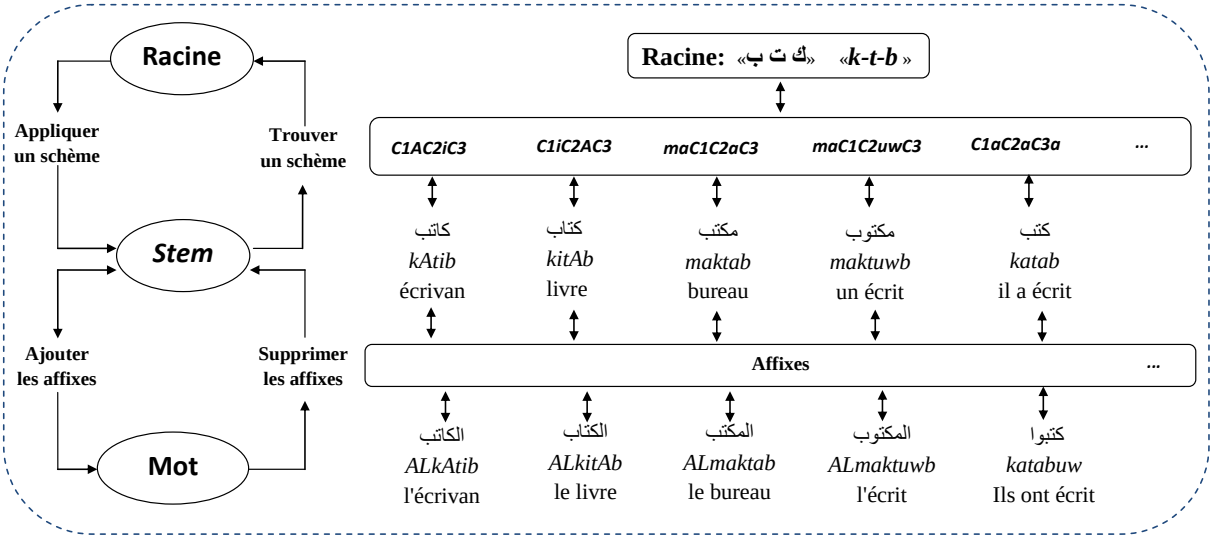


FIGURE 1.2: Un exemple du système de dérivation de l'arabe

La plupart des noms de la langue arabe sont dérivés à partir des verbes, appelé noms déverbaux, en utilisant des schèmes, ce qui introduit un changement de l'étiquète grammatical (Part of Speech Tag) [Habash 2009].

- **Nom verbale مصدر :** est un nom abstrait qui se comporte comme un verbe à l'infinitif, exprime le même sens que le verbe à partir duquel est dérivé. Néanmoins, il n'y'a pas une règle générale pour dériver les noms verbaux à partir des verbes. Par exemple, les noms verbaux de نام nAm (il a dormi) et كتب $kataba$ (il a écrit) sont respectivement نوم $nawm$ et كتابة $/kitAbah$.
- **Participe actif إسم الفاعل :** désigne l'agent du verbe d'action (transitif ou intransitif), dérivé à partir du verbe en utilisant le schème $C1AC2iC3$. Par exemple, le participe actif du verbe نام est نائم $nAim$.
- **Participe passif إسم المفعول :** est un associé à un verbe d'action transitif, désigne celui qui subit l'action et dérivé à partir du verbe transitif en utilisant le schème $maC1C2uwC3$. Par exemple, le participe passif du verbe كتب $kataba$ est مكتوب $maktuwb$ (un écrit).
- **Nom de lieu et nom de temps إسم المكان والزمان :** sont dérivés à partir du

verbe pour désigner le lieu et le temps en utilisant respectivement les schèmes $maC1C2aC3$ et $maC1C2iC3$. Par exemple le nom de lieu du verbe كتب *kataba* est مكتب *maktab*.

- **Nom d'instrument إسم الآلة :** désigne l'instrument qu'on utilise pour faire l'action du verbe. Il existe plusieurs schèmes pour dériver les noms d'instruments à partir des verbes. Par exemple, le schème $miC1C2AC3$ est utilisé pour dérivé le nom منشار *minšAr* (scie) à partir du verbe نشر *našar* (scier).

Il existe d'autres formes de noms qui sont dérivés à partir des verbes moins utilisés tels que le nom d'une fois, le nom de manière et le nom de diminutives إسم التصغير.

Morphologie flexionnelle

La morphologie de l'arabe est aussi flexionnelle, où un ensemble de flexions s'attachent à la fin de la forme de base des mots (*stem*), sans affecter ni leur sens ni leurs catégories grammaticales (POS). On distingue deux types de flexions : les flexions des verbes et les flexions des noms.

- **Flexion des verbes :** la flexion des verbes est dite régulière et elle suit un nombre de règles limité, sert à marquer des traits, d'aspect (accomplit, inaccompli, et impératif), de mode (indicatif, subjonctif et l'apocopé), de sujet, et de voix (passive ou active). Le sujet du verbe est spécifié, par affixation selon l'aspect et le mode, en utilisant trois caractéristiques : la personne, le genre (masculin et féminin) et le nombre (singulier, duel et pluriel). Les sujets se notent en tant que des suffixes dans l'accomplit et des préfixes et suffixes dans le cas d'aspect inaccompli. La Table 1.2 présente l'ensemble des affixes marquant le sujet du verbe.

Tableau 1.2: Les affixes de sujet du verbe. 1,2 et 3 représentent celui qui parle المتكلم, le destinataire المخاطب et l'absent الغائب respectivement. (Source [Habash 2009], page 53)

	Accompli			Inaccompli (Indicatif, Subjonctif, Apocopé)		
	Singulier	Duel	Pluriel	Singulier	Duel	Pluriel
1	+tu	+nA		+ +(u, a, .)	n+ +(u,a,.)	
2	+ta +ti	+tumA	+tum +tun∼a	t+ +(u, a, .) t+ +(iyn, iy, iy)	t+ +(Ani, A, A)	t+ +(uwna, uwA, uwA) t+ +na
3	+a +at	+A +atA	+uwA +na	y+ +(u, a, .) t+ +(u, a, .)	y+ +(Ani, A, A) t+ +(Ani, A, A)	y+ +(uwna, uwA, uwA) y+ +na

- **Flexion des noms :** la flexion des noms est plus complexe que celle des verbes. Ces flexions servent à marquer les traits, de genre, de nombre, d'état et de cas. Le genre et le nombre partagent un ensemble de morphèmes, ces deux traits sont désignés également par des combinaisons des morphèmes de cet ensemble. Pour 80% de noms, leur forme est consistante avec les morphèmes qui désignent le genre et le nombre (cas de pluriel régulier الجمع السالم). Cependant, pour 20% des noms

arabes, le genre et le nombre sont inconsistants avec les morphèmes qui marque le genre et le nombre (cas de pluriel irrégulier **جمع التكسير**) [Maamouri *et al.* 2004]. D'autres inconsistances s'ajoutent au problème du pluriel irrégulier telles que la disparité du genre et le pluriel collectif [Habash 2009]. L'état du mot peut être défini indéfini et construit. Le mot est défini lorsque l'article défini **ال** qui apparait au début de sa forme de base. L'état est construit lorsque le mot se comporte en tant que tête d'annexion **الإضافة**. Dans **كتاب الطالب** *kitAb AlTAlibi* (le livre d'étudiant), l'état du mot **كتاب** est construit. Les noms peuvent prendre trois leur cas : nominatif **مرفوع** $marfuw'$, accusatif **منصوب** *manSuwb* et génitif **مجرور** *majruwr*.

1.3.2.5 Syntaxe de l'arabe

Tandis que la morphologie s'intéresse à l'étude de la structure interne et la forme des mots, la syntaxe en revanche s'intéresse à la façon dont les mots se combinent pour former des phrases et des énoncés. Les phrases de l'arabe se subdivisent en deux classes : la phrase verbale (**الجملة الفعلية**) et la phrase nominale (**الجلة الإسمية**).

Phrase verbale

La phrase verbale est constituée d'un sujet, d'un verbe et d'un complément. Ce type de phrases est caractérisé par l'ordre flexible de ses constituants, sans changer le sens. Par exemple la phrase en français "Le garçon est allé à l'école" peut prendre les formes suivantes :

- Verbe + Sujet + Complément : **ذهب الوِلدُ إلى المدرسة** (Est allé le garçon à l'école)
- Sujet + verbe + complément : **الولدُ ذهب إلى المدرسة** (Le garçon est allé à l'école)
- Complément + Verbe + Sujet : **إلى المدرسة ذهب الولدُ** (A l'école est allé le garçon)

La forme de base d'une phrase verbale est constituée d'un seul verbe avec l'omission du pronom sujet (pro-drop), où le verbe exprime le genre et le nombre. Par exemple, **كَتَب** *kataba* (il a écrit) désigne une phrase où le sujet est masculin singulier et le sujet pronominal est caché (**ضمير مستتر**). Ainsi, les compléments pronominaux peuvent apparaitre en tant que des suffixes verbaux que ce soit le sujet est pronominale ou non-pronominale : **كَتَبها** $kataba + hA$ (il l'a écrit) ou **كَتَبَها الولد** (le garçon l'a écrit).

Phrase nominale

Il s'agit d'une phrase constituée de deux parties, un sujet (ou un groupe de sujet) et un prédicat (ou un groupe de prédicats) (**مبتدأ و خبر**). Le sujet peut être un nom défini, un nom propre ou un pronom. Le prédicat , en revanche, peut être un nom indéfini, un nom propre, un adjectif qui s'accorde en genre et en nombre avec le sujet, une phrase propositionnelle ou encore une phrase nominale.

- Sujet (Nom) + Prédicat (Nom) : **الكِتَاب جَدِيدٌ** (le livre est nouveau) et dans le cas d'accord en genre et en nombre **الكِتابَان جديَدان** (les deux livres sont nouveaux) ;

- Sujet (pronom) + Prédicat (nom propre) : هي سلمى *hiya salmaý* (elle est Salma) ;
- Sujet (Nom) + Prédicat (phrase propositionnelle) : الولدُ في البيْت $Al + rajul + u$ $[fiy\ Albayt + i]$ (le garçon est à la maison) ;
- Sujet (Nom) + Prédicat (phrase nominale) : البَيْت بَابُه جَديد $Al + bayt + u$ $[bAb + u + hu\ jadiyd + un]$(la maison, [sa porte est neuve])

La structure de la phrase nominale est aussi caractérisée par l'ordre flexible, lorsque le sujet est indéfini ou le prédicat est constitué d'une phrase verbale.

- Prédicat (Phrase propositionnelle) + Sujet (Nom) : في البيت رجل $[fiy\ Al + bayti]$ $rajul + \acute{u}$ ([dans la maison] un homme : un homme est à la maison) ;
- Sujet (Nom) + Prédicat [Verbe + Sujet (Nom) + Complément] : الأولادُ كَتَبُوا القصص $Al + AwlAd + u$ $[katab + uwA\ Al + qiSaS + a]$ (les garçons, [ils ont écrit des récits]) équivalent de la phrase verbale (Verbe + Sujet + Complément) كَتَبَ الأولادُ القصص ;

La phrase nominale de base est constituée d'un nom ou d'une adjective et peut avoir plusieurs types de modificateurs, nous notons principalement :

- **l'adjectif :** Les adjectives et les noms s'accordent toujours en termes d'article défini et de cas. De plus les adjectives de noms rationnels s'accordent en genre et en nombre. Cependant les adjectives des noms irrationnels s'accorde en genre et en nombre seulement pour le singulier et le duel. Par exemple, dans la phrase : مكاتبٌ جديدةٌ *makAtibu jadiydahũ* (nouveaux bureaux), l'adjectif جديدةٌ (au féminin singulier) ne s'accorde pas au nom مكاتبٌ (masculin pluriel).
- **l'annexion** (الإضافة) : est une construction possessive/genitive permet de lier deux noms : le premier est le $muDAf$ المضاف et le deuxième est appelé $muDAf\ Ailayhi$ المضاف إليه . Cette construction peut être réalisée par l'insertion d'un nom, d'une adjective ou d'un groupe nominal. Par exemple, la phrase مفاتيح السيارة $mafAtiyHu\ Al + sayArati$ (les clés de la voiture).
- Íl existe aussi des phrases nominales où les noms verbaux jouent le rôle de sujet. Par exemple, dans la phrase تحليل البيانات (analyse de donnés), le nom verbal تحليل (verbe حلّل analyser) est le sujet de la phrase.

1.3.3 Techniques de TAL arabe

1.3.3.1 Tokenisation

La tokenisation consiste à identifier les mots, appelés *tokens*, à partir des séquences de caractères d'un document textuel. La plupart des méthodes de tokenisation des textes arabes reposent sur un certain nombre de séparateurs de mots, telles que les signes de ponctuation, les espaces blancs et les chiffres. Dans un tel contexte, les traitements spécifiques de la langue sont effectués par d'autres blocs (systèmes de racinisation, analyseurs morphologiques, etc.). D'autres méthodes reposent sur une analyse rigoureuse, répondant à la complexité de la morphologie de l'arabe, des textes arabes pour segmenter les mots. Elles consistent à segmenter les clitiques : les conjonctions, les propositions, les clitiques

pronominaux, le trait de futur et l'article défini [Diab 2009]. Donc la forme de base du mot est séparée de l'ensemble des clitiques. Par exemple la segmentation du mot وسيكتبونها *wasayaktubuwnahA* (et ils vont l'écrire) est $wa + sa + y + [aktub] + uwna + hA$.

1.3.3.2 Normalization orthographique

Cette étape consiste à traiter les variations orthographiques des mots. Dans l'arabe écrit, les voyelles (signes diacritiques) ne sont pas souvent notées. Parfois, les auteurs notent quelques voyelles courtes, principalement la voyelle *cheddah* " ّ ", dans leurs textes pour faciliter la lecture et la compréhension. De plus, certaines lettres subissent quelques modifications légères dans leurs écritures, sans affecter le sens des mots où elles sont apparues, mais leur encodage change. Par ailleurs, les mots peuvent être allongés par l'insertion du caractère *kashida*. Ces lettres sont أ, آ , إ, ي, ئ, ة. Les variations orthographiques sont faciles à traiter :

- remplacer les lettres أ, آ et إ par ا ;
- remplacer les lettres ي et ئ par ى ;
- remplacer la la lettre ة par ه ;
- éliminer le caractère *kashida* (-) ;
- éliminer les voyelles courtes ;

1.3.3.3 Elimination des mots vides

Les mots vides sont les mots non importants qui ne véhiculent pas un sens particulier dans le texte et ils apparaissent à peu près dans tous les documents. Ces mots sont éliminés afin de réduire la taille de lexique d'indexation et de considérer juste les termes importants ou représentatifs du contenu.

L'élimination des mots vides repose principalement sur des dictionnaires des mots vides ou des méthodes statistiques. Plusieurs listes de mots vides ont été utilisées pour le TAL arabe. Khoja & Garside [1999] ont proposé une liste de 168 mots vides qui sont utilisés par leur racineur de l'arabe. Chen et Gey [Chen & Gey 2001] ont crée une liste de mots vides pour l'arabe en traduisant les mots vides de l'anglais. De plus, ils ont ajouté les mots les plus fréquents de la collection TREC 2001. Dans un autre travail, Abu El-Khair [2006] a proposé trois listes de mots vides. Une liste de mots vides générale crée en se basant sur la structure de la langue arabe et une deuxième liste construite à partir des mots dont leur fréquence dans le corpus est supérieure à 25.000. La troisième est une combinaison des deux listes. Bouzoubaa et al. [2009] ont proposé une structure standard d'un dictionnaire de mots vides en s'appuyant sur les ressources existantes. La plupart des études ont montré que l'élimination des mots vides augmente considérablement la performance des SRI [Darwish & Magdy 2014].

1.3.3.4 Techniques de racinisation et racinisation légère

Vu la complexité et les défis que présentent la morphologie de l'arabe en particulier la nature flexionnelle et le problème de l'agglutination, l'analyse morphologique constitue l'aspect le plus étudié dans la RI et les applications de TAL de cette langue [Abu El-Khair 2007, Nwesri 2008, Darwish & Magdy 2014]. Dans la littérature, on distingue deux approches principales pour le traitement de la morphologie de l'arabe[Hadni *et al.* 2012] :

- l'approche basée-racine : les techniques de cette approche consistent à réduire l'ensemble des variantes morphologiques du mot à leur racine (morphème irréductible formé de 2 à 4 consonnes discontinues). La plupart des techniques de cette approche procèdent par la suppression des affixes, en vérifiant à chaque suppression qu'on n'a pas enlevé une partie de la racine à partir de la forme réduite du mot. Ensuite, ils utilisent un dictionnaire de schèmes et un autre de racines pour déduire la racine à partir de la forme réduite du mot [Khoja & Garside 1999]. Pour éviter toute sorte de confusion, nous désignons par le terme **racinisation** l'ensemble des techniques visant à extraire la racine à partir des forme fléchies des mots.
- l'approche basée-*stem* : les techniques de cette approche consistent à réduire les différentes variantes morphologiques du mot à leur *stem*. Ces techniques reposent principalement sur l'utilisation d'une liste de suffixes et une autre de préfixes pour réduire les formes fléchies des mots à leurs *stems*. En langue arabe, le *stem* correspond à la forme de base du mot obtenu par l'application d'un schème morphologique à la racine(Section 1.3.2.4). Nous désignons par le terme **racinisation légère** l'ensemble des techniques de l'approche basée-*stem*.

Nous passons ensuite en revue des travaux qui s'intéressent à la racinisation et la racinisation légère pour la RI en langue arabe.

Racinisation (*Stemming*)

Les premiers travaux proposés pour la RI en langue arabe ont été influencés par la nature dérivationnelle de sa morphologie, ils ont utilisé principalement des techniques permettant de réduire l'ensemble des mots de la collection à leurs racines. Cela consiste à représenter l'ensemble des mots dérivés par leurs racines. Ces techniques reposent sur l'utilisation d'une liste de schèmes et une autre de racines.

Al-Fedaghi & Al-Anzi [1998] ont proposé un système permettant d'extraire la racine formée de trois lettres des mots arabes. Ils ont utilisé une liste étendue de schèmes avec leurs possibles affixes et un dictionnaire de racines. Dans un premier temps le mot est comparé avec les schèmes de même longueur pour extraire la racine correspondante. Cette racine est retournée dans le cas où elle figure dans la liste des racines valides. Selon les auteurs, l'algorithme à réussi à extraire les racines correctes pour plus de 80% des mots en utilisant un petit corpus. Cet algorithme est étendu par Al-Shalabi & Evens [1998] pour extraire des racines formées de quatre lettres. L'algorithme est amélioré en élimant les préfixes possibles, puis il extrait la racine en prenant les cinq premières lettres du

mot, en supposant que la racine de chaque mot doit apparaître toujours dans les cinq premières lettres de celui-ci. Ensuite, ils ont utilisé un dictionnaire de schèmes et un autre de racines pour en déduire la racine du mot. Ces deux algorithmes ne sont pas évalués pour la RI en langue arabe, mais ils constituent le point de départ pour plusieurs techniques de racinisation.

Khoja & Graside [1999] ont introduit un autre algorithme plus élaboré pour l'extraction des racines. En effet, c'est l'algorithme le plus utilisé pour l'extraction des racines des mots de la langue arabe. Avant de procéder à la comparaison du mot avec la liste de racine, l'algorithme élimine le plus long préfixe et suffixe. Ensuite, il procède à l'extraction de la racine à partir de la forme réduite en vérifiant à chaque suppression qu'on n'a pas enlevé une partie de la racine. La dernière étape consiste à comparer la forme réduite du mot avec la liste de schèmes des verbes et des noms de même longueur, pour extraire la racine la plus appropriée. Dans le cas où l'algorithme n'a pas trouvé la racine du mot, il retourne le mot sous sa forme originale sans suppression des affixes.

Les premiers travaux en RI ont montré que la racinisation donne de meilleures performances que l'indexation à base des mots ou des stems. En effet, Al-Kharashi et Evens [1994a] ont comparé la performance d'indexation manuelle à base de racine, de stems et des mots en utilisant une petite collection de 355 documents et 29 requêtes avec leur jugement de pertinence préétablies. Ils ont utilisé un dictionnaire créé manuellement à partir de la collection de tests contenant 1126 mots, 725 stems, et 526 racines. Les résultats des évaluations ont montré que l'indexation à base de racine donne une meilleure performance que celles à base de mots bruts ou des stems. Dans un autre travail similaire, Abu-Salem et al. [1999a] ont confirmé la même conclusion sur une autre collection de 120 documents et 32 requêtes. Les résultats ont montré que l'indexation à base de racine a amélioré significativement la performance de RI par rapport à l'utilisation des stems ou des mots.

Hmeidi et al. [1997] ont étudié l'indexation automatique et l'indexation manuelle en utilisant les mots, les racines et les stems. La collection de test est constituée de 242 résumés et 60 requêtes avec leurs jugements de pertinence. Leurs résultats montrent que l'indexation manuelle en utilisant des racines donne de meilleurs résultats par rapport à l'utilisation de mots et de stems. Dans un autre travail qui se rapporte à l'indexation automatique, Darwish et al. [2005] ont utilisé l'outil *Sebawai* [Darwish 2002] pour indexer les documents de la collection standard TREC 2002 en utilisant les racines et les stems extraits par cet outil. Les résultats ont montré que les performances d'indexation à base des racines et celle à base des stems sont comparables. En utilisant la collection TREC 2001 [Darwish & Oard 2007], les résultats ont montré que l'indexation à base des racines est à peu près comparable avec l'indexation à base des mots. Tandis que l'indexation à base de stem a donnée la meilleur performance.

Racinisation légère (*light stemming*)

La racinisation légère consiste à éliminer l'ensemble des suffixes et des préfixes de la forme du mot. Les techniques de racinisation légère reposent principalement sur l'utilisation

d'une liste de suffixes et une autre de préfixes pour réduire les mots fléchis à leurs *stems*.

Aljlayl & Frieder [2002] ont proposé une méthode de racinisation légère. Avant de procéder à la suppression des affixes (voir Tableau 1.3), l'algorithme commence par l'élimination des voyelles courte et la normalisation des mots. Dans un premier temps l'algorithme supprime la conjonction و *wa*, l'article défini ال et les préposions qu'il précèdent. Puis, il supprime les suffixes tout en commençant par ceux qui sont plus longs. Cet algorithme a été comparé avec la méthode de racinisation de Khoja [Khoja & Garside 1999] en utilisant la collection de test TREC 2001. Les résultats d'évaluation ont montré que cette méthode de racinisation légère améliore significativement la performance par rapport à l'algorithme de Khoja pour la RI ad hoc et l'expansion de requêtes.

Larkey & Connell [2002] ont proposé une multitude de techniques de racinisation légère :light1, light2, light3 et light8. La plupart entre eux partagent la même technique de normalisation. Ces techniques varient selon le nombre et la profondeur des préfixes et des suffixes à éliminer. Ces techniques consistent à supprimer les affixes dans un ordre spécifique de la gauche à la droite de la forme des mots (voir Tableau 1.3). La dernière version est la plus performante, appelée light10 [Larkey *et al.* 2007], consiste à combiner les autres versions. L'évaluation a été effectuée en utilisant les deux collections de tests TREC 2001 et TREC 2002. Les résultats ont montré que l'algorithme light10 donne de meilleurs résultats par rapport à la méthode de racinisation de Khoja.

Chen & Gey [2002] ont proposé deux techniques pour la racinisation légère des textes arabes. La première consiste à grouper les mots selon leur traduction en anglais. Les mots qui ont la même traduction sont représentés par le mot dont la forme est la plus courte. La deuxième technique consiste à supprimer les préfixes et les suffixes (voir Tableau 1.3). La liste de 26 préfixes et 22 suffixes sont construites selon leur rôle grammatical et leurs fréquences d'apparition dans les mots distincts de la collection TREC 2001. Les deux algorithmes sont comparés avec l'algorithme de Berkeley. Ce dernier a donné une meilleure performance par rapport aux deux techniques proposés.

Darwish & Orad [2003] ont proposé un autre algorithme de racinisation légère, appelé Al-Stem. L'algorithme utilise une liste de 24 préfixes et une autre de 23 suffixes (voir Tableau 1.3). Cet algorithme a été comparé avec sa version modifiée et l'algorithme light8 de Larkey en utilisant les collections TREC 2001 et TREC 2002. Les résultats ont montré que la version modifiée de Al-Stem a donné une performance supérieure à celles de l'algorithme de base et l'algorithme light8.

Kadri & Nie [2008] ont proposé une méthode linguistique pour la racinisation légère. Ils ont utilisé le corpus TREC 2001 pour construire l'ensemble de stems possibles pour chaque terme de la collection. L'algorithme procède par la suppression des préfixes et suffixes selon leurs fréquences d'apparition dans les mots du corpus utilisé. Cet algorithme a été comparé avec light10 en utilisant les deux collections de test TREC 2001 et TREC 2002. Les résultats obtenus par cet algorithme donne de meilleures performances par rapport à l'algorithme light10.

Abdelali et al. [2016] ont proposé un outil pour la segmentation des mots arabes en utilisant SVM-rank. Cet outil a été évalué pour plusieurs applications de TAL et,

Tableau 1.3: L'ensemble des affixes utilisé par les méthodes de racinisation légère : Aljlayl, Light10, Al-Stem et Chen

Méthode	Préfixes	Suffixes
Aljlayl	و ، ال ، وال ، كال ، ست ، سي ، ل، ب ، ت ، ي ، لل ، ال	ين ، ون ، ات ، ة ، ان ، ي ، هم ، هن
Light10	لل، ال ، وال، بال، كال ، فال، و	ها ، ان ، ات ، ون ، ين ، يه ، يةه، ة، ي
AL-Stem	وال، بال، فال، بت، يت، لت، ، مت، وت، ست، لم، بم نت، وم، كم، فم، ال، لل، وي، لي، في، وا، فا، لا، با	ات، وا ، ونوه، ان، ، تي، ته، تم، كم، هم هن ، ها، ية، تك، نا ، ين، يه، ة، ه، ي، ا
Chen	وال، بال، فال، كال ، ولل، مالما ال، لال، فا ، كا، ول، وي، وس ، سي، سال لا، وب ، وت، وم، لل، با، و ، ب، ل	ها، ية ،هم، ن، ما، و، يا ، ني كن، تم، تن، ين، يا، ه، كم ان، ات، ون، ة، ه، ي، ت

particulièrement la RI en lange arabe. Cet outil a été comparé avec le segmenteur de Stanford et MADAMIRA [Pasha *et al.*]. L'idée sous-jacente consiste à éliminer tous les affixes segmentés. Les résultats sur la collection TREC 2002 ont montré que le segmenteur Farasa donne une meilleure performance par rapport au sgementeur de Stanford et MADAMIRA.

Dans nos évaluations, nous utilisons le segmenteur Farasa [Abdelali *et al.* 2016], l'algorithme light10 [Larkey *et al.* 2007] et le racineur Khoja [Khoja & Garside 1999] (chapitres 3 et 4).

1.3.3.5 Étiquetage morpho-syntaxique

L'étiquetage morho-syntaxique (*Part of Speech Tagging*) consiste à affecter à chaque mot, selon son contexte, une étiquète (*tag*) correspondant à sa catégorie morpho-syntaxique, son genre, son nombre, etc. Il joue un rôle central dans la lemmatisation, l'analyse syntaxique, l'extraction de l'information (terminologie, entités nommées, etc.) et également la RI [Lioma & Blanco 2009, Boulaknadel *et al.* 2008a, Kanaan *et al.* 2005].

En plus de l'analyse de la morphologie de l'arabe, l'étiquetage morpho-syntaxique devient particulièrement important à cause des ambigüités lexicales des mots [Hadni *et al.* 2013]. En effet, plusieurs étiqueteurs morpho-syntaxiques sont proposés pour la langue arabe. Ces étiqueteurs se diffèrent selon l'approche adoptée (à base de règles ou par apprentissage automatique) et l'ensemble des étiquettes (*tagset*).

Khoja [2001] a proposé un étiqueteur hybride en adaptant le système BNC, utilisé pour l'anglais. Cet étiqueteur combine les donnés statistiques des unités lexicales ainsi un ensemble de règles qui sont dérivées de la théorie traditionnelle de la grammaire arabe. Pour entrainer cet étiqueteur, elle a utilisé un corpus de 50.000 mots et un ensemble de 131 étiquettes. L'évaluation a montré que la précision du système est de 90% lorsque la

FIGURE 1.3: Un exemple de texte arabe étiqueté en utilisant le système AMIRA

désambiguïsation des mots est utilisée.

Diab et al. [2004a] ont adapté le système proposé pour traiter l'anglais *Yamcha* qui utilise le modèle SVM (Support Vector Machine). Pour l'apprentissage, ils ont utilisé le corpus TreeBank arabe, où 4000 phrases sont utilisées pour l'apprentissage, 119 phrases pour le développement et 400 sont utilisées pour le test. La précision obtenue est autour de 94%. En analysant les résultats, ils ont remarqué que la majorité des erreurs sont liées à la confusion des adjectifs et des noms et les erreurs de segmentation de l'article défini. Dans un autre travail plus important, Diab [2009] a proposé un nouveau système, appelé AMIRA [11] successeur de son premier étiqueteur [Diab *et al.* 2004], plus rapide pour la segmentation des textes (mots et des phrases) et l'étiquetage morphosyntaxique. Le système propose deux ensembles des étiquettes : 25 étiquettes PATB (Pen Arabic TreeBank) et ERTS pour annoter les traits morphologiques des mots telle que le genre, le nombre, l'article défini, etc. Pour les deux ensemble des étiquettes, les précisions du système sont autour de 96%. La Figure 1.3 présente un exemple de texte annoté en utilisant le système AMIRA.

D'autres étiqueteurs sont également proposés, mais leurs performances restent comparables, voire inférieures de celle de AMIRA. Nous avons utilisé ce dernier système pour annoter les textes arabes.

1.4 Conclusion

Dans ce chapitre, nous avons présenté les concepts fondamentaux de la RI. Nous avons introduit les concepts de base de la RI telle que la requête, la collection de documents, la notion de pertinence et le processus général de la RI. Nous avons décrit également la phase d'indexation ainsi que les techniques et les niveaux de TAL (morphologie, syntaxe, sémantique) qui se mettent en jeu pour construire une meilleure représentation des

11. http ://nlp.ldeo.columbia.edu/amira/

documents et des requêtes. Puis, nous avons passé en revue des différents modèles de RI qui sont utilisés pour l'appariement des documents et des requêtes. Nous avons aussi présenté les techniques de reformulation des requêtes. Enfin, nous avons présenté les méthodes utilisées pour l'évaluation des SRI.

Dans la deuxième partie du chapitre, nous avons examiné les caractéristiques de la langue arabe et les techniques de TAL proposées pour en faire face. Dans un premier temps, nous avons présenté les propriétés orthographiques. Puis, nous avons décrit ces propriétés morphologiques telles que la nature agglutinante, la morphologie dérivationnelle et flexionnelle. Nous avons aussi discuté ces propriétés syntaxiques telles que les deux types de phrases (verbale et nominale) ainsi leurs constituants. Ensuite, nous avons passé en revue des différentes techniques de TAL proposées pour traiter la morphologie de l'arabe telles que la racinisation et la racinisation légère. Enfin, nous avons introduit quelques méthodes de l'étiquetage morphosyntaxique et particulièrement l'outil AMIRA que nous avons utilisé pour l'annotation des textes arabe.

Dans le chapitre suivant, nous allons présenter notre méthode hybride proposée pour l'extraction des termes complexes.

Chapitre 2

Extraction des termes complexes

Sommaire

2.1 Introduction

L'extraction des termes complexes (TCs) est une tâche importante pour plusieurs applications de TAL telles que l'acquisition de terminologie, l'extraction de l'information, le résumé automatique et la RI [Liu *et al.* 2004b, Jacquemin *et al.* 1997, Zhang *et al.* 2007, Boulaknadel *et al.* 2008a]. Le terme complexe est une unité monoréférentielle, composée au moins de deux unités lexicales simples liées syntaxiquement, qui représente une notion univoque, et qui dénomme un concept appartenant à un domaine de spécialité [Collet 2000]. En RI par exemple, ces termes permettent d'exploiter des connaissances syntaxiques dans l'indexation et l'appariement des documents et des requêtes et d'aller au-delà de la représentation par sac de mots simples. L'objectif principal d'extraction automatique des TCs consiste à extraire des termes spécifiques qui sont plus représentatifs du contenu sémantique de textes [Korkontzelos *et al.* 2008]. En effet, ces TCs sont des constructions syntaxiques moins ambigües et moins polysémiques que les termes simples isolés [Daille 1994, Jacquemin *et al.* 1997, Zhang *et al.* 2008]. Par exemple, le terme complexe **ذهب أسود** se réfère au "pétrole", mais le terme **ذهب** peut se référer au verbe "aller"

(“il est allé”) ou le métal ”or” et le terme أسود peut se référer à la couleur “noir” ou le pluriel de “lion”.

Trois approches ont été introduites pour l'extraction des TCs : approche linguistique, approche statistique et l'approche hybride ou mixte. La première approche utilise des filtres linguistiques pour l'extraction des TCs telle que les patrons syntaxiques. L'approche statistique repose sur l'exploitation des mesures statistiques ou mesures d'association pour bien déterminer les frontières d'un TC. Ces mesures peuvent être classées en deux catégories : celles qui permettent de mesurer le degré de la relation des TCs aux concepts du domaine spécifique ou degré de spécificité (*Termhood*) et celles qui permettent de quantifier le degré de stabilité d'une combinaison syntagmatique ou degré d'unité (*Unithood*) [Kageura & Umino 1996]. La dernière approche consiste à combiner les filtres linguistiques et statistiques pour l'extraction des TCs, où les termes candidats sont identifiés par les filtres linguistiques et leurs degrés d'importance sont calculés par les mesures d'association.

Dans ce chapitre, nous examinons les approches principales d'extraction des termes complexes dans la Section 2.2. Dans la Section 2.3, nous passons en revue des méthodes proposées pour l'extraction de terminologie de la langue arabe. Puis, nous introduisons aussi notre méthode hybride proposée pour l'extraction des TCs dans la Section 2.4.Dans la Section 2.5, nous présentons la méthode d'évaluation et les résultats obtenus. Enfin, nous terminons ce chapitre par une conclusion (Section 2.6).

2.2 Approches d'extraction des termes complexes

Plusieurs travaux de recherche ont été proposés pour l'extraction des TCs. Ces travaux sont classés selon l'approche adoptée : linguistique, statistique ou hybride. Les études récentes ont montré que les méthodes hybrides donnent de meilleurs résultats par rapport aux méthodes linguistiques et statistiques [Tadić & Šojat 2003].

Dans cette Section, nous passons en revue des approches principales d'extraction des termes complexes.

2.2.1 Approche linguistique

Les méthodes de cette approche reposent sur l'exploitation des connaissances linguistiques et la structure de la langue traitée (syntaxe, morphologie, etc.) pour le repérage et l'identification des TCs. La plupart des méthodes adoptant cette approche reposent sur l'utilisation des patrons syntaxiques et peu de travaux utilisent les frontières de TCs. Les méthodes utilisant les partons syntaxiques reposent sur une analyse complète des phrases pour identifier les syntagmes nominaux susceptibles d'être des TCs. Les autres procèdent par une analyse de surface de la phrase pour les repérer [Bounhas & Slimani 2009, Boulaknadel *et al.* 2008b]. Nous présentons par la suite quelques outils proposés pour l'acquisition des terminologies.

2.2.1.1 TERMINO

TERMINO est le premier système proposé pour l'extraction de terminologie à partir des textes en utilisant des patrons morpho-syntaxiques [David & Plante 1990]. Les versions récentes de ce système sont distribuées sous le nom NOMINO. Il permet d'extraire les termes candidats, appelés *synapsies*, en procédant par l'identification des syntagmes nominaux du corpus. Le prétraitement du corpus s'effectue en utilisant une base de données lexicale et des règles de désambigüisation lexico-syntaxiques. Les termes candidats sont repérés à partir des différentes expansions des noms. Le système retourne également une liste de termes qui sont jugés valides.

2.2.1.2 LEXTER

Le système LEXTER a été proposé dans [Bourigault 1994] en adoptant l'approche linguistique, où les syntagmes nominaux (susceptibles d'être des termes) sont repérés en utilisant la méthode de frontières. Puis, ces syntagmes sont décomposés en tête et expansion afin de les proposer en tant que des candidats termes. Pour le repérage, il définit plusieurs signes permettant de séparer les syntagmes nominaux à partir des autres constituants de la phrase :

- les signes de ponctuation ;
- les verbes ;
- les pronoms ;
- les déterminants précédés d'un verbe ou d'un signe de ponctuation ;
- etc.

2.2.1.3 FASTER

L'outil FASTER [1] est proposé dans [Jacquemin 1997] pour l'extraction de terminologie. Cet outil utilise une liste de référence de termes valides et permet d'identifier l'ensemble des variantes de ces termes. Ces variantes sont identifiées en utilisant un ensemble de méta-règles qui opèrent à plusieurs niveaux : morpho-syntaxique, syntaxique ou syntaxico-sémantique.

2.2.1.4 SYNTEX

Le SYNTEX est un analyseur syntaxique proposé pour l'extraction des syntagmes nominaux, l'extraction des terminologies et la construction des ontologies [Bourigault & Fabre 2000, Bourigault *et al.* 2005]. Le système reçoit comme entrée un corpus étiqueté préalablement. Il procède par une analyse de dépendance pour reconnaitre les différentes relations syntaxiques telles que le sujet, l'objet direct, le complément prépositionnel (de nom, de verbe et d'adjectif), l'antécédence relative, la modification adjectivale (épithète, attribut) et la subordination [Bourigault *et al.* 2005]. Ces relations permettent de construire un réseau terminologique, ou chaque syntagme est lié à sa tête et ses expansions.

1. https ://perso.limsi.fr/jacquemi/FASTR/

2.2.2 Approche statistique

Le but principal de l'utilisation des méthodes statistiques pour l'extraction des TCs consiste à ordonner les termes candidats selon une mesure d'association particulière qui donne des scores plus élevés aux «bons» termes candidats. Les termes candidats au-dessus d'un seuil particulier sont considérés en tant que des TCs valides. L'idée sous-jacente est que les termes candidats qui sont fréquents ont tendance d'être des TCs valides et de représenter des concepts importants du domaine en question. Cependant, la fréquence seule ne compte que le nombre d'occurrences d'un terme candidat dans le texte, mais ne donne aucune information sur le degré d'association entre les mots qui le composent. Par conséquent, la plupart des approches statistiques visent à extraire les TCs à partir d'un corpus en utilisant des mesures d'association qui se concentrent sur le degré d'unité et/ou le degré de spécificité [Kageura & Umino 1996]. Ces mesures sont basées sur des informations de fréquences, de co-occurrence et de contexte telles que le T-score [Church *et al.* 1991], la loglikelihood (LLR) [Dunning 1993], le C/NC-Value [Frantzi *et al.* 2000], etc. Nous présentons brièvement par la suite quelques systèmes qui reposent sur des méthodes statistiques.

2.2.2.1 MANTEX

Le système MANTEX repose sur la méthode de segment répété pour l'extraction de la terminologie [OUESLATI 1999]. Les segments non séparés par des délimiteurs et dont le nombre d'occurrences est supérieur à un seuil particulier sont considérés comme des termes candidats. Ces délimiteurs sont les signes de ponctuation, des verbes, des pronoms, etc. Le repérage s'effectue en indexant les mots ainsi leurs positions en utilisant des fenêtres de un à dix mots de la même phrase. Une étape de filtrage est nécessaire pour considérer que les meilleurs termes candidats.

2.2.3 Approche hybride

Les approches linguistiques se concentrent sur les structures syntaxiques de la langue et les méthodes statistiques se concentrent sur les caractéristiques récurrentes de TCs. Les méthodes hybrides consistent à combiner les deux approches pour bien bénéficier des avantages des deux [Daille 1994]. En effet, les méthodes linguistiques effectuent une analyse plus fine de la langue pour assurer un meilleur découpage des termes du corpus. Les méthodes statistiques permettent de filtrer les termes candidats importants en vue de les considérer comme des termes valides.

Nous présentons brièvement par la suite quelques systèmes qui reposent sur des méthodes hybrides.

2.2.3.1 ACABIT

Le système ACABIT adopte une méthode hybride pour l'extraction de terminologie [Daille 1994, Daille 1996]. Le filtre linguistique consiste à extraire des termes candidats en utilisant des patrons syntaxiques sur un texte préalablement étiqueté et lemmatisé. De plus, il permet d'identifier les variantes des termes candidats appartenant aux niveaux morpho-syntaxique, syntaxique et sémantique. Pour le filtrage statistique, ce système utilise la mesure d'association LLR [Dunning 1993] pour mesurer le degré d'unité des termes. La liste des termes retournée est ordonnée selon la valeur attribuée par le filtre statistique. Les termes ayant une grande valeur de LLR sont supposés des termes valides représentant le contenu des documents du corpus.

2.2.3.2 TERMS

Le système TERMS adopte aussi une méthode mixte pour l'extraction de terminologie à partir de corpus de domaine spécifique [Justeson & Katz 1995]. L'idée sous-jacente est que le nombre d'occurrences des termes dans un corpus technique est élevé par rapport aux syntagmes non terminologiques. De plus, ils ont des structures différentes de ces syntagmes non terminologiques. À partir d'un corpus étiqueté, ce système utilise des patrons syntaxiques construits en étudiant des structures syntaxiques des dictionnaires terminologiques. Les termes sont filtrés par la suite en utilisant leurs fréquences pour en éliminer ceux qui sont moins fréquents.

2.3 Travaux reliés à l'arabe

La plupart des méthodes proposées pour l'extraction de terminologie de la langue arabe adoptent des méthodes hybrides. En effet, les méthodes d'acquisition de terminologie doivent prendre en considération les caractéristiques morphologiques (agglutination, flexions, etc.), morpho-syntaxiques et syntaxiques (structure complexe de la phrase nominale et ordre flexible des constituants de la phrase). D'où la nécessité de reposer sur une analyse linguistique pour repérer les TCs et traiter leurs variantes. Vu la richesse syntaxique de la langue arabe, où l'utilisation des phrases nominales est très récurrente, le filtrage statistique devient particulièrement important pour ne considérer que les termes désirables.

Boulaknadel et al. [2008] ont introduit une méthode hybride pour l'acquisition des terminologies, inspirées des travaux proposés pour l'anglais et le français (système ACABIT). Le filtrage linguistique s'effectue sur un corpus étiqueté en utilisant l'étiqueteur morpho-syntaxique de Diab et al. [2004b]. Dans la première étape, les termes candidats sont identifiés en utilisant trois patrons syntaxiques : Nom_1 $Nom2$, Nom ADJ et Nom_1 $Prep$ Nom_2. Puis, le filtre linguistique identifie les variantes orthographiques, morpho-syntaxiques et syntaxiques pour chaque terme identifié dans l'étape précédente. Ensuite, la liste des termes candidats est filtrée en utilisant des mesures d'association telle que LLR [Dunning 1993],

l'information mutuelle (IM), T-score et la mesure FLR [Nakagawa & Mori 2003]. Enfin, le système retourne une liste ordonnée des termes candidats. Les résultats obtenus sur un corpus du domaine d'environnent ont montré que la mesure LLR donne une meilleure performance par rapport aux autres mesures.

En plus de l'identification de la structure des TCs, Bounhass et al. [2009] ont introduit une méthode hybride pour l'extraction des TCs. Cette méthode permet d'identifier le rôle des constituants d'un terme candidat, dans l'optique d'utiliser cette méthode pour la construction d'ontologie. Cette méthode repose sur six composants. Le corpus est annoté en utilisant l'étiqueteur de Diab et al. [2004b] afin d'identifier les termes candidats à l'aide des patrons syntaxiques. Ils ont utilisé Aramorph pour traiter les ambigüités morphologiques. Un composant appelé morpho-POS est développé pour la mise en correspondance d'étiquette morpho-syntaxique et l'analyse morphologique la plus probable. Après avoir identifié les séquences des termes candidats, ils ont utilisé un analyseur syntaxique pour identifier la typologie des termes candidats (annexion, postposition, substitution, etc.). Le filtrage statistique est effectué en utilisant la mesure LLR.

Al Khatib & Badarneh [2010] ont proposé une méthode similaire à la méthode introduite par Boulaknadel et al. [2008]. Pour étiqueter le corpus, ils ont utilisé l'étiqueteur à base de règles proposé par [Al-Taani & Al-Rub 2009]. Le filtrage statistique est effectué en utilisant la mesure d'unité LLR (*loglikelihood*) [Dunning 1993] et la mesure de spécificité C-value [Frantzi *et al.* 2000] ainsi la combinaison des deux. Les résultats ont montré que la combinaison des deux mesures donne une meilleure performance.

La plupart des méthodes hybrides présentées précédemment ont été évaluées pour les 100 premiers termes candidats et traitent les bigrammes seulement (c'est-à-dire des termes candidats de longueur 2). En outre, ils s'appuient sur la mesure LRR ou la combinaison de LRR et C-value [Al Khatib & Badarneh 2010] et ignorent l'information contextuelle dans l'étape de filtrage statistique. Cependant, la phrase nominale de la langue arabe est sujette d'insertion de plusieurs types de modificateurs (adjectif, annexion, etc.). Donc, un terme candidat plus long (trigramme) peut être plus représentatif que ses variantes de longueur 2 (bigramme). Pour surmonter ces limitations, nous introduisons une nouvelle mesure d'association qui intègre l'information contextuelle pour bien quantifier l'apport d'insertion des modificateurs et combiner les degrés d'unité et de spécificité des termes candidats.

2.4 Méthode proposée pour l'extraction des termes complexes

Dans cette section, nous présentons notre méthode hybride proposée pour l'extraction des TCs [El Mahdaouy *et al.* 2013]. Cette méthode consiste à combiner le filtrage linguistique et le filtrage statistique pour l'extraction des termes complexes. Le filtre linguistique procède par l'identification des termes candidats en utilisant des patrons syntaxiques à partir d'un corpus étiqueté. Ce filtre traite également les variantes graphiques,

flexionnelles, morpho-syntaxiques et syntaxiques des termes candidats identifiés. Pour le filtrage statique, nous avons introduit une mesure d'association qui consiste à combiner le degré de spécificité [Frantzi *et al.* 2000], le degré d'unité [Dunning 1993] et l'information contextuelle [Frantzi *et al.* 2000].

2.4.1 Filtre linguistique

Notre filtre linguistique repose sur l'utilisation des patrons syntaxiques pour repérer les termes candidats à partir d'un texte étiqueté. Pour annoter le corpus, nous avons utilisé l'étiqueteur morpho-syntaxique AMIRA [Diab 2009]. Ce choix est motivé par la performance de cet étiqueteur (96,13% pour l'ensemble des étiquettes ERTS de PATB) et la richesse des étiquettes morpho-syntaxiques (article défini, genre, nombre, etc.). La Figure 2.1 présente le schéma global de notre filtre linguistique.

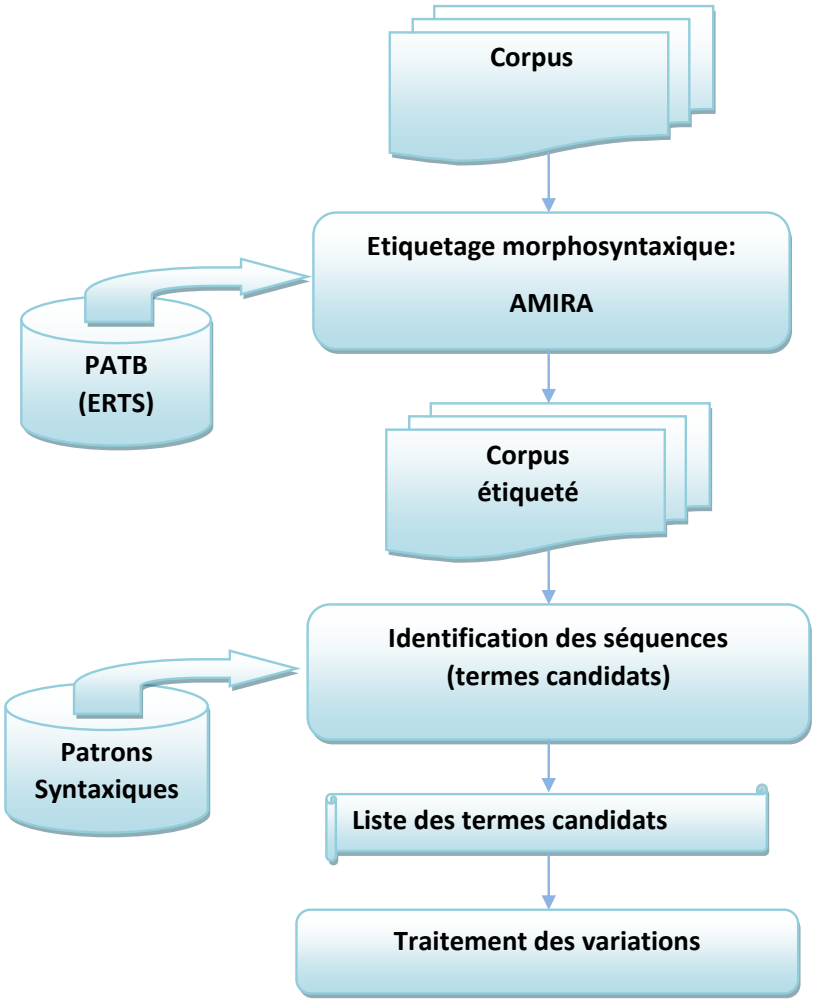


FIGURE 2.1: Schéma global du filtre linguistique

Dans un premier temps, le corpus est annoté en utilisant l'outil AMIRA pour associer

à chaque mot du corpus une étiquette morpho-syntaxique. Puis, les séquences des termes candidats sont repérées en utilisant un ensemble de patrons syntaxique. Nous avons également étendu la liste de patrons syntaxiques utilisés par Boulaknadel et al. [2008] pour l'acquisition des termes plus longs :

- $(Nom + (Nom|ADJ) + |(Nom|ADJ) + |(Nom|ADJ))$
- $Nom\ Prep\ Nom$

La deuxième étape de notre filtre linguistique consiste à identifier et traiter les variantes de chaque terme candidat. Dans cette étape, les variantes graphiques, flexionnelles, morpho-syntaxiques et syntaxiques sont traitées.

2.4.1.1 Variations des termes complexes

Variations graphiques

Le traitement de ce type de variations concerne la normalisation des variantes orthographiques des mots. Dans l'arabe écrite, quelques lettres (أ, آ , إ, ي, ئ, ة) subissent de légères modifications sans affecter le sens des mots. Le traitement de ces variations repose sur la normalisation orthographique des mots (voir Section 1.3.3.2). Par exemple, le terme "التلوث الكيميائي" (la pollution chimique) dont la structure $Nom\ ADJ$, le filtre linguistique a trouvé comme variante graphique "التلوث الكيميائى" ($Nom\ ADJ$) qui est le résultat de la substitution de la lettre ي par ى.

Variations flexionnelles

Ce type de variation des termes candidats est relié à la nature flexionnelle de la morphologie de la langue arabe (voir Section 1.3.2.4), où les mots possèdent différentes formes fléchies pour marquer des traits de genre, de nombre et l'article défini. Les flexions des noms s'attachent à la fin de la forme de base et l'article défini au début de celle-ci (nature agglutinante). Le traitement de ce type de variation est effectué par une suppression des flexions (liste des affixes de la méthode de racinisation légère light10 Tableau 1.3) guidée par les étiquettes morpho-syntaxiques des constituants du terme candidat (ces étiquettes marquent le genre, le nombre et l'article défini). L'ordre et les catégories grammaticales des constituants du terme candidat restent inchangés pour ce type de variation.

Par exemple, la variante du terme "تلوث المحيط" (pollution d'océan) dont la structure syntaxique NN DET_NN est "تلوث المحيطات" (pollution des océans) dont la structure NN DET_NNS_FP (où DET_NNS_FP désigne nom féminin pluriel défini). Les variantes de ce terme sont représentées par la forme non fléchie "تلوث محيط" (NN NN).

Variations morpho-syntaxiques

Ce type de variation est lié principalement à la morphologie dérivationnelle, où le prédicat du sujet de la phrase nominale peut être un nom, un adjectif ou une phrase propositionnelle (Section 1.3.2.5).

- $Nom_1\, Nom_2 \Leftrightarrow Noun_1\, Adj$: "تلوث الهواء" et "التلوث الهوائي" ("pollution de l'air").
- $Noun1\, Adj \Leftrightarrow Nom_1\, Prep\, Nom_2$: "برميل نفطي" et "برميل من النفط" ("baril de pétrole").

Ce type de variation est traité par la suppression de l'article défini, les prépositions et les suffixes de l'adjectif relationnel ية (féminin) et ي (masculin). La variante la plus fréquente du terme candidat dans le corpus est considérée comme étant la structure la plus représentative de ce terme.

Variations syntaxiques

Les variantes syntaxiques modifient la structure du terme candidat par l'insertion d'un nom, un adjectif,une phrase propositionnelle sans affecter les catégories grammaticales de leurs constituants (Section 1.3.2.5). Les adjectifs peuvent être insérés à l'intérieur de la structure interne du terme candidat ou par postposition. Cependant, les noms sont insérés par coordination de tête ou d'expansion. Le Tableau 2.1 illustre quelques exemples des variantes syntaxiques trouvées.

Tableau 2.1: Exemples de variantes syntaxiques

Type	sous-type	terme	Variante syntaxique
Modification	Insertion	النباتات السامة $Nom_1\ ADJ_1$ les plantes toxiques	النباتات البرية السامة $Nom_1\ ADJ_2\ ADJ_1$ les Plantes toxiques sauvages
	Postposition	تدوير النفايات $Nom_1\ Nom_2$ recyclage des déchets	تدوير النفايات الإلكترونية $Nom_1\ Nom_2\ ADJ$ recyclage des déchets éléctroniques
Coordination	Coordination d'expansion	تلوث الهواء $Nom_1\ Nom_2$ pollution de l'air	تلوث الهواء والماء $Nom_1\ Nom_2\ Prep\ Nom_3$ Pollution de l'air et de l'eau
	Coordination de tête	تدوير النفايات $Nom_1\ Nom_2$ Recyclage des déchets	تدوير وتثمين النفايات $Nom_1\ Prep\ Nom_3\ Nom_2$ Recyclage et valorisation des déchets

Ces variantes sont considérées comme des variantes imbriquées du terme candidat, voire des termes candidats lors du filtrage statistique. De plus, les termes insérés dans la structure du terme candidat de base sont considérés comme des contextes.

2.4.2 Filtre statistique

Le filtrage statistique consiste à ordonner les termes candidats selon des mesures statistiques leur attribuant des poids qui correspondent aux degrés d'unité et/ou de spécificité. De plus, ces mesures statistiques permettent de mesurer la stabilité d'un

terme candidat par rapport à ses variantes imbriquées. Notre filtre statistique consiste à combiner les deux mesures d'unité LLR [Dunning 1993] et de spécificité ainsi l'information contextuelle pour trier la liste des termes candidats (NC-value) [Frantzi *et al.* 2000].

2.4.2.1 Mesure C-value

La mesure d'association C-value permet d'affecter des scores aux termes candidats (degré de spécificité *termhood*) selon leurs fréquences, longueurs et leurs variantes imbriquées [Frantzi *et al.* 2000]. Le score d'un terme candidat est donné par l'équation 2.1 :

$$C\text{-Value}(a) = \begin{cases} \log_2(|a|) \cdot f(a) & \text{si } a \text{ est non imbriqué,} \\ \log_2(|a|) \cdot (f(a) - g(a)) & \text{sinon} \end{cases} \tag{2.1}$$

où $|a|$ est la longueur ou le nombre de constituants du terme candidat a, $f(a)$ est le nombre d'occurrences de a dans le corpus et :

$$g(a) = \frac{1}{|T_a|} \sum_{b \in T_a} f(b)$$

où $T(a)$ désigne l'ensemble des termes imbriqués où a est apparu ($|T(a)|$ est le cardinal de cet ensemble).

D'après l'équation 2.1 on peut noter que si le terme candidat n'est pas imbriqué, son score est uniquement calculé à base de son nombre d'occurrences et de sa longueur. S'il est imbriqué, son nombre d'occurrences est corrigé par le nombre d'occurrences des termes dans lesquels il apparait (fréquence des variantes imbriquées du terme candidat). L'idée sous-jacente est que les termes candidats qui tendent d'apparaitre le plus souvent dans des termes imbriqués, ne forment pas probablement un terme. Un terme imbriqué qui apparait dans plusieurs séquences plus longues, il s'agit probablement d'un terme. De plus, la fréquence d'un terme est normalisée par sa longueur, car les termes plus longs apparient moins fréquemment que les autres de petites longueurs.

2.4.2.2 Mesure NC-value

La mesure NC-value, proposé aussi par Frantzi et al. [2000] , consiste à combiner l'information contextuelle et la mesure C-Value. L'information contextuelle ou le poids des mots apparus en tant que contexte (ou mots collocatifs) d'un terme candidat est calculé par la mesure Nvalue :

$$N\text{value (a)} = \sum_{b \in C_a} f_a(b) \cdot \frac{|T(b)|}{n} \tag{2.2}$$

où C_a désigne l'ensemble des mots distincts apparus en tant que contexte de a, $f_a(b)$ représente le nombre d'occurrences où le mot b apparu en tant que contexte de a et n le nombre total des termes candidats. L'idée sous-jacente est de mesurer l'importance ou le

degré de relation d'un mot de contexte au terme candidat. La mesure NC-value consiste à combiner Nvalue avec C-Value :

$$NC - \text{value(a)} = 0.8 \cdot C - \text{value(a)} + 0.2 \cdot N\text{value(a)} \tag{2.3}$$

2.4.2.3 Mesure NTC-value

Le but principal de la mesure NTC-value [Matthijs & Radlinski 2011] est d'intégré le degré d'unité T-score dans la mesure NC-value. La mesure T-score permet de mesurer le degré de stabilité ou d'unité d'une combinaison syntaxique par rapport aux nombres d'occurrences des constituants de cette combinaison dans le corpus :

$$Ts(w_i, w_j) = \frac{p(w_i, w_j) - p(w_i) \cdot p(w_j)}{\sqrt{\frac{p(w_i, w_j)}{N}}} \tag{2.4}$$

où $p(w_i, w_j)$ désigne la probabilité d'observer le bigramme w_i, w_j dans le corpus ; $p(w_i)$ est la probabilité d'observer w_i dans le corpus et correspond à la probabilité marginale $p(w_i, w)$. La mesure T-score est intégrée dans les mesures C-value et NC-value par repondération de la fréquence du terme candidat lorsque son degré d'unité (Ts) est positif :

$$F(a) = \begin{cases} f(a) & \text{si } \min(Ts(w_i, w_{i+1})) \leq 0 \text{ avec } 0 \leq i < |a| \\ f(a) \ln(2 + \min(Ts(w_i, w_{i+1})) & \text{sinon} \end{cases} \tag{2.5}$$

où $\min(Ts(w_i, w_{i+1}))$ correspond à la valeur minimale de T-score des bigrammes w_i w_{i+1} de a. La substitution de $f(a)$ par $F(a)$ dans l'équation 2.1 permet d'obtenir TC-value, qui sera combiné par la suite avec Nvalue afin d'obtenir la mesure d'association NTC-value :

$$NTC\text{-value(a)} = 0.8 \cdot TC\text{value(a)} + 0.2 \cdot N\text{value(a)} \tag{2.6}$$

Ceci permet de prendre en considération les degrés d'unité et de spécificité ainsi l'information contextuelle dans la mesure résultante (Équation 2.6).

2.4.2.4 Mesure NLC-value

Pour développer cette mesure d'association, nous avons suivi la même méthode utilisée par Matthijs & Radlinski [2011] pour combiner les différentes mesures statistiques (NC-value et T-score). En effet, notre mesure s'appuie sur l'utilisation de la mesure du degré d'unité LLR [Dunning 1993] pour mesurer l'importance d'une combinaison syntaxique. La mesure sous-jacente est calculée pour les bigrammes par la formule suivante :

$$\begin{aligned} LLR(w_j, w_j) &= a \log(a) + b \log(b) + c \log(c) \\ &+ d \log(d) - (a + b) \log(a + b) \\ &- (a + c) \log(a + c) - (b + d) \log(b + d) \\ &- (c + d) \log(c + d) + N \log(N) \end{aligned}$$

avec :
a : nombre de termes où w_i et w_j sont apparus ;
b : nombre de termes où w_i est apparu avec d'autres mots que w_j ;
c : nombre de termes où w_j est apparu avec d'autres mots que w_i ;
d : Nombre de termes dans lesquels ni w_i ni w_j sont apparus ;
N : le nombre total de termes extraits.

Pour les termes les plus longs, la mesure LLR est calculée à base de la valeur minimale des bigrammes du terme candidat. Par la suite le nombre d'occurrences d'un terme candidat est pondéré par la valeur minimale de LLR : $FL(a) = f(a) \cdot \ln(2 + \min(LLR(w_i, w_{i+1})))$ avec $0 \leq i < |a|$. En remplaçant $f(a)$ par $FL(a)$ dans la mesure C-value (Équation 2.1), nous obtenons la mesure LC-value qui combine le degré d'unité LLR et le degré de spécificité C-value :

$$L\text{C-value(a)} = \begin{cases} \log_2(|a|) \cdot FL(a) & \text{si } a \text{ est non imbriqué,} \\ \log_2(|a|) \cdot (FL(a) - GL(a)) & \text{sinon} \end{cases} \tag{2.7}$$

avec $GL(a) = \frac{1}{|T_a|} \sum_{b \in T_a} FL(b)$

Cette mesure est combinée par la suite avec la mesure Nvalue pour la prise en considération de l'information contextuelle. La mesure résultante, appelée NLC-value, consiste à combiner l'information sur le contexte, le degré d'unité (*unithood*) et le degré de spécificité (*termhood*) :

$$NLC\text{-value(a)} = 0.8 \cdot LC\text{-value(a)} + 0.2 \cdot N\text{value(a)} \tag{2.8}$$

2.5 Expérimentations et résultats

2.5.1 Corpus d'évaluation

À notre connaissance, il n'y'a pas un corpus standard d'un domaine spécifique pour l'évaluation des méthodes d'extraction de terminologie pour la langue arabe. En effet, la plupart des chercheurs ont construit leurs propres corpus afin d'évaluer leurs méthodes. Pour cela, nous avons collecté un corpus du domaine d'environnement dont les propriétés similaires aux corpus déjà utilisés dans l'état de l'art d'extraction des termes complexes de la langue arabe [Boulaknadel 2008, Bounhas & Slimani 2009, Al Khatib & Badarneh 2010].

Notre corpus d'évaluation contient 1.666 documents contenant 53.569 mots distingues qui sont extraits du site web "Al-Khat Alakhdar" [2]. Ce corpus couvre divers sujets environnementaux tels que la pollution, les effets du bruit, la purification de l'eau, la dégradation du sol, la préservation des forêts, les changements climatiques et les catastrophes naturelles.

2. http ://www.greenline.com.kw

2.5.2 Méthode d'évaluation et résultats

Généralement, l'évaluation des méthodes d'extraction de terminologie repose sur l'utilisation d'une liste de référence des termes d'un domaine. Lorsqu'aucune liste de références n'est disponible pour le domaine ou la langue retenue, on peut d'abord traduire les termes candidats (en utilisant un système de traduction automatique ou un dictionnaire bilingue) et utiliser une liste de références disponible pour une autre langue.

Pour évaluer notre méthode, nous avons constitué automatiquement une liste de référence de tous les termes candidats de la langue arabe qui sont disponibles dans la dernière version du thésaurus AGROVOC [3] et ensuite comparer les termes candidats retenus avec les termes de cette liste. Dans le cas où le terme candidat n'est pas trouvé, nous procédons par une recherche d'une variante de la même longueur du terme candidat dans la liste de référence. Pour les termes candidats non retrouvés, nous traduisons et considérons ces termes comme pertinents si leurs traductions sont contenues dans la base terminologique européenne IATE [4]. Enfin, la précision est calculée en utilisant le nombre de termes candidats attestés et le nombre de termes considérés. En effet, nous avons utilisé plusieurs mesures statistiques pour classer les termes candidats (LLR, C-value, NC-value, NTC-value, LLR + C-value, NLC-value) et retenir de chaque classement produit les k-premiers candidats, avec k allant de $100 - 300$ à des intervalles de 100. Le Tableau 2.2 présente les résultats obtenus pour chaque mesure d'association.

Tableau 2.2: Résultats obtenus en termes de précision pour les différentes mesures statistiques

	Top termes candidats considérés		
mesure statistique	**P@100**	**P@200**	**P@300**
LLR	75,0%	70,5%	64,3%
C-value	71,0%	69,0%	65,0%
NC-value	74,0%	70,0%	68,3%
NTC-value	80,0%	71,5%	69,7%
LLR + C-value	73,0%	72,0%	68,3%
NLC-value	82,0%	75,5%	73,0%

Les résultats expérimentaux illustrés dans le tableau 2.2 montrent que notre méthode NLC-value assure une meilleure précision par rapport aux autres mesures d'association. De plus, l'intégration de l'information contextuelle dans la mesure C-value (mesure statistique NC-value) améliore la performance d'extraction des TCs. En concordance avec les études citées précédemment dans le contexte d'extraction de terminologie de l'arabe, le degré d'unité (mesure LLR) donne une meilleure performance par rapport aux mesures C/NC-value lorsque le nombre des TCs retenus n'est pas assez grand (voir Figure 2.2). D'où la nécessité de combiner le degré d'unité, le degré de spécificité et

3. www.fao.org/agrovoc/
4. http ://iate.europa.eu/iatediff

l'information contextuelle pour augmenter la précision du filtrage statistique. En effet, la combinaison de ces informations a montré que l'intégration de T-score dans les mesures C/NC-value (mesure NTC-value) assure une meilleure performance par rapport à ces deux dernières mesures (voir Figure 2.3). Enfin, la combinaison des mesures d'unité et

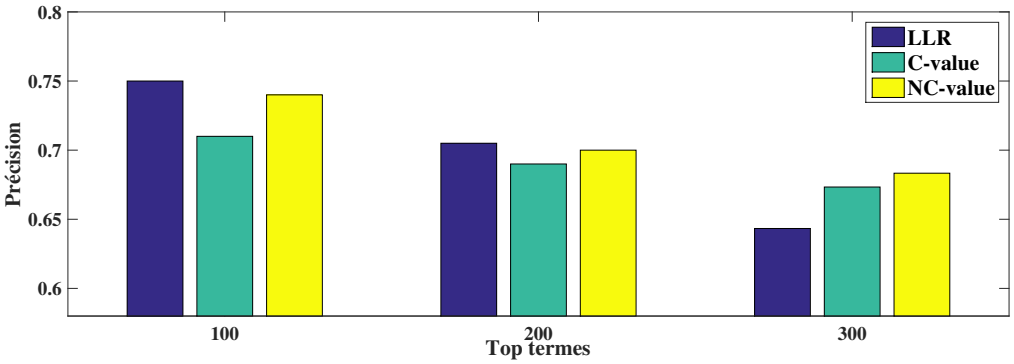


FIGURE 2.2: Precision obtained for the LLR and the C/NC-value

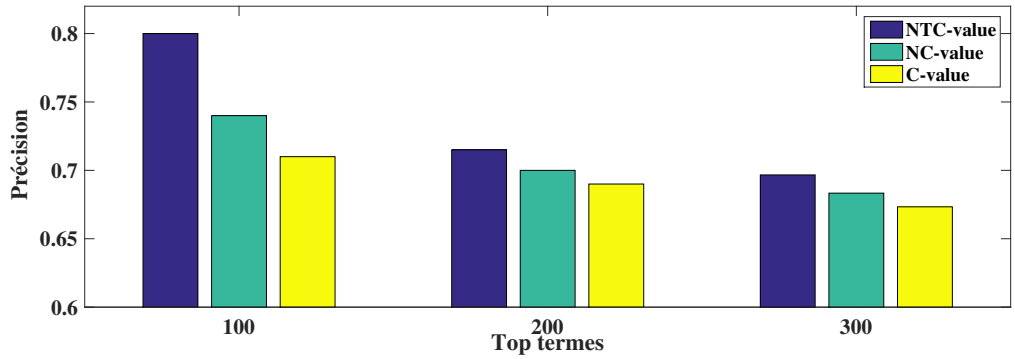


FIGURE 2.3: Precision obtained for the C/NC-value and the NTC-value

de spécificité (NTC-value, $LLR + C$-value, NLC-value) est essentielle pour assurer une meilleure performance, puisque toutes les mesures basées sur ces combinaisons ont obtenu une meilleure performance par rapport à celles qui sont basées sur le degré d'unité ou le degré de spécificité seul (C-value, NC-value, LLR). Nous notons que la mesure statistique que nous proposons, NLC-value, donne une meilleure performance par rapport aux autres combinaisons. La mesure NLC-value prend les avantages des travaux précédents proposés dans [Matthijs & Radlinski 2011] et [Al Khatib & Badarneh 2010] pour la prise en compte de l'information contextuelle et la mesure d'unité LLR et de spécificité C/NC-value. La Figure 2.4 présente une comparaison de la précision obtenue pour ces différentes mesures statistiques. Le nombre total des termes candidats distincts évalués pour les six mesures statistiques est 1.095 de 1800 termes. Les différentes mesures statistiques partagent une liste de 141 termes candidats. Tableau 2.3 et Tableau 2.4 représentent le nombre de termes retrouvé dans le thésaurus AGROVOC et la base IATE respectivement.

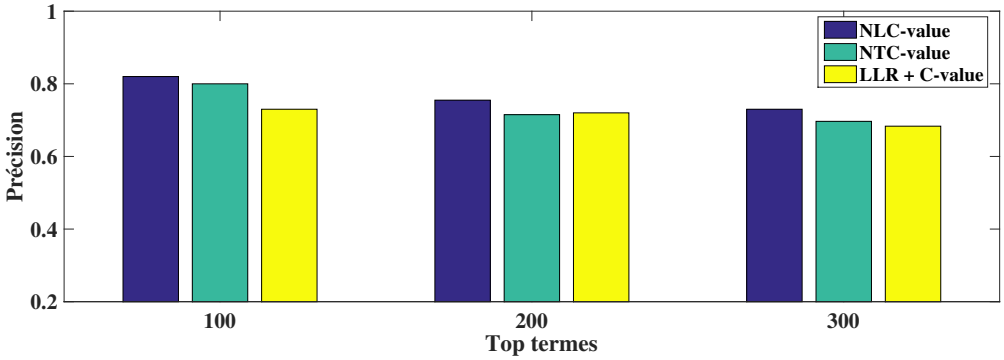


FIGURE 2.4: Performances obtenues pour les mesures statistiques qui combinent le degré d'unité, de spécificité et l'information contextuelle

Tableau 2.3: Nombre total de termes candidats trouvés dans AGROVOC pour chaque mesure statistique

	Top termes candidats considérés		
Mesure statistique	**100**	**200**	**300**
LLR	35	60	80
C-value	27	59	82
NC-value	32	62	82
NTC-value	35	60	83
LLR + C-value	34	60	84
NLC-value	41	65	86

Tableau 2.4: Nombre total de termes candidats trouvé dans la base IATE pour chaque mesure statistique

	Top termes candidats considérés		
Mesure statistique	**100**	**200**	**300**
LLR	40	81	113
C-value	44	79	113
NC-value	42	78	123
NTC-value	45	83	126
LLR + C-value	39	84	121
NLC-value	41	86	133

2.6 Conclusion

Dans ce chapitre, nous avons passé en revu les différentes approches proposées pour l'extraction des termes complexes : approche linguistique, approche statistique et l'approche hybride ou mixte. Nous avons aussi décrit quelque outil d'extraction de terminologie selon l'approche adoptée. Puis, nous avons présenté les méthodes proposées pour la langue arabe. Ensuite, nous avons décrit notre méthode hybride introduite pour l'extraction des termes

complexes de la langue arabe. En effet, notre méthode repose sur un filtre linguistique permettant d'extraire les termes candidats, dont la structure syntaxique respecte, un ensemble de patrons syntaxiques pré-établis. Par la suite, ce filtre traite les variantes graphiques, flexionnelles, morpho-syntaxiques et syntaxiques de ces termes candidats. Notre filtre statistique repose principalement sur la mesure NLC-value, que nous avons proposée afin de combiner le degré d'unité LLR, le degré de spécificité C-value et l'information contextuelle N-value, pour extraire les termes pertinents. Enfin, nous avons présenté notre méthode d'évaluation, où nous avons comparé plusieurs mesures statistiques ainsi leurs combinaisons pour montrer l'intérêt de combiner ces mesures. Les résultats expérimentaux ont montré que notre mesure statistique donne une meilleure performance par rapport aux autres mesures.

Dans le chapitre suivant, nous nous intéressons à l'intégration de ces termes complexes dans la représentation (indexation) et l'appariement des documents et des requêtes pour la RI en langue arabe.

Chapitre 3

Apport des dépendances explicites et implicites pour la RI en langue arabe

Sommaire

3.1 Introduction

Dans le chapitre précédent, nous avons introduit notre méthode hybride d'extraction des termes complexes. L'idée sous-jacente repose sur l'utilisation d'un filtre linguistique pour la prise en compte principalement des propriétés morphologiques et syntaxiques de la langue arabe ainsi qu'un filtre statistique pour extraire les termes complexes pertinents. Ces termes complexes sont moins ambigus et moins polysémiques que les termes simples isolés. De plus, plusieurs concepts sont représentés par des séquences de termes (principalement des termes complexes). Ainsi, ce chapitre étudie les apports d'intégration de ces termes complexes dans les modèles probabilistes de RI, dans l'optique d'améliorer le processus d'appariement des documents et des requêtes en langue arabe.

Les modèles traditionnels de RI sont basés sur l'hypothèse d'indépendance de termes et adoptent la représentation en sac de mots pour représenter le contenu des documents et des

requêtes. Par conséquent, l'estimation du score de pertinence d'un document par rapport à une requête repose sur un appariement exact des termes simples partagés entre eux (Section 1.2.3 du Chapitre 1). Cependant, la représentation par des termes simples n'est pas assez précise et ne véhicule que relativement le contenu sémantique des documents et des requêtes. En effet, cette représentation ne prend en considération ni les associations entre les termes ni leur ordre d'apparition dans les documents et les requêtes. Afin de remédier à ces limitations en RI, plusieurs méthodes et approches ont été proposées pour aller au-delà de la représentation en sac de mots (Section 1.2.2.2, Section 1.2.2.3 et Section 1.2.4 du Chapitre 1). Dans ce chapitre, nous nous intéressons particulièrement à l'intégration des dépendances entre les termes, à savoir les termes complexes et les dépendances implicites (modèles de proximité) pour la RI en langue arabe.

Le reste de ce chapitre est organisé comme suit : Section 3.2 présente le cadre dans lequel s'inscrit notre étude, où nous introduisons la problématique et les motivations de notre travail. La Section 3.3 décrit notre méthode d'indexation des termes complexes. L'intégration des dépendances explicites à base des termes complexes et des dépendances implicites est présentée dans la Section 3.4. La Section 3.5 introduit une contrainte sur l'intégration des dépendances de termes dans les modèles de RI. La méthode d'évaluation et les résultats obtenus sont décrits dans la Section 3.6. Ce chapitre se termine par une discussion des résultats obtenus (Section 3.7) et une brève conclusion. (Section 3.8).

3.2 Problématique et motivations

Face à la morphologie riche et complexe, la plupart des travaux proposés pour la RI en langue arabe ont mis l'accent sur l'élaboration et l'évaluation des techniques de racinisation [Abu El-Khair 2007, Mustafa *et al.* 2008, Darwish & Magdy 2014]. L'appariement des documents et des requêtes est, donc, effectué à base des racines ou des *stems* partagés entre eux. Ces travaux peuvent être classés selon le niveau d'analyse des mots : approche basée-racine (racinisation) [Khoja & Garside 1999] et approche basée-*stem* (racinisation légère) [Larkey *et al.* 2002]. Malgré que les travaux pionniers ont montré que les techniques de racinisation sont plus efficaces pour traiter la morphologie de l'arabe dans le contexte de la RI [Al-Kharashi & Evens 1994, Hmeidi *et al.* 1997, Abu-Salem *et al.* 1999], les travaux récents ont montré l'efficacité des techniques de racinisation légère [Larkey *et al.* 2002, Larkey *et al.* 2007, Goweder *et al.* 2004, Abdelali *et al.* 2016]. L'inconvénient majeur des techniques de racinisation réside dans le fait de regrouper des mots sémantiquement différents dans la même racine, car chaque racine peut générer, par application des schèmes morphologiques, une centaine des mots distincts [Beesley 1996]. La plupart des techniques de racinisation légère, en revanche, ne permettent pas de discriminer les conjonctions et les prépositions de forme de base des mots [Nwesri *et al.* 2005, Darwish & Mubarak 2016]. De plus, elles ne traitent pas le pluriel irrégulier des mots. Par conséquent, les techniques de racinisation légère peuvent regrouper des mots sémantiquement similaires à des *stems* distincts. De toute évidence, les deux approches de racinisation introduisent

des ambigüités au niveau de la représentation des textes [El Mahdaouy *et al.* 2014]. En outre, d'autres niveaux d'ambigüité présentent des défis importants pour les applications de TAL de l'arabe [Maamouri & Bies 2010]. En particulier, l'absence de représentation des diacritiques (voyelles courtes) dans les textes augmente considérablement le nombre d'ambigüités. [Farghaly 2004] a souligné que, pour la plupart des langues, le taux moyen des ambigüités pour un mot est d'ordre 2.3, alors qu'il atteint 19.2 pour l'ASM (Arabe Standard Moderne).

Bien que la RI en arabe ait connu des progrès tangibles, l'intégration de dépendances entre les termes (termes complexes et dépendances de proximité) demeure toutefois largement sous-exploré. À notre connaissance, il n'y a qu'un seul travail qui a étudié la RI en langue arabe à base de termes complexes [Boulaknadel *et al.* 2008a], où l'évaluation a été effectuée à l'aide d'un petit corpus du domaine de l'environnement (1062 documents contenant 475148 mots). Dans ce chapitre, nous étudions le problème d'indexation et de recherche de documents à base des termes complexes en utilisant une grande collection standard pour la RI en l'arabe. Ces termes complexes sont extraits en utilisant notre méthode hybride qui combine un filtre linguistique complexe pour la prise en compte de leurs propriétés morphologiques et syntaxiques ainsi qu'un filtre statistique plus élaboré qui consiste à combiner l'information contextuelle avec le degré de spécificité et le d'unité [El Mahdaouy *et al.* 2013]. En outre, nous explorons un large éventail de modèles de proximité pour la RI en langue arabe, en utilisant trois algorithmes de racinisation, introduit respectivement par [Khoja & Garside 1999, Larkey *et al.* 2007, Abdelali *et al.* 2016]. Notre objectif est d'évaluer l'impact de la prise en compte des dépendances (proximité) entre les termes de la requête sur la performance de RI. Pour ce faire, nous comparons les différents niveaux d'analyse morphologique des mots pour les termes complexes (dépendances explicites) et les dépendances de proximité (dépendances implicites), afin d'aller au-delà de la représentation en sac de mots pour la RI en langue arabe [El Mahdaouy *et al.* 2018c]. Les questions que nous abordons sont les suivantes :

- les modèles de proximité et l'utilisation des termes complexes, peuvent-ils améliorer la performance de RI, lorsqu'on utilise différents niveaux d'analyse morphologique des mots ?
- les termes complexes qui sont extraits en utilisant un pipeline complexe (filtrage linguistique et statistique), peuvent-ils améliorer significativement la performance par rapport aux modèles de proximité ?

En plus de ces points, à notre connaissance, cette étude est la première qui fournit (a) une extension complète des termes croisés pour les modèles standards de RI, termes proches dans la requête et dont les fonctions de densité chevauchent [Zhao *et al.* 2011], (b) une comparaison complète des modèles de RI les plus importants en intégrant les dépendances des termes (18 modèles sont comparés dans nos évaluations), dans le contexte de RI en langue arabe, et (c) une contrainte heuristique qui permet de caractériser les différents modèles pour l'intégration des dépendances.

Cette étude est principalement axée sur les apports morphologiques et syntaxiques en

RI pour plusieurs raisons : (a) l'arabe est caractérisée par sa morphologie riche et il n'y'a pas de consensus sur la meilleure technique de racinisation pour la RI ; nous abordons ce problème en effectuant une comparaison étendue de différentes approches de racinisation et racinisation légère, y compris la technique récente de racinsation légère Farasa, couplées avec cinq modèles de RI ; (b) la langue arabe est riche en termes complexes, où leur production repose sur les deux compositions *Nom Nom* et *Nom préposition Nom* ; (c) contrairement aux autres langues telles que l'anglais et le français, à notre connaissance, il n'y'a pas d'étude complète consacrée à l'évaluation de l'impact des termes complexes et les dépendances de proximité pour la RI en langue arabe ; le but principal de cette étude est précisément d'évaluer cet impact.

3.3 Méthode d'indexation des termes complexes

Pour indexer les documents à base des termes simples et composés, nous avons utilisé trois modules. La Figure 3.1 présente le processus d'indexation. Dans un premier temps, le premier module d'indexation consiste à annoter la collection des documents en utilisant l'étiqueteur morpho-syntaxique AMIRA 2.0 [Diab 2009]. Puis, nous utilisons un deuxième module qui permet d'extraire la liste globale des termes complexes de la collection. Ce module fait appel à notre méthode d'acquisition des termes complexes, présentée dans le chapitre précédent. Ensuite, le troisième module porte sur l'extraction des descripteurs d'index de chaque document de la collection en utilisant un pipeline pour les termes simple et un autre pour les termes complexes. Le premier pipeline effectue la tokenisation des mots du document, la normalisation de leurs formes graphiques et le traitement des leurs variations morphologiques en utilisant une technique de racinisation ou de racinisation légère. Le deuxième pipeline, en revanche, fait appel au filtre linguistique pour extraire la liste des termes complexes qui correspondent aux patrons syntaxiques prédéfinis. En plus du traitement des variations des termes complexes, ce pipeline sélectionne également ceux qui ont une valeur de $NLC-$value dans la liste des termes complexes de la collection supérieure à un seuil afin de ne considérer que les termes pertinents. Ce seuil est varié entre 0 et 30 et fixé expérimentalement à 5 à base de la meilleure valeur de mesure MAP (*Mean Average Precision*). Enfin, chaque document de la collection est indexé en utilisant les deux listes de termes simples et complexes issues du troisième module.

3.4 Intégration des dépendances de termes

Dans cette section, nous passons en revue des extensions des modèles de base de RI utilisés pour la prise en compte des dépendances explicites (termes complexes) et des dépendances implicites (de proximité).

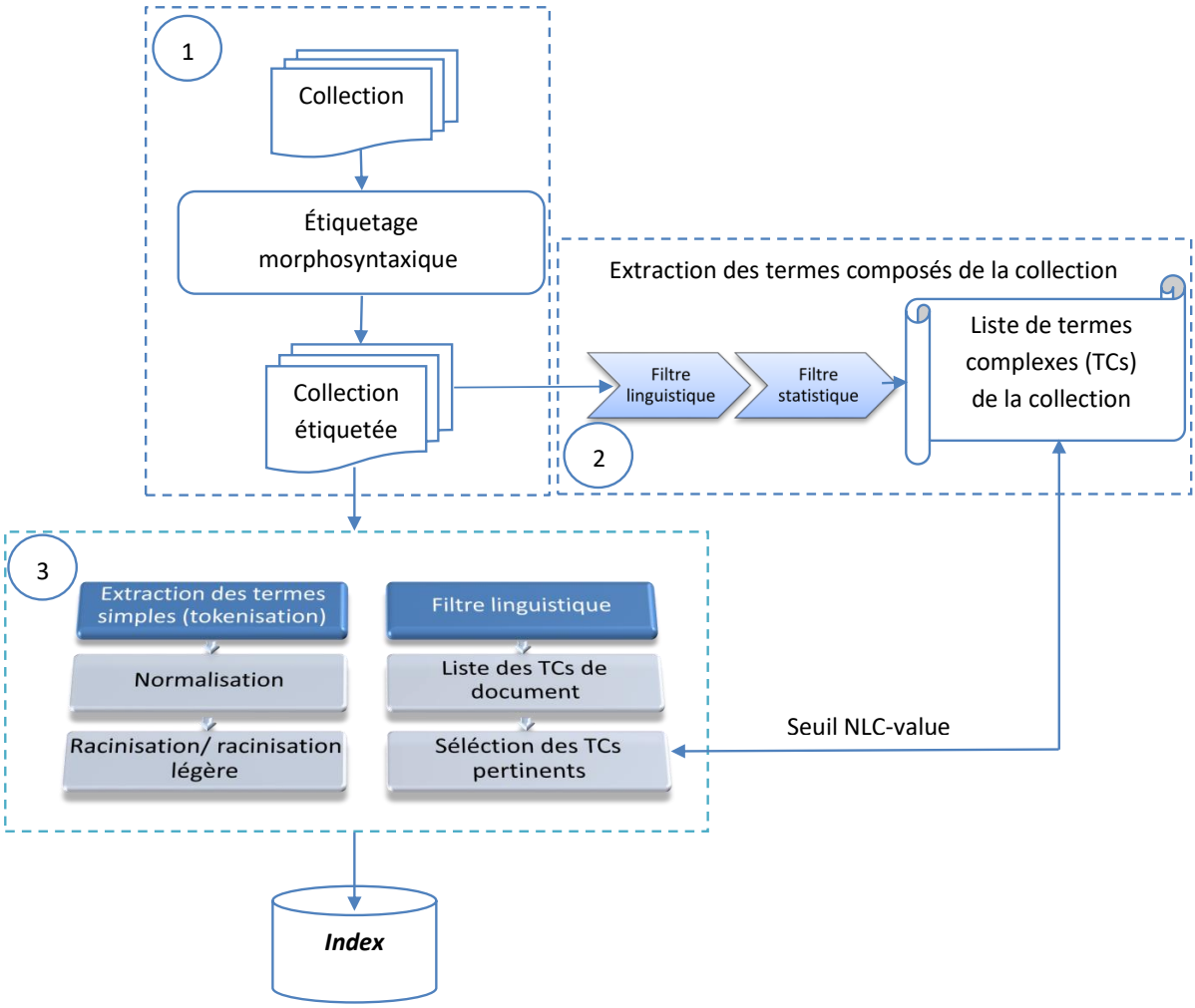


FIGURE 3.1: Processus d'indexation des termes simples et complexes

3.4.1 Extension pour les termes complexes

L'extraction et l'indexation des termes complexes conduisent à une nouvelle représentation des documents et des requêtes. En effet, en plus de la représentation par des termes simples (représentation en sac de mots), on peut également utiliser de la même façon la représentation par des termes complexes (sac de termes complexes). Par conséquent, chaque requête et chaque document peuvent être représentés par un ensemble de termes simples et un autre ensemble de termes complexes, notés respectivement st et mwt : $q = \{q_{st}, q_{mwt}\}$. En outre, Il est possible de reposer sur les mêmes statistiques (fréquence de termes et la fréquence documentaire inverse), utilisées pour pondérer les termes simples, pour les termes complexes.

Dans cette optique, une intégration directe des termes complexes dans les modèles de RI s'effectue tout simplement par une combinaison linéaire des contributions des deux

représentations (termes simples et termes complexes) :

$$RSV(q,d) = (1-\lambda) \cdot RSV(q_{st}, d_{st}) + \lambda \cdot RSV(q_{mwt}, d_{mwt}) \tag{3.1}$$

où le paramètre λ contrôle l'influence de chaque représentation. l'Équation 3.1 représente le score obtenu pour un document d par rapport à une requête q selon les deux représentations utilisées. Cette approche a été utilisée pour intégrer des dépendances de termes telles que les phrases et les dépendances de proximité dans les modèles de RI [Shi & Nie 2009, Metzler & Croft 2005, Zhao *et al.* 2011].

Tous les modèles de base que nous avons présentés dans la Section 1.2.3.3 du Chapitre 1 (BM25, LM, PL2, LGD et SPL) peuvent être étendus directement par l'Équation 3.1, menant aux modèles BM25_MWT, LM_MWT, PL2_MWT, LGD_MWT et SPL_MWT.

3.4.2 Modèle CRTER

Le modèle CRTER (*CRoss TErm Retrieval*) ou modèle des termes croisés, proposé par Zhao et al. [2011], consiste à introduire des dépendances de termes, appelés termes croisés CT (*Cross Terms*), pour modéliser la proximité des termes des documents et améliorer la performance de RI. L'idée derrière ce modèle est que l'occurrence d'un terme de requête a un impact sur ses termes voisins qui diminue progressivement avec l'augmentation de la distance entre eux. Un terme croisé se produit lorsque deux termes de requête apparaissent à proximité et que leurs fonctions de densité ont une intersection. Pour faciliter l'intégration de ces nouveaux termes dans les modèles de RI, ils ont défini (1) la fréquence dans le document (2), la fréquence inverse des documents et (3) la fréquence dans la requête des termes croisés.

(1) La fréquence $x^d_{p_{i,j}}$ d'un CT dans un document d est l'accumulation des valeurs des différences des positions de ces constituants : $x^d_{p_{i,j}} = \sum_{k_1=1}^{x^d_{w_i}} \sum_{k_2=1}^{x^d_{w_j}} \text{Kernel}(\frac{1}{2}|pos_{k_1,i} - pos_{k_2,j}|)$ où $Kernel$ est une fonction de densité (*Kernel Function*).

(2) La fréquence inverse des documents ($N_{p_{i,j}}$) repose sur le nombre de documents dans lesquels le CT est apparu ($x^d_{p_{i,j}} \neq 0$) :

$$N_{p_{i,j}} = \sum_{d \in index} \mathbf{1}_{x^d_{p_{i,j}} \neq 0} \tag{3.2}$$

(3)La fréquence du CT dans la requête est obtenue en supposant que les termes de la requête sont adjacents et en considérant toutes les paires possibles formées de termes de la requête :

$$x^q_{p_{i,j}} = Kernel(\frac{1}{2}) \cdot \min(x^q_{w_i}, x^q_{w_j}) \tag{3.3}$$

Plusieurs fonctions de densité ont été utilisées :

— Le noyau gaussien : $\text{Kernel}(u) = \exp(\frac{-u^2}{2\sigma^2})$
— Le noyau triangulaire : $\text{Kernel}(u) = (1 - \frac{u}{\sigma}) \cdot \mathbf{1}_{u \leq \sigma}$
— Le noyau circulaire : $\text{Kernel}(u) = \sqrt{1 - \frac{u}{\sigma}^2} \cdot \mathbf{1}_{u \leq \sigma}$

— Le noyau cosinus : $\text{Kernel}(u) = \frac{1}{2}[1 + \cos(\frac{u\pi}{\sigma})] \cdot \mathbf{1}_{u \leq \sigma}$

où μ est la distance entre deux termes de la requête et σ est un paramètre à optimiser, qui contrôle la propagation des courbes du noyau. La fonction d'appariement est donnée par (Équation 3.4) :

$$CRTER(d, q) = (1 - \lambda) \sum_{w \in q \cap d} \omega(x_w, d) + \lambda \sum_{1 \leq i \leq j \leq K} \omega(x_{p_{i,j}}, d) \tag{3.4}$$

où le paramètre λ contrôle l'influence des termes simples et des termes croisés et ω la fonction d'appariement du modèle BM25 [Robertson *et al.* 1994]. Elle peut cependant être remplacée par n'importe quelle fonction d'appariement des modèles de base (présentés dans la Section 1.2.3.3 du Chapitre 1).

3.4.3 Extensions spécifiques au modèle de langue

Dans cette Section, nous allons passer en revue des extensions spécifique à la famille de modèle de langue.

3.4.3.1 Modèle MRF

Le modèle MRF (*Markov Random Field*) [Metzler & Croft 2005] est une généralisation du modèle de langue de base pour la prise en compte des dépendances de termes, via le formalisme de champs aléatoire de Markov. Le modèle considère trois niveaux de dépendance entre les termes : (1) indépendance totale des termes (FI), basée sur l'appariement des termes simples, équivalent au modèle de langue de base ; (2) dépendance séquentielle (SD) pour la prise en compte des phrases ordonnées dans l'appariement ; (3) dépendance totale (FD) repose sur les occurrences des phrases non ordonnées. Ce modèle vise à construire un graphe G à partir des termes de la requête et d'un document d. Les différentes configurations possibles permettent de considérer différentes hypothèses de dépendances. Le score de chaque document est estimé en utilisant la distribution jointe sur l'ensemble des variables de G par le biais des fonctions potentielles sur l'ensemble de configurations de cliques associées aux différents types de dépendances (termes simples (FI), phrases ordonnées (SD) et phrase non-ordonnées (FD)). La fonction d'appariement est donnée par l'Équation 3.5 :

$$RSV(d, q) = \sum_{c \in T} \lambda_T f_T(c) + \sum_{c \in O} \lambda_O f_O(c) + \sum_{c \in O \cup U} \lambda_U f_U(c) \tag{3.5}$$

où λ_T est le poids des termes simples, λ_O est le poids des phrases ordonnées et λ_U est le poids des phrases non-ordonnées. T est défini comme étant un ensemble de 2-cliques impliquant un terme de la requête et un document d, O est un ensemble de cliques contenant le noeud du document et au moins deux termes de la requête qui sont apparus de façon contigüe, et U est l'ensemble de cliques contenant le noeud de document et au moins deux termes de la requête qui sont apparus de façon non-contigüe. Les fonctions potentielles,

associées a chaque type ou niveau de dépendance, sont estimées par la méthode de lissage de Dirichlet [Zhai & Lafferty 2001a]. Pour les termes simples w, la fonction potentielle est donnée par $f_T(c=(w;d)) = \log[(1-\alpha_d)\frac{x_w^d}{l_d} + \alpha_d \frac{x_w^c}{|C|}]$ où α_d est le paramètre de lissage. Les fonctions potentielles $f_O()$ et $f_U()$ sont obtenus par la généralisation de $f_T()$ pour l'intégration des dépendances SD et FD dans la fonction d'appariement. Pour les phrases ordonnées ou encore les dépendances séquentielles (SD), la fonction potentielle est donnée par la formule suivante :

$$f_O(c=({w_i}^q,...,{w_{i+k}}^q;d)) = \log[(1-\alpha_d)\frac{x^d_{(w_i,...,w_{i+k})}}{l_d} + \alpha_d \frac{x^C_{w_i,...,w_{i+k})}}{|C|}] \tag{3.6}$$

où $({w_i}^q,...,{w_{i+k}}^q)$ est une phrase ordonnée de la requête. $x^d_{(w_i,...,w_{i+k})}$ et $x^C_{(w_i,...,w_{i+k})}$ sont les nombres d'occurrences de $({w_i}^q,...,w_{i+k})$ dans un document d et la collection C respectivement. Pour les dépendances FD (phrases non-contigües), la fonction potentielle est donnée par :

$$f_U(c=({w_i}^q,...,{w_j}^q;d)) = \log[(1-\alpha_d)\frac{x^d_{N(w_i,...,w_j)}}{l_d} + \alpha_d \frac{x^C_{N(w_i,...,w_j)}}{|C|}] \tag{3.7}$$

où $({w_i}^q,...,{w_j}^q)$ est une phrase non contigüe de la requête. $x^d_{N(w_i,...,w_j)}$ et $x^C_{N(w_i,...,w_j)}$ sont les nombres d'occurrences de $({w_i}^q,...,{w_j}^q)$, apparait ordonné ou non dans une fenêtre de longueur fixe N dans un document d et la collection C respectivement.

3.4.3.2 Modèle PLM

Le modèle PLM (*Positional Language Model*) a été introduit par Lv & Zhai [2009b], dans le but d'unifier la proximité de termes et les heuristiques de passage au sein du même modèle. L'idée de base de ce modèle consiste à estimer un modèle de langue pour chaque position d'un document et classer ce document à base des scores obtenus pour chaque position. Un document virtuel est construit pour chaque position, où l'importance du terme augmente lorsque ce terme apparait proche de cette position. Formellement, le modèle de langue à la position i d'un document d est donné par :

$$p(w|d,i) = \frac{c'(w,i)}{\sum_{w' \in v} c'(w',i)} \tag{3.8}$$

avec $c'(w,i) = \sum_{j=1}^{d_l} x_w^{d,j} \dot{K}(i,j)$ est la fréquence virtuelle du terme w à la position i obtenu par la propagation des occurrences de w dans toutes les positions du document. $x_w^{d,j}$ est le nombre d'occurrences de w dans la position j du document d qui prend la valeur 0 si w n'a pas apparu à la position j dans d et 1 sinon. $K(i,j)$ est la propagation des occurrences de w de la position j à i, estimé à l'aide d'une fonction de densité. Par exemple, pour le noyau gaussien : $K(i,j) = \exp[\frac{-(i-j)^2}{2\sigma^2}]$.

Le score d'appariement du PLM à la position i d'un document d par rapport à une requête q est obtenu en utilisant la divergence de Kullback-Leibler (Équation 3.9) :

$$S(q,d,i) = -\sum_{w \in V} p(w|q) \log \frac{p(w|q)}{p(w|d,i)} \tag{3.9}$$

Le score final du document est obtenu vis-à-vis une requête est obtenu par différentes stratégies : meilleur position, multi-positions, et la stratégie multi-σ. Dans cette étude, nous utilisons la stratégie meilleure position d'un terme de la requête, le noyau gaussien pour estimer la propagation des occurrences d'une position à une autre et le paramètre σ est fixé entre 25 et 300 en utilisant la validation croisée [Lv & Zhai 2009b].

3.4.3.3 Modèle QLM

Le modèle QLM (*Quantum Language Model*) a été proposé par Sordoni et al. [2013] pour éviter le problème de normalisation des poids introduits par la prise en compte de la contribution des termes simple et celle des dépendances de termes. L'idée sous-jacente est de ne considérer les dépendances des termes qu'au niveau de la phase d'estimation des densités matricielles des documents et des requêtes. Les dépendances des termes sont considérées comme étant une superposition (état quantique) des évènements d'apparition de leurs constituants. Les termes simples sont donc représentés par un ensemble de projecteurs (évènement quantique représentant l'occurrence d'un terme de la requête) dans la base standard : $\mathcal{X} = \{|e_i\rangle\langle e_i|\}_{i=1}^{n}$, i.e $|e_i\rangle = (\delta_{1i}, ..., \delta_{ni})^T$, appelée vecteur *ket*, et $\langle e_i| = (\delta_{1i}, ..., \delta_{ni})$, appelée vecteur *bra*, où $\delta_{ij} = 1$ ssi $i = j$. Les termes simples sont représentés par l'évènement $X_w = m\{(x_w)\} = |e_{x_w}\rangle\langle e_{x_w}|$ qui consiste à associer w à l'opérateur $|e_{x_w}\rangle\langle e_{x_w}|$. Chaque dépendance de termes $k = \{x_{w_1}, x_{w_2}, ..., x_{w_k}\}$ est associée également à un opérateur $X_k = m(\{x_{w_1}, x_{w_2}, ..., x_{w_k}\}) = |k\rangle\langle k|$ telle que $|k\rangle = \sum_{i=1}^{k} \sigma_i |e_{x_{w_i}}\rangle$. L'opérateur $|k\rangle\langle k|$ est un opérateur de superposition représentant l'évènement d'observer k ; σ_i sont des coefficients réels et $\sum_{i=1}^{k} \sigma_i^2 = 1$ pour assurer la normalisation de $|k\rangle$. L'évènement $|k\rangle\langle k|$ ajoute une fraction d'occurrence aux évènements d'apparition de ses constituants $|e_{x_w}\rangle\langle e_{x_w}|$.

Dans un premier temps, le modèle QLM construit l'ensemble des opérateurs représentant les termes simples et les dépendances de termes. Puis, il procède par une étape d'estimation des matrices de densité des documents et de la requête en utilisant l'estimateur du maximum de vraisemblance. Soit $\mathcal{X}_d = \{X_1, ..., X_M\}$ l'ensemble des opérateurs construit dans la première étape pour un document d. La vraisemblance de la densité matricielle est donnée par :

$$\mathcal{L}_{\mathcal{X}_d}(\rho) = \prod_{i=1}^{M} tr(\rho X_i) \tag{3.10}$$

où $tr(\rho X_i)$ est la probabilité d'observer X_i. Pour un certain nombre d'itérations, la maximisation de la densité ρ est approximée par l'algorithme $R\rho R$ [Lvovsky 2004] qui consiste à résoudre l'inéquation suivante :

$$\begin{cases} \underset{\rho}{maximiser} \ \log \mathcal{L}_{\mathcal{X}_d}(\rho) \\ R(\rho) = \sum_{i=1}^{M} \frac{1}{tr(\rho X_i)} X_i \\ \hat{\rho}(k+1) = \frac{1}{Z} R(\hat{\rho}(k))\hat{\rho}(k)R(\hat{\rho}(k)) \text{ où } Z = tr(R(\hat{\rho}(k))\hat{\rho}(k)R(\hat{\rho}(k))) \end{cases} \tag{3.11}$$

où la matrice de densité $R(\rho)$ sert à trouver l'ensemble $\hat{\rho}$ qui maximise le log de la vraisemblance et Z est un facteur de normalisation, utilisé pour assurer une trace unitaire de la matrice de densité. La convergence est assurée grâce à l'amortissement de densité lorsque le maximum de vraisemblance diminue. Par exemple, si la vraisemblance diminue à l'itération $k+1$, la densité $\tilde{\rho}(k+1)$ est définie par $\tilde{\rho}(k+1) = (1-\gamma)\hat{\rho}(k) + \gamma\hat{\rho}(k+1)$ où $\gamma \varepsilon [0,1)$ est un paramètre qui contrôle l'amortissement de densité. En outre, l'algorithme d'estimation commence par les matrices initiales pour un document $\rho(0)_d = diag(\frac{x^d_{w1,1}}{l_d}, \frac{x^d_{w2,2}}{l_d}, ..., \frac{x^q_{w_{l_q},l_q}}{l_d}, \frac{l_d - \sum_{i=1}^{l_q} x^d_{w_{i,i}}}{l_d})$, une requête $\rho(0)_q = diag(\frac{x^q_{w1,1}}{l_q}, \frac{x^q_{w2,2}}{l_q}, ..., \frac{x^q_{w_{l_q},l_q}}{l_q}, 0)$, et la collection $\rho(0)_C = diag(\frac{x^C_{w1,1}}{N}, \frac{x^C_{w2,2}}{|C|}, ..., \frac{x^C_{w_{l_q},l_q}}{|C|}, \frac{|C| - \sum_{i=1}^{l_q} x^C_{w_{i,i}}}{|C|})$. La dimension des densités matricielles est $l_q + 1$, où la dimension supplémentaire représente la probabilité des autres termes du vocabulaire.

Après avoir terminé la phase d'estimation, le modèle QLM repose aussi sur le lissage de la densité du document pour éviter le problème de probabilité nulle en utilisant la formule : $\rho_d = (1-\alpha_d)\hat{\rho}_d + \alpha_d\hat{\rho}_C$, où $\alpha_d = \frac{\mu}{\mu+M}$ est le paramètre de lissage. La fonction d'appariement est donnée par la divergence négative de Von-Neumann donnée par la formule suivante :

$$\begin{aligned} RSV(q,D) &= -\Delta_{VN}(\rho_q || \rho_d) \\ &\overset{rank}{=} tr(\rho_q \log \rho_d) \\ &\overset{rank}{=} \sum_i \lambda_{qi} \sum_j \log \lambda_{dj} \langle q_i | d_j \rangle^2 \end{aligned} \tag{3.12}$$

où $\rho_q = \sum_i \lambda_{d_i} |q_i\rangle\langle q_i|$ et $\rho_d = \sum_j \lambda_{d_i} |d_i\rangle\langle d_i|$ sont les décompositions en éléments propres des matrices de densité ρ_d et ρ_q respectivement.

3.4.4 Modèle de dépendance DFR

Le modèle de dépendance DFR [Peng *et al.* 2007] consiste à intégrer les dépendances des termes dans la famille de modèle de déviation à l'aléatoire [Amati & Van Rijsbergen 2002]. Ce modèle permet d'affecter des scores pour chaque paire de termes de la requête ainsi les termes simples. La fonction d'appariement générale est donnée par :

$$RSV(d,q) = \lambda_1 \cdot \sum_{w \in q} score(w,d) \; + \; \lambda_2 \cdot \sum_{p \in q_2} score(p,d) \tag{3.13}$$

où $score(w,d)$ est le score d'un terme simple w du document d, p correspond à une paire de termes de la requête, $score(p,d)$ est le score affecté à p pour le document d, et q_2 est la requête formée par l'ensemble des paires de termes de q. Le score $score(w,d)$ peut être obtenu en utilisant n'importe quel modèle de base DFR. Dans ce travail, nous utilisons le modèle PL2 [Amati & Van Rijsbergen 2002]. Pour l'hypothèse d'indépendance totale (FI), les dépendances sont ignorées, *i.e* $\lambda_1 = 1$ et $\lambda_2 = 0$. Pour la prise en considération des dépendances séquentielles et des dépendances totales (phrase non-contigüe), les paramètres

peuvent prendre les valeurs $\lambda_1 = 1$ et $\lambda_2 = 1$. Le modèle de proximité DFR calcule le score pour une paire de terme $score(p, d)$ sans reposer sur la fréquence de cette paire dans la collection. En particulier, il est basé sur le modèle DFR binomial donné par l'Équation 3.2 :

$$\begin{aligned} score(d,p) =& \frac{1}{t_p^d + 1} \cdot (-\log_2(l_d - 1)! + \log_2 t_p^d! \\ &+ \log_2(l_d - 1 - t_p^d)! - t_p^d log_2(p_p)! \\ &- (-l_d - 1 - t_p^d)\log_2(p'_p)) \end{aligned} \tag{3.14}$$

où $p_p = \frac{1}{l_d - 1}$ et $p'_p = 1 - p_p$, et t_p^d est la fréquence normalisée du paire de termes p qui est obtenue en utilisant le principe de normalisation 2 [Amati & Van Rijsbergen 2002] : $t_p^d = x_p^d \cdot \log_2(1 + c\frac{l_{avg}-1}{l_d-1})$. Dans cette normalisation, c est un paramètre de normalisation de la longueur du document et x_p^d est le nombre d'occurrences de p dans le document d.

En considérant des paires de termes ordonnées ou non-ordonnées en plus de termes simples, on retrouve les mêmes dépendances de termes utilisés dans le modèle MRF (FI, SD et FD).

3.4.5 Récapitulatif

Les extensions présentées dans la section précédente montrent la diversité des méthodes pour l'intégration des dépendances en RI. En commençant par les familles de modèles de RI introduit dans la Section 1.2.3.3 du Chapitre 1, nous nous retrouverons avec les modèles suivants pour la prise en considération des dépendances de termes :

1. Pour la famille de modèles de langue : LM_MWT, LM_CT, MRF, PLM etQLM
2. Pour la famille DFR : LGD_MWT, LGD_CT, SPL_MWT, SPL_CT, PL2_MWT, PL2_CT et DFR_TD
3. Pour le modèle BM25 : BM25_MWT et BM25_CT

où _MWT désigne l'extension à base des termes complexes (Section 3.4.1), _CT l'extension de termes croisés (Section 3.4.2) et _TD le modèle de dépendance DFR (Section 3.4.4).

3.5 Contrainte des dépendances des termes

Conformément à l'esprit des approches axiomatiques pour la RI [Fang *et al.* 2004, Clinchant & Gaussier 2011, Fang & Zhai 2014], nous introduisons une contrainte formelle que les modèles de RI doivent satisfaire pour la prise en compte de dépendance de termes de manière adéquate. Selon cette condition, pour une requête constituée de deux termes dépendants, formant une dépendance de termes, donc un document qui contient plus d'occurrences de la dépendance sous-jacente doit avoir un score plus élevé par rapport aux documents contenant moins d'occurrences de cette dépendance.

Condition 1 *Soit* $q = \{w_1, w_2\}$ *une requête constituée d'un seul terme complexe* $p = \{w_1, w_2\}$*, et* d_1 *et* d_2 *sont deux documents de la même longueur telle que* $x_{w_1}^{d_1} = x_{w_1}^{d_2}$, $x_{w_2}^{d_1} = x_{w_2}^{d_2}$*. Si* $x_p^{d_1} > x_p^{d_2}$*, donc* $RSV(q, d_1) > RSV(q, d_2)$*.*

Il est facile de noter que les extensions des termes complexes (MWT) et ceux des termes croisés (CT) considérés satisfont cette condition. Cela est dû au fait que ces extensions sont basées sur des combinaisons linéaires de la contribution des termes simples et celle de dépendances (MWT et CT), et que leurs modèles de base satisfont la contrainte de la fréquence de termes (Condition TF [Fang *et al.* 2004]). Le même raisonnement s'applique aux extensions DFR_TD et MRF. Dans le contexte du modèle QLM, la représentation des dépendances de termes à base des opérateurs de superposition ajoute une fraction d'occurrence à leurs termes constituants. Donc ce modèle satisfait la condition 1.

Pour le modèle PLM, la situation est plus complexe. Pour toutes les stratégies utilisées par ce modèle pour combiner les modèles de langue virtuels des positions, le score d'appariement augmente en fonction de la proximité des termes de la requête. Dans la mesure où la dépendance de termes est assurée par leurs proximités, le modèle PLM a tendance de satisfaire la condition 1. Cela signifie que ce comportement n'est pas toujours garanti ; il est donc possible que les deux termes qui constituent la dépendance soient séparés par des mots (par exemple l'insertion d'adjectifs ou une séquence Nom-Nom) dans un document, alors qu'ils peuvent être plus proches dans un autre document sans qu'ils forment une dépendance de termes (par exemple les deux termes sont séparés par une virgule et appartiennent à deux propositions différentes). Il est possible de construire de telles instances afin que la différence dans le nombre d'occurrences ne dépasse pas le facteur de proximité.

En définitive, à l'exception du modèle PLM, tous les modèles considérés satisfont la condition 1. Dans cette perspective, ils sont, à l'exception de PLM, des modèles valides pour traiter les dépendances de termes en RI. Comme nous le verrons dans la section suivante, PLM aboutit à une performance inférieure par rapport aux autres extensions.

3.6 Expérimentations

3.6.1 Collection de test et méthode d'évaluation

Pour évaluer la performance des modèles présentés dans les sections précédentes, nous avons effectué nos évaluations en utilisant la collection standard TREC de la langue arabe. Nous avons également utilisé les requêtes et les jugements de pertinence des deux collections TREC-2001 et TREC-2002. Pour avoir un nombre suffisant de requêtes afin d'optimiser les paramètres des modèles en utilisant la validation croisée, les deux ensembles de requêtes TREC-2001 (25 requêtes) et TREC-2002 (50 requêtes) sont fusionnés dans TREC-2002/2001. Le Tableau 3.1 présente les collections de tests utilisées.

Les évaluations sont effectuées en étendant la plateforme de RI Terrier [1] 3.5. Toutefois,

1. www.terrier.org

Tableau 3.1: Description du corpus

Corpus	Collection de test	Requêtes	Champs des requêtes	#Documents
LDC2001T55	TREC 2001	1–25	titre, titre-description	383872
	TREC 2002	26–75	titre, titre-description	
	TREC 2002/2001	1–75	titre, titre-description	

Tableau 3.2: Les valeurs utilisées des paramètres pour la validation croisée

Modèle	**Paramètre**	**Valeurs**
LGD et ces extensions **SPL et ces extensions** **PL2/DFR_TD**	c	0.1, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0 4.5, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 20
LM/MRF **QLM/PLM** **QLM_MWT**	μ	10, 25, 50, 75, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500, 3000, 4000, 5000
DFR_TD **extensions MWT** **extensions CT**	λ	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
BM25 et ces extensions	b	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8, 0.9, 1.0 1.25, 1.5, 1.75, 2.0, 2.25, 2.5, 2.75, 3.0

nous avons utilisé les implémentations des modèles déjà existants dans cette plateforme ; néanmoins, nous avons implémenté les modèles PLM, CRTER et QLM ainsi que les extensions des termes complexes. Tous les modèles et leurs extensions sont évalués en utilisant la précision moyenne (MAP) et la précision au niveau de 10 premiers documents (P10). Les meilleures performances sont désignées en gras et gras-italique pour les valeurs de MAP et P10 respectivement. De plus, nous avons effectué des tests de significativité des résultats obtenus en utilisant le test bilatéral de Student et nous avons attaché $\uparrow$ à la valeur de MAP lorsque le test passe à 90% de confiance (risque $\alpha = 10\%$). Le Tableau 3.2 présente les modèles de RI, leurs paramètres, et les valeurs utilisés pour la validation croisée.

3.6.2 Résultats obtenus

3.6.2.1 Apport des techniques de racinisation

Pour évaluer la performance des modèles de RI pour la langue arabe, nous avons utilisé trois techniques : la technique de racinisation de Khoja et les deux techniques de racinisation légère Light10 et Farasa. Le but principal de ces évaluations est de répondre à la question : *Quels sont les modèles standards de RI et les approches de racinisation appropriés pour la RI en langue arabe ?*

Les Tableau 3.3 et Tableau 3.4 résument les résultats obtenus pour les requêtes titre

Tableau 3.3: Résultats obtenus pour les modèles standards de RI (représentation en sac de mots) en utilisant les approches de racinisation Farasa, Light10 et Khoja pour les requêtes titre. Pour le test de significativité, f = meilleur que Farasa, l = meilleur que Light10, et h = meilleur que Khoja

Approche		Farasa			Light10			Khoja		
Model	TERC	2002/2001	2001	2002	2002/2001	2001	2002	2002/2001	2001	2002
LGD	MAP	$30.09^{l,h}$	$33.94^{l,h}$	$28.17^{l,h}$	27.24^{h}	32.99^{h}	23.79	23.79	26.40	22.49
	P10	44.13	54.80	38.80	37.33	49.20	37.87	37.87	52.00	30.80
SPL	MAP	$30.48^{l,h}$	$34.64^{l,h}$	$28.39^{l,h}$	27.13^{h}	32.90^{h}	24.24	**24.92**	27.25	**23.74**
	P10	47.07	***62.40***	39.40	***42.53***	***58.72***	34.40	***41.87***	***54.80***	34.20
PL2	MAP	$30.58^{l,h}$	$34.60^{l,h}$	$28.57^{l,h}$	27.23^{h}	32.54^{h}	24.59^{h}	24.34	**27.47**	22.78
	P10	***47.33***	62.00	40.00	41.33	55.60	35.80	41.47	54.40	***35.00***
BM25	MAP	$\mathbf{31.50}^{l,h}$	$\mathbf{35.84}^{l,h}$	$\mathbf{29.32}^{l,h}$	$\mathbf{27.65}^{h}$	$\mathbf{33.22}^{h}$	24.86^{h}	23.67	26.63	22.19
	P10	47.07	60.80	***40.20***	40.13	52.80	33.80	37.20	50.80	30.40
LM	MAP	$29.67^{l,h}$	$32.73^{l,h}$	$28.14^{l,h}$	27.05^{h}	31.25^{h}	$\mathbf{24.95}^{h}$	23.68	26.34	22.35
	P10	44.93	54.40	***40.20***	42.40	52.40	***37.40***	39.87	52.80	33.40

Tableau 3.4: Résultats obtenus pour les modèles standards de RI (représentation en sac de mots) en utilisant les approches de racinisation Farasa, Light10 et Khoja pour les requêtes titre-description. Pour le test de significativité, f = meilleur que Farasa, l = meilleur que Light10, et h = meilleur que Khoja

Approche		Farasa			Light10			Khoja		
Model	TERC	2002/2001	2001	2002	2002/2001	2001	2002	2002/2001	2001	2002
LGD	MAP	$32.42^{l,h}$	$35.92^{l,h}$	$31.77^{l,h}$	28.94^{h}	31.87^{h}	$\mathbf{27.47}^{h}$	24.97	27.34	23.78
	P10	47.33	58.40	***45.00***	***44.20***	53.60	37.40	41.07	54.00	34.60
SPL	MAP	$\mathbf{33.51}^{l,h}$	$\mathbf{36.30}^{l,h}$	$\mathbf{32.12}^{l,h}$	28.72^{h}	32.32^{h}	26.93^{h}	25.28	26.45	**24.70**
	P10	***50.67***	***63.60***	44.20	44.80	***62.40***	36.00	***45.73***	55.20	***41.00***
PL2	MAP	$33.22^{l,h}$	$36.10^{l,h}$	$31.77^{l,h}$	$\mathbf{28.95}^{h}$	32.91^{h}	26.98^{h}	**25.86**	**28.37**	24.61
	P10	50.53	61.60	***45.00***	42.80	57.60	35.40	44.13	***56.80***	37.80
BM25	MAP	$33.42^{l,h}$	$36.32^{l,h}$	$31.96^{l,h}$	28.93^{h}	$\mathbf{33.21}^{h}$	26.78^{h}	25.17	28.14	23.68
	P10	49.60	60.40	44.20	42.93	58.80	35.00	44.40	56.40	38.40
LM	MAP	$31.15^{l,h}$	$33.11^{l,h}$	$30.18^{l,h}$	27.85^{h}	30.22^{h}	26.66^{h}	25.22	27.56	24.05
	P10	46.93	56.00	42.40	43.07	52.80	***38.20***	43.87	55.60	38.00

et titre-description respectivement. Les résultats montrent que l'approche de racinisation Farasa assure des améliorations significatives par rapport aux autres approches classiques de racinisation. Cela s'explique par la performance élevée de cette approche dans la segmentation des mots [Darwish & Mubarak 2016]. Conformément à ce qui ressort de travaux antérieurs, l'approche de racinisation légère Light10 surpasse de façon significative l'ap-

proche de racinisation Khoja. La performance faible de l'approche de racinisation (Khoja) s'explique par le fait que les approches basées-racine regroupent des mots sémantiquement différents dans la même racine. Pour les requêtes titre, de petites améliorations sont obtenues en utilisant les approches de racinisation légère Farasa et Light10 pour le modèle BM25 en comparaison avec les autres modèles de RI. Pour l'approche basée-racine, les modèles basés sur le contenu informatif des termes de requête dans les documents sont plus efficaces que les autres modèles. Par conséquent, la meilleure performance est obtenue en utilisant les deux modèles SPL et PL2. En ce qui concerne les requêtes titre-description, des améliorations légères ont été obtenues en utilisant le modèle SPL, couplé avec l'approche de racinisation légère Farasa. En outre, les résultats globaux de comparaison des trois approches de racinisation montrent que les modèles SPL, PL2 et BM25 donnent de meilleures performances par rapport aux modèles LGD et LM.

3.6.2.2 Apport des dépendances de proximité

Dans cette section, nous comparons la performance des modèles de proximité (en utilisant les extensions des termes croisés ainsi que les extensions spécifiques au modèle DFR, DFR_TD, et au modèle de langue, PLM, MRF et QLM) avec leurs modèles de base pour la RI en langue arabe, couplés avec les trois approches de racinisation. Les Tableau 3.5 et Tableau 3.6 présentent les résultats obtenus pour les requêtes titre et titre-description respectivement. Pour évaluer l'apport des modèles de proximité, nous avons évalué divers modèles, y compris le modèle DFR de dépendance (noté DFR_TD), les modèles MRF, PLM et QLM de la famille des modèles de langue, ainsi que le modèle des termes croisés CRTER pour intégrer les dépendances dans le modèle BM25. De plus, nous avons intégré les termes croisés CTs dans les modèles d'information, le modèle PL2 de la famille DFR et le modèle de langue. Pour les extensions basées termes croisés, nous utilisons le noyau gaussien. Le paramètre σ est optimisé en utilisant la validation croisée sur l'ensemble de valeurs $2, 5, 10, 15, 20, 25, 50, 75, 100$.

Les résultats obtenus pour requêtes titre montrent que, à l'exception du modèle PLM, les modèles de proximité surperforment de façon significative la performance de leurs modèles de base pour toutes les approches de racinisation. Par conséquent, la proximité des termes de la requête est un facteur très utile pour améliorer la performance de RI en langue arabe. De plus, l'intégration des termes croisés (extension _CT) dans les modèles de RI mène à des améliorations significatives par rapport à leurs modèles de base. Pour l'approche de racinisation légère Farasa, les meilleures performances sont obtenues par l'intégration des termes croisés (CT) dans les modèles BM25 (noté CRTER), PL2, SPL et LGD sur toutes les collections de tests. Dans le contexte de la famille de modèle de langue, l'extension LM_CT donne une performance légèrement supérieure par rapport aux autres extensions, y compris les modèles PLM, MRF et QLM, sur toutes les collections de tests. Pour l'autre approche de racinisation légère Light10, les meilleurs résultats sont obtenus en utilisant les extensions DFR_TD, PL2_CT, CRTER , SPL_CT et PL2_CT sur les collections TREC-2002/2001 et TREC-2002. En outre, sur la collection TREC-2001, les meilleures

Tableau 3.5: Résultats de comparaison des modèles de proximité avec leurs modèles de base en utilisant les trois approches de racinisation pour les requêtes titre

Approche		Farasa			Light10			Khoja		
Model	TREC	2002/2001	2001	2002	2002/2001	2001	2002	2002/2001	2001	2002
LGD	MAP	30.09	33.94	28.17	27.24	32.99	23.79	23.79	26.40	22.49
	P10	44.13	54.80	38.80	37.33	49.20	***37.87***	37.87	52.00	30.80
LGD_CT	MAP	**32.47**↑	**36.04**↑	**30.68**↑	**28.44**↑	**34.12**↑	**25.60**↑	**24.49**	**26.80**	**23.33**
	P10	***45.60***	***56.00***	***40.40***	***42.67***	***55.20***	36.40	***39.73***	***54.00***	***32.60***
SPL	MAP	30.48	34.64	28.39	27.13	32.90	24.24	24.92	27.25	23.74
	P10	47.07	62.40	39.40	42.53	***58.72***	34.40	41.87	54.80	34.20
SPL_CT	MAP	**32.05**↑	**36.47**↑	**29.84**↑	**29.59**↑	**34.59**↑	**27.09**↑	**25.61**	**28.32**↑	**24.26**
	P10	***48.13***	***64.40***	***40.00***	***45.60***	58.40	***39.20***	***42.93***	***56.80***	***36.00***
PL2	MAP	30.58	34.60	28.57	27.23	32.54	24.59	24.34	27.47	22.78
	P10	47.33	62.00	40.00	41.33	55.60	35.80	41.47	54.40	35.00
PL2_CT	MAP	**32.41**↑	**36.69**↑	**30.27**↑	**29.67**↑	**35.39**↑	**26.81**↑	**25.44**↑	**28.50**↑	**23.85**↑
	P10	48.27	63.20	***40.80***	***45.07***	***57.20***	***39.00***	***43.50***	***56.00***	***37.25***
DFR_TD	MAP	32.00↑	36.20↑	29.90↑	29.59↑	35.21↑	26.78↑	25.36↑	28.33↑	**23.87**↑
	P10	***48.40***	***63.60***	***40.80***	44.27	56.00	38.40	43.07	56.80	36.20
BM25	MAP	31.50	35.84	29.32	27.65	33.22	24.86	23.67	26.63	22.19
	P10	47.07	60.80	40.20	40.13	52.80	33.80	37.20	50.80	30.40
CRTER	MAP	**33.31**↑	**37.96**↑	**30.99**↑	**29.61**↑	**35.41**↑	**26.71**↑	**24.68**↑	**27.24**↑	**23.40**↑
	P10	***48.93***	***63.20***	***41.80***	***43.65***	***53,35***	***38.80***	***40.93***	***55.20***	***33.80***
LM	MAP	29.67	32.73	28.14	27.05	31.25	24.95	23.68	26.34	22.35
	P10	44.93	54.40	40.20	42.40	52.40	37.40	39.87	52.80	33.40
LM_CT	MAP	**31.90**↑	**34.45**↑	**30.63**↑	**28.50**↑	33.32↑	**26.10**↑	**25.36**↑	**28.40**↑	**23.84**↑
	P10	***46.00***	***56.00***	***41.00***	***43.33***	52.40	**38.80**	***42.02***	***56.05***	35.00
PLM	MAP	29.97	32.96	28.47	27.38	32.02	25.07	24.04	27.40	22.36
	P10	45.07	54.80	40.20	42.13	50.80	37.80	40.27	53.00	33.90
MRF	MAP	31.38↑	33.95↑	30.10↑	28.02	32.50	25.78	25.24↑	28.01↑	23.86↑
	P10	45.87	***56.00***	40.80	42.67	52.40	37.80	41.60	55.60	34.60
QLM	MAP	31.50↑	34.03↑	30.23↑	28.29↑	**34.52**↑	25.18	24.66↑	28.01↑	23.15↑
	P10	***46.00***	***56.00***	***41.00***	41.47	***54.80***	35.60	41.87	54.81	***35.40***

performances sont obtenues par les modèles PL2_CT, DFR_TD et BM25_CT. Concernant la famille de modèles de langue, l'extension LM_CT donne une performance supérieure à celles des autres extensions sur les collections TREC-2002/2001 et TREC-2002, mais le modèle QLM a montré une meilleure performance sur la collection TREC-2001. Pour l'approche basée-racine, des améliorations importantes ont été atteintes par DFR_TD, SPL_CT, MRF et LM_CT sur les collections TREC-2002/2001 et TREC-2001. De plus, le modèle SPL_CT a donné la meilleure performance sur TREC-2002. En outre, l'intégration des termes croisés (CTs) dans les deux modèles SPL et PL2 donne une meilleure performance par rapport au modèle CRTER pour l'approche basée-racine.

Conformément aux résultats obtenus pour les requêtes titre, les modèles de proximité

Tableau 3.6: Résultats de comparaison des modèles de proximité avec leurs modèles de base en utilisant les trois approches de racinisation pour les requêtes titre-description

Approche		Farasa			Light10			Khoja		
Model	TREC	2002/2001	2001	2002	2002/2001	2001	2002	2002/2001	2001	2002
LGD	MAP	32.42	35.92	31.77	28.94	31.87	27.47	24.97	27.34	23.78
	P10	47.33	58.40	**45.00**	***44.20***	53.60	***37.40***	41.07	54.00	34.60
LGD_CT	MAP	**34.23**↑	**38.34**↑	**32.18**↑	**29.92**	**34.06**↑	**27.85**	**27.48**↑	**30.51**↑	**25.96**↑
	P10	***48.27***	***60.80***	42.00	44.00	59.20	36.40	***43.73***	***56.40***	***37.40***
SPL	MAP	33.51	36.30	32.12	28.72	32.32	26.93	25.28	26.45	24.70
	P10	50.67	63.60	44.20	44.80	***62.40***	36.00	***45.73***	55.20	***41.00***
SPL_CT	MAP	**35.28**↑	**39.12**↑	**33.36**↑	**31.66**↑	**35.23**↑	**29.87**↑	**27.82**↑	**29.70**↑	**26.87**↑
	P10	***50.93***	***63.60***	***44.60***	***47.07***	***62.40***	***39.40***	45.60	***56.40***	40.20
PL2	MAP	33.22	36.10	31.77	28.95	32.91	26.98	25.86	28.37	24.61
	P10	50.53	61.60	45.00	42.80	57.60	35.40	44.13	*56.80*	37.80
PL2_CT	MAP	**34.99**↑	**38.40**↑	**33.29**↑	**31.24**↑	34.76↑	**29.48**↑	**28.29**↑	**31.31**↑	26.79↑
	P10	51.20	***63.20***	***45.20***	46.67	***62.40***	***38.80***	***46.13***	***58.40***	***40.00***
DFR_TD	MAP	34.24	37.45	32.63	30.90↑	**35.20**↑	28.75↑	28.13↑	31.08↑	**26.65**↑
	P10	***50.80***	62.00	***45.20***	***47.07***	60.80	40.20	46.00	58.00	***40.00***
BM25	MAP	33.42	36.32	31.96	28.93	33.21	26.78	25.17	28.14	23.68
	P10	49.60	60.40	44.20	42.93	58.80	35.00	***44.40***	56.40	38.40
CRTER	MAP	**35.02**↑	**37.96**↑	**33.54**↑	**31.56**↑	**35.55**↑	**29.56**↑	**27.12**↑	**31.01**↑	**25.17**↑
	P10	***50.93***	***61.20***	***45.80***	***47.73***	***62.00***	***40.60***	43.47	***56.80***	***36.80***
LM	MAP	31.15	33.11	30.18	27.85	30.22	26.66	25.22	27.56	24.05
	P10	46.93	56.00	***42.40***	43.07	52.80	38.20	43.87	55.60	38.00
LM_CT	MAP	**32.85**↑	**34.91**↑	**31.82**↑	29.33↑	32.09↑	**27.95**	**27.41**↑	30.13↑	**26.05**↑
	P10	***47.47***	***57.60***	***42.40***	***45.07***	***58.00***	38.60	***44.53***	56.40	***38.60***
PLM	MAP	31.66	33.65	30.66	28.49	31.03	27.23	26.05	28.93	24.61
	P10	46.93	56.80	42.00	43.73	54.80	38.20	44.27	***57.20***	37.80
MRF	MAP	32.04	33.95	31.09	29.37↑	32.72↑	27.69	27.27↑	**30.15**↑	25.83↑
	P10	47.07	56.80	42.20	44.13	56.80	37.80	***44.53***	56.80	38.40
QLM	MAP	32.54↑	34.58↑	31.52↑	**29.44**↑	**32.22**↑	28.05↑	27.22↑	29.92↑	25.88↑
	P10	47.07	56.80	42.20	44.67	55.60	***39.20***	44.40	56.40	38.40

améliorent aussi de façon significative la performance de leurs modèles de base pour les requêtes titre-description sur toutes les collections de tests. De plus, l'intégration des termes croisés dans les modèles de RI (extensions _CT) mène à des améliorations statistiquement significatives par rapport aux modèles de base. En effet, pour toutes les approches de racinisation utilisées , l'incorporation des termes croisés dans les modèles SPL, PL2 et BM25 a montré la meilleure performance. Enfin, si les modèles de proximité améliorent significativement leurs modèles de base (modèle d'indépendance de termes) pour toutes les collections de tests et les deux types de requêtes, leurs performances sont plus élevées pour l'approche de racinisation légère Farasa par rapport aux autres approches de racinisation (Light10 et Khoja).

3.6.2.3 Apport des termes complexes

Dans cette section, nous étudions l'intégration des termes complexes dans les familles de modèles de RI considérées. Pour la racinisation des textes, nous avons sélectionné les deux approches de racinisation Farasa et Light10 puisque ces deux approches donnent de meilleures performances. Pour le filtrage statistique des termes complexes, nous avons varié le seuil de la mesure d'association *NLC-value* entre 0 et 30. Ce seuil est fixé expérimentalement à 5 en se basant sur la meilleure valeur de MAP. Au niveau de la requête, les termes complexes sont extraits en utilisant seulement le filtrage linguistique ; par exemple, pour la première requête titre q_1 = {**فنون العرض و المؤسسات الاسلامية في العالم العربي** }, les termes extraits sont **فنون العرض** (arts de la scène.), **المؤسسات الاسلامية** (les institutions islamiques), et **العالم العربي** (Le monde arabe). Les Tableau 3.7 et Tableau 3.8 présentent les résultats obtenus pour les extensions des termes complexes et leurs modèles de base pour les requêtes titre et titre-description respectivement.

Tableau 3.7: Résultats de comparaison des extensions de termes complexes avec leurs modèles de base en utilisant les deux approches de racinisation pour les requêtes titre

Approche	Farasa						Light10					
TREC	2002/2001		2001		2002		2002/2001		2001		2002	
Model/Metric	MAP	P10	MAP	P10	MAP	P10	MAP	P10	MAP	P10	MAP	P10
LGD	30.09	44.13	33.94	54.80	28.17	38.80	27.24	37.33	32.99	49.20	24.37	31.40
LGD_MWT	**33.09**↑	***46.53***	**36.58**↑	***56.80***	**31.31**↑	***42.20***	**28.90**↑	**41.73**	**34.63**↑	***52.80***	**26.03**↑	***36.20***
SPL	30.48	47.07	34.64	62.40	28.39	39.40	27.13	42.53	32.90	58.72	24.24	34.40
SPL_MWT	**32.34**↑	***48.13***	**36.80**↑	***64.80***	**30.11**↑	***39.80***	**29.96**↑	***45.20***	**36.18**↑	***60.80***	**28.85**↑	***37.40***
PL2	30.58	47.33	34.60	62.00	28.57	40.00	27.23	41.33	32.54	55.60	24.59	34.40
PL2_MWT	**32.74**↑	***48.67***	**37.11**↑	***63.60***	**30.55**↑	***41.20***	**29.56**↑	***43.60***	**35.50**↑	***56.80***	**26.59**↑	***37.00***
BM25	31.50	47.07	35.84	60.80	29.32	40.20	27.65	40.13	33.22	52.80	24.86	33.80
BM25_MWT	**33.73**↑	***49.33***	**38.58**↑	***63.60***	**31.31**↑	***42.20***	**30.50**↑	***44.27***	**36.85**↑	***57.20***	**27.32**↑	***37.40***
LM	29.67	44.93	32.73	54.40	28.14	40.20	27.05	***42.40***	31.25	52.40	24.95	***37.40***
LM_MWT	**31.63**↑	***46.00***	**34.14**↑	***56.00***	**30.38**↑	***41.00***	**28.12**↑	41.27	**33.18**↑	***54.50***	**25.59**	36.20

Les résultats montrent également que l'intégration des termes complexes dans les modèles de RI améliore significativement la performance de RI en langue arabe. Conformément à l'intégration des termes croisés (extensions _CT) dans les modèles de RI, les modèles BM25_MWT, SPL_MWT et PL2_MWT donnent une meilleure performance par rapport aux modèles LGD_MWT et LM_MWT pour toutes les collections de tests. Bien que les extensions des termes complexes surpérforment de façon significative leurs modèles de base en utilisant les deux approches de racinisation, leurs performances sont plus élevées avec l'approche de racinisation Farasa qu'avec l'approche de racinisation légère Light10.

Tableau 3.8: Résultats de comparaison des extensions de termes complexes avec leurs modèles de base en utilisant les deux approches de racinisation pour les requêtes titre-description

Approche	Farasa						Light10					
TREC	2002/2001		2001		2002		2002/2001		2001		2002	
Model/Metric	MAP	P10	MAP	P10	MAP	P10	MAP	P10	MAP	P10	MAP	P10
LGD	32.42	47.33	35.92	58.40	31.77	45.00	28.94	44.20	31.87	53.60	27.47	37.40
LGD_MWT	**34.96**↑	***48.67***	**39.23**↑	***61.20***	**32.82**↑	***42.40***	**30.56**↑	**46.67**	**34.81**↑	***63.60***	**28.43**	***38.20***
SPL	33.51	50.67	36.30	63.60	32.12	44.20	28.72	44.80	32.32	62.40	26.93	36.00
SPL_MWT	**35.88**↑	***51.87***	**39.92**↑	***65.60***	**33.86**↑	***45.00***	**31.83**↑	***49.33***	**36.14**↑	***65.60***	**29.68**↑	***41.20***
PL2	33.22	50.53	36.10	61.60	31.77	45.00	28.95	42.80	32.91	57.60	26.98	35.40
PL2_MWT	**35.55**↑	***51.20***	**39.25**↑	***62.80***	**33.70**↑	***45.40***	**31.50**↑	***47.87***	**35.85**↑	***61.60***	**29.32**↑	***41.00***
BM25	33.42	49.60	36.32	60.40	31.96	44.20	28.93	42.93	33.21	58.80	26.78	35.00
BM25_MWT	**35.56**↑	***51.47***	**38.60**↑	***61.20***	**34.04**↑	***46.60***	**31.86**↑	***48.93***	**36.94**↑	***65.20***	**29.32**↑	***40.80***
LM	31.15	46.93	33.11	56.00	30.18	42.40	27.85	43.07	30.22	52.80	26.66	38.20
LM_MWT	**33.32**↑	***47.87***	**35.19**↑	***58.00***	**32.38**↑	***42.80***	**29.62**↑	46.00	**32.90**↑	***58.00***	**27.98**↑	***40.00***

3.6.2.4 Comparaison avec des modèles de proximité et modèles termes complexes

Pour comparer les deux approches d'intégration de dépendances de termes, nous étudions les modèles de proximité et les extensions des termes complexes pour la RI en langue arabe. Pour cela, nous avons sélectionné les résultats obtenus par les modèles de proximité et les extensions de termes complexes pour les deux approches de racinisation légère Farasa et Light10. Les Tableau 3.9 et Tableau 3.10 présentent les résultats de comparaison en utilisant les requêtes titre et les requêtes titre-description.

Les résultats des comparaisons montrent que l'intégration des termes complexes (extensions _MWT) dans les modèles de RI, à l'exception du modèle de langue, mène à des légères améliorations par rapport aux modèles de proximité. Pour l'approche de racinisation légère Farasa, les meilleures performances de MAP ont été obtenues par les modèles LGD_MWT et BM25_MWT pour toutes les collections de tests. En outre, les extensions du modèle SPL (SPL_MWT et SPL_CT) ont obtenu la meilleure performance en termes de P10. Malgré que les modèles SPL_MWT et BM25_MWT surpassent de façon significative les extensions de proximité sur la collection de tests TREC-2001, les résultats de comparaison globale montrent que la différence entre les extensions des termes complexes et les termes croisés n'est pas statistiquement significative. Contrairement au modèle de langue, l'incorporation des termes composés dans les modèles BM25, SPL, PL2 et LGD améliore légèrement la performance pour l'approche de racinisation légère Light10. Par ailleurs, les résultats obtenus pour les requêtes titre-description montrent que toutes les extensions des termes complexes donnent de bonnes performances par rapport aux modèles de proximité sur

Tableau 3.9: Comparaison de la performance des modèles de proximité et des extensions de termes complexes pour les requêtes titre

Approche	Farasa						Light10					
TREC	2002/2001		2001		2002		2002/2001		2001		2002	
Metric	MAP	P10	MAP	P10	MAP	P10	MAP	P10	MAP	P10	MAP	P10
LGD_CT	32.47	45.60	36.04	56.00	30.68	40.40	28.44	***42.67***	34.12	***55.20***	25.60	***36.40***
LGD_MWT	**33.09**	***46.53***	**36.58**	***56.80***	**31.31**	***42.20***	**28.90**	41.73	**34.63**	52.80	**26.03**	36.20
SPL_CT	32.05	48.13	36.47	64.40	29.84	40.00	29.59	45.60	34.59	58.40	27.09	***39.20***
SPL_MWT	**32.34**	***48.13***	**36.80**	***64.80***	**30.11**	***39.80***	**29.96**	***45.20***	**36.18**↑	***60.80***	**28.85**	37.40
PL2_CT	32.41	48.27	36.69	63.20	30.27	40.80	**29.67**	***45.07***	35.39	***57.20***	**26.81**	***39.00***
DFR_TD	32.00	48.40	36.20	63.60	29.90	40.80	29.59	44.27	35.21	56.00	26.78	38.40
PL2_MWT	**32.74**	***48.67***	**37.11**	***63.60***	**30.55**	***41.20***	29.56	43.60	**35.50**	56.80	26.59	37.00
CRTER	33.31	*48.93*	37.96	*63.20*	30.99	*41.80*	29.61	*43.65*	35.41	*53,35*	26.71	***38.80***
BM25_MWT	**33.73**	***49.33***	**38.58**	***63.60***	**31.31**	***42.20***	**30.50**	***44.27***	**36.85**↑	***57.20***	**27.32**	37.40
LM_CT	**31.90**	***46.00***	**34.45**	***56.00***	**30.63**	***41.00***	**28.50**	***43.33***	**33.32**	52.40	**26.10**	***38.80***
PLM	29.97	45.07	32.96	54.80	28.47	40.20	27.38	42.13	32.02	50.80	25.07	37.80
MRF	31.38	45.87	33.95	***56.00***	30.10	40.80	28.02	42.67	32.50	52.40	25.78	37.80
QLM	31.50	***46.00***	34.03	***56.00***	30.23	***41.00***	28.29	41.47	34.52	54.80	25.18	35.60
LM_MWT	31.63	***46.00***	34.14	***56.00***	30.38	***41.00***	28.12	41.27	33.18	***54.50***	25.59	36.20

Tableau 3.10: Comparaison de la performance des modèles de proximité et des extensions de termes complexes pour les requêtes titre-description

Approche	Farasa						Light10					
TREC	2002/2001		2001		2002		2002/2001		2001		2002	
Metric	MAP	P10	MAP	P10	MAP	P10	MAP	P10	MAP	P10	MAP	P10
LGD_CT	34.23	48.27	38.34	60.80	32.18	42.00	29.92	44.00	34.06	59.20	27.85	36.40
LGD_MWT	**34.96**	***48.67***	**39.23**	***61.20***	**32.82**	***42.40***	**30.56**	**46.67**	**34.81**	***63.60***	**28.43**	***38.20***
SPL_CT	35.28	50.93	39.12	63.60	33.36	44.60	31.66	47.07	35.23	62.40	29.87	39.40
SPL_MWT	**35.88**	***51.87***	**39.92**	***65.60***	**33.86**	***45.00***	**31.83**	***49.33***	**36.14**	***65.60***	**29.68**	***41.20***
PL2_CT	34.99	51.20	38.40	63.20	33.29	45.20	31.24	46.67	34.76	62.40	29.48	38.80
DFR_TD	34.24	50.80	37.45	62.00	32.63	45.20	30.90	47.07	35.20	60.80	28.75	40.20
PL2_MWT	**35.55**	***51.20***	**39.25**	***62.80***	**33.70**	***45.40***	**31.50**	***47.87***	**35.85**	***61.60***	**29.32**	***41.00***
CRTER	35.02	50.93	37.96	61.20	33.54	45.80	31.56	47.73	35.55	62.00	29.56	40.60
BM25_MWT	**35.56**	***51.47***	**38.60**	***61.20***	**34.04**	***46.60***	**31.86**	***48.93***	**36.94**↑	***65.20***	**29.32**	***40.80***
LM_CT	32.85	47.47	34.91	57.60	31.82	42.40	29.33	45.07	32.09	58.00	27.95	38.60
PLM	31.66	46.93	33.65	56.80	30.66	42.00	28.49	43.73	31.03	54.80	27.23	38.20
MRF	32.04	47.07	33.95	56.80	31.09	42.20	29.37	44.13	32.72	56.80	27.69	37.80
QLM	32.54	47.07	34.58	56.80	31.52	42.20	29.44	44.67	32.22	55.60	28.05	39.20
LM_MWT	**33.32**	***47.87***	**35.19**	***58.00***	**32.38**	***42.80***	**29.62**	***46.00***	**32.90**	***58.00***	**27.98**	***40.00***

toutes les collections de tests. Conformément aux résultats obtenus pour les requêtes titre, la différence entre les extensions des termes complexes et les modèles de proximité n'est pas statistiquement significative.

3.7 Discussion

Les résultats montrent que le niveau d'analyse des mots a un effet majeur sur la performance de RI en langue arabe pour tous les modèles évalués. Concernant les approches de racinisation et conformément à l'étude récente introduite par [Abdelali *et al.* 2016], l'utilisation de l'approche Farasa mène à des améliorations statistiquement significatives de la performance de RI par rapport à l'approche de racinisation Khoja et l'approche de racinisation légère Light10. Ceci s'explique notamment par la performance de Farasa dans la segmentation des affixes [Darwish & Mubarak 2016] et le fait que l'approche de racinisation légère Light10 ne parvient pas à discriminer la plupart des prépositions et des conjonctions de la forme de base du mot. Dans la lignée des précédentes études [Larkey *et al.* 2002, Goweder *et al.* 2004], l'approche Light10 de racinisation légère surpasse de façon significative l'approche de racinisation Khoja. La performance faible de cette dernière approche réside dans le fait de grouper des mots sémantiquement différents dans la même racine [Kadri 2008, Froud *et al.* 2012]. Les résultats des études antérieures, qui ont montré que les approches basées-racine sont plus efficaces que les approches de racinisation légère (approches basées-*stem*) pour la RI en langue arabe, ont été principalement obtenus sur des corpus relativement petits ; sur de tels corpus, la représentation des documents à base des racines augmente la probabilité d'appariement des termes de la requête aux termes du document [Abu El-Khair 2007]. De plus, les modèles de RI qui sont basés sur le contenu informatif des termes (LGD, SPL and PL2) sont plus efficace pour l'approche basée-racine.

Concernant l'intégration des dépendances de termes, les deux approches d'incorporation des dépendances explicites à base des termes complexes et les dépendances implicites à base des modèles de proximité (particulièrement les termes croisés) améliorent significativement la performance de RI en langue arabe pour les trois approches de racinisation. Par conséquent, les modèles de dépendances sont très utiles pour la recherche du contenu en langue arabe, où les techniques de racinisation introduisent une certaine quantité de bruit dans la représentation dans documents. Ces constatations confirment que les dépendances de termes ou la proximité des termes sont très utiles pour améliorer la performance de RI sur une représentation bruitée du contenu [Ye *et al.* 2013]. Bien que la comparaison des modèles de dépendances explicites (extensions _MWT) et de dépendance implicite (modèle de proximité) a montré que les meilleures performances (en termes de MAP et P10) sont obtenues par l'intégration des termes complexes dans les modèles SPL et BM25, la différence entre les deux approches de dépendance n'est pas statistiquement significative pour la plupart des collections de tests. La performance des extensions de termes croisés (_CT) s'explique par le fait que ces termes peuvent capturer les dépendances distantes

où leurs importances augmentent graduellement avec la diminution de la distance entre les termes de la requête. Toutefois, leur principal inconvénient réside dans le fait que pour chaque requête, le SRI doit identifier ces termes et calculer leurs fréquences dans les documents et la collection. En revanche, l'extraction des termes complexes (MWT) en utilisant des paramètres linguistiques et statistiques conduit à de meilleures représentations de documents et de requêtes. L'inconvénient d'utiliser les termes complexes en tant que dépendances repose sur l'indexation de termes supplémentaires, qui augmente la taille de l'index et ajoute des traitements hors ligne (étiquetage morpho-syntaxique du corpus et l'extraction des termes complexes).

Conformément à la condition introduite dans la Section 3.5, la performance la plus faible des modèles de dépendances est obtenue par le modèle PLM. Ce dernier est le seul modèle qui ne satisfait pas la condition d'intégration des dépendances.

3.8 Conclusion

Dans ce chapitre, nous avons étudié principalement les apports de proximité et des termes complexes pour la RI en langue arabe à base de dépendances de termes en utilisant trois approches de racinisation. Notre analyse nous a amené à conclure que :

1. L'approche Farasa améliore significativement la performance de RI par rapport aux approches classiques de racinisation : Light10 et Khoja. Par conséquent, Farasa est l'approche la plus appropriée pour la racinisation des textes arabes et le traitement de sa morphologie riche et complexe dans le contexte de RI ;
2. L'intégration des termes croisés et des termes complexes dans les modèles de RI (LM, BM25, LGD, SPL) mène à des améliorations significatives ; il convient toutefois de nuancer ces bons résultats avec l'absence d'amélioration significative de la performance par rapport aux modèles PLM, MRF et QLM de la famille de modèle de langue.
3. Les meilleurs résultats sont obtenus par l'intégration des termes complexes dans les modèles SPL et BM25. Le modèle CRTER est particulièrement intéressant sur les collections arabes utilisées dans cette étude. En effet, si l'intégration des termes complexes entraine des améliorations légères de la performance par rapport à l'utilisation de termes croisés, la différence n'est pas significative dans la plupart des cas. Par conséquent, le choix d'une méthode par rapport à l'autre dépend d'autres considérations que la performance en RI tout simplement.

Dans le chapitre suivant, nous allons proposer d'aller plus loin dans l'intégration des dépendances pour la RI en langue arabe par la prise en considération des dépendances sémantiques entre les termes, grâce à l'exploitation des représentations distribuées des vecteurs de mots.

Chapitre 4

RI à base des représentations distribuées des mots

Sommaire

4.1 Introduction

Dans le chapitre précédent, nous avons étudié les apports d'intégration des dépendances implicites (dépendances de proximités) et explicites (termes complexes) pour la RI en langue arabe. Le raisonnement intuitif sur lequel repose l'intégration des dépendances de termes dans la RI est que les termes complexes et les opérateurs de proximité sont moins ambigus que les termes simples, isolés de leurs contextes. L'exploitation des dépendances de termes permet de construire une meilleure représentation du contenu des documents et des requêtes et injecter de la sémantique dans leurs appariements.

Cependant, les dépendances de termes (opérateurs de proximité et termes complexes) ne permettent pas de remédier au problème de disparité des termes (*Term Mismatch*). Ce

problème est dû principalement au fait que les concepts ne sont pas toujours exprimés en utilisant les mêmes termes. Un document pertinent peut, donc, ne pas partager les mêmes termes, utilisés par un utilisateur pour exprimer son besoin en information, avec la requête. De ce fait, l'appariement exact des termes heurte la performance de RI. C'est pourquoi le passage vers l'appariement sémantique des documents et des requêtes est nécessaire pour faire face à ce problème [Fang & Zhai 2006, Li & Xu 2014].

Nous proposons dans ce chapitre deux méthodes pour faire face au problème de disparité des termes en RI en langue arabe. Ces deux méthodes reposent sur l'intégration des Représentations Distribuées des Mots (RDM) dans les modèles de RI et les modèles de Rétro-Pertinence PRF (*Pseudo-Relevance Feedback*) :

1. Méthode d'intégration des similarités acquises des RDMs dans les modèles de RI [El Mahdaouy *et al.* 2018a] : consiste à étendre les modèles probabilistes en utilisant les RDMs pour la prise en compte des termes sémantiquement similaires lors de l'appariement des documents et des requêtes. Cette méthode repose sur la sélection des termes similaires à ceux de la requête à partir de la collection ou pour chaque document, ainsi leur pondération et intégration dans les modèles de RI.
2. Méthode d'intégration des similarités acquises des RDMs dans les modèles PRF consiste à intégrer la similarité entre les termes d'expansion et la requête initiale dans les modèles PRF afin d'améliorer le processus d'expansion de la requête [El Mahdaouy *et al.* 2018b]. L'idée principale consiste à combiner les poids des termes candidats d'expansion avec leurs poids de similarités à la requête initiale pour la pondération et la sélection des termes d'expansion.

Le reste de ce chapitre est organisé comme suit : Section 4.2 présente le cadre dans lequel s'inscrit notre étude, où nous introduisons la problématique et les motivations de notre travail. Dans la Section 4.3, nous passons en revue des modèles et méthodes proposés pour l'injection de la sémantique dans la RI en langue arabe et des modèles de l'état de l'art basés sur les RDMs. La Section 4.3 présente les modèles de RDMs utilisés dans ce chapitre. Dans la Section 4.4, nous introduisons notre méthode proposée pour l'extension des modèles de RI en utilisant les RDMs et leur validation théorique. La Section 4.6 présente notre méthode proposée pour l'extension des modèles PRF en utilisant les RDMs. La méthode d'évaluation et les résultats obtenus sont décrits dans la Section 4.7. Nous terminons ce chapitre par une conclusion (Section 4.8).

4.2 Motivations

De récents progrès dans les modèles de langue neuronale ont introduit des méthodes efficaces pour l'apprentissage des représentations distribuées des mots, appelé *Word Embedding* (WE) [Mikolov *et al.* 2013, Pennington *et al.* 2014]. L'idée de base consiste à représenter chaque mot du corpus, en utilisant son contexte, par un vecteur sémantiquement informatif, dans un espace vectoriel de dimension réduite. Ces représentations permettent de capturer des similitudes entre des mots, des phrases ou des documents à

travers des architectures neuronales simples. Ces similitudes appartiennent principalement aux niveaux morphologiques et sémantiques. En effet, de nombreux travaux de recherche ont montré l'efficacité des WEs pour les tâches de similarité et en tant qu'une base de représentation pour les applications de TAL telles que la classification de textes [Ma *et al.* 2016, EL Mahdaouy *et al.* 2017] et la RI [Ganguly *et al.* 2015, Vulić & Moens 2015, Zuccon *et al.* 2015, EL Mahdaouy *et al.* 2016]. De plus, une récente évaluation a montré l'efficacité des représentations distribuées des mots basées contexte pour plusieurs tâches de similarité par rapport aux méthodes traditionnelles de co-occurrence [Baroni *et al.* 2014], telles que l'Information Mutuelle spécifique (*Pointwise Mutual Information PMI*), la décomposition en valeurs singulières (*Singular Value Decomposition SVD*) et la méthode de factorisation par matrices non négatives (*Non-negative Matrix Factorization NMF*). L'atout principal des WE réside dans leur faible sensibilité aux paramètres d'apprentissage ainsi que leurs facilité d'adaptation à n'importe quel domaine où un corpus suffisamment grand est disponible. En outre, ils ont montré des résultats prometteurs sur plusieurs tâches de similarité par rapport aux méthodes basées sur l'ontologie WordNet[Lofi 2015].

Le problème de disparité des termes en RI en langue arabe n'est pas dû seulement au fait que les concepts ne sont pas toujours exprimés en utilisant les mêmes termes, mais aussi à la nature riche et complexe de la morphologie arabe. En effet, les méthodes de racinisation légère, dont l'efficacité est éprouvée pour traiter la morphologie de l'arabe dans la RI, ne parviennent pas à discriminer la plupart des prépositions et des conjonctions de la forme de base du mot [Nwesri *et al.* 2005, Darwish & Mubarak 2016], où des mots sémantiquement similaires sont réduits à de différents stems [Kadri & Nie 2006]. De plus, le pluriel irrégulier et les variantes orthographiques des mots demeurent des défis majeurs pour les techniques de racinisation légère. Donc, le problème de disparité des termes pour la RI en langue arabe devient plus particulièrement crucial, où l'appariement exact des termes diminue dramatiquement la performance de RI en langue arabe.

Les méthodes proposées dans ce chapitre partent de l'hypothèse que les représentations distribuées des mots permettent de capturer des similitudes de niveaux morphologique et sémantique. Donc, ces représentations seront exploitées pour remédier à la limitation de disparité des termes dans la RI en langue arabe. L'idée sous-jacente est que les termes sémantiquement similaires, ainsi les termes qui ont le même stem auront des vecteurs proches dans l'espace vectoriel. Pour illustrer ceci, nous avons utilisé la projection à deux dimensions des vecteurs obtenues pour le terme **تعليم** (enseignement) et ses 100 termes les plus similaires en utilisant l'Analyse en Composante Principale ACP (Principal Component Analysis). Pour l'apprentissage de la représentation distribuée des mots, nous avons utilisé le modèle de sac de mots continus CBOW (*Continuous Bag of Words model*), où le prétraitement (racinisation légère) du corpus d'apprentissage (voir Section 4.7.1) est effectué en utilisant l'outil Farasa [Abdelali *et al.* 2016]. Le résultat de la projection est donné par la Figure 4.1. L'utilisation de la racinisation légère avant l'apprentissage des représentations des mots est motivée par le fait que les méthodes de racinisation légère améliorent significativement la performance de RI en langue arabe par rapport aux

autres techniques de prétraitement de textes (racinisation et normalisation de textes). En concordance avec l'hypothèse de départ, la figure montre que non seulement des mots similaires apparaissent proches les uns des autres dans l'espace vectoriel, mais aussi des mots qui doivent être regroupés dans le même stem. Nous citons à titre d'exemple :

* Termes similaires : **تعليم** (enseignement), **تدريس** (enseignement), **تنشء** (éducation), etc ;
* Pluriel irrégulier :
  - **تلميذ** (élève) : **تلامذ** et **تلاميذ** (élèves) ;
  - **درس** (leçon) : **دروس** (leçons)
  - **استاذ** (enseignant) : **اساتذه** (enseignants) ;
  - etc. ;
* Variantes orthographiques (arabisation) :
  - **بكلوريا** et **باكلوريا** (baccalauréat) ;
  - **بكالريوس** et **بكالوريوس** (licence) ;
  - etc. ;

Dans ce chapitre nous proposons :

1. Une méthode pour intégrer la similarité entre les termes dans les modèles de RI en utilisant les représentations distribuées des mots. Pour permettre aux termes similaires à ceux de la requête à contribuer au score de pertinence, notre méthode est basée principalement sur les modèles de translations proposés par Li & Gaussier [2012], où l'ensemble de traduction possibles d'un terme est remplacé par l'ensemble de ces termes similaires. L'intégration de la similarité entre les termes dans les modèles de RI s'effectue en introduisant une fonction permettant de normaliser la relation entre un terme de la requête et ces termes similaires. Pour chaque terme de la requête, un ensemble de termes similaires peut être sélectionné en utilisant un seuil sur la distance cosinus entre le vecteur du terme de la requête et les vecteurs des termes du vocabulaire de la collection ou des termes du document. Pour cela, nous avons étendu les modèles SPL et LGD de la famille des modèles d'information [Clinchant & Gaussier 2010], le modèle BM25 [Robertson *et al.* 1994], et le modèle de langue [Ponte & Croft 1998] en utilisant la méthode de lissage de Dirichlet [Zhai & Lafferty 2001b]. Les extensions proposées sont validées théoriquement en utilisant les contraintes introduites dans le cadre de l'approche axiomatique pour l'appariement sémantique de termes [Fang & Zhai 2006]. Ces extensions sont comparées avec leurs modèles de base, l'approche d'indexation sémantique en utilisant l'ontologie ArabicWrodNet, et trois extensions du modèle de langue basées sur les représentations distribuées des mots.

2. Une méthode pour intégrer la similarité entre les termes dans les modèles PRF en utilisant les représentations distribuées des mots. L'idée de base consiste à intégrer la similarité entre les termes candidats d'expansion et la requête initiale pour améliorer le processus de sélection des termes d'expansion. La pondération des termes candidats d'expansion s'effectue en combinant leur poids dans l'ensemble des documents

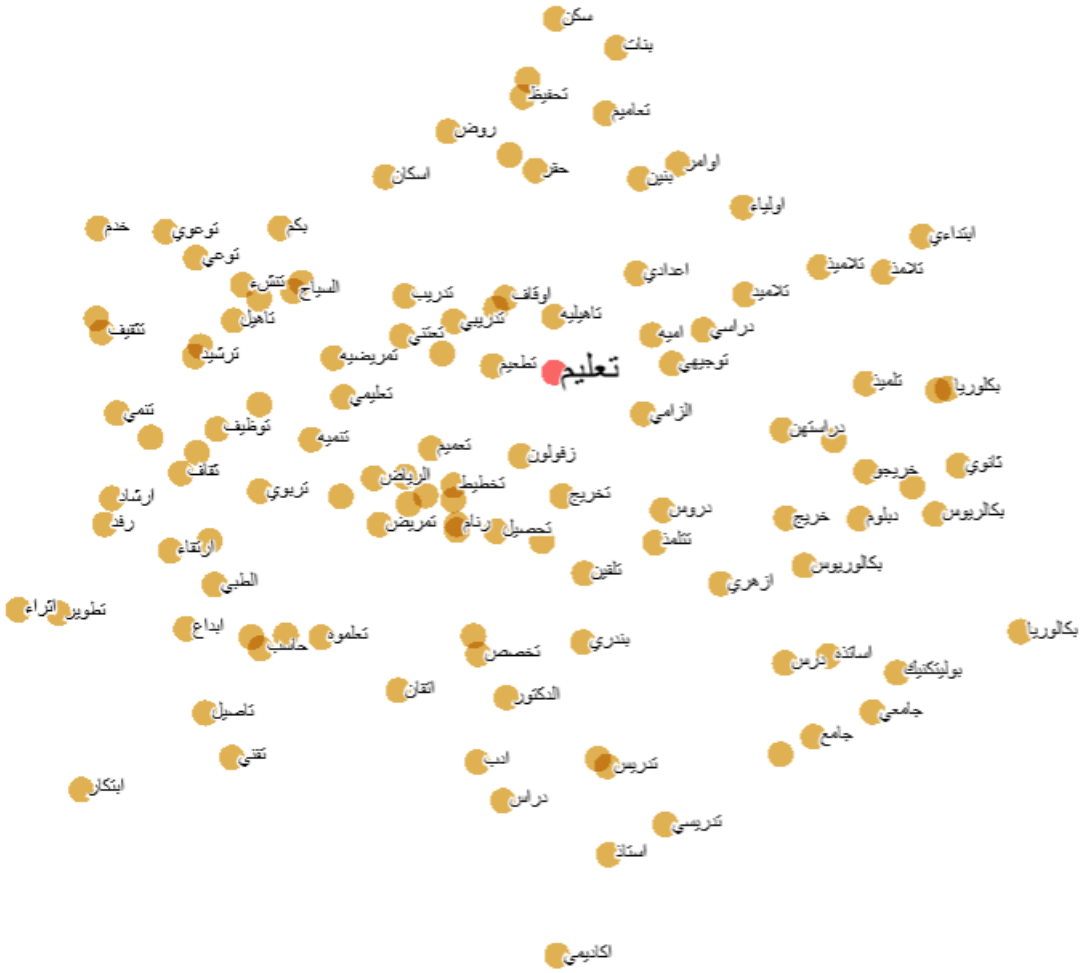


FIGURE 4.1: Projection à deux dimensions du terme **تعليم** (enseignement) et ces 100 plus proches voisins en utilisant l'ACP

d'expansion et leurs similarités à la requête initiale. Pour ce faire, nous avons intégré la similarité entre les termes dans le modèle KLD [Carpineto *et al.* 2001], le modèle Bo2 [Amati & Van Rijsbergen 2002], et le modèle LL [Clinchant & Gaussier 2013]. Étant basées sur les représentations distribuées des mots, les extensions des modèles PRF sont comparées à la méthode d'expansion de requête VEXP [ALMasri *et al.* 2016].

Les deux méthodes proposées sont évaluées en utilisant trois modèles de représentation distribuée des mots, y compris le modèle Glove, le modèle CBOW et le modèle Skip-gram introduit dans *Word2Vec*. Bien que nos méthodes proposées pour l'exploitation des représentations distribuées dans la RI sont indépendantes de la langue utilisée, nous nous concentrons sur l'arabe pour plusieurs raisons : (a) la langue arabe est caractérisée par sa morphologie riche et complexe, où la disparité des termes présente un intérêt particulier (erreurs de racinisation, pluriel irrégulier, variantes orthographiques, etc.). (b) Les représentations distribuées sont émergées en tant qu'une représentation plus viable

pour les applications de TAL et notamment la RI. Contrairement à la représentation en sac de mots, elles permettent de capturer des similitudes de nivaux morphologique et sémantique. (c) l'exploitation de ces représentations dans le contexte de la RI en langue arabe demeure toutefois sous-explorée.

4.3 Travaux reliés

Au cours des dernières années, un effort considérable a été fait en vue de remédier au problème de disparité des termes dans le contexte de la RI en langue arabe. Shaalan et al. [2012] ont introduit une méthode pour intégrer la similarité sémantique dans l'expansion de la requête en utilisant l'algorithme Espérance-Maximisation EM (*Expectation-Maximization*). L'algorithme EM est utilisé pour sélectionner les termes d'expansion à partir des top documents retrouvés pour la requête initiale. L'évaluation est effectuée en utilisant la collection INFILE de la campagne d'évaluation CLEF 2009. Les résultats obtenus ont montré une amélioration de rappel par rapport au modèle de base. Dans un autre travail [Mahgoub *et al.* 2014], une technique pour l'expansion sémantique des requêtes en utilisant une ontologie construite à partir de Wikipedia a été proposée pour améliorer la RI en langue arabe. Les résultats d'évaluation ont montré que cette méthode donne de meilleurs résultats par rapport au modèle standard de RI. Une autre méthode pour l'expansion automatique et interactive de requêtes basées sur l'ontologie ArabicWordNet (AWN) ont été introduites par Belalem et al. [2014]. Les termes d'expansion sont sélectionnés à partir des synonymes extraits de l'AWN en utilisant les étiquettes grammaticales des termes. Récemment, Atwan et al. [2016] ont introduit une méthode PRF basée sur l'AWN et l'information mutuelle afin de sélectionner des termes d'expansion. Les résultats obtenus sur la collection TREC 2001 ont montré une amélioration significative par rapport au modèle de base en utilisant la technique de racinisation légère Light10 [Larkey *et al.* 2007]. Dans un autre travail qui se rapporte à l'appariement sémantique pour la RI en langue arabe, Abderrahim et al. [2016] ont proposé une méthode d'indexation sémantique en utilisant l'AWN et l'algorithme LESK pour la désambigüisation sémantique. Pour chaque terme de la collection, la méthode procède par la recherche des concepts à partir de l'AWN en utilisant l'élimination des affixes et l'extraction des racines. L'algorithme LESK est utilisé pour déterminer le meilleur sens du mot en utilisant son contexte. Pour les termes non trouvés dans l'AWN, la méthode utilise leurs racines en tant que descripteur d'index.

Récemment, les représentations distribuées des mots ont suscité beaucoup d'intérêt pour la communauté de recherche du domaine de RI. Vulic et al. [2015] ont introduit un cadre unifié pour exploiter les représentations distribuées des mots dans la RI et la RI bilingue. Le modèle proposé dans ce dernier cadre consiste à représenter les documents et les requêtes en utilisant une méthode de composition basée sur la distribution des termes, ainsi leurs vecteurs appris en utilisant le modèle Skip-gram. La fonction d'appariement globale consiste à combiner le score du document, obtenu en utilisant le modèle de langue, avec la similarité entre le vecteur du document et celui de la requête. Les résultats ont

montré que leur méthode améliore significativement la performance du modèle de base (modèle de langue) et les modèles de thème LDA (*Latent Dirichlet Allocation*), pour la RI et la RI bilingue. Dans un autre travail, Ganguly et al. [2015] ont proposé un modèle de langue généralisé GLM (*Generalized Language Model*) basé sur les représentations distribuées des mots. Le modèle GLM consiste à estimer des probabilités d'observer un terme de la requête, ainsi leurs termes similaires. Ce modèle repose principalement sur trois probabilités de transformation : la probabilité d'observer un terme de la requête (modèle de base), la probabilité d'observer un terme similaire dans le document, et la probabilité d'observer un terme similaire dans la collection. Ces trois probabilités sont linéairement combinées dans le cadre du modèle GLM. L'évaluation de ce dernier modèle a montré des améliorations significatives par rapport au modèle de base et le modèle LM-LDA. Dans un autre travail similaire, Zuccon et al. [2016] ont proposé un modèle de translation neuronal NTLM (*Neural Translation Model*) pour intégrer les similarités acquises des représentations distribuées des mots dans le modèle de langue de RI. Ce modèle consiste à estimer une probabilité de translation entre un terme de la requête et ces termes similaires en utilisant la similarité entre leurs vecteurs. Les résultats expérimentaux ont montré des améliorations significatives par rapport au modèle de base, ainsi le modèle de translation de base.

Pour l'amélioration des méthodes d'expansion de requêtes en utilisant les représentations distribuées des vecteurs des mots, Zamani et al. [2016] ont étendu le modèle de pertinence (RM) (*Relevance Model*) pour la prise en compte des similarités entre les termes d'expansion et la requête. Les résultats obtenus sur des collections TREC ont montré des améliorations significatives de la performance du modèle RM. Dans un travail plus similaire, Kuzi et al. [2016] ont proposé une multitude de méthodes d'expansion de requêtes en utilisant le modèle de sac de mots continu (CBOW). En outre, étant basé RDMs, [Roy *et al.* 2016] ont exploré les techniques d'expansion de requête. Pour ce faire, ils ont utilisé la méthode des k plus proches termes similaires à ceux de la requête. Ces termes similaires sont sélectionnés en utilisant deux stratégies. La première consiste à extraire les termes similaires au vecteur de la requête à partir du vocabulaire de la collection. En revanche, la deuxième stratégie consiste à extraire les termes similaires au vecteur de la requête à partir des n documents d'expansion (document de feedback). Les expérimentations sont effectuées en utilisant quatre collections TREC et la collection WT10G. Les résultats obtenus montrent que les deux stratégies d'expansion de requêtes basées RDMs assurent des améliorations significatives par rapport au modèle de langue de base. Cependant, le modèle standard RM3 surpasse significativement les deux stratégies d'expansion de requêtes basées RDMs. Dans un autre travail, ALMasri et al. [2016] ont comparé plusieurs méthodes d'expansion de requêtes, y compris la méthode VEXP basée sur les RDMs, la méthode d'expansion basée sur l'information mutuelle [Hu *et al.* 2006] et le modèle RM de la famille de modèle PRF [Lavrenko & Croft 2001]. La méthode d'expansion de requête VEXP consiste à sélectionner, à partir du vocabulaire de la collection, les k plus proches termes similaires à ceux de la requête pour l'expansion de celle-ci. L'idée consiste à sélectionner un ensemble de termes similaires pour chaque terme de la requête afin d'étendre la requête initiale

en utilisant le modèle Skip-gram du word2vec [Mikolov *et al.* 2013]. Les évaluations sont effectuées en utilisant quatre collections CLEF. Les résultats globaux montrent que la méthode VEXP aboutit à des améliorations statistiquement significatives par rapport au modèle de langue de RI et le modèle standard RM [Lavrenko & Croft 2001]. De plus, la méthode VEXP a atteint de meilleures performances par rapport à la méthode d'expansion basée sur l'information mutuelle [Hu *et al.* 2006]. Dans le contexte de la langue arabe, [Zahran *et al.* 2015] ont évalué la performance d'expansion de requête en utilisant les représentations distribuées des mots. Les termes d'expansion sont sélectionnés à base de leur similarité aux termes de la requête. Les résultats de l'évaluation effectuée en utilisant la collection standard TREC 2002 ont montré des améliorations par rapport à la RI sans expansion et la méthode d'expansion sémantique de requête basée sur des ontologies, proposée par Mahgoub et al. [2014].

Dans la section suivante, nous allons présenter les représentations distribuées des mots. Les applications de TAL nécessitent une étape de représentation des mots et pourraient bénéficier d'une représentation qui reflète des similitudes et dissimilitudes entre eux, plutôt que les traiter comme des symboles indépendants. Par conséquent, plusieurs travaux de recherche ont été introduits pour représenter les mots par des vecteurs denses dans un espace vectoriel de dimension réduite, obtenues en utilisant diverses méthodes d'apprentissage automatique inspirées des modèles de langue neuronaux. L'estimation de ces vecteurs repose sur l'idée que les mots qui apparaissent dans les mêmes contextes sont sémantiquement proches.

4.3.1 Modèle CBOW

Dans le modèle CBOW (Continuous Bag of Words) [Mikolov *et al.* 2013], les contextes de chaque mot sont construits en utilisant des fenêtres symétriques. La représentation vectorielle du mot est construite en maximisant le log de la probabilité de prédire le mot cible étant donné ces contextes. Le modèle CBOW utilise une architecture simplifiée du modèle de langue neuronal [Bengio *et al.* 2003], où la couche cachée est supprimée et la couche de projection est partagée pour tous les mots. Pour chaque mot w_t du corpus et son contexte $\{w_{t-c}, ..., w_{t-1}, w_{t+1}, ..., w_{t+c}\}$, le modèle CBOW maximise l'Équation 4.1 :

$$\frac{1}{|C|}\sum_{t=1}^{|C|} log[P(w_t|w_{t-c}, ..., w_{t-1}, w_{t+1}, ..., w_{t+c})] \tag{4.1}$$

où $|C|$ est le nombre de mots dans la collection et c est la taille du contexte dynamique du mot w_t.

4.3.2 Modèle Skip-gram

Au lieu de prédire le mot actuel étant donné son contexte, le modèle Skip-gram utilise une architecture similaire à celle du modèle CBOW en inversant l'entrée et la sortie du

réseau de neurones [Mikolov *et al.* 2013]. La fonction de coût consiste à maximiser le log de la probabilité de prédire le contexte étant donné le mot qui se trouve au milieu de la fenêtre symétrique. Étant donné une séquence de mots $\{w_{t-c}, ..., w_{t+c}\}$ qui représente le contexte à gauche et le contexte à droite du mot w_t, le modèle Skip-gram maximise l'Équation 4.2 :

$$\frac{1}{|C|}\sum_{t=1}^{|C|}\sum_{j=t-c, j\neq t}^{t+c} log[P(w_j|w_t)] \tag{4.2}$$

où $|C|$ est le nombre de mots dans la collection et c est la taille du contexte dynamique du mot w_t.

4.3.3 Modèle Glove

Le modèle Glove est un modèle de régression qui consiste à combiner les avantages des méthodes de factorisation matricielle et les méthodes basées sur le contexte local du mot [Pennington *et al.* 2014]. L'apprentissage s'effectue sur les entrées non nulles de la matrice globale de co-occurrence. Ce modèle procède par la construction de la matrice de co-occurrence X, où l'élément X_{ij} représente le nombre d'occurrences du mot W_j apparu en tant que contexte du mot W_i. Pour chaque paire de mots, le modèle Glove définit une contrainte sur les éléments de la matrice de co-occurrence : $w_i^T w_j + b_i + b_j = log(X_{ij})$, où w_i et w_j sont les vecteurs W_i et W_j respectivement. Les biais b_i et b_j sont ajoutés pour restaurer la symétrie. La fonction de coût est donnée par l'Équation 4.3 :

$$J = \sum_{i=1}^{|C|}\sum_{j=1}^{|C|} f(X_{ij})(w_i^T w_j + b_i + b_j - \log X_{ij})^2 \tag{4.3}$$

où f est une fonction de pondération, utilisée pour pondérer les occurrences des mots :

$$f(X_{ij}) = \begin{cases} (\frac{X_{ij}}{x_{max}})^{\alpha} & \text{si } X_{ij} < x_{max} \\ 1 & \text{sinon} \end{cases} \tag{4.4}$$

où x_{max} et $alpha$ sont fixés expérimentalement à 100 et $\frac{3}{4}$ pour traiter les paires rares des mots.

4.4 Intégration des RDMs dans les modèles de RI

4.5 Représentations distribuées des mots

4.5.1 Extensions des modèles de RI

Les modèles traditionnels de RI sont basés sur l'hypothèse d'indépendance de termes et adoptent la représentation en sac de mots pour représenter le contenu des documents

et des requêtes. Par conséquent, l'estimation du score de pertinence d'un document par rapport à une requête repose sur un appariement exact des termes simples partagés entre eux. Cependant, il est plus raisonnable de considérer les termes des documents qui sont sémantiquement proches à ceux de la requête. Pour ce faire, nous proposons une méthode pour exploiter les RDMs dans les modèles de RI. L'idée principale consiste à intégrer la similarité entre les vecteurs de termes, acquise des RDMs, dans les modèles de RI pour la prise en compte des termes similaires dans l'appariement des documents et des requêtes.

Nos extensions des modèles de RI reposent sur les termes similaires à ceux de la requête pour estimer les scores de pertinences des documents. La similarité entre les termes est calculée en utilisant la distance cosine entre leurs vecteurs, obtenus par apprentissage de RDMs en utilisant les modèles CBOW, Skip-gram, ou Glove. L'ensemble des termes similaires au terme de la requête w est noté $\mathcal{S}(w)$. Par définition, $w \in \mathcal{S}(w)$. Pour sélectionner les termes similaires à partir de document ou à partir du vocabulaire de la collection, nous reposons également sur un seuil, θ_s, et une limite supérieure k :

$$\mathcal{S}_d(w) = Topk(\{w' \in d, cos(w, w') \geq \theta_s\})$$

$$\mathcal{S}_C(w) = Topk(\{w' \in C, cos(w, w') \geq \theta_s\})$$

où $Topk$ est un opérateur qui retourne les k meilleurs termes similaires selon leurs similarité cosine. A noter que nous utilisons w pour designer le terme et son vecteur également. La distinction entre les deux devrait être clair à partir du contexte. L'utilisation conjointe d'un seuil et une limite supérieure sur le nombre de mots conservés permet d'être mieux contrôler l'ensemble de mots similaires à un terme donné et pour éviter d'être trop restrictif ou permissif si seulement un seuil ou une est utilisée limite supérieure.

chaque terme similaire w' peut être pondéré selon sa similarité à w comme suite :

$$\mathcal{A}(w, w', d) = \begin{cases} 1 & \text{if } w = w' \\ \lambda_d \cdot \frac{cos(w,w')}{\sum_{w'' \in \mathcal{S}_d(w)} cos(w,w'')} & \text{else} \end{cases} \tag{4.5}$$

$$\mathcal{A}(w, w', C) = \begin{cases} 1 & \text{if } w = w' \\ \lambda_C \cdot \frac{cos(w,w')}{\sum_{w'' \in \mathcal{S}_c(w)} cos(w,w'')} & \text{else} \end{cases} \tag{4.6}$$

où λ_d et λ_C sont des paramètres qui contrôlent l'importance des termes similaire par rapport l'appariement exacte de termes. Ces paramètre peuvent être optimisés en utilisant la validation croisée (voir Section 4.7.1). Dans ce qui suit, nous utilisons $\mathcal{S}$ pour désigner soit $\mathcal{S}_d$ ou $\mathcal{S}_C$, $\mathcal{A}(w, w', \cdot)$ pour noter soit $\mathcal{A}(w, w', d)$ ou $\mathcal{A}(w, w', C)$, et λ pour désigner soit λ_d ou λ_C.

En adoptant l'approche proposée par [Li & Gaussier 2012] dans le contexte de la RI translinguistique, chaque modèle de RI peut être étendu en considérant les termes similaires par intégration de la contribution pondérée de tous les termes similaires à ceux de la requête dans la RSV :

$$RSV(q, d) = \sum_{w \in q} A(w, q) \sum_{w' \in \mathcal{S}(w) \cap d} \mathcal{A}(w, w', \cdot) B(w', d, C) \tag{4.7}$$

Tableau 4.1: Exemple des ensembles de termes similaire obtenus pour la requête 1 ($\theta_s = 0.4$, $k = 4$ pour $\mathcal{S}_d$ et $k = 15$ pour $\mathcal{S}_C$). Le modèle CBOW est utilisé pour obtenir les vecteurs des termes.

(a) Termes similaires de document 19990908_AFP_ARB.0085.

terme w_q	$\mathcal{S}_d(w_q)$			
فنون arts	ثقاف culture			
عرض exhibition	عروض exhibitions			
موءسس institution	معهد institut			
اسلامي islamique				
عالم monde				
عربي arabe	مصر Egypte	سوري Syrien		
اثر effet	ناجم a abouti à			
رقص danse	راقص danseur	غناء chant	انغام mélodies	مسرح théâtre
موسيقي musique	اغاني chansons	غناءي musicale	مسرحي théâtral	كلاسيكي classique

(b) Termes similaires de la collection.

terme w_q	$\mathcal{S}_c(w_q)$			
فنون arts	تشكيلي beaux-arts	فلكلور folklore	رواءع Chefs-d'œuvre	فولكلور folklore
عرض exhibition	عروض exhibitions	يعرض il montre	استعراض exhibition	
موءسس institution	موءسسه institution	جمعي association	هيء organisation	مراكز centres
اسلامي islamique	دين religion	سلفيه Salafiste	وهابيه Wahhabisme	اخوانيه Frères musulmans
عالم monde	عالمي international	اوروبا Europe	اسيا Asie	خليج Golf
عربي arabe	عربيه arabe	عرب arabes	مغاربي Maghrébin	مصر Egypte
اثر effet	اعقاب séquelle	اعقب suivi	نتج a abouti à	جراء à la suite
رقص danse	راقص danseur	غناء chant	دبكه Dabke	انغام mélodies
موسيقي musique	موسيقا musique	اغاني chansons	معزوف morceau de musique	عازف musicien

Tableau 4.2: Exemple des ensembles de termes similaire obtenus pour $\mathcal{S}_d$, sélectionnés du document 19990908_AFP_ARB.0085 pour les termes **رقص** et **موسيقي** de la requête 1 en utilisant le modèle CBOW ($\theta_s = 0.4$, $k = 4$ and $\lambda = 0.4$).

terme w_q	رقص				موسيقي			
	w'	traduction	$cos(w_q, w')$	$\mathcal{A}(w_q, w', d)$	w'	traduction	$cos(w_q, w')$	$\mathcal{A}(w_q, w', d)$
$\mathcal{S}_d(w_q)$	رقص	danse	1	1	موسيقي	music	1	1
	راقص	danseur	0.6813	0.0801	اغاني	chansons	0.6557	0.0844
	غناء	chant	0.6538	0.0768	غناءي	musicale	0.5967	0.0768
	انغام	mélodies	0.6035	0.0709	مسرحي	théâtral	0.4289	0.0552
	مسرح	théâtre	0.4648	0.0546	كلاسيكي	classique	0.4246	0.0547

Tableau 4.3: Exemple des ensembles de termes similaire obtenus pour $\mathcal{S}_d$, sélectionnés à partir de la collection pour les termes **رقص** et **موسيقي** de la requête 1 en utilisant le modèle CBOW ($\theta_s = 0.4$, $k = 15$ et $\lambda = 1$).

terme w_q	رقص				موسيقي			
	w'	traduction	$cos(w_q, w')$	$\mathcal{A}(w_q, w', C)$	w'	traduction	$cos(w_q, w')$	$\mathcal{A}(w_q, w', C)$
$\mathcal{S}_c(w_q)$	رقص	danse	1	1	موسيقي	musique	1	1
	راقص	danseur	0.6813	0.0711	موسيقا	musique	0.6586	0.0654
	غناء	chant	0.6538	0.0682	اغاني	chansons	0.6557	0.0652
	دبكه	Debka	0.6129	0.0640	معزوف	morceau de musique	0.6376	0.0633
	انغام	mélodies	0.6035	0.0630	عازف	musicien	0.6355	0.0631

ou $A(w, q)$ et $B(w', d, C)$ sont données par les équation de chaque modèle (Équations 1.8, 1.11, 1.18, et 1.19).

Nous présentons une illustration de $\mathcal{S}_d$ et $\mathcal{S}_C$ sur la première requête de la collection de test (voir Figure 4.2) en utilisant le modèle CBOW. Tableau 4.1 illustre les termes sélectionnés pour cette requête avec θ_s et k fixés respectivement à 0.4 et 4 pour $\mathcal{S}_d$ et k à 15 pour $\mathcal{S}_C$. Comme prévu, les ensembles de termes similaires contiennent non seulement des termes sémantiquement liés à ceux de la requête, mais aussi leurs variantes morphologiques, en l'ocurrence le erreurs de racinisation légère, le pluriel irrégulier, et les variantes orthographiques (problème d'arabisation) :

- Erreurs de racinisation légère : **موءسس** et **موءسسه** (institution) ;
- Pluriel irrégulier de **عرض** : **عروض** (exhibitions) ;
- Variantes orthographiques : **فلكلور** et **فولكلور** (folklore) ;

Les Tableau 4.2 et Tableau 4.3 illustrent des exemples de normalisation de la relation de similarité entre les termes de la requête **رقص** (stem de danse) et **موسيقي** (stem de musique) avec leurs termes similaires en utilisant les stratégies de construction de ces derniers pour chaque document et à partir du vocabulaire de la collection respectivement. Les deux tableaux montrent que les valeurs prises par les fonctions $\mathcal{A}(w_q, w', d)$ et $\mathcal{A}(w_q, w', C)$ diffèrent selon la stratégie de construction de l'ensemble de termes similaires. Cela est expliqué par le fait que la plupart des documents pertinents apparient juste un sous-

ensemble des termes de $\mathcal{S}_c(w)$. Pour cela, la valeur du paramètre λ doit être plus grande dans $\mathcal{A}(w_q, w', C)$ par rapport à $\mathcal{A}(w_q, w', d)$.

Les extensions définies par Équation 4.7 ont pour but l'intégration, dans les modèles de RI de l'état-de-l'art, les relations sémantiques entre les termes. Nous étudions ci-dessous si ces extensions sont "valides" selon l'approche axiomatique d'appariement sémantique de termes (SMTCs), proposée [Fang & Zhai 2006] pour l'intégration des relations sémantiques dans les modèles de RI.

4.5.2 Validation théorique

L'approche axiomatique à la recherche d'information fournit un cadre théorique pour trouver un modèle optimal de RI dans l'espace des fonctions de RI. L'idée fondamentale de ce cadre axiomatique est de définir un ensemble de contraintes qui doivent être satisfaites par une fonction de RI pour être empiriquement efficaces [Fang *et al.* 2004, Clinchant & Gaussier 2011]. Cette approche consiste à lier la pertinence directement aux termes à travers des contraintes heuristiques qui sont utilisées pour évaluer l'optimalité d'un modèle de RI. Puisque l'intégration des similarités entre les termes est effectuée sans modifier la pondération du terme dans la requête et le document, nos extensions satisfont toutes les contraintes validées par leur modèle de base. Cependant, l'injection de la sémantique dans les modèles de RI repose sur les contraintes d'appariement sémantiques des termes STMC (*Semantic Term Matching Constraints*), proposées par Fang & Zhai [2006].

L'intégration de la similarité entre deux termes u et v repose sur une fonction symétrique de similarité $s(u, v) \in [0, +\infty[$. La fonction $s(u, v)$ doit donner une grande valeur aux termes plus similaires. Cette fonction de similarité doit vérifier les conditions : un terme w est plus similaire à u qu'à v si et seulement si $s(w, v) > s(w, u)$ et $\forall v \neq w \; s(w, w) > s(w, v)$. Dans ce travail, nous utilisons la distance cosinus dans l'espace positif ($[0, 1]$) pour mesurer la similarité entre les vecteurs des mots.

La contrainte SMTC1 nécessite une fonction qui permet de donner un score plus élevé au document contenant un terme plus similaire au terme de la requête. la SMTC1 est définie comme suit :

Contrainte 1 - SMTC1 *"Soit $q = w_q$ une requête contenant un seul terme w_q. Soient $d_1 = w_{d_1}$ et $d_2 = w_{d_2}$ deux documents formés d'un seul terme, où $w_q \neq w_{d_1}$ et $w_q \neq w_{d_2}$. Si $s(w_q, w_{d_1}) > s(w_q, w_{d_2})$, alors $RSV(q, d_1) > RSV(q, d_2)$."* [[Fang & Zhai 2006, page 117]].

Cette contrainte est satisfaite par toutes nos extensions puisque $cos(w_q, w_{d_1}) > cos(w_q, w_{d_2})$ ce qui implique $\mathcal{A}(w_q, w_{d_1}, d1) > \mathcal{A}(w_q, w_{d_2}, d_2)$ et $\mathcal{A}(w_q, w_{d_1}, C) > \mathcal{A}(w_q, w_{d_2}, C)$ et donc $RSV(q, d_1) > RSV(q, d_2)$ pour toutes les extensions considérées.

Selon la condition STMC2, la contribution d'un terme de la requête w_q au score de pertinence doit être plus importante que celle d'un terme similaire w. Quel que soit le nombre d'occurrences de ce dernier (w) dans le document. Cette contrainte est définie comme suit :

Contrainte 2 - SMTC2 *"Soient $q = w_q$ une requête constituée d'un seul terme et w*

est un terme tel que $s(w_q, w) > 0$. *Si* d_1 *et* d_2 *sont deux documents, avec* $l_{d_1} = 1$, $x_{w_q}^{d_1} = 1$, $l_{d_2} = k$, *et* $x_w^{d_2} = k$ $(k \geq 1)$, *alors* $RSV(q, d_1) > RSV(q, d_2)$*."*[[Fang & Zhai 2006, page 117]]

Selon la SMTC2, $B(w_q, d) = B(w, d)$ pour tous les modèles. De plus, puisque la somme dans la RSV de chaque modèle de RI implique un seul terme, on a :

$$\begin{aligned} & RSV(q, d_1) > RSV(q, d_2) \\ \rightleftarrows & 1 > \lambda \frac{cos(w, w_q)}{1 + cos(w, w_q)} \\ \rightleftarrows & \lambda < \frac{1 + cos(w, w_q)}{cos(w, w_q)} \end{aligned}$$

pour $\lambda = \lambda_d$ et $\lambda = \lambda_C$. Donc, cette contrainte donne une borne supérieure au paramètres λ_d et λ_C :

$$\lambda < \min_{q, w_q \in q, w \in S(w_q)} \frac{1 + cos(w, w_q)}{cos(w, w_q)} \tag{4.8}$$

où λ (respectivement S) désigne soit λ_d ou λ_C (respectivement S_d ou S_C).

La troisième contrainte indique que, *mutatis mutandis*, les documents qui "couvrent" plusieurs termes de la requête, même si par des termes similaires, devraient obtenir un score plus élevé que les documents couvrant moins de termes de la requête. Il est défini comme suit :

Contrainte 3 - SMTC3 *"Soient* $q = \{w_1, w_2\}$ *une requête contenant deux termes de même importance et* w_3 *un terme similaire telle que* $s(w_3, w_2) > 0$. *Soient* d_1 *et* d_2 *deux documents. Si* $l_{d_1} = l_{d_2} > 1$, $s(w_1, w_1) = s(w_2, w_2)$, $x_{w_1}^{d_1} = l_{d_1}$, *et* $x_{w_1}^{d_2} = l_{d_2} - 1$, $x_{w_3}^{d_2} = 1$, *alors* $RSV(q, d_1) < RSV(q, d_2)$. *"*[[Fang & Zhai 2006, page 117]]

Selon la SMTC3, $\mathcal{S}_{d_2}(w_2) = \{w_2, w_3\}$, $\mathcal{S}_C(w_2) = \{w_2, w_3\}$. Pour chaque extension, on a :

$$\text{RSV(q,d}_1) = B(w_1, d_1), \quad RSV(q, d_2) = B(w_1, d_2) + \lambda \frac{cos(w_2, w_3)}{1 + cos(w_2, w_3)}$$

$RSV(q, d_1) < RSV(q, d_2)$ ce qui est équivalent à :

$$\lambda > \frac{B(w_1, d_1) - B(w_1, d_2)}{B(w_3, d_2)} \frac{1 + cos(w_2, w_3)}{cos(w_2, w_3)}$$

La valeur de la borne inférieure ci-dessus sur λ dépend de l'extension considérée, des termes de la requête et de leurs nombres d'occurrences dans les documents. Nous ne connaissons aucune borne inférieure raisonnable qui peut être dérivée de l'équation ci-dessus et recourir à la validation croisée dans nos expériences pour déterminer la meilleure valeur de λ, à partir de 0 jusqu'à la limite supérieure fournie par SMTC2.

Pour résumer, les extensions proposées satisfont la contrainte SMTC1 inconditionnellement alors que la contrainte SMTC2 fournit une borne supérieure sur le paramètre λ. La

situation est moins claire avec la contrainte SMTC3 car aucune limite inférieure générale ne peut être dérivée. Nous avons recours à la validation croisée dans nos expériences pour définir λ, avec la garantie que SMTC2 est satisfait. Nous ne vérifions cependant pas que SMTC3 est satisfait car cela prendrait trop de temps.

4.6 Intégration des RDMs dans les modèles PRF

Pour améliorer la performance d'expansion de la requête pour la RI en langue arabe, nous proposons une méthode basée sur l'exploitation des RDMs dans les modèles PRF (Pseudo-Relevance Feedback). L'idée principale de notre méthode consiste à combiner la distribution des termes d'expansion dans l'ensemble des documents d'expansion avec leurs similarités à la requête initiale. Le processus de notre méthode se compose de 7 étapes principales :

1. Sélection de l'ensemble des documents de feedback (document d'expansion) $F = \{d_1, ..., d_k\}$ du top k documents retrouvés. Les termes de F sont appelés des termes candidats d'expansion ;
2. Pour chaque terme de F, nous estimons sa similarité à la requête initiale en utilisant les RDMs. Les similarités calculées sont transformées en des probabilités pour les intégrer dans les modèles PRF ;
3. Pour chaque terme de F, nous calculons sa distribution (poids) dans l'ensemble F en utilisant un modèle standard de PRF ;
4. Pour chaque terme de F, nous calculons son poids modifié en multipliant ces deux poids estimés dans l'étape 1 et l'étape 3 ;
5. Sélection des meilleurs n termes (termes d'expansion) de F selon leurs poids modifiés (obtenus dans l'étape 4) ;
6. Pondération des termes d'expansion et modification de la requête ;
7. Classement des documents en utilisant la requête modifiée ;

Après avoir sélectionné l'ensemble de top k documents d'expansion $F = \{d_1, ..., d_k\}$, notre méthode repose sur deux techniques pour calculer la similarité entre un terme candidat d'expansion et la requête initiale. La première technique consiste à estimer le vecteur de la requête à partir des vecteurs des termes constituants. Cette technique est basée sur l'idée que l'addition des vecteurs des termes donne une composition sémantique des termes constituants des documents et des requêtes [Vulić & Moens 2015]. Ce vecteur est donné par :

$$\vec{q} = \sum_{w_q \in q} \frac{x_{w_q}}{l_q} \cdot \vec{w}_q \tag{4.9}$$

où w_q est un terme de la requête, x_{w_q} est le poids de w_q dans la requête q, et l_q est le nombre de termes de q. Donc, la similarité entre un terme candidat d'expansion w et la

requête q en utilisant la distance cosinus entre leurs vecteurs est donnée par :

$$Sim_{comp}(\vec{w}, \vec{q}) = cos(\vec{w}, \vec{q}) \tag{4.10}$$

La deuxième technique consiste à calculer la similarité moyenne d'un terme candidat d'expansion avec les termes de la requête. Cette similarité est obtenue en utilisant la formule suivante :

$$Sim_{avg}(w, q) = \frac{1}{l_q} \sum_{w_q \in q} cos(\vec{w}, \cdot\vec{w}_q) \tag{4.11}$$

où l_q est le nombre de termes de la requête q. Ensuite, nous utilisons la fonction *softmax* pour normaliser et transformer les similarités obtenues en des probabilités pour faciliter leur intégration dans les modèles PRF. La probabilité d'un terme candidat d'expansion, étant donné la requête q et l'ensemble F, est obtenue par :

$$P(w|q, F) = \frac{exp(Sim(w, q))}{\sum_{w \in F} exp(Sim(w, q))} \tag{4.12}$$

où $Sim(w, q)$ est la similarité entre un terme candidat d'expansion w et la requête q, calculée en utilisant l'Équation 4.10 ou l'Équation 4.11.

Le poids modifié d'un terme candidat d'expansion est obtenu en multipliant sa probabilité $P(w|q, F)$ par sa distribution dans l'ensemble F. Cette dernière est obtenue en utilisant un modèle standard de PRF. Donc, le poids modifié d'un terme candidat d'expansion w est donné par :

$$F_s(w) = F(w) \cdot P(w|q, F) \tag{4.13}$$

où $F(w)$ est la distribution de w dans F, obtenue en utilisant un modèle standard de PRF (voir Section 1.2.5.1 du Chapitre 1). Pour calculer $F(w)$ nous avons utilisé le modèle de divergence de Kullback-Liebler KLD [Carpineto *et al.* 2001], Bo2 de la famille DFR [Amati 2003], et le modèle Log-Logistique LL de la famille des modèles de l'information [Clinchant & Gaussier 2013].

Après la sélection des meilleurs n termes d'expansion selon leurs poids obtenus en utilisant l'Équation 4.13. Les poids finaux des termes de requête et des termes d'expansion sont calculés en utilisant l'équation suivante :

$$x'_{w_q} = \frac{x_{w_q}}{\max_w x_{w_q}} + \beta \cdot \frac{F_s(w)}{\max_w F_s(w)} \tag{4.14}$$

où x'_{w_q} est le poids du de w dans la nouvelle requête, β est un paramètre qui contrôle l'importance des termes d'expansion, et $F_s(w)$ est la pondération que nous avons introduit pour intégrer la similarité dans les modèles PRF (Équation 4.13). Pour la prise en compte de la similarité seule dans le processus d'expansion, nous substituons $F_s(w)$ par $P(w|q, F)$ dans l'Équation 4.14 :

$$x'_{w_q} = \frac{x_{w_q}}{\max_w x_{w_q}} + \beta \cdot \frac{P(w|q, F)}{\max_w P(w|q, F)} \tag{4.15}$$

La dernière étape de notre méthode consiste à retrouver les documents pour la nouvelle requête en utilisant un modèle de RI.

4.7 Evaluations et résultats

4.7.1 Méthode d'évaluation

Pour évaluer la performance des deux méthodes proposées pour l'exploitation des RDMs dans les modèles de RI et les modèles PRF, les évaluations sont effectuées sur la collection standard TREC 2002/2001 en utilisant la plateforme Terrier 3.5 [1]. Pour les 75 requêtes de la collection de test, nous avons utilisé les champs titre et description (voir Figure 4.2). En plus des mesures MAP et P10, utilisées pour évaluer les extensions de modèles de RI et leurs modèles de base, nous avons utilisé l'indice de robustesse RI (*Robustness Index*) pour évaluer les extensions des modèles PRF. L'indice RI est défini par $\frac{Q_+ - Q_-}{|Q|} \in [-1, 1]$ où $|Q|$ est le nombre total des requêtes, Q_+ et Q_- désignent le nombre de requêtes dont la performance est améliorée et le nombre de requêtes dont la performance, est diminuée, respectivement. Pour toutes les évaluations, nous avons utilisé la méthode de racinisation légère Farasa [Abdelali *et al.* 2016].

<top>
<num> 1
<title>
فنون العرض و المؤسسات الاسلامية في العالم العربي
<desc>

ما هو اثر المؤسسات الاسلامية على فنون العرض مثل الرقص و الموسيقى في العالم العربي؟
<narr>
المقالات المتعلقة بالفنون الرياضية او التشكيلية و بفنون العرض خارج العالم العربي او
بالسلوكيات الدينية خارج اطار فنون العرض او بالديون و القروض المالية لا علاقة لها بالموضوع
</top>

FIGURE 4.2: Exemple de requête de la collection TREC 2002/2001

Pour les extensions des modèles de RI, les ensembles de termes similaires sont construits en sélectionnant les $k = \{2, 4, 10, 15, 20\}$ termes similaires pour chaque terme de la requête et θ_s est fixé à 0.4. Puisque le classement des documents à base des extensions qui reposent sur la sélection des termes similaires pour chaque document est coûteux, nous avons reclassés, les 2000 documents retournés par le modèle de base, pour chaque requête en utilisant ces extensions.

Pour les extensions des modèles PRF et leurs modèles standards, le nombre de documents d'expansion k et le nombre de termes d'expansion sont variés parmi $\{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$. Pour comparer les extensions des modèles PRF avec les méthodes d'expansion de requêtes basées RDMs, nous avons sélectionné la méthode VEXP, proposée par ALMasri et al. [2016]. Pour cela, nous avons varié le nombre de termes d'expansion parmi

1. http ://terrier.org/download/

$\{2, 5, 10, 15, 20, 30, 40, 50\}$ et le paramètre de pondération des termes similaires entre 0.1 et 1. Les valeurs optimales sont sélectionnées selon la meilleure valeur de MAP.

De plus, nous avons effectué des tests de significativité des résultats obtenus en utilisant le test bilatéral de Student et nous avons attaché un caractère à la valeur de MAP lorsque la *p-valeur*< 0.1. Le Tableau 4.4 présente les modèles de RI, leurs paramètres et les valeurs utilisés pour la validation croisée.

Tableau 4.4: Les valeurs des paramètres utilisées pour la validation croisée.

Modèle	Paramètre	Valeurs
LGD **SPL**	c	$0.1, 0.5, 0.7, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0,$ $4.5, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 20.0$
LM	μ	$10, 25, 50, 75, 100, 200, 300, 400, 500, 600, 700,$ $800, 900, 1000, 1500, 2000, 2500, 3000, 4000, 5000$
BM25	b	$0.1, 0.2, 0.3, 0.35, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8, 0.9$ $1.0, 1.25, 1.5, 1.75, 2.0, 2.25, 2.5, 2.75, 3.0$
IR extentions	λ	$0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, ..., 1.5$
PRF models	β	$0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, ..., 2$

Pour les modèles de RI basés RDMs, nous avons fixé leurs paramètres aux valeurs qui optimisent la mesure MAP :

- $c = 0.7$ et $c = 1$ pour les extensions des modèles LGD et SPL respectivement ;
- $\mu = 200$ pour les extensions du modèle LM ;
- $b = 0.35$ pour les extensions du modèle BM25 ;
- les valeurs optimales données par la validation croisée du paramètre λ dépendent du modèle de RI et de la stratégie de sélection des termes similaire :
  - $\lambda_d \in [0.3, 0.5]$, la valeur exacte dépends du modèle de RI ;
  - $\lambda_C \in [0.6, 1.2]$, la valeur exacte dépends également du modèle de RI ;
  - $k = 4$ pour $\mathcal{S}_d$ et 15 pour $\mathcal{S}_C$.

Pour l'apprentissage des RDMs, nous avons collecté un grand corpus contenant 216M mots à partir des collections arabes disponibles sur Internet, y compris les corpus de classification de textes BBC, CNN, et OSAC [2], le corpus de la collection standard TREC 2002/2001 [3], et WORTSHATZ [4]. Pour les trois modèles utilisés pour l'apprentissage des RDMs (CBOW, Skip-gram [5], et Glove [6] modèles), la taille du contexte et la dimension des vecteurs sont fixées à 10 et 300 respectivement.

2. https ://sourceforge.net/projects/ar-text-mining/files/Arabic-Corpora/
3. catalogue de LDC numéro $LDC2001T55$
4. http ://www.cls.informatik.uni-leipzig.de/langs/ara
5. https ://code.google.com/archive/p/word2vec/
6. https ://nlp.stanford.edu/projects/glove/

4.7.2 Résultats obtenus pour les modèles de RI

4.7.2.1 Évaluation des modèles de RI basés RDMs

Pour évaluer la performance des extensions proposées pour l'intégration des RDMs dans les modèles de RI, nous avons utilisé les deux stratégies de construction de l'ensemble des termes similaires. Ces extensions sont comparées avec leurs modèles de base en utilisant les trois modèles de RDMs (CBOW, Skip-gram et Glove).

Le Tableau 4.7 présente les résultats de comparaison des extensions des modèles de RI avec leurs modèles de base. D'après les résultats obtenus, les extensions proposées améliorent significativement la performance de leurs modèles de base, pour les deux stratégies de construction de l'ensemble des termes similaires et les trois modèles de RDMs. De plus, la différence en termes de performance entre les modèles de RDMs (SKIP-gram, CBOW, et Glove) n'est pas statistiquement significative et ils ont atteint des performances comparables. La stratégie de construction de l'ensemble des termes similaires pour chaque document aboutit de légères améliorations par rapport à sa construction à partir du vocabulaire de la collection. La meilleure performance est obtenue par l'intégration de la similarité entre les termes dans le modèle SPL.

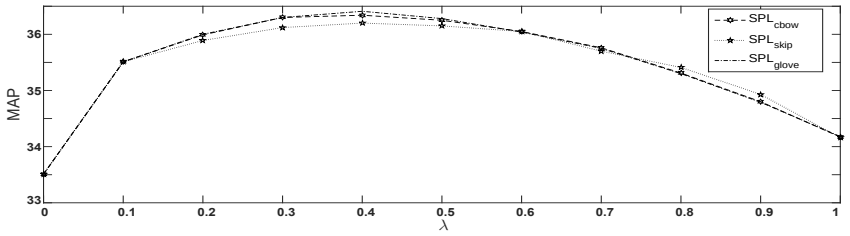
Tableau 4.5: Résultats obtenus pour les extensions basées sur les RDMs et leurs modèles de base. Pour le test de significativité, b = meilleur que le modèle de base, c = meilleur que CBOW, s = meilleur que Skip-gram, et g = meilleur que Glove.

Modèle	Modèle de base		$\mathcal{S}_d$						$\mathcal{S}_C$					
			CBOW		SKIP-gram		Glove		CBOW		SKIP-gram		Glove	
Mesure	MAP	P10	MAP	P10	MAP	P10	MAP	P10	MAP	P10	MAP	P10	MAP	P10
LGD	32.42	47.33	34.63^b	49.60	34.36^b	47.87	34.15^b	49.20	34.09^b	49.87	$33{,}98^b$	49.47	34.25^b	50.53
SPL	33.51	50.67	36.34^b	51.60	36.2^b	51.47	$36{,}41^b$	52.3	36.15^b	52.53	$36{,}09^b$	52.00	$36{,}27^b$	51.87
BM25	33.42	49.60	35.47^b	51.20	35.38^b	51.60	$35{,}59^b$	52.00	$35{,}16^b$	51.20	35.12^b	51.73	$35{,}13^b$	51.72
LM	31.15	46.39	33.65^b	48.53	$33{,}51^b$	47.60	33.47^b	50.00	$33{,}3^b$	48.07	33.37^b	48.11	$33{,}4^b$	48.53

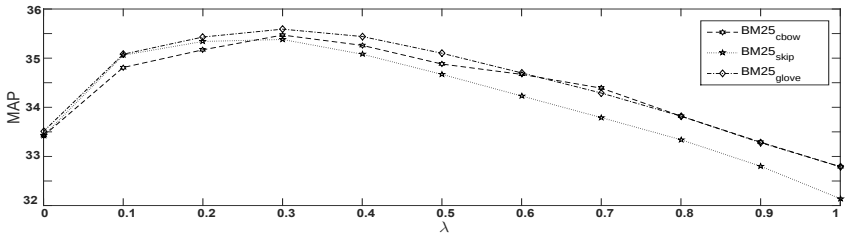
4.7.2.2 Sensibilité au paramètre λ

Pour étudier la sensibilité de la performance des extensions de modèles de RI au paramètre λ, nous avons tracé la courbe de performance de MAP des extensions des modèles SPL et BM25 en utilisant les deux stratégies de construction de l'ensemble des termes similaires. Le but de cette étude est de vérifier si les valeurs optimales du paramètre λ sont en concordance avec les bornes déterminées lors de la validation théorique.

Les Figure 4.3 et Figure 4.4 présentent les courbes de sensibilité au paramètre λ pour les extensions basées sur la construction de l'ensemble des termes similaires pour chaque document et à partir du vocabulaire de la collection respectivement. Les deux figures montrent que la performance des extensions proposées est affectée significativement par les valeurs du paramètre λ. En concordance avec les résultats de la validation théoriques des extensions des modèles de RI (voir Section 4.5.2), les figures montrent que les valeurs

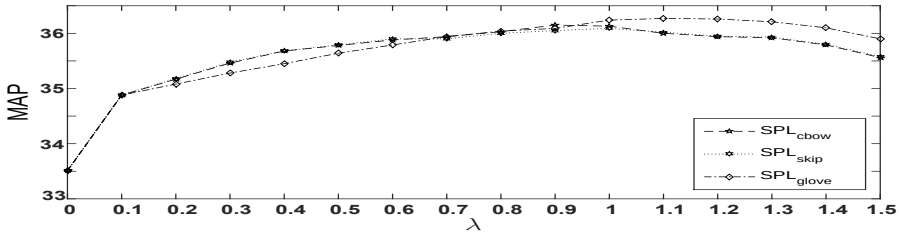


(a) Extensions du modèle SPL.

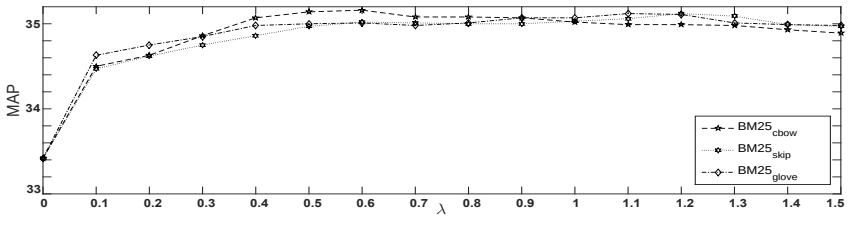


(b) Extensions du modèle BM25.

FIGURE 4.3: Effet du paramètre λ_d sur la performance de MAP pour les extensions de modèles SPL et BM25 en utilisant $\mathcal{S}_d$.



(a) Extensions du modèle SPL.



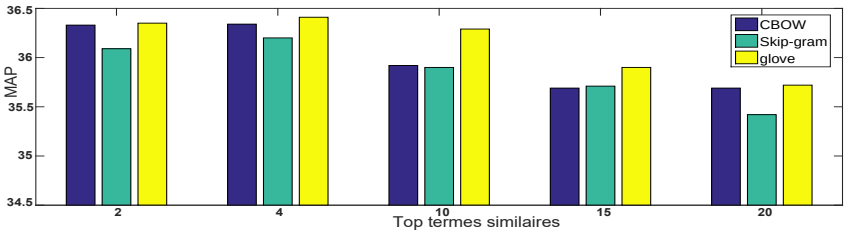
(b) Extensions du modèle BM25.

FIGURE 4.4: Effet du paramètre λ_d sur la performance de MAP pour les extensions de modèles SPL et BM25 en utilisant $\mathcal{S}_C$.

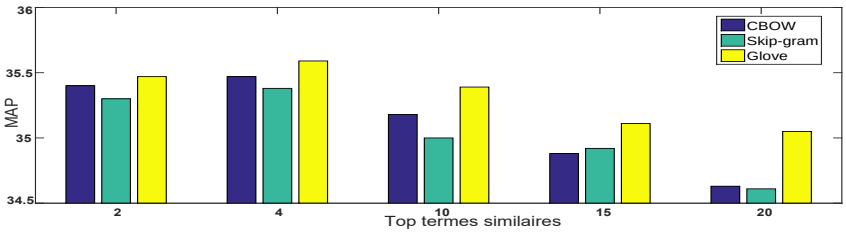
optimales du paramètre λ dépendent de la pondération du terme dans le document (utilisé par le modèle de RI), et l'ensemble des termes similaires. En outre, les valeurs optimales du paramètre λ diffèrent selon la stratégie de construction des termes similaires.

4.7.2.3 Effet de la taille de l'ensemble des termes similaires

Pour étudier la sensibilité des extensions proposées à la taille de l'ensemble des termes similaires, nous avons varié le nombre de termes similaires utilisés parmi $\{2, 4, 10, 15, 20\}$ pour les deux stratégies de construction de leurs ensembles. Les valeurs de MAP sont sélectionnées à base de la valeur optimale du paramètre λ pour chaque valeur du nombre de termes similaires utilisés.



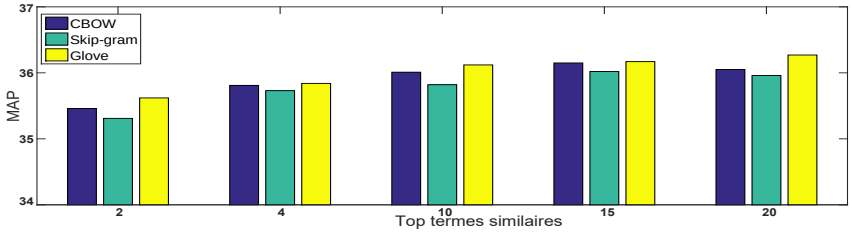
(a) Extensions du modèle SPL.



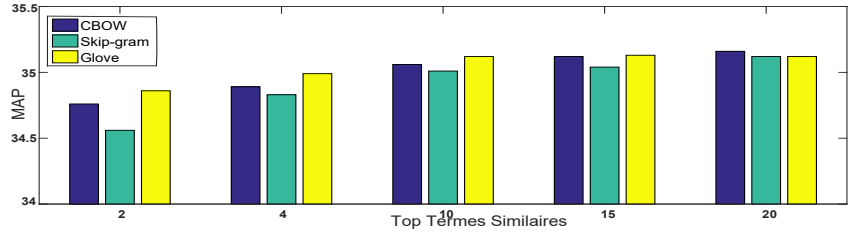
(b) Extensions du modèle BM25.

FIGURE 4.5: Effet de la taille de l'ensemble de termes similaire sur la performance de MAP des extension de modèles SPL et BM25 en utilisant $\mathcal{S}_d$.

Les Figure 4.5 et Figure 4.6 illustrent l'effet de la taille des termes similaires pour la stratégie de construction de leurs ensembles pour chaque document et à partir du vocabulaire de la collection respectivement. Les deux figures montrent que la valeur optimale de la taille de l'ensemble des termes similaires dépend de la stratégie de leur sélection. La stratégie de sélection de l'ensemble des termes similaires à partir du vocabulaire de la collection requiert plus de termes similaires par rapport à la stratégie de sélection des termes similaires pour chaque document. Cela est expliqué par le fait que les documents apparient juste un sous-ensemble de termes similaires sélectionnés à partir du vocabulaire de la collection. De plus, l'utilisation de seuils sur la similarité et le filtrage des termes les plus similaires ont un effet positif sur la performance des extensions pour les deux stratégies de construction de l'ensemble des termes similaires. En outre, de légères améliorations sont obtenues par les extensions basées sur le modèle Glove et CBOW pour la plupart des nombres de termes similaires sélectionnés.



(a) Extensions du modèle SPL.



(b) Extensions du modèle BM25.

FIGURE 4.6: Effet de la taille de l'ensemble de termes similaire sur la performance de MAP des extension de modèles SPL et BM25 en utilisant $\mathcal{S}_C$.

4.7.2.4 Comparaison des extensions avec l'état de l'art

Pour évaluer nos extensions avec les modèles et les méthodes d'injection de la sémantique dans la RI en langue arabe, nous avons sélectionné l'approche d'indexation sémantique basée sur l'ontologie ArabicWordNet (SI) [Abderrahim *et al.* 2016] et les trois modèles de langue basée sur les RDMs, y compris le modèle de langue basé Skip-gram de RI monolingue, noté LM+WE [Vulić & Moens 2015], le modèle neuronal de langue de translation NTLM [7] [Zuccon *et al.* 2015], et le modèle de langue généralisé GLM [8] [Ganguly *et al.* 2015]. Toutes les extensions des modèles de RI (nos extensions, LM+WE, NTLM, et GLM) sont évaluées en utilisant le modèle Skip-gram. Pour l'approche d'indexation sémantique (SI), nous avons évalué la méthode proposée par Abderrahim et al. [2016] en utilisant une version récente de l'ontologie ArabicWordNet AWN [9] [Abouenour *et al.* 2013]. Au lieu d'indexer les racines des termes, dont les concepts non trouvés dans l'AWN, nous avons indexé leurs stems qui sont produits en utilisant Farasa.

Les paramètres des modèles de l'état de l'art sont fixés aux valeurs qui optimisent la mesure MAP :

- le paramètre μ est fixé à 200 pour les modèles LM+WE et NTLM ;
- le paramètre $\lambda \in [0.1, 0.9]$ est fixé à 0.2 pour le modèle LM+WE ;
- les paramètres λ, α, et β du modèle GLM sont variés entre $[0.1, 0.4]$ sous la condition $\lambda + \alpha + \beta < 1$. Les valeurs optimales sont 0.2, 0.2, et 0.3 pour λ, α, et β respectivement ;

7. https ://github.com/ielab/adcs2015-NTLM
8. https ://github.com/gdebasis/wvlm/
9. http ://arabic.emi.ac.ma/ibtikarat/ ?q=Resources

- pour l'approche d'indexation sémantique SI, nous avons optimisé les paramètres des modèles de RI en utilisant la validation croisée (Tableau 4.4).

Tableau 4.6: Résumé des résultats de comparaison de nos extensions avec l'approche SI et les modèles LM+WE, NTLM, et GLM. Pour le test de significativité, b = meilleur que le modèle de base, s = meilleur que l'approche SI, et w = meilleur que les extensions RDMs du modèle de langue (LM+WE, NTLM, et GLM).

Modèle	Modèle de base		SI		LM+WE		NTLM		GLM		$\mathcal{S}_d$		$\mathcal{S}_C$	
											SKIP-gram		SKIP-gram	
Mesure	MAP	P10	MAP	P10	MAP	P10	MAP	P10	MAP	P10	MAP	P10	MAP	P10
LGD	32.42^{s}	47.33	29.59	45.60	−−	−−	−−	−−	−−	−−	$34.36^{b,s,w}$	47.87	$33{,}98^{b,s,w}$	49.47
SPL	33.51^{s}	50.67	28.98	45.73	−−	−−	−−	−−	−−	−−	$36.2^{b,s,w}$	51.47	$36{,}09^{b,s,w}$	52.00
BM25	33.42^{s}	49.60	30.46	47.73	−−	−−	−−	−−	−−	−−	$35.38^{b,s,w}$	51.60	$35.12^{b,s,w}$	51.73
LM	31.15^{s}	46.39	28.05	45.73	31.97^{s}	46.67	32.35^{s}	47.13	32.27^{s}	46.83	$33{,}51^{b,s}$	47.60	$33.37^{b,s}$	48.11

Le Tableau 4.6 présente les résultats de comparaison de nos extensions avec les modèles de l'état de l'art pour l'intégration des RDMs et l'approche d'indexation sémantique pour la RI en langue arabe. Les résultats obtenus montrent que les modèles de base et leurs extensions basées sur les RDMs améliorent significativement la performance de l'approche d'indexation sémantique basée sur l'AWN. Ceci est expliqué par le fait que le AWN est très limitée. En effet, elle contient seulement 11.269 synsets couvrant 23.481 mots (noms, verbes, adjectifs et adverbes). En outre, nos extensions des modèles LGD, SPL, et BM25 montrent des améliorations significatives par rapport aux modèles de langue basés RDMs de l'état de l'art (LM+WE, NTLM, et GLM). Bien que nos extensions du modèle de langue donnent de meilleurs résultats par rapport aux modèles LM+WE, NTLM, et GLM, la différence en termes de performances de leurs MAP n'est pas statistiquement significative. Par ailleurs, les modèles LM+WE, NTLM et GLM n'ont pas entrainé une amélioration substantielle par rapport au modèle de langue de base (LM). Ce dernier résultat peut s'expliquer par le fait que, dans le contexte de LM+WE, la composition des vecteurs des documents et des requêtes entraine une perte d'information sur les statistiques de termes qui sont utilisés comme des signaux de pertinence. Pour le modèle NTLM, les probabilités de traduction sont calculées sur le vocabulaire de la collection ce qui conduit à un faible poids de traduction des termes de requête originaux les termes similaires. Dans le modèle GLM, tous les termes du document sont autorisés à contribuer à son score de pertinence.

4.7.3 Résultats obtenus pour les modèles PRF

4.7.3.1 Évaluation des extensions PRF basées RDMs

Pour évaluer l'impact de l'intégration de la similarité entre les termes d'expansion et la requête initiale, nous avons comparé les extensions proposées en utilisant les deux fonctions de similarités Sim_{comp} et Sim_{avg}, avec leurs modèles PRF standards. Pour le modèle de RDMs, nous avons sélectionné le modèle Glove. Les autres modèles de RDMs

(CBOW et Skip-gram) aboutissent à des performances comparables à celles du modèle Glove (Tableau 4.8). Les valeurs de MAP sont sélectionnées selon les valeurs optimales du nombre de documents d'expansion (parmi $\{10, 20\}$) et le nombre de termes d'expansion (parmi $\{60, 100\}$). Pour le modèle de base de RI, nous avons sélectionné le modèle SPL de la famille de modèle d'information. Pour la méthode d'expansion VEXP, les valeurs optimales du nombre de termes d'expansion et celle du paramètre de leur pondération sont parmi $\{5, 10\}$ et entre $[0.1, 0.3]$ respectivement.

Tableau 4.7: Résultats de comparaison des extensions PRF proposées avec leurs modèles PRF de base, la méthode d'expansion VEXP et le modèle de base SPL. Les exposants 1, 2, 3 et 4 désignent une amélioration significative par rapport aux modèles de base SPL, méthode VEXP, les méthodes Sim_{comp} et Sim_{avg} et aux modèles PRF de base respectivement.

Modèle \ Mesure	MAP	P10	RI
Modèle de base (SPL)	33.51	50.67	−−
VEXP	35.67^{1}	51.47	0.34
Sim_{comp}	$38.57^{1,2}$	52.57	0.57
Sim_{avg}	$38.65^{1,2}$	52.72	0.57
KLD	$38.76^{1,2}$	52.27	0.57
KLD_Sim_{comp}	$40.45^{1,2,3,4}$	54.03	0.65
KLD_Sim_{avg}	$40.62^{1,2,3,4}$	54.03	0.65
Bo2	$39.7^{1,2}$	52.8	0.59
Bo2_Sim_{comp}	$41{,}03^{1,2,3,4}$	54,8	**0,68**
Bo2_Sim_{avg}	$\mathbf{41{,}11}^{1,2,3,4}$	**55,07**	**0,68**
LL	$40.71^{1,2,3}$	53.07	0.62
LL_Sim_{comp}	$41.01^{1,2,3}$	54.00	**0.68**
LL_Sim_{avg}	$41.04^{1,2,3}$	54.40	**0.68**

Le Tableau 4.7 présente les résultats obtenus pour les extensions PRF, leurs modèles de base, et le modèle de base de RI SPL. Les résultats obtenus montrent que nos extensions des modèles PRF et leurs modèles de base améliorent significativement le modèle de base de RI (SPL). En concordance avec les résultats de [Roy *et al.* 2016], les modèles PRF standards, leurs extensions basées RDMs, les deux modèles Sim_{comp} et Sim_{avg} (pondération des termes d'expansion à base de leur similarité seulement) aboutissent à des améliorations significatives par rapport à la méthode d'expansion de requête VEXP [ALMasri *et al.* 2016]. Les modèles standards PRF assurent des meilleures performances par rapport aux deux modèles Sim_{comp} et Sim_{avg}, en particulier le modèle de base LL a montré une amélioration statistiquement significative par rapport aux deux. En outre, des améliorations statistiquement significatives sont obtenues par nos extensions Sim_{comp} et Sim_{avg} des modèles KLD et Bo2 par rapport à leurs modèles PRF de base. Cependant, les meilleurs résultats sont obtenus par les extensions des modèles Bo2 et LL. Ces deux

dernières extensions améliorent la performance du modèle de base (SPL) par 22% et 68% pour la mesure MAP et l'indice RI respectivement. Malgré que les deux extensions (Sim_{comp} et Sim_{avg}) du modèle LL améliorent leur modèle de base, la différence en termes de performance n'est pas statistiquement significative. En outre, le calcul de la similarité moyenne entre les termes d'expansion et les termes de la requête initiale donne une performance légèrement supérieure que l'utilisation de composition du vecteur de la requête.

4.7.3.2 Impact du nombre de termes d'expansion

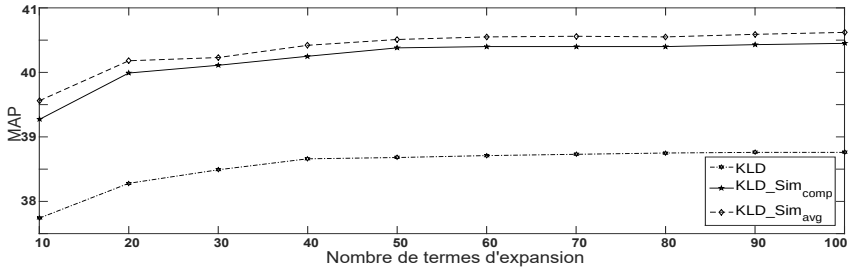
Pour étudier l'impact du nombre de termes d'expansion sur la performance des extensions des modèles PRF et les deux modèles Sim_{comp} et Sim_{avg}, nous avons fixé le nombre des documents d'expansion k à 10 et nous avons varié le nombre de termes d'expansion n dans l'ensemble $\{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$.

La Figure 4.7 présente la sensibilité au nombre des termes d'expansion des extensions PRF proposées et leurs modèles de base. La figure montre que la performance des extensions PRF et leurs modèles de base s'améliorent en augmentant le nombre de termes d'expansion. En outre, pour toutes les valeurs choisies de n (nombre de termes d'expansion), les extensions proposées surpassent leurs modèles de base. Les extensions Sim_{comp} et Sim_{avg} aboutissent à des performances comparables, en particulier lorsque le nombre de termes d'expansion est grand (> 10). Pour la plupart des extensions PRF et leurs modèles PRF de base, la performance reste stable pour un nombre de termes d'expansion $n \geq 60$. Par ailleurs, de légères améliorations ont été obtenues pour les extensions du modèle LL lorsque le nombre de termes d'expansion est suffisamment grand ($n \geq 60$). Cette dernière constatation s'explique par le fait que le modèle LL satisfait la plupart des contraintes qu'un modèle PRF devrait satisfaire pour être empiriquement efficace [Clinchant & Gaussier 2013].

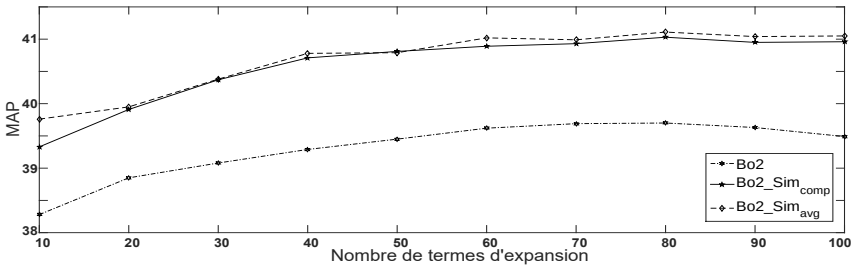
4.7.3.3 Comparaison des modèles de RDMs pour l'extension des modèles PRF

Pour déterminer le meilleur modèle de RDMs en termes de performance, nous avons comparé les modèles CBOW, Skip-gram, et Glove pour l'intégration de similarité dans les modèles PRF (KLD, Bo2, et LL) et la méthode d'expansion de requête VEXP.

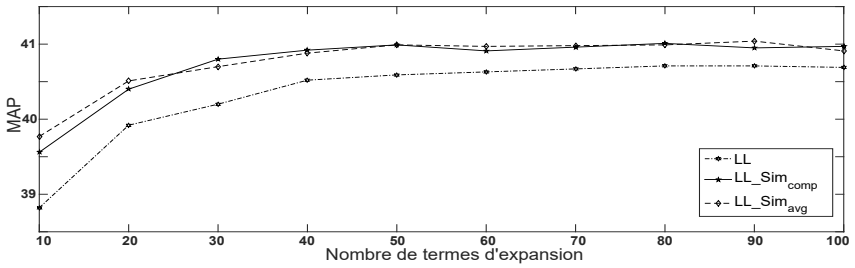
Le Tableau 4.8 illustre les résultats de comparaison des modèles CBOW, Skip-gram, et Glove pour l'intégration de similarité dans les modèles PRF. Globalement, les résultats de comparaison montrent que, pour chaque extension PRF et la méthode d'expansion VEXP, la différence en termes de performance de MAP entre les trois modèles de RDMs n'est pas statistiquement significative. Bien que les meilleures valeurs de MAP et de P10 sont obtenues par l'intégration de similarité basée Skip-gram dans le modèle Bo2_Sim_{avg}, les extensions PRF atteignent des performances similaires pour les trois modèles de RDMs. Par ailleurs, la différence en termes de performance de MAP entre les extensions Sim_{avg} et Sim_{comp} n'est pas statistiquement significative.



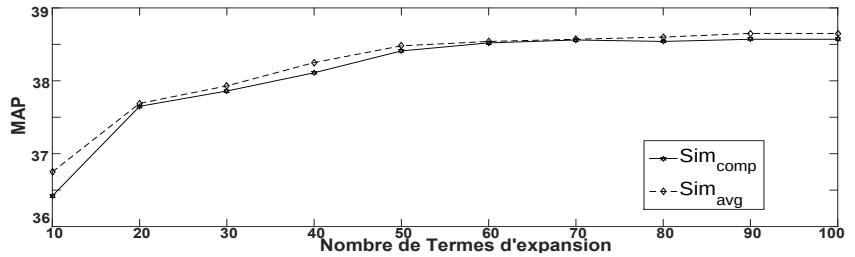
(a) Modèle KLD.



(b) Modèle Bo2.



(c) Modèle LL.



(d) Modèles Sim_{comp} et Sim_{avg} .

FIGURE 4.7: Impact du nombre de termes d'expansion sur la performance de MAP pour les extensions des modèles PRFs, leurs modèles de base et les deux modèles Sim_{comp} et Sim_{avg}.

Tableau 4.8: Résultats de comparaison de performance des modèles Glove, Skip-gram et CBOW pour les extensions des modèles PRF et la méthode d'expansion requête VEXP.

RDM	Glove		Skip-gram		CBOW	
Modèle \ Mesure	MAP	P10	MAP	P10	MAP	P10
VEXP	35.67	51.47	35.79	51.47	35.82	51.33
Sim_{comp}	38.57	52.57	38.64	52.65	38.54	52.65
Sim_{avg}	38.65	52.72	38.70	52.67	38.63	52.65
KLD_Sim_{comp}	40.45	54.03	40.41	53.47	40.16	53.73
KLD_Sim_{avg}	40.62	54.03	40.71	53.73	40.67	53.73
Bo2_Sim_{comp}	41.03	54.8	41.18	54.40	41.09	54.67
Bo2_Sim_{avg}	**41.11**	**55.07**	**41.26**	54.67	**41.21**	**55.07**
LL_Sim_{comp}	41.01	54.00	40.95	53.73	41.05	54.33
LL_Sim_{avg}	41.04	54.40	41.04	**54.93**	41.12	54.80

4.7.4 Comparaison des extensions de modèles de RI et de PRF

Pour comparer les extensions des modèles de RI et celles des modèles PRF, nous avons évalué le modèle de base SPL, tous les modèles PRF, l'extension SPL_Glove basé sur la construction de l'ensemble de termes similaires pour chaque document, et les modèles de dépendance de termes SPL_CT (termes croisés) et SPL_MWT (termes complexes). Pour toutes les extensions évaluées, nous avons utilisé le modèle Glove pour la RDMs. Le but principal de cette évaluation est de comparer les modèles basés sur l'intégration des similarités entre les termes et les modèles de dépendance de termes étudiés dans le chapitre précédent.

Le Tableau 4.9 présente les résultats de comparaison des extensions des modèles PRF, celles des modèles de RI pour l'intégration de la similarité entre les termes en utilisant les RDMs, et les extensions basées sur l'intégration des dépendances de termes. Les résultats obtenus montent que :

- L'extension basée sur l'intégration de similarité entre les termes en utilisant les RDMs (SPL_Glove) aboutissent à de meilleurs résultats que celle basée sur l'intégration des termes croisés (SPL_CT) et des termes complexes (SPL_MWT). Donc, l'exploitation des architectures simples pour la représentation des termes (RDMs) dans les modèles de RI assure de meilleurs résultats que ceux qui sont basés sur pipeline complexes pour l'extraction et l'indexation des termes complexes.
- Les modèles de PRF de base et leurs extensions améliorent significativement les extensions SPL_Glove, SPL_CT, et SPL_MWT.
- Les meilleurs résultats sont atteints par les extensions des modèles PRF, en particulier les extensions des modèles Bo2 et LL où une amélioration de MAP à l'ordre de 22% est obtenue.

Tableau 4.9: Résultats de comparaison des extensions de modèles PRF et des extensions de modèles de RI. Les exposants 1, 2, 3 et 4 désignent une amélioration significative par rapport aux modèles de base SPL et aux modèles de dépendance (SPL_CT et SPL_MWT), l'extension basée RDMs (SPL_Glove), et les modèles PRF de base respectivement.

Modèle \ Mesure	MAP	P10	taux d'amélioration de MAP
Modèle de base (SPL)	33.51	50.67	−−
SPL_CT	35.28^{1}	50.93	5.3%
SPL_MWT	35.88^{1}	51.87	7%
SPL_Glove	$36{,}41^{1}$	52.3	8.6%
KLD	$38.76^{1,2,3}$	52.27	15.6%
KLD_Sim_{comp}	$40.45^{1,2,3,4}$	54.03	20.7%
KLD_Sim_{avg}	$40.62^{1,2,3,4}$	54.03	21%
Bo2	$39.7^{1,2,3}$	52.8	18.5%
Bo2_Sim_{comp}	$41{,}03^{1,2,3,4}$	54,8	22.5%
Bo2_Sim_{avg}	$\mathbf{41{,}11}^{1,2,3,4}$	**55,07**	**22.6%**
LL	$40.71^{1,2,3}$	53.07	21.5%
LL_Sim_{comp}	$41.01^{1,2,3}$	54.00	22.4%
LL_Sim_{avg}	$41.04^{1,2,3}$	54.40	22.5%

4.8 Conclusion

Pour remédier au problème de disparité des termes dans le contexte de la RI en langue arabe, nous avons proposé :

1. une méthode basée sur les RDMs pour l'intégration de similarité entre les termes dans les modèles de RI (LM, LGD, SPL, et BM25). Cette méthode repose sur la construction de l'ensemble des termes similaires à ceux de la requête et une fonction primitive permettant de normaliser leurs relations de similarité aux termes de la requête. La fonction de normalisation est utilisée pour la pondération et l'intégration des termes similaires dans les fonctions d'appariement des modèles de RI.
2. une méthode basée sur les RDMs pour l'intégration de similarité entre les termes d'expansion et requête dans les modèles PRF (KLD, Bo2, et LL). L'idée principale de cette méthode consiste à combiner la distribution des termes candidats d'expansion dans les documents d'expansion avec leurs similarités aux termes de la requête initiale.

Les extensions proposées dans le cadre de ces deux méthodes sont évaluées en utilisant la collection standard TREC 2002/2001. En outre, toutes les extensions de modèles de RI et de modèles PRF sont évaluées en utilisant trois modèles de RDMs, en l'occurrence le modèle CBOW, Skip-gram, et Glove. Pour l'apprentissage des RDMs, nous avons collecté un grand corpus contentant 216M mots. Les résultats des évaluations ont montré que :

- Les extensions basées RDMs, que nous avons proposées, améliorent significativement leurs modèles de base, l'approche d'indexation sémantique (SI) basée sur l'AWN, et trois modèles de langue basés RDMs de l'état de l'art (LM+WE, NTLM, et GLM) ;
- L'étude de la sensibilité de nos extensions au paramètre λ, utilisé pour contrôler l'importance des termes similaires, est en concordance aux bornes déterminées dans la validation théorique de ces extensions ;
- Nos extensions basées RDMs des modèles PRF améliorent significativement leurs modèles standards PRF, la méthode d'expansion globale basée RDMs (VEXP) et le modèle standard de RI ;
- L'utilisation de la similarité seule entre les termes d'expansion et la requête initiale (i.e, Sim_{comp} et Sim_{avg}) a atteint une performance comparable au modèle standard KLD, en particulier pour un grand nombre de termes d'expansion : $n \geq 60$. Cependant, les modèles standards Bo2 et LL aboutissent à de meilleures performances par rapport aux modèles Sim_{comp} et Sim_{avg}.
- Des améliorations d'ordre de 22% ont été obtenues par nos extensions PRF pour la mesure MAP et l'indice RI respectivement ;
- Nos extensions basées RDMs de modèle de PRF, ainsi leurs modèles PRF standards surpassent significativement les extensions utilisées pour l'intégration des dépendances entre les termes (SPL_CT et SPL_MWT). De plus, l'extension SPL_Glove assure une meilleure performance par rapport aux modèles SPL_CT et SPL_MWT ;

Conclusion générale

Bien que les techniques et les modèles classiques de RI ont montré leur performance, la représentation par sac de mots du contenu textuel est à l'origine des problèmes d'ambigüité et de disparité des termes. Les travaux présentés dans ce mémoire rentrent dans le cadre de la recherche d'information en langue arabe, en particulier l'intégration des dépendances statistiques, syntaxiques et sémantiques entre les termes pour aller au-delà de la représentation par sac-de-mots. En effet, nous nous sommes intéressés à apporter des solutions permettant, d'une part, de mieux représenter le contenu informatif des documents pour remédier au problème d'ambigüité des termes simples et, d'autre part, de pallier le problème de disparité des termes.

Les contributions apportées dans ce mémoire sont structurées en deux volets importants :

- **indexation :** l'amélioration de la représentation du contenu des documents et des requêtes par la prise en compte des termes complexes et de proximité entre les termes, dans le but de remédier au problème d'ambigüité des termes simples ;
- **interrogation :** l'amélioration du processus d'appariement par l'extension des modèles de RI pour la prise en compte des termes similaires à ceux de la requête et l'intégration de similarité entre les termes dans les modèles PRF, dans le but de remédier au problème de disparité des termes et d'améliorer l'expression des requêtes ;

D'une manière explicite, nous avons proposé une méthode hybride pour l'extraction des termes complexes de la langue arabe. Dans un premier temps, le corpus est étiqueté en utilisant l'étiqueteur morpho-syntaxique AMIRA. Puis, notre filtre linguistique identifie les séquences des unités lexicales susceptibles de représenter des termes complexes en utilisant une liste des patrons syntaxiques. Ensuite, il procède par le traitement des variantes graphiques, flexionnelles, morpho-syntaxiques et syntaxiques des termes candidats. Pour le filtrage statistique, nous avons proposé une mesure d'association, appelée NLC-value, qui consiste à combiner les degrés de spécificité et d'unité avec l'information contextuelle afin de ne considérer que les termes pertinents. La validation de cette méthode est effectuée sur un corpus de domaine d'environnement. Les résultats expérimentaux ont montré que notre mesure statistique assure une meilleure performance par rapport aux autres mesures évaluées.

Nous avons également étudié le problème d'indexation et de recherche de documents à base des termes complexes. Pour intégrer ces termes complexes dans le processus d'appariement, nous avons utilisé une méthode simple qui consiste à combiner les scores des termes complexes et ceux des termes simples qui sont obtenus en utilisant les modèles standards de RI. En outre, nous avons exploré plusieurs modèles de proximité, et en particulier les termes croisés. Pour étudier l'impact de ces dépendances sur la performance de RI en langue arabe, les deux approches d'intégration de dépendances sont évaluées en utilisant différents nivaux d'analyse morphologique. De plus, nous avons introduit

une condition qui permet de caractériser les modèles de RI pour la prise en compte des dépendances de termes. De plus, nous avons comparé l'approche basée sur les termes complexes et l'approche basée proximité pour la prise en compte des dépendances dans la RI en langue arabe. Les résultats obtenus sur la collection standard TREC 2002/2001 ont montré que les termes complexes et les dépendances de proximité améliorent significativement la performance de RI pour toutes les approches de racinisation/racinisation légère. Ce qui montre que l'intégration de dépendance dans la RI en langue arabe mène à une représentation plus précise du contenu textuel. En revanche, la différence entre les extensions des termes complexes et les modèles de proximité n'est pas statistiquement significative. Les extensions basées sur l'intégration des termes croisés sont relativement meilleures par rapport aux autres modèles de proximité.

Sur le volet de l'interrogation, nous avons proposé une méthode basée sur les représentations distribuées des mots pour l'intégration de la similarité entre les termes dans les modèles de RI. Ce choix est motivé par le fait que les représentations distribuées des mots permettent de capturer des similitudes de niveaux morphologiques et sémantiques entre les termes. De ce fait, les vecteurs des termes sémantiquement similaires et les vecteurs des variantes morphologiques seront proches du vecteur du terme de la requête dans l'espace vectoriel. La méthode proposée consiste à étendre les modèles probabilistes de RI pour la prise en compte des termes distincts, mais similaires à ceux de la requête dans l'appariement des documents et des requêtes. Les extensions introduites dans le cadre de cette méthode sont validées en utilisant les contraintes heuristiques de l'appariement sémantique des termes. En outre, ces extensions sont comparées à quatre méthodes de l'état de l'art, y compris la méthode d'indexation sémantique à base de l'ontologie Arabic-WordNet [Abderrahim *et al.* 2016] et trois modèles de langue basés sur les représentations distribuées des mots [Vulić & Moens 2015, Zuccon *et al.* 2015, Ganguly *et al.* 2015]. Les résultats obtenus ont montré que nos extensions améliorent significativement leurs modèles de base et l'approche d'indexation sémantique. En outre, nos extensions du modèle BM25 et des modèles d'information (LGD et SPL) aboutissent à des améliorations significatives par rapport aux extensions basées sur les représentations distribuées des mots du modèle de langue. De plus, nos extensions basées RDMs ont donné des résultats prometteurs par rapport aux deux méthodes d'intégration de dépendances de termes dans les modèles de RI.

Pour améliorer davantage la performance de RI, nous avons proposé une méthode d'expansion des requêtes basée sur l'intégration des représentations distribuées des mots dans les modèles de PRF. Cette méthode consiste à intégrer la similarité entre les termes candidats d'expansion et ceux de la requête dans les modèles standards PRF. De ce fait, la pondération des termes candidats d'expansion s'effectue en combinant leurs poids dans l'ensemble des documents d'expansion et leurs similarités à la requête initiale. D'après les résultats obtenus, des améliorations significatives ont été obtenues par l'extension des modèles PRF KLD [Carpineto *et al.* 2001] et Bo2 [Amati & Van Rijsbergen 2002]. Cependant, une légère amélioration a été obtenue par rapport au modèle LL [Clinchant & Gaussier 2013]. En outre, les modèles PRF standards ont atteint des performances statistiquement signifi-

catives par rapport à toutes les extensions des modèles de RI, y compris les extensions basées RDMs et les extensions basées dépendances de termes.

Les différentes pistes explorées dans le cadre de cette thèse ont soulevé plusieurs perspectives. Dans ce qui suit, nous citons celles que nous prévoyons explorer.

- Exploitation des représentations distribuées des mots pour la prise en compte des variantes sémantiques lors de l'extraction et de l'indexation des termes complexes. Ceci peut être effectué en utilisant l'approche compositionnelle des vecteurs des mots qui constituent le terme complexe.
- Étude de l'impact des paramètres d'apprentissage des représentations distribuées des mots sur la performance de la RI, tel que la dimension des vecteurs des mots et la taille du contexte. En effet, pour les paramètres utilisés dans les évaluations, les trois modèles CBOW, Skip-gram et Glove ont abouti à des performances comparables pour les extensions de modèles de RI et PRF.
- Évaluation des extensions des modèles de RI et PRF pour la recherche d'information translinguistique sur les documents arabes à travers l'apprentissage des représentations distribuées multilingues des mots.
- Exploitation des représentations distribuées des sens des mots pour mieux considérer l'ambigüité et la polysémie dans la RI. Le principal inconvénient des représentations utilisées dans ce travail réside dans le fait que chaque mot est représenté par un seul vecteur. De toute évidence, pour mieux considérer l'ambigüité des mots, il est plus réaliste de représenter chaque sens de ces mots par un vecteur à part.

Liste de publications de la thèse

Articles

- Abdelkader El Mahdaouy, Saïd Ouatik El Alaoui, et Eric Gaussier, *"Improving Arabic Information Retrieval Using Word Embedding Similarities"*. International Journal of speech and Technology, Jan 2018.
- Abdelkader El Mahdaouy, Eric Gaussier, et Saïd Ouatik El Alaoui, *"Should one Use Term Proximity or Multi-Word Terms for Arabic Information Retrieval ?*, Computer Speech and Language (**soumis le 02/08/2016, révisé 31/01/2017**).
- Abdelkader El Mahdaouy, Saïd Ouatik El Alaoui, et Eric Gaussier, *"Word Embedding-Based Pseudo-Relevance Feedback for Arabic Information Retrieval"*, **soumis le 09/06/2017** au Journal of Information Science (JIS).

Chapitre d'ouvrage

- Abdelkader El Mahdaouy, Eric Gaussier, et Saïd Ouatik El Alaoui. *"Arabic Text Classification Based on Word and Document Embeddings"*, Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2016. AISI 2016. Advances in Intelligent Systems and Computing, vol 533, 2017, 32-41. Springer

Conférences internationales

- Abdelkader El Mahdaouy, Saïd Ouatik El Alaoui, et Eric Gaussier, *"Semantically enhanced term frequency based on word embeddings for Arabic information retrieval"*, 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt), Tangier, 2016, pp. 385-389.
- Abdelkader EL Mahdaouy, Eric Gaussier, et Saïd Ouatik El Alaoui, *"Exploring term proximity statistic for Arabic information retrieval"*, 2014 Third IEEE International Colloquium in Information Science and Technology (CIST), Tetouan, 2014, pp. 272-277.
- Abdelkader El Mahdaouy, Saïd Ouatik El Alaoui, et Eric Gaussier, *"A Study of Association Measures and their Combination for Arabic MWT Extraction,"* Terminology and Artificial Intelligence, Paris, France, Oct 2013, pp.45-52.

Références bibliographiques

[Abdelali *et al.* 2016] Ahmed Abdelali, Kareem Darwish, Nadir Durrani et Hamdy Mubarak. *Farasa : A Fast and Furious Segmenter for Arabic.* In Proceedings of the Demonstrations Session, NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, San Diego California, USA, June 12-17, 2016, pages 11–16, 2016.

[Abderrahim *et al.* 2016] Mohammed Alaeddine Abderrahim, Mohammed Dib, Mohammed El-Amine Abderrahim et Mohammed Amine Chikh. *Semantic Indexing of Arabic Texts for Information Retrieval System.* Int. J. Speech Technol., vol. 19, no. 2, pages 229–236, Juin 2016.

[Abouenour *et al.* 2013] Lahsen Abouenour, Karim Bouzoubaa et Paolo Rosso. *On the evaluation and improvement of Arabic WordNet coverage and usability.* Language Resources and Evaluation, vol. 47, no. 3, pages 891–917, 2013.

[Abu El-Khair 2007] Ibrahim Abu El-Khair. *Arabic information retrieval.* Annual review of information science and technology, vol. 41, no. 1, pages 505–533, 2007.

[Abu-Salem *et al.* 1999] Hani Abu-Salem, Mahmoud Al-Omari et Martha W. Evens. *Stemming methodologies over individual query words for an Arabic Information Retrieval System.* Journal of the American Society for Information Science, vol. 50, no. 6, pages 524–529, 1999.

[Al-Kharashi & Evens 1994] Ibrahim A. Al-Kharashi et Martha W. Evens. *Comparing words, stems, and roots as index terms in an Arabic Information Retrieval system.* Journal of the American Society for Information Science, vol. 45, no. 8, pages 548–560, 1994.

[Al Khatib & Badarneh 2010] Khalid Al Khatib et Amer Badarneh. *Automatic extraction of Arabic multi-word terms.* In Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on, pages 411–418. IEEE, 2010.

[Al-Taani & Al-Rub 2009] Ahmad T Al-Taani et Salah Abu Al-Rub. *A rule-based approach for tagging non-vocalized Arabic words.* Int. Arab J. Inf. Technol., vol. 6, no. 3, pages 320–328, 2009.

[ALMasri *et al.* 2016] Mohannad ALMasri, Catherine Berrut et Jean-Pierre Chevallet. *A comparison of deep learning based query expansion with pseudo-relevance feedback and mutual information.* In European Conference on Information Retrieval, pages 709–715. Springer, 2016.

[Amati & Van Rijsbergen 2002] Gianni Amati et Cornelis Joost Van Rijsbergen. *Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness.* ACM Trans. Inf. Syst., vol. 20, no. 4, pages 357–389, Octobre 2002.

[Amati 2003] Giambattista Amati. *Probability models for information retrieval based on divergence from randomness*. PhD thesis, University of Glasgow, 2003.

[Attia 2008] Mohammed A Attia. *Handling Arabic morphological and syntactic ambiguity within the LFG framework with a view to machine translation*. PhD thesis, University of Manchester, 2008.

[Attia 2012] Mohammed Attia. Ambiguity in arabic computational morphology and syntax : A study within the lexical functional grammar framework. LAP Lambert Academic Publishing, 2012.

[Atwan *et al.* 2016] Jaffar Atwan, Masnizah Mohd, Hasan Rashaideh et Ghassan Kanaan. *Semantically Enhanced Pseudo Relevance Feedback for Arabic Information Retrieval*. J. Inf. Sci., vol. 42, no. 2, pages 246–260, Avril 2016.

[Baeza-Yates *et al.* 1999] R. Baeza-Yates, B. de Araujo Ribeiro-Neto et B. Ribeiro-Neto. Modern information retrieval. ACM Press books. ACM Press, 1999.

[Baroni *et al.* 2014] Marco Baroni, Georgiana Dinu et Germán Kruszewski. *Don't count, predict ! A systematic comparison of context-counting vs. context-predicting semantic vectors*. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 238–247, Baltimore, Maryland, Juin 2014. ACL.

[Baziz 2005] Mustapha Baziz. *Indexation conceptuelle guidée par ontologie pour la recherche d'information*. PhD thesis, 2005. Thèse de doctorat dirigée par Boughanem, Mohand et Aussenac-Gilles, Nathalie Informatique Toulouse 3 2005.

[Beesley 1996] Kenneth R. Beesley. *Arabic Finite-state Morphological Analysis and Generation*. In Proceedings of the 16th Conference on Computational Linguistics - Volume 1, COLING '96, pages 89–94, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics.

[Beesley 1998] Kenneth R Beesley. *Consonant spreading in Arabic stems*. In Proceedings of the 17th international conference on Computational linguistics-Volume 1, pages 117–123. Association for Computational Linguistics, 1998.

[Belalem *et al.* 2014] Ghalem Belalem, Ahmed Abbache, Fatiha Barigou et Fatma Zohra Belkredim. *The Use of Arabic WordNet in Arabic Information Retrieval*. Int. J. Inf. Retr. Res., vol. 4, no. 3, pages 54–65, Juillet 2014.

[Belguith *et al.* 2005] L Belguith, Leila Baccour et Ghassan Mourad. *Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules*. In Actes de la 12éme Conférence annuelle sur le Traitement Automatique des Langues Naturelles, pages 451–456, 2005.

[Ben Guirat *et al.* 2016] Souheila Ben Guirat, Ibrahim Bounhas et Yahya Slimani. *Combining Indexing Units for Arabic Information Retrieval*. Int. J. Softw. Innov., vol. 4, no. 4, pages 1–14, Octobre 2016.

[Bengio *et al.* 2003] Yoshua Bengio, Réjean Ducharme, Pascal Vincent et Christian Jauvin. *A neural probabilistic language model*. Journal of machine learning research, vol. 3, no. Feb, pages 1137–1155, 2003.

[Berger & Lafferty 1999] Adam Berger et John Lafferty. *Information Retrieval As Statistical Translation*. In Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '99, pages 222–229, New York, NY, USA, 1999. ACM.

[Borlund 2003] Pia Borlund. *The Concept of Relevance in IR*. J. Am. Soc. Inf. Sci. Technol., vol. 54, no. 10, pages 913–925, Aot 2003.

[Boubekeur 2008] Fatiha Boubekeur. *Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets*. Theses, Université Paul Sabatier - Toulouse III, Juillet 2008. Thèse de doctorat co-dirigée par Mohand Boughanem et Lynda Tamine-Lechani.

[Boudchiche *et al.* 2016] Mohamed Boudchiche, Azzeddine Mazroui, Mohamed Ould Abdallahi Ould Bebah, Abdelhak Lakhouaja et Abderrahim Boudlal. *AlKhalil Morpho Sys 2 : A robust Arabic morpho-syntactic analyzer*. Journal of King Saud University - Computer and Information Sciences, pages –, 2016.

[Boulaknadel *et al.* 2008a] S. Boulaknadel, B. daille et A. driss. *Multi-word term indexing for Arabic document retrieval*. In Computers and Communications, 2008. ISCC 2008. IEEE Symposium on, pages 869–873, July 2008.

[Boulaknadel *et al.* 2008b] Siham Boulaknadel, Béatrice Daille et Driss Aboutajdine. *A Multi-Word Term Extraction Program for Arabic Language.* In LREC, pages 380–383. European Language Resources Association, 2008.

[Boulaknadel 2008] Siham Boulaknadel. *Impact of Term-Indexing for Arabic Document Retrieval*. In Natural Language and Information Systems, 13th International Conference on Applications of Natural Language to Information Systems, NLDB 2008, London, UK, June 24-27, 2008, Proceedings, pages 380–383, 2008.

[Bounhas & Slimani 2009] Ibrahim Bounhas et Yahya Slimani. *A hybrid approach for Arabic multi-word term extraction*. In Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009. International Conference on, pages 1–8. IEEE, 2009.

[Bourigault & Fabre 2000] Didier Bourigault et Cécile Fabre. *Approche linguistique pour l'analyse syntaxique de corpus*. Cahiers de grammaire, no. 25, pages 131–151, 2000.

[Bourigault *et al.* 2005] Didier Bourigault, Cécile Fabre, Cécile Frérot, Marie-Paule Jacques et Sylwia Ozdowska. *Syntex, analyseur syntaxique de corpus*. In Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles, 2005.

[Bourigault 1994] Didier Bourigault. *Lexter : un Logiciel d'EXtraction de TERminologie : application à l'acquisition des connaissances à partir de textes*. PhD thesis, EHESS, 1994.

[Bouzoubaa *et al.* 2009] Karim Bouzoubaa, Hicham Baidouri, Taoufik Loukili et Taoufik El Yazidi. *Arabic Stop Words : Towards a Generalisation and Standardisation*. In the 13th International Business Information Management Association Conference IBIMA, 2009.

[Carpineto & Romano 2012] Claudio Carpineto et Giovanni Romano. *A Survey of Automatic Query Expansion in Information Retrieval*. ACM Comput. Surv., vol. 44, no. 1, pages 1 :1–1 :50, Janvier 2012.

[Carpineto *et al.* 2001] Claudio Carpineto, Renato de Mori, Giovanni Romano et Brigitte Bigi. *An Information-theoretic Approach to Automatic Query Expansion*. ACM Trans. Inf. Syst., vol. 19, no. 1, pages 1–27, Janvier 2001.

[Chen & Gey 2001] Aitao Chen et Fredric C Gey. *Translation Term Weighting and Combining Translation Resources in Cross-Language Retrieval.* In TREC, 2001.

[Church *et al.* 1991] Kenneth Church, William Gale, Patrick Hanks et Donald Hindle. *Using statistics in lexical analysis*. In Lexical Acquisition : Exploiting On-Line Resources to Build a Lexicon, pages 115–164. Erlbaum, 1991.

[Cleverdon *et al.* 1966] C.W. Cleverdon, J. Mills et M. Keen. Factors determining the performance of indexing systems : Text. Numéro vol. 1,p. 1. Aslib Cranfield research project, 1966.

[Cleverdon 1962] Cyril W Cleverdon. *Aslib Cranfield research project : report on the testing and analysis of an investigation into the comparative efficiency of indexing systems*. Rapport technique, 1962.

[Clinchant & Gaussier 2010] Stéphane Clinchant et Eric Gaussier. *Information-based Models for Ad Hoc IR*. In Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '10, pages 234–241, New York, NY, USA, 2010. ACM.

[Clinchant & Gaussier 2011] Stéphane Clinchant et Eric Gaussier. *Retrieval Constraints and Word Frequency Distributions a Log-logistic Model for IR*. Inf. Retr., vol. 14, no. 1, pages 5–25, Fier 2011.

[Clinchant & Gaussier 2013] Stéphane Clinchant et Eric Gaussier. *A Theoretical Analysis of Pseudo-Relevance Feedback Models*. In Proceedings of the 2013 Conference on the Theory of Information Retrieval, ICTIR '13, pages 6 :6–6 :13, New York, NY, USA, 2013. ACM.

[Collet 2000] Tanja Collet. *La réduction des unités terminologiques complexes de type syntagmatique*. PhD thesis, Université de Montréal, Département de linguistique et de traduction., 2000.

[Cooper 1971] W.S. Cooper. *A definition of relevance for information retrieval*. Information Storage and Retrieval, vol. 7, no. 1, pages 19 – 37, 1971.

[Croft *et al.* 1991] W. Bruce Croft, Howard R. Turtle et David D. Lewis. *The Use of Phrases and Structured Queries in Information Retrieval*. In Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '91, pages 32–45, New York, NY, USA, 1991. ACM.

[Croft *et al.* 2011] W. Bruce Croft, Michael Bendersky, Hang Li et Gu Xu. *Query Representation and Understanding Workshop*. SIGIR Forum, vol. 44, no. 2, pages 48–53, Janvier 2011.

[Daille *et al.* 1994] Béatrice Daille, Eric Gaussier et Jean-Marc Langé. *Towards Automatic Extraction of Monolingual and Bilingual Terminology*. In Proceedings of the 15th Conference on Computational Linguistics - Volume 1, COLING '94, pages 515–521, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[Daille 1994] Béatrice Daille. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. PhD thesis, Paris 7, 1994.

[Daille 1996] Beatrice Daille. *Lexicomatique et Dictionnairiques*. chapitre ACABIT : une maquette d'aide à la construction automatique de banques terminologiques monolingues ou bilingues, pages 123–136. FMA, Beyrouth, 1996.

[Darwish & Magdy 2014] Kareem Darwish et Walid Magdy. *Arabic Information Retrieval*. Found. Trends Inf. Retr., vol. 7, no. 4, pages 239–342, feb 2014.

[Darwish & Mubarak 2016] Kareem Darwish et Hamdy Mubarak. *Farasa : A New Fast and Accurate Arabic Word Segmenter*. In Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016., 2016.

[Darwish & Oard 2007] Kareem Darwish et Douglas W. Oard. Adapting morphology for arabic information retrieval*, pages 245–262. Springer Netherlands, Dordrecht, 2007.

[Darwish 2002] Kareem Darwish. *Building a shallow Arabic morphological analyzer in one day*. In Proceedings of the ACL-02 workshop on Computational approaches to semitic languages, pages 1–8. Association for Computational Linguistics, 2002.

[David & Plante 1990] S. David et P. Plante. *De la nécessité d'une approche morphosyntaxique dans l'analyse de textes*. ICO, vol. 2, no. 3, pages 140–154, 1990.

[Debili & Achour 1998] Fathi Debili et Hadhemi Achour. *Voyellation automatique de l'arabe*. In Proceedings of the Workshop on Computational Approaches to Semitic Languages, pages 42–49. Association for Computational Linguistics, 1998.

[Diab *et al.* 2004] Mona Diab, Kadri Hacioglu et Daniel Jurafsky. *Automatic tagging of Arabic text : From raw text to base phrase chunks*. In Proceedings of HLT-NAACL 2004 : Short papers, pages 149–152. Association for Computational Linguistics, 2004.

[Diab *et al.* 2007] Mona Diab, Kadri Hacioglu et Daniel Jurafsky. *Automatic processing of modern standard Arabic text*. In Arabic Computational Morphology, pages 159–179. Springer, 2007.

[Diab 2009] Mona Diab. *Second Generation Tools (AMIRA 2.0) : Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking*. In Khalid Choukri et Bente Maegaard, editeurs, Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt, April 2009. The MEDAR Consortium.

[Dichy & Farghaly] Joseph Dichy et Ali Farghaly. *Roots & Patterns vs. Stems plus Grammar-Lexis Specifications : on what basis should a multilingual lexical database centred on Arabic be built.*

[Dominich 2001] S. Dominich. Mathematical foundations of information retrieval. Library of Public Policy and Public Administration. Springer Netherlands, 2001.

[Dragoni *et al.* 2012] Mauro Dragoni, Célia da Costa Pereira et Andrea G.B. Tettamanzi. *A conceptual representation of documents and queries for information retrieval systems by using light ontologies.* Expert Systems with Applications, vol. 39, no. 12, pages 10376 – 10388, 2012.

[Dunning 1993] Ted Dunning. *Accurate Methods for the Statistics of Surprise and Coincidence.* Comput. Linguist., vol. 19, no. 1, pages 61–74, Mars 1993.

[El-Khair 2006] Ibrahim Abu El-Khair. *Effects of stop words elimination for Arabic information retrieval : a comparative study.* International Journal of Computing & Information Sciences, vol. 4, no. 3, pages 119–133, 2006.

[El Mahdaouy *et al.* 2013] Abdelkader El Mahdaouy, Saïd El Alaoui Ouatik et Eric Gaussier. *A Study of Association Measures and their Combination for Arabic MWT Extraction.* In Terminology and Artificial Intelligence, pages 45–52, Paris, France, Octobre 2013.

[El Mahdaouy *et al.* 2014] Abdelkader El Mahdaouy, Éric Gaussier et Saïd Ouatik El Alaoui. *Exploring term proximity statistic for Arabic information retrieval.* In 2014 Third IEEE International Colloquium in Information Science and Technology (CIST), pages 272–277, Oct 2014.

[EL Mahdaouy *et al.* 2016] Abdelkader EL Mahdaouy, Saïd Ouatik El Alaoui et Eric Gaussier. *Semantically enhanced term frequency based on word embeddings for Arabic information retrieval.* In 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt), pages 385–389, Oct 2016.

[EL Mahdaouy *et al.* 2017] Abdelkader EL Mahdaouy, Eric Gaussier et Saïd Ouatik El Alaoui. Arabic text classification based on word and document embeddings, pages 32–41. Springer International Publishing, Cham, 2017.

[El Mahdaouy *et al.* 2018a] Abdelkader El Mahdaouy, Saïd Ouatik El Alaoui et Eric Gaussier. *Improving Arabic information retrieval using word embedding similarities.* International Journal of Speech Technology, Jan 2018.

[El Mahdaouy *et al.* 2018b] Abdelkader El Mahdaouy, Saïd Ouatik El Alaoui et Eric Gaussier. *Word Embedding-Based Pseudo-Relevance Feedback for Arabic Information Retrieval.* **Submitted on 08/06/2017 to** Journal of Information Science (JIS), 2018.

[El Mahdaouy *et al.* 2018c] Abdelkader El Mahdaouy, Eric Gaussier et Saïd Ouatik El Alaoui. *Should one Use Term Proximity or Multi-Word Terms for Arabic Information Retrieval ?* **Submitted on 02/08/2016 to** Computer Speech & Language, 2018.

[Fagan 1987] J. Fagan. *Automatic Phrase Indexing for Document Retrieval.* In Proceedings of the 10th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '87, pages 91–101, New York, NY, USA, 1987. ACM.

[Fang & Zhai 2006] Hui Fang et ChengXiang Zhai. *Semantic Term Matching in Axiomatic Approaches to Information Retrieval.* In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06, pages 115–122, New York, NY, USA, 2006. ACM.

[Fang & Zhai 2014] Hui Fang et ChengXiang Zhai. *Axiomatic Analysis and Optimization of Information Retrieval Models.* In Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '14, pages 1288–1288, New York, NY, USA, 2014. ACM.

[Fang *et al.* 2004] Hui Fang, Tao Tao et ChengXiang Zhai. *A Formal Study of Information Retrieval Heuristics.* In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '04, pages 49–56, New York, NY, USA, 2004. ACM.

[Fang 2008] Hui Fang. *A Re-examination of Query Expansion Using Lexical Resources.* Citeseer, 2008.

[Farghaly 2004] Ali Farghaly. *Computer Processing of Arabic Script-based Languages. Current State and Future Directions.* In Ali Farghaly et Karine Megerdoomian, editeurs, COLING 2004 Computational Approaches to Arabic Script-based Languages, pages 1–1, Geneva, Switzerland, August 28th 2004. COLING.

[Fehri 1993] A.F. Fehri. Issues in the structure of arabic clauses and words. Studies in Natural Language and Linguistic Theory. Springer Netherlands, 1993.

[Fernández *et al.* 2011] Miriam Fernández, Iván Cantador, Vanesa López, David Vallet, Pablo Castells et Enrico Motta. *Semantically enhanced Information Retrieval : An ontology-based approach.* Web Semantics : Science, Services and Agents on the World Wide Web, vol. 9, no. 4, pages 434 – 452, 2011. {JWS} special issue on Semantic Search.

[Frantzi *et al.* 2000] Katerina Frantzi, Sophia Ananiadou et Hideki Mima. *Automatic recognition of multi-word terms :. the C-value/NC-value method.* International Journal on Digital Libraries, vol. 3, no. 2, pages 115–130, 2000.

[Froud *et al.* 2012] Hanane Froud, Abdelmonaime Lachkar et Saïd Ouatik El Alaoui. *Stemming versus Light Stemming for measuring the simitilarity between Arabic Words with Latent Semantic Analysis model.* In 2012 Colloquium in Information Science and Technology, pages 69–73, Oct 2012.

[Ganguly *et al.* 2015] Debasis Ganguly, Dwaipayan Roy, Mandar Mitra et Gareth J.F. Jones. *Word Embedding Based Generalized Language Model for Information Retrieval.* In Proceedings of the 38th International ACM SIGIR Conference on Research

and Development in Information Retrieval, SIGIR '15, pages 795–798, New York, NY, USA, 2015. ACM.

[Gaussier & Stéfanini 2003] Eric Gaussier et M.H. Stéfanini. Assistance intelligente à la recherche d'informations. Traité des sciences et techniques de l'information. Hermès science publications, 2003.

[Gaussier & Yvon 2012] Eric Gaussier et François Yvon. Textual Information Access : Statistical Models. John Wiley & Sons, Avril 2012.

[Gaussier *et al.* 2000] Eric Gaussier, Gregory Grefenstette, David Hull et Claude Roux. *Recherche d'information en français et traitement automatique des langues.* TAL. Traitement automatique des langues, vol. 41, no. 2, pages 473–493, 2000.

[Gey & Oard 2001] Fredric C Gey et Douglas W Oard. *The TREC-2001 Cross-Language Information Retrieval Track : Searching Arabic Using English, French or Arabic Queries.* In TREC, pages 16–26, 2001.

[Goweder *et al.* 2004] Abduelbaset Goweder, Massimo Poesio et Anne De Roeck. *Broken Plural Detection for Arabic Information Retrieval.* In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '04, pages 566–567, New York, NY, USA, 2004. ACM.

[Gövert & Kazai] Norbert Gövert et Gabriella Kazai. *Overview of the Initiative for the Evaluation of XML retrieval (INEX) 2002.*

[Habash 2009] Nizar Y. Habash. Arabic natural language processing. Synthesis lectures on human language technologies. Morgan & Claypool Publishers, 2009.

[Haddad 2002] Mohamed Hatem Haddad. *Extraction et Impact des connaissances sur les performances des Systèmes de Recherche d'Information.* PhD thesis, Université Joseph-Fourier-Grenoble I, 2002.

[Haddad 2003] Hatem Haddad. *French Noun Phrase Indexing and Mining for an Information Retrieval System.* In MarioA. Nascimento, EdlenoS. de Moura et ArlindoL. Oliveira, editeurs, String Processing and Information Retrieval, volume 2857 of *Lecture Notes in Computer Science*, pages 277–286. Springer Berlin Heidelberg, 2003.

[Hadni *et al.* 2012] Meryem Hadni, Abdelmonaime Lachkar et Saïd Ouatik El Alaoui. *A new and efficient stemming technique for Arabic Text Categorization.* In 2012 International Conference on Multimedia Computing and Systems, pages 791–796, May 2012.

[Hadni *et al.* 2013] Meryeme Hadni, Said Alaoui Ouatik, Abdelmonaime Lachkar et Mohammed Meknassi. *Hybrid Part-Of-Speech Tagger for Non-Vocalized Arabic Text.* International Journal on Natural Language Computing (IJNLC) Vol, vol. 2, 2013.

[Hammache *et al.* 2014] Arezki Hammache, Mohand Boughanem et Rachid Ahmed-Ouamer. *Combining compound and single terms under language model framework.* Knowl. Inf. Syst., vol. 39, no. 2, pages 329–349, 2014.

[Harman & Voorhees 2006] Donna K. Harman et Ellen M. Voorhees. *TREC : An Overview*. Annual Rev. Info. Sci & Technol., vol. 40, no. 1, pages 113–155, Dmbre 2006.

[Harman 1991] Donna Harman. *How effective is suffixing ?* Journal of the American Society for Information Science, vol. 42, no. 1, pages 7–15, 1991.

[Harman 1992] Donna Harman. *Evaluation Issues in Information Retrieval*. Inf. Process. Manage., vol. 28, no. 4, pages 439–440, Mars 1992.

[He *et al.* 2011] Ben He, Jimmy Xiangji Huang et Xiaofeng Zhou. *Modeling Term Proximity for Probabilistic Information Retrieval Models*. Inf. Sci., vol. 181, no. 14, pages 3017–3031, Juillet 2011.

[Hliaoutakis *et al.* 2006] Angelos Hliaoutakis, Giannis Varelas, Epimenidis Voutsakis, Euripides G. M. Petrakis et Evangelos E. Milios. *Information Retrieval by Semantic Similarity*. Int. J. Semantic Web Inf. Syst., vol. 2, no. 3, pages 55–73, 2006.

[Hmeidi *et al.* 1997] Ismail Hmeidi, Ghassan Kanaan et Martha Evens. *Design and implementation of automatic indexing for information retrieval with Arabic documents*. Journal of the American Society for Information Science, vol. 48, no. 10, pages 867–881, 1997.

[Hofmann 1999] Thomas Hofmann. *Probabilistic Latent Semantic Indexing*. In Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM.

[Hu *et al.* 2006] Jiani Hu, Weihong Deng et Jun Guo. *Improving Retrieval Performance by Global Analysis*. In 18th International Conference on Pattern Recognition (ICPR'06), volume 2, pages 703–706, 2006.

[Jacquemin *et al.* 1997] Christian Jacquemin, Judith L. Klavans et Evelyne Tzoukermann. *Expansion of Multi-word Terms for Indexing and Retrieval Using Morphology and Syntax*. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, ACL '98, pages 24–31, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics.

[Jacquemin 1997] Christian Jacquemin. *Variation terminologique : reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus thèse*. PhD thesis, HDR, Université de Nante, France, 1997.

[Jing & Croft 1994] Yufeng Jing et W. Bruce Croft. *An Association Thesaurus for Information Retrieval*. In Intelligent Multimedia Information Retrieval Systems and Management - Volume 1, RIAO '94, pages 146–160, Paris, France, France, 1994. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.

[Justeson & Katz 1995] John S. Justeson et Slava M. Katz. *Technical terminology : some linguistic properties and an algorithm for identification in text*. Natural Language Engineering, vol. 1, no. 1, page 9–27, 1995.

[Kadri & Nie 2006] Y. Kadri et Jian-Yun Nie. *Effective stemming for Arabic information retrieval*. In The Challenge of Arabic for NLP/MT, International Conf. at the British Computer Society (BCS), pages 68–74, 2006.

[Kadri 2008] Youssef Kadri. *Recherche d'information translinguistique sur les documents en arabe*. PhD thesis, Université de Montréal, Canada, 2008.

[Kageura & Umino 1996] Kyo Kageura et Bin Umino. *Methods of automatic term recognition : a review*. Terminology, vol. 3, no. 2, pages 259–289, 1996.

[Kanaan *et al.* 2005] Ghassan Kanaan, Riyad Al-Shalabi et Majdi Sawalha. *Improving Arabic information retrieval systems using part of speech tagging*. Information Technology Journal, vol. 4, no. 1, pages 32–37, 2005.

[Karimzadehgan & Zhai 2010] Maryam Karimzadehgan et ChengXiang Zhai. *Estimation of Statistical Translation Models Based on Mutual Information for Ad Hoc Information Retrieval*. In Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '10, pages 323–330, New York, NY, USA, 2010. ACM.

[Khoja & Garside 1999] S. Khoja et R. Garside. *Stemming Arabic Text*. In Computing Department. Lancaster University, 1999.

[Korkontzelos *et al.* 2008] Ioannis Korkontzelos, Ioannis P Klapaftis et Suresh Manandhar. *Reviewing and evaluating automatic term recognition techniques*. In Advances in Natural Language Processing, pages 248–259. Springer, 2008.

[Kouloughli 1994] D.E. Kouloughli. Grammaire de l'arabe d'aujourd'hui. Langues pour tous. Pocket, 1994.

[Kraft & Buell 1983] Donald H Kraft et Duncan A Buell. *Fuzzy sets and generalized Boolean retrieval systems*. International journal of man-machine studies, vol. 19, no. 1, pages 45–56, 1983.

[Krovetz & Croft 1992] Robert Krovetz et W. Bruce Croft. *Lexical Ambiguity and Information Retrieval*. ACM Trans. Inf. Syst., vol. 10, no. 2, pages 115–141, Avril 1992.

[Krovetz 1997] Robert Krovetz. *Homonymy and Polysemy in Information Retrieval*. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, ACL '98, pages 72–79, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics.

[Kuzi *et al.* 2016] Saar Kuzi, Anna Shtok et Oren Kurland. *Query Expansion Using Word Embeddings*. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16, pages 1929–1932, New York, NY, USA, 2016. ACM.

[Larkey *et al.* 2002] Leah S. Larkey, Lisa Ballesteros et Margaret E. Connell. *Improving Stemming for Arabic Information Retrieval : Light Stemming and Co-occurrence*

Analysis. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '02, pages 275–282, New York, NY, USA, 2002.

[Larkey *et al.* 2007] LeahS. Larkey, Lisa Ballesteros et MargaretE. Connell. *Light Stemming for Arabic Information Retrieval*. In Abdelhadi Soudi, Antalvan den Bosch et Günter Neumann, editeurs, Arabic Computational Morphology, volume 38 of *Text, Speech and Language Technology*, pages 221–243. Springer Netherlands, 2007.

[Lavrenko & Croft 2001] Victor Lavrenko et W Bruce Croft. *Relevance based language models*. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pages 120–127. ACM, 2001.

[Lefévre 2000] P. Lefévre. La recherche d'informations : du texte intégral au thésaurus. Hermes Science, 2000.

[Lesk 1969] Michael E Lesk. *Word-word associations in document retrieval systems*. American documentation, vol. 20, no. 1, pages 27–38, 1969.

[Lewis *et al.* 1989] David D. Lewis, W. Bruce Croft et Nehru Bhandaru. *Language-oriented information retrieval*. International Journal of Intelligent Systems, vol. 4, no. 3, pages 285–318, 1989.

[Li & Gaussier 2012] Bo Li et Éric Gaussier. *An Information-Based Cross-Language Information Retrieval Model*. In 34th European Conference on IR Research, ECIR 2012, volume 7224 of *Lecture Notes in Computer Science (LNCS)*, pages 281–292, Barcelone, Spain, April 2012. Springer.

[Li & Xu 2014] Hang Li et Jun Xu. *Semantic Matching in Search*. Found. Trends Inf. Retr., vol. 7, no. 5, pages 343–469, Juin 2014.

[Lioma & Blanco 2009] Christina Lioma et Roi Blanco. *Part of speech based term weighting for information retrieval*. In European Conference on Information Retrieval, pages 412–423. Springer, 2009.

[Liu *et al.* 2004a] Shuang Liu, Fang Liu, Clement Yu et Weiyi Meng. *An Effective Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases*. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '04, pages 266–272, New York, NY, USA, 2004. ACM.

[Liu *et al.* 2004b] Shuang Liu, Fang Liu, Clement Yu et Weiyi Meng. *An effective approach to document retrieval via utilizing WordNet and recognizing phrases*. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pages 266–272. ACM, 2004.

[Lofi 2015] Christoph Lofi. *Measuring Semantic Similarity and Relatedness with Distributional and Knowledge-based Approaches*. Information and Media Technologies, vol. 10, no. 3, pages 493–501, 2015.

[Lv & Zhai 2009a] Yuanhua Lv et ChengXiang Zhai. *A comparative study of methods for estimating query language models with pseudo feedback*. In Proceedings of the 18th ACM conference on Information and knowledge management, pages 1895–1898. ACM, 2009.

[Lv & Zhai 2009b] Yuanhua Lv et ChengXiang Zhai. *Positional Language Models for Information Retrieval*. In Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09, pages 299–306, New York, NY, USA, 2009.

[Lvovsky 2004] A I Lvovsky. *Iterative maximum-likelihood reconstruction in quantum homodyne tomography*. Journal of Optics B : Quantum and Semiclassical Optics, vol. 6, no. 6, pages 556–559, 2004.

[Ma *et al.* 2016] Chenglong Ma, Qingwei Zhao, Jielin Pan et Yonghong Yan. *Short Text Classification Based on Distributional Representations of Words*. IEICE Transactions, vol. 99-D, no. 10, pages 2562–2565, 2016.

[Maamouri & Bies 2010] M. Maamouri et A. Bies. *The Penn Arabic Treebank*. In Ali Farghali, editeur, Arabic Computational Linguistics, pages 103–135. CSLI studies in Computational Linguistics, Stanford, CA, 2010.

[Maamouri *et al.* 2004] Mohamed Maamouri, Ann Bies, Tim Buckwalter et Wigdan Mekki. *The Penn Arabic Treebank : Building a Large-Scale Annotated Arabic Corpus*. In NEMLAR Conference on Arabic Language Resources and Tools, 2004.

[Mahgoub *et al.* 2014] Ashraf Y Mahgoub, Mohsen A Rashwan, Hazem Raafat, Mohamed A Zahran et Magda B Fayek. *Semantic query expansion for Arabic information retrieval*. ANLP 2014, pages 87–92, 2014.

[Manning *et al.* 2008] Christopher D. Manning, Prabhakar Raghavan et Hinrich Schütze. Introduction to information retrieval. Cambridge University Press, New York, NY, USA, 2008.

[Maron & Kuhns 1960] Melvin Earl Maron et John L Kuhns. *On relevance, probabilistic indexing and information retrieval*. Journal of the ACM (JACM), vol. 7, no. 3, pages 216–244, 1960.

[Matthijs & Radlinski 2011] Nicolaas Matthijs et Filip Radlinski. *Personalizing web search using long term browsing history*. In Proceedings of the fourth ACM international conference on Web search and data mining, pages 25–34. ACM, 2011.

[Metzler & Croft 2005] Donald Metzler et W. Bruce Croft. *A Markov Random Field Model for Term Dependencies*. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '05, pages 472–479, New York, NY, USA, 2005. ACM.

[Mikolov *et al.* 2013] Tomas Mikolov, Kai Chen, Greg Corrado et Jeffrey Dean. *Efficient Estimation of Word Representations in Vector Space*. In Proceedings of International Conference on Learning Representations, ICLR '13, 2013.

[Mitra *et al.* 1997] Mandar Mitra, Chris Buckley, Amit Singhal et Claire Cardie. *An Analysis of Statistical and Syntactic Phrases.* In Proceedings of RIAO, pages 200–214, 1997.

[Mizzaro 1997] Stefano Mizzaro. *Relevance : The whole history.* Journal of the American Society for Information Science, vol. 48, no. 9, pages 810–832, 1997.

[Moreau & Sébillot 2005] Fabienne Moreau et Pascale Sébillot. *Contributions des techniques du traitement automatique des langues à la recherche d'information.* Research Report RR-5484, INRIA, 2005.

[Mustafa *et al.* 2008] Mohammed Mustafa, Hisham AbdAlla et Hussein Suleman. Current approaches in arabic ir : A survey, pages 406–407. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[Nakagawa & Mori 2003] Hiroshi Nakagawa et Tatsunori Mori. *Automatic term recognition based on statistics of compound nouns and their components.* Terminology, vol. 9, no. 2, pages 201–219, 2003.

[Nie & Dufort 2002] Jian-Yun Nie et Jean-François Dufort. *Combining words and compound terms for monolingual and cross-language information retrieval.* Proceedings of Information 2002, pages 453–458, 2002.

[Nwesri *et al.* 2005] AbdusalamF.A. Nwesri, S.M.M. Tahaghoghi et Falk Scholer. *Stemming Arabic Conjunctions and Prepositions.* In Mariano Consens et Gonzalo Navarro, editeurs, String Processing and Information Retrieval, volume 3772 of *Lecture Notes in Computer Science*, pages 206–217. Springer Berlin Heidelberg, 2005.

[Nwesri 2008] Abdusalam F Ahmed Nwesri. *Effective retrieval techniques for Arabic text.* PhD thesis, School of Computer Science and Information Technology of Melbourne, 2008.

[OUESLATI 1999] ROCHDI OUESLATI. *Aide a l'acquisition de connaissances a partir de corpus.* PhD thesis, Strasbourg 1, 1999.

[Pasha *et al.*] Arfath Pasha, Mohamed Al-Badrashiny, Mona T Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow et Ryan Roth. *MADAMIRA : A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic.*

[Peat & Willett 1991] Helen J Peat et Peter Willett. *The limitations of term co-occurrence data for query expansion in document retrieval systems.* Journal of the american society for information science, vol. 42, no. 5, page 378, 1991.

[Peng *et al.* 2007] Jie Peng, Craig Macdonald, Ben He, Vassilis Plachouras et Iadh Ounis. *Incorporating Term Dependency in the DFR Framework.* In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07, pages 843–844, New York, NY, USA, 2007. ACM.

[Pennington *et al.* 2014] Jeffrey Pennington, Richard Socher et Christopher Manning. *Glove : Global Vectors for Word Representation*. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar, Octobre 2014. Association for Computational Linguistics.

[Peters & Braschler 2001] Carol Peters et Martin Braschler. *European research letter : Cross-language system evaluation : The CLEF campaigns*. Journal of the Association for Information Science and Technology, vol. 52, no. 12, pages 1067–1072, 2001.

[Ponte & Croft 1998] Jay M. Ponte et W. Bruce Croft. *A Language Modeling Approach to Information Retrieval*. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '98, pages 275–281, New York, NY, USA, 1998. ACM.

[Porter 1980] Martin F Porter. *An algorithm for suffix stripping*. Program, vol. 14, no. 3, pages 130–137, 1980.

[Rau & Jacobs 1989] Lisa F. Rau et Paul S. Jacobs. *NL ∩ IR : Natural language for information retrieval*. International Journal of Intelligent Systems, vol. 4, no. 3, pages 319–343, 1989.

[Rijsbergen 1979] C. J. Van Rijsbergen. Information retrieval. Butterworth-Heinemann, Newton, MA, USA, 2nd édition, 1979.

[Robertson & Jones 1976] Stephen E Robertson et K Sparck Jones. *Relevance weighting of search terms*. Journal of the American Society for Information science, vol. 27, no. 3, pages 129–146, 1976.

[Robertson & Walker 1997] S. E. Robertson et S. Walker. *On Relevance Weights with Little Relevance Information*. In Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '97, pages 16–24, New York, NY, USA, 1997. ACM.

[Robertson *et al.* 1994] "S E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu et M. Gatford". *Okapi at TREC-3*. In TREC'94, pages 109–126, 1994.

[Robertson 1977] Stephen E Robertson. *The probability ranking principle in IR*. Journal of documentation, vol. 33, no. 4, pages 294–304, 1977.

[Roelleke 2013] Thomas Roelleke. Information retrieval models : Foundations and relationships. Morgan & Claypool Publishers, 1st édition, 2013.

[Roy *et al.* 2016] Dwaipayan Roy, Debjyoti Paul, Mandar Mitra et Utpal Garain. *Using Word Embeddings for Automatic Query Expansion*. CoRR, vol. abs/1606.07608, 2016.

[Salton & Buckley 1988] Gerard Salton et Christopher Buckley. *Term-weighting approaches in automatic text retrieval*. Information processing & management, vol. 24, no. 5, pages 513–523, 1988.

[Salton & McGill 1986] Gerard Salton et Michael J. McGill. Introduction to modern information retrieval. McGraw-Hill, Inc., New York, NY, USA, 1986.

[Salton *et al.* 1983] Gerard Salton, Edward A Fox et Harry Wu. *Extended Boolean information retrieval*. Communications of the ACM, vol. 26, no. 11, pages 1022–1036, 1983.

[Salton 1969] Gerard Salton. *A comparison between manual and automatic indexing methods*. American Documentation, vol. 20, no. 1, pages 61–71, 1969.

[Salton 1971a] G. Salton. The smart retrieval system—experiments in automatic document processing. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.

[Salton 1971b] Gerard Salton, editeur. Relevance feedback in information retrieval. Prentice Hall, Englewood, Cliffs, New Jersey, 1971.

[Salton 1989] Gerard Salton. Automatic text processing : The transformation, analysis, and retrieval of information by computer. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[Salton 1997] Gerard Salton. *Improving retrieval performance by relevance feedback*. Readings in information retrieval, vol. 24, no. 5, pages 355–363, 1997.

[SanJuan & Ibekwe-SanJuan 2010] Eric SanJuan et Fidelia Ibekwe-SanJuan. *Multi Word Term Queries for Focused Information Retrieval*. In Alexander Gelbukh, editeur, Computational Linguistics and Intelligent Text Processing, volume 6008 of *Lecture Notes in Computer Science*, pages 590–601. Springer Berlin Heidelberg, 2010.

[Saracevic 1975] Tefko Saracevic. *Relevance : A review of and a framework for the thinking on the notion in information science*. Journal of the American Society for Information Science, vol. 26, no. 6, pages 321–343, 1975.

[Saracevic 1996] Tefko Saracevic. *Relevance reconsidered*. In Proceedings of the second conference on conceptions of library and information science (CoLIS 2), pages 201–2018, 1996.

[Schamber *et al.* 1990] Linda Schamber, Michael B. Eisenberg et Michael S. Nilan. *A re-examination of relevance : toward a dynamic, situational definition*. Information Processing Management, vol. 26, no. 6, pages 755 – 776, 1990.

[Shaalan *et al.* 2012] Khaled F. Shaalan, Sinan Al-Sheikh et Farhad Oroumchian. *Query Expansion Based-on Similarity of Terms for Improving Arabic Information Retrieval*. In Zhongzhi Shi, David B. Leake et Sunil Vadera, editeurs, Intelligent Information Processing, volume 385 of *IFIP Advances in Information and Communication Technology*, pages 167–176. Springer, 2012.

[Shi & Nie 2009] Lixin Shi et Jian-Yun Nie. *Integrating Phrase Inseparability in Phrase-based Model*. In Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09, pages 708–709, New York, NY, USA, 2009. ACM.

[Shi & Nie 2010] Lixin Shi et Jian-Yun Nie. *Using Various Term Dependencies According to Their Utilities*. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10, pages 1493–1496, New York, NY, USA, 2010. ACM.

[Singhal *et al.* 1996] Amit Singhal, Chris Buckley et Mandar Mitra. *Pivoted document length normalization*. In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, pages 21–29. ACM, 1996.

[Sordoni *et al.* 2013] Alessandro Sordoni, Jian-Yun Nie et Yoshua Bengio. *Modeling Term Dependencies with Quantum Language Models for IR*. In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13, pages 653–662, New York, NY, USA, 2013. ACM.

[Sparck Jones 1972] Karen Sparck Jones. *A statistical interpretation of term specificity and its application in retrieval*. Journal of documentation, vol. 28, no. 1, pages 11–21, 1972.

[Sparck Jones 1974] Kareen Sparck Jones. *Automatique indexing*. Journal of Documentation, vol. 30, no. 4, pages 393–432, 1974.

[Strzalkowski 1999] T. Strzalkowski. Natural language information retrieval. Text, Speech and Language Technology. Springer Netherlands, 1999.

[Tadić & Šojat 2003] Marko Tadić et Krešimir Šojat. *Finding Multiword Term Candidates in Croatian*. In Information Extraction for Slavic Languages 2003 Workshop, 2003.

[Tannier 2006] Xavier Tannier. *Extraction et recherche d'information en langage naturel dans les documents semi-structurées*. Theses, Ecole Nationale Supérieure des Mines de Saint-Etienne, Sep 2006.

[Tayli & Al-Salamah 1990] Murat Tayli et Abdulah I Al-Salamah. *Building bilingual microcomputer systems*. Communications of the ACM, vol. 33, no. 5, pages 495–504, 1990.

[Turing 1950] Alan M Turing. *Computing machinery and intelligence*. Mind, vol. 59, no. 236, pages 433–460, 1950.

[Voorhees 1994] Ellen M. Voorhees. *Query Expansion Using Lexical-semantic Relations*. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '94, pages 61–69, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[Voorhees 2001] Ellen M Voorhees. *Overview of TREC 2001*. In TREC, 2001.

[Vulić & Moens 2015] Ivan Vulić et Marie-Francine Moens. *Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings*. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '15, pages 363–372, New York, NY, USA, 2015. ACM.

[Waller & Kraft 1979] WG Waller et Donald H Kraft. *A mathematical model of a weighted Boolean retrieval system*. Information Processing & Management, vol. 15, no. 5, pages 235–245, 1979.

[Watson 2007] J.C.E. Watson. The phonology and morphology of arabic. The Phonology of the World's Languages. OUP Oxford, 2007.

[Wei & Croft 2006] Xing Wei et W. Bruce Croft. *LDA-based Document Models for Ad-hoc Retrieval*. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06, pages 178–185, New York, NY, USA, 2006. ACM.

[Wong & Raghavan 1984] SK Michael Wong et Vijay V Raghavan. *Vector space model of information retrieval : a reevaluation*. In Proceedings of the 7th annual international ACM SIGIR conference on Research and development in information retrieval, pages 167–185. British Computer Society, 1984.

[Ye *et al.* 2013] Zheng Ye, Ben He, Lifeng Wang et Tiejian Luo. *Utilizing term proximity for blog post retrieval*. Journal of the American Society for Information Science and Technology, vol. 64, no. 11, pages 2278–2298, 2013.

[Yvon 2010] François Yvon. *Une petite introduction au traitement automatique des langues naturelles*. In Knowledge discovery and data mining, pages 27–36, 2010.

[Zahran *et al.* 2015] Mohamed A. Zahran, Ahmed Magooda, Ashraf Y. Mahgoub, Hazem Raafat, Mohsen Rashwan et Amir Atyia. Word representations in vector space and their applications for arabic, pages 430–443. Springer International Publishing, Cham, 2015.

[Zamani & Croft 2016] Hamed Zamani et W. Bruce Croft. *Embedding-based Query Language Models*. In Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval, ICTIR '16, pages 147–156, New York, NY, USA, 2016. ACM.

[Zhai & Lafferty 2001a] Chengxiang Zhai et John Lafferty. *A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval*. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '01, pages 334–342, New York, NY, USA, 2001. ACM.

[Zhai & Lafferty 2001b] Chengxiang Zhai et John Lafferty. *A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval*. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '01, pages 334–342, New York, NY, USA, 2001. ACM.

[Zhang *et al.* 2007] Wen Zhang, Taketoshi Yoshida et Xijin Tang. *Text classification using multi-word features*. In 2007 IEEE International Conference on Systems, Man and Cybernetics, pages 3519–3524. IEEE, 2007.

[Zhang *et al.* 2008] Wen Zhang, Taketoshi Yoshida et Xijin Tang. *Advanced Web and NetworkTechnologies, and Applications*. chapitre A Study on Multi-word Extraction from Chinese Documents, pages 42–53. Springer-Verlag, Berlin, Heidelberg, 2008.

[Zhang *et al.* 2009] Jiuling Zhang, Beixing Deng et Xing Li. *Concept Based Query Expansion Using WordNet*. In Proceedings of the 2009 International e-Conference on Advanced Science and Technology, AST '09, pages 52–55, Washington, DC, USA, 2009. IEEE Computer Society.

[Zhang *et al.* 2011] Wen Zhang, Taketoshi Yoshida et Xijin Tang. *A Comparative Study of TF*IDF, LSI and Multi-words for Text Classification*. Expert Syst. Appl., vol. 38, no. 3, pages 2758–2765, Mars 2011.

[Zhao *et al.* 2011] Jiashu Zhao, Jimmy Xiangji Huang et Ben He. *CRTER : Using Cross Terms to Enhance Probabilistic Information Retrieval*. In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '11, pages 155–164, New York, NY, USA, 2011. ACM.

[Zhu *et al.* 2012] Yadong Zhu, Yuanhai Xue, Jiafeng Guo, Yanyan Lan, Xueqi Cheng et Xiaoming Yu. *Exploring and Exploiting Proximity Statistic for Information Retrieval Model*. In Yuexian Hou, Jian-Yun Nie, Le Sun, Bo Wang et Peng Zhang, editeurs, Information Retrieval Technology, volume 7675 of *Lecture Notes in Computer Science*, pages 1–13. Springer Berlin Heidelberg, 2012.

[Zitouni 2014] I. Zitouni. Natural language processing of semitic languages. Theory and Applications of Natural Language Processing. Springer Berlin Heidelberg, 2014.

[Zuccon *et al.* 2015] Guido Zuccon, Bevan Koopman, Peter Bruza et Leif Azzopardi. *Integrating and Evaluating Neural Word Embeddings in Information Retrieval*. In Proceedings of the 20th Australasian Document Computing Symposium, ADCS '15, pages 12 :1–12 :8, New York, NY, USA, 2015. ACM.