Université de Montréal

Recherche d'Information Translinguistique sur les Documents en Arabe

par
Youssef Kadri

Département d' informatique et de recherche opérationnelle
Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)
en informatique

Septembre, 2008

Your file Votre référence
ISBN: 978-0-494-53542-4
Our file Notre référence
ISBN: 978-0-494-53542-4

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée:

Recherche d'Information Translinguistique sur les Documents en Arabe

présentée par:

Youssef Kadri

a été évaluée par un jury composé des personnes suivantes:

Guy Lapalme,	président-rapporteur
Jian-Yun Nie,	directeur de recherche
Nadia El-Mabrouk,	membre du jury
Jacques Savoy,	examinateur externe
Guy Lapalme,	représentant du doyen de la FES

Thèse acceptée le: .....

RÉSUMÉ

La Recherche d'information Translinguistique (RIT) traite du problème de la recherche de documents écrits dans une langue différente de celle des requêtes. Avec la popularisation de l'Internet qui a permis l'accès à des langues moins connues telle que l'arabe, cette recherche est de plus en plus en demande sur le Web. Toutefois, une langue comme l'arabe, qui est très riche morphologiquement et complexe, présente de nombreux défis à la Recherche d'Information (RI) et au traitement automatique des langages naturels en général. Cette thèse explore les problématiques de la RI monolingue en arabe et de la RIT anglais-arabe.

En RI monolingue arabe, un des défis principaux soulevés est le traitement morphologique qui vise à déterminer une forme appropriée d'index à partir des mots. Dans cette thèse, nous cherchons à identifier la meilleure technique de lemmatisation des mots arabes. Pour ce faire, nous proposons une approche qui essaye de déterminer le noyau d'un mot selon des règles linguistiques appuyées par des statistiques de corpus. Cette approche est comparée à l'approche traditionnelle, qui opère quelques légères troncatures sur un mot aux deux extrémités. Les deux méthodes sont testées et comparées sur une large collection de test. Les résultats montrent que la nouvelle méthode proposée aboutie à une meilleure performance que la méthode traditionnelle.

En RIT, le défi majeur est la traduction de la requête vers la langue des documents. La traduction automatique plein texte est peu adaptée à traduire des requêtes, puisque les requêtes sont rarement des phrases et plus souvent juste une séquence de mots sans structure syntaxique. Dans ce contexte, l'utilisation des dictionnaires bilingues et les corpus parallèles deviennent des alternatifs intéressants, surtout que la traduction d'une requête vise à suggérer de bons termes pour trouver des documents et non à produire une phrase compréhensible par l'humain. Cependant les ressources en arabe sont limitées. Il existe des dictionnaires bilingues électroniques, mais il y a peu de corpus parallèles comme le Hansard. Ainsi, nous avons d'abord exploité le Web afin de construire automatiquement un corpus de

pages Web parallèles anglais-arabe. Sur la base de ce corpus, un modèle de traduction statistique est entraîné spécifiquement pour la RIT anglais-arabe.

Dans un contexte où les ressources sont limitées, il est souvent avantageux de combiner plusieurs ressources disponibles pour la traduction. En effet, la combinaison de plusieurs ressources de traduction permettrait de combler le manque de couverture d'une ressource pour certains termes de requête, et de bénéficier de l'effet d'expansion de requête, fortement souhaité en RI. Dans cette thèse, deux techniques de combinaison sont étudiées. La première méthode est traditionnelle : elle effectue une combinaison linéaire des ressources permettant de regrouper différentes traductions suggérées par différentes ressources pour un même mot en attribuant un poids de confiance pour chaque ressource globalement. La deuxième méthode utilise des facteurs de confiance associés à chaque traduction. Cette nouvelle méthode de combinaison de ressources reconsidère toutes les traductions candidates proposées par les différentes ressources et, en introduisant des attributs additionnels, elle les réévalue pour déterminer un nouveau score. Ces deux méthodes sont expérimentées sur deux collections de RIT anglais-arabe et les résultats ont montré que la méthode des facteurs de confiance est plus performante que la méthode traditionnelle.

Cette thèse apporte deux contributions. D'une part, elle propose une méthode de lemmatisation de mots en arabe, mieux adaptée pour la RI. D'autre part, elle propose une nouvelle méthode de combinaison de ressources de traduction en utilisant les facteur de confiance. À notre connaissance, c'est la première fois que les facteurs de confiance ont été utilisés dans le contexte de RIT.

Mots clés : Recherche d'information arabe, recherche d'information translinguistique, lemmatisation, traduction de requête, combinaison de ressources de traduction, facteurs de confiance.

ABSTRACT

Cross Language Information Retrieval (CLIR) deals with the problem of retrieving documents written in a language different from that of the query. The popularization of the Internet has enabled access to less known languages such as Arabic, and made the CLIR with Arabic language increasingly in demand. However, a language such as Arabic, which is morphologically very rich and complex, presents many challenges for Information Retrieval (IR) and natural language processing in general. This thesis investigates the problems of Arabic monolingual information retrieval and English-Arabic CLIR.

In Arabic monolingual IR, one of the main challenges is the morphological processing which aims to determine an appropriate form of index from words. In this thesis, we try to identify the best stemming technique for Arabic words. To this end, we propose a new approach which tries to determine the core of a word according to linguistic rules and corpus statistics. This approach is compared to the traditional approach, which operates some light truncations on both extremities of a word. Both methods are tested and compared on a large test collection. The results show that the proposed new method leads to a higher effectiveness than the traditional method.

In CLIR, the major challenge is the translation of the query to the document language. Full text machine translation is not fully adapted to query translation, since queries are rarely sentences and more often just a sequence of words without syntactic structure. In this context, the use of bilingual dictionaries and parallel corpora become interesting alternatives. This is especially the case given the fact that query translation aims to suggest good terms to retrieve documents and not to produce a human readable sentence. However, the resources in Arabic are limited. There are machine-readable dictionaries, but there are few parallel corpora such as Hansard. So, we first exploited the Web to automatically build a corpus of English-Arabic parallel Web pages. From this corpus, a statistical translation model is trained for English-Arabic CLIR.

In a context where the resources are limited, it is often advantageous to combine several available resources for the translation. Indeed, the combination of several translation resources would allow to improve the coverage of a resource for query terms, and to benefit from the query expansion effect, which is strongly desired in IR. In this thesis, two techniques of combination are studied. The first method is traditional: it makes a linear combination of resources, which groups translations suggested by different resources for the same word by assigning a global confidence to each resource. The second method uses confidence factors associated with each translation. This new method of combination of resources reconsiders all the translation candidates proposed by the different resources and, by introducing additional features, it re-evaluates them to determine a new score. These two methods have been tested on two English-Arabic CLIR collections and the results show that the method using confidence factors performs better than the traditional method.

This thesis made two contributions. On the one hand, it proposes a new method for stemming Arabic words, better suited for IR. On the other hand, it proposes a new method for the combination of translation resources using confidence factors. To our knowledge, it is the first time that confidence factors are used in the context of CLIR.

Keywords : Arabic information retrieval, Cross-language information retrieval, stemming, query translation, combination of translation resources, confidence factors.

REMERCIEMENTS

J'adresse tout particulièrement ma reconnaissance à mon directeur de recherche, Jian-Yun Nie, pour le soutien qui m'a apporté tout au long de la réalisation de cette thèse. Il a su me faire profiter de ses nombreuses connaissances et compétences dans le domaine de la recherche d'information.

Je tiens à remercier Monsieur Guy Lapalme d'avoir accepté de présider ce jury de thèse. Qu'il soit assuré de mon plus profond respect. Mes remerciements vont également aux professeurs Jacques Savoy et Nadia El-Mabrouk qui ont accepté d'examiner ce travail. Je leur suis pleinement reconnaissant pour leur participation à ce jury.

Je remercie les membres du laboratoire RALI : Elliott Macklovitch, Philippe Langlais, Fabrizio Gotti, Guihong Cao et Hugo Larochelle du laboratoire LISA, pour leur nombreuses aides. J'ai trouvé au RALI une ambiance particulièrement favorable pour mener à bien ce travail.

Enfin, je tiens à exprimer mes remerciements ainsi que mon affection à mes parents et à ma famille pour leur irremplaçable et inconditionnel soutien. Cette thèse est un peu la leur, aussi.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ANNEXES

CHAPITRE 1

INTRODUCTION

1.1 Contexte de recherche

L'invention de l'ordinateur au milieu des années 1940 a permis de numériser l'information et de la rendre disponible sous forme électronique. Peu après, le développement de supports de stockage a facilité l'archivage de grandes quantités d'information. Cet essor de l'informatique a conduit rapidement à une explosion de l'information. Naturellement, naîtra le concept de "recherche d'information" (information retrieval), proposé pour la première fois par Calvin N. Mooers en 1948, pour désigner l'automatisation de la recherche des informations qui dépassait les capacités de l'être humain. La Recherche d'Information (RI) est un domaine de recherche complexe et multidisciplinaire qui tire profit principalement de la science de l'information, de l'informatique, de la science cognitive et du traitement automatique des langages naturels. Elle se préoccupe de l'étude de l'ensemble des méthodes, des procédures et des techniques ayant pour objet d'extraire l'information pertinente à partir d'un ensemble de documents. Initialement, ses applications étaient liées à la bibliothéconomie : retrouver des références bibliographiques dans des bibliothèques. Peu après, des applications se sont étendues à d'autres domaines. Surtout, depuis la naissance de l'ère Internet, cette science connaît un regain d'intérêt.

Avec l'avènement de l'Internet au début des années 1990, l'information électronique a proliféré exponentiellement. Le constat de cette révolution est que les utilisateurs du Web se sont retrouvés entourés et parfois étouffés par cette grande quantité d'informations. Ces internautes se sont vite rendus compte qu'ils sont incapables d'utiliser efficacement cet univers d'informations. En conséquence, une demande croissante s'est manifestée pour les outils de recherche d'information.

C'est dans ce contexte que la RI a retrouvé sa vocation dans presque toute application sur le Web et s'est propulsée au premier plan des technologies de l'information. Récemment avec la popularisation de l'Internet, l'information multilingue est devenue de plus en plus disponible sur le World Wide Web. L'utilisation de langues différentes vient ajouter un autre défi à la recherche d'information. Avec des méthodes traditionnelles, pour une requête en français, on ne peut retrouver que des documents en français. Or en réalité, des documents en anglais peuvent aussi satisfaire le besoin d'un utilisateur s'exprimant en français. Ainsi, la pertinence d'un document est souvent indépendante de la langue utilisée. Naturellement pour franchir la barrière linguistique, un besoin alarmant d'outils spécifiques pour la Recherche d'Information Translinguistique (RIT) s'est fait sentir dans les divers secteurs de l'industrie. Le grand problème à traiter dans cette branche de la recherche d'information est notamment la traduction de la requête (ou la traduction du document).

1.2 Motivation et problématique

Dans la plupart des recherches précédentes en RI, les pionniers dans ce domaine ont concentré leurs efforts au développement d'outils de RI sur des collections en anglais. Ensuite graduellement, ils se sont intéressés à étudier les langues européennes et les langues asiatiques, notamment le chinois, le japonais et le coréen. Néanmoins, toute une famille de langues telle que l'arabe, n'a connu que peu d'intérêt par la communauté de recherche d'information. Parallèlement, la vulgarisation de l'Internet a de plus en plus permis l'accès à d'autres langues moins connues comme l'arabe. C'est dans cette optique que nous avons trouvé l'intérêt de notre travail de recherche dont l'objectif est d'explorer la langue arabe et de lui proposer des traitements spécifiques répondant aux besoins de la recherche d'information.

La langue arabe présente plusieurs défis au traitement automatique des lan-

gages naturels, en grande partie, dus à sa morphologie très riche et variable. Dans cette langue, le traitement morphologique devient particulièrement important pour la recherche d'information, parce que la RI doit déterminer une forme appropriée d'index à partir des mots. La plupart des études faites dans le contexte de la lemmatisation concluent que l'utilisation des termes obtenus à partir d'une analyse morphologique est plus efficace que l'utilisation des mots sans transformation [30] [60] [66]. L'arabe, de son coté, n'a pas échappé à ce fait. La lemmatisation des mots arabes a été une problématique majeure pour plusieurs travaux dans la RI arabe [37] [48] [18] [14]. Dans ces travaux, des approches pour lemmatiser les mots arabes sont proposées, certaines plus souples et d'autres plus sévères. Malgré ces études, il est encore peu clair quel type de lemmatisation est approprié pour la recherche d'information arabe. D'une part, une lemmatisation assouplie peut empêcher de grouper deux mots différents; mais elle court également le risque de ne pas grouper deux mots sémantiquement semblables, menant à un rappel plus faible. D'autre part, une lemmatisation plus sévère peut grouper incorrectement des mots sémantiquement non similaires dans un même index, menant à une précision plus faible. Une recherche plus poussée des effets de la lemmatisation sur l'efficacité de la RI est donc nécessaire. Ceci constitue le premier volet de cette thèse.

Un deuxième volet de cette thèse concerne la Recherche d'Information Translinguistique (RIT). La RIT est un cas particulier de la RI. La RIT s'intéresse à la recherche de documents écrits dans un langage T en utilisant des requêtes écrites dans un autre langage S. Comme nous avons mentionné précédemment, l'information représentée dans une langue autre que l'anglais est de plus en plus disponible et s'accroît rapidement. De ce fait, un utilisateur ne peut pas satisfaire toujours son besoin d'information par un système de RI monolingue. Il est possible que l'information pertinente qu'il cherche à repérer puisse être disponible, mais rédigée dans une langue autre que celle de la requête. Dans ce cas, un système conventionnel de RI monolingue ne peut pas répondre à cette question. Ainsi, plusieurs systèmes de RIT ont été développés, traitant le passage entre les dif-

férentes langues européennes [40] [53] [67]. Peu après, d'autres travaux en RIT ont étudié le passage de l'anglais vers les langues asiatiques [43] [55]. Cependant, d'autres langues comme l'arabe, n'ont pas connu le même intérêt de recherche en RIT qu'ont connu les langues européennes et asiatiques. Par rapport à d'autres paires de langues, nous disposons de beaucoup moins de ressources. Ainsi, nous allons étudier le problème de la RIT avec des ressources limitées.

Les principales approches de la traduction des requêtes sont la traduction automatique, les dictionnaires bilingues et les corpus parallèles [56]. Les systèmes de traduction automatique sont peu adaptées à traduire des requêtes, puisque les requêtes sont rarement des phrases et plus souvent juste une séquence de mots [80]. Les dictionnaires bilingues, malgré le fait qu'ils soient largement utilisés en RIT [7] [27], présentent quelques problèmes pour la traduction des requêtes. Ces problèmes sont reliés particulièrement au manque de couverture de traduction pour certains termes de requêtes tels que les noms propres ainsi que l'effet de polysémie[1] qui engendre parfois des traductions ambigües. Ajoutons aussi à ces inconvénients la difficulté d'obtenir de telles ressources pour la paire de langues anglais-arabe. Les corpus parallèles contiennent des informations utiles pour la traduction des mots dans des domaines particuliers. Cette technique a été utilisée en RIT pour traduire des requêtes de l'anglais vers le français et le chinois et a révélée de bonnes performances [52] [53]. Cependant, nous n'avons pas de corpus parallèles de très grande taille et leur constitution est coûteuse.

Quand plusieurs ressources de traduction limitées sont disponibles, il est évident de penser à leur combinaison pour bénéficier des avantages qu'offre chacune des ressources. Des méthodes simples ont souvent été employées pour ce genre de combinaison [53] [79] [19]. Ainsi, une question naturelle est comment ces ressources limitées peuvent être combinées de façon à maximiser leur effet collectif.

[1] Un mot qui possède plus d'un sens

1.3 Pistes de solution et approches

1.3.1 Recherche d'information arabe monolingue

Dans l'état de l'art actuel en RI, les mots sont souvent utilisés comme des index (éléments de base dans la représentation de documents et de requêtes). Ces index peuvent être créés de différentes façons. Dans cette étude, nous évaluons l'efficacité des différentes façons d'indexer les mots pour le compte de la recherche d'information monolingue arabe (les documents et les requêtes sont écrits en arabe). Plus particulièrement, par cette évaluation on cherche à identifier la meilleure technique de lemmatisation des mots arabes pour avoir des performances raisonnables en recherche monolingue.

Le principe de la lemmatisation est de représenter les termes ayant un sens similaire avec de petites différences sur la forme morphologique par une forme standard appelée racine ou lemme. Cependant, la morphologique de l'arabe est assez complexe due en majeure partie à sa variation orthographique et au phénomène d'agglutination. En arabe, les mots sont constitués à partir d'une racine ou un lemme linguistique concaténé à des affixes. Quand plusieurs de ces affixes sont présents dans un mot, il est difficile d'extraire le lemme de ce mot. Ajoutons à ce problème l'ambiguïté que peuvent créer ces affixes i.e. une séquence particulière de lettres peut ou non jouer un rôle d'affixe, selon le mot.

Dans cette thèse, nous étudions deux méthodes de lemmatisation pour les mots arabes. Une première méthode assouplie, opère quelques légères troncatures sur un mot aux deux extrémités. La décision de tronquer ou non un segment d'un mot est faite selon des statistiques sur l'ensemble des affixes tirés d'une grande collection de documents arabes. Cette méthode est similaire à celles proposées dans la littérature. L'autre méthode, nouvelle et linguistiquement motivée, essaye de déterminer le noyau d'un mot selon des règles linguistiques appuyées par des statistiques de corpus. Cette deuxième approche opère plusieurs décompositions

possibles sur un mot, produit d'abord un ensemble de lemmes candidats, et ensuite, grâce aux statistiques de corpus, choisit le lemme le plus utilisé dans le corpus. Nos expérimentations montrent que la nouvelle méthode est meilleure que la méthode traditionnelle.

1.3.2 Recherche d'information translinguistique anglais-arabe

Comme nous l'avons mentionné précédemment, beaucoup moins de ressources sont disponibles pour la traduction anglais-arabe. Dans notre cas, nous avons deux dictionnaires bilingues. En l'absence des corpus parallèles comme le Hansard, nous avons exploité le Web pour construire un corpus de textes parallèles pour la paire anglais-arabe, à l'aide d'un système de fouille automatique des pages parallèles. Sur la base de ce corpus, un modèle de traduction statistique est entraîné spécifiquement pour la RIT anglais-arabe [32]. De plus, un autre modèle de traduction statistique entraîné sur un corpus parallèle des nations unies [22] nous est aussi rendu disponible.

Une fois les ressources de traduction identifiées, la question qui se pose naturellement est : quelle est la meilleure façon de les combiner adéquatement pour tirer profit des avantages de chaque ressource afin d'augmenter la performance de recherche des documents pertinents ? Pour répondre à cette question, deux techniques de combinaison sont étudiées [35], l'une traditionnelle et l'autre nouvelle. La première méthode est une combinaison linéaire des ressources permettant de regrouper différentes traductions suggérées par différentes ressources pour un même mot en attribuant un poids de confiance pour chaque ressource.

Nous proposons une deuxième méthode, qui utilise des facteurs de confiance associés à chaque candidat de traduction. Cette nouvelle méthode de combinaison de ressources reconsidère toutes les traductions candidates proposées par les différentes ressources et, en introduisant des attributs additionnels, elle les réévalue plus radicalement que dans la combinaison linéaire. Une étude comparative

et expérimentale entre les deux méthodes de combinaison des ressources de traduction est réalisée sur différentes collections de RIT [36]. Nous allons montrer que notre nouvelle méthode aboutit à une meilleure performance que la méthode traditionnelle.

1.4 Contributions

Dans cette thèse, deux problématiques majeures sont soulevées et étudiées : la recherche d'information monolingue arabe et la recherche d'information translinguistique anglais-arabe. Plus particulièrement dans la première partie, notre défi est l'identification de la meilleure technique de lemmatisation des mots arabes pour avoir de bonnes performances en RI arabe monolingue. Dans la partie translinguistique, le problème de la traduction des requêtes de l'anglais vers l'arabe est abordé. Notre objectif est la proposition des méthodes de traduction de requêtes efficaces pour la recherche d'information translinguistique. Nous dressons nos principales contributions apportées à la résolution de ces deux problématiques :

- Nous proposons des traitements morphologiques spécifiques pour la langue arabe. En plus des prétraitements relatifs à la normalisation morphologique de certains caractères, ces traitements touchent principalement la lemmatisation des mots arabes. Une nouvelle méthode de lemmatisation est conçue. Cette méthode basée sur des concepts linguistiques essaye de déterminer le noyau d'un mot selon des règles linguistiques confirmées par des statistiques de corpus. L'originalité de cette méthode est qu'elle élimine tous les affixes d'un mot et les lemmes obtenus encodent la sémantique de base dans la langue arabe.

- Pour renforcer la traduction des requêtes et bénéficier des avantages de différentes ressources de traduction, une nouvelle méthode de combinaison des ressources est introduite en RIT. Cette approche est basée sur l'estimation de confiance d'une traduction. Cette mesure de confiance peut être vue comme un mécanisme général pour combiner d'une manière efficace des ressources

de traduction différentes et non homogènes. C'est la première fois que cette méthode est utilisée dans le domaine de la RIT [34].

1.5 Organisation de la thèse

Cette thèse est structurée en six chapitres. Le deuxième chapitre présente l'état de l'art de la recherche d'information, les concepts clés ainsi que les modèles de base sur lesquels repose la recherche d'information. Les techniques de traduction de requête d'une langue source vers la langue cible de documents pour le besoin de la recherche d'information translinguistique font l'objet du troisième chapitre. Le quatrième chapitre étudie la problématique de la recherche d'information arabe monolingue. Plus particulièrement, on présente dans ce chapitre les traitements morphologiques ainsi que la question de lemmatisation pour la RI arabe. Le cinquième chapitre présente notre méthode pour résoudre le problème de la RIT anglais-arabe. Différentes techniques de traduction ainsi que leur combinaison sont étudiées. Enfin, le dernier chapitre récapitule nos principales contributions apportées à la problématique de recherche d'information monolingue et translinguistique sur les documents en arabe et identifie quelques avenues pour des travaux futurs.

CHAPITRE 2

RECHERCHE D'INFORMATION

2.1 Introduction

Historiquement, la Recherche d'Information (RI) est un domaine lié aux sciences de l'information et à la bibliothéconomie qui ont toujours eu le souci d'établir des représentations des documents dans le but d'en récupérer des informations, à travers la construction d'index. L'informatique a permis le développement d'outils pour traiter l'information et établir la représentation des documents au moment de leur indexation, ainsi que pour rechercher l'information. Classiquement, lorsqu'on demande à un bibliothécaire de trouver des documents pour nos besoins d'information, ce dernier va manuellement chercher et trouver les documents qu'on lui a demandés, c'est à dire les documents pertinents. Le défi de la recherche d'information est d'imiter cette interaction en remplaçant le bibliothécaire par un système automatique [72]. De prime abord, la tâche semble trop compliquée pour être réalisée automatiquement dans la mesure où la compréhension automatique du langage naturel est toujours une problématique de recherche non résolue. Les pionniers de la recherche d'information se sont investis à modéliser le processus de repérage de documents pertinents répondant à une requête en simulant le comportement du bibliothécaire pour rechercher les documents demandés par les usagers.

Plus précisément, un système de recherche d'information est un système qui permet de retrouver automatiquement les documents pertinents à partir d'une grande collection de documents pour une requête donnée. De cette définition, on dégage quelques éléments clefs dans un système de RI : "document", "requête", et une relation entre eux - "pertinence". La requête exprime le besoin d'information d'un utilisateur. Elle est généralement formulée de quelques mots clefs. Les documents sont les entités de la collection. Traditionnellement, la RI est souvent perçue

comme synonyme de recherche des documents ou recherche des textes [21], bien que plusieurs systèmes de RI peuvent chercher des images, des séquences audio ou vidéo. Dans cette thèse, on s'intéresse seulement à l'information textuelle et par conséquent la recherche portera sur les documents textes. Finalement, la pertinence exprime la correspondance entre un document et une requête. Un document est jugé pertinent pour une requête si le contenu de ce document répond au besoin de l'utilisateur. Ce jugement est généralement fait par l'utilisateur, ou dans le contexte d'expérimentation, par un évaluateur humain. La pertinence reste toujours un concept flou, qui est difficile à caractériser précisément [74]. C'est pourquoi la RI est une tâche très difficile.

Pour mieux cerner le contexte de notre travail, nous proposons un petit corpus de trois documents et une requête qui serviront de projection de toutes les étapes de la RI que nous allons présenter dans les sections qui suivent. Ces trois documents contiennent respectivement les phrases suivantes :

1. Google est un moteur de recherche d'information bien connu sur le Web.
2. Le gouvernement alloue plus de budget à la recherche scientifique.
3. Le Web est devenu la source d'information la plus populaire.

Nous voulons chercher le (ou les) document (s) pertinent (s) parmi les trois documents cités pour la requête suivante : "recherche d'information sur le Web".

La recherche des documents est un processus renfermant deux procédures principales : indexation et recherche, qui sont en étroite relation l'une avec l'autre. L'indexation fait référence à la façon dont les documents sont restructurés pour être recherchés, et à la manière dont les requêtes (besoins d'information des utilisateurs) sont représentées [71]. La recherche est le procédé de comparaison des représentations respectives du document et de la requête pour estimer la pertinence du document à l'égard de la requête. La manière typique de cette comparaison est

une mesure de similarité entre les représentations du document et de la requête.

Sur notre petit corpus, il est évident que notre requête de "recherche d'information sur le Web" va retourner le document 1 comme document le plus pertinent dans le corpus. Ce document est le seul parmi les trois qui contient tous les termes (mots) de la requête. Pour comprendre comment le mécanisme de recherche nous conduit à classer le document 1 au premier rang dans la liste des documents retrouvés, il est utile de définir d'abord quelques concepts clés de la RI et de décrire les modèles de base sur lesquels repose la RI classique.

2.2 Notion de pertinence

La pertinence peut être définie [65] comme la relation entre un document désiré et une requête. Elle est aussi vue comme une mesure d'informativité et de topicalité (i.e. la relation avec le sujet demandé) du document à la requête. Cette notion de pertinence est le pivot de la RI car toutes les évaluations tournent autour d'elle. Cependant ce concept reste vague et subjectif car la définition même de la pertinence est imprécise. De plus les utilisateurs d'un système de RI ont des besoins très variés et des critères assez différents pour juger si un document est pertinent. Ainsi, quand on évalue un système de RI, on doit utiliser une collection de test dans laquelle les jugements de pertinence sont sujet de discussions et de consensus de plusieurs évaluateurs humains.

2.3 Evaluation

Les systèmes de recherche d'information sont toujours évalués en fonction de la pertinence des documents retrouvés. Cette tâche est très compliquée si on prend en considération que le jugement de pertinence doit être fait par un humain pour tout document retrouvé. Afin de procéder à des évaluations automatiques, nous avons besoin de corpus de test "standard". Chaque corpus de test contient les trois ensembles de données suivants :

- Un ensemble de documents (corpus de documents ou collection de documents);
- Un ensemble de requêtes de test;
- Une liste de documents pertinents pour chaque requête (liste de référence).

Ainsi, un système de RI peut être utilisé pour trouver des documents pour les requêtes données, et nous pouvons comparer ces documents retrouvés avec la liste de documents pertinents pour évaluer la qualité du système. Toutefois, il est primordial que l'évaluation d'un système ne repose pas sur une seule requête, mais sur un ensemble (typiquement 50 ou plus) de requêtes. L'évaluation d'un ensemble de requêtes est assez objective surtout si ces requêtes traitent des sujets variés. Ainsi, l'évaluation du système doit tenir compte des réponses du système pour tout l'ensemble des requêtes. Et bien sûr, le troisième ensemble - la liste de documents pertinents - doit contenir les réponses idéales pour l'utilisateur pour chaque requête.

Le TREC (Text REtrieval Conference) est organisé depuis 1992 pour évaluer les systèmes de RI sur différentes collections. Cette série de conférences met à la disposition des participants un certain nombre de ressources pour l'évaluation des systèmes de recherche de textes. Les corpus de documents sont choisis parmi des collections d'ordre général, souvent de type journalistique comme le "Wall Street Journal" ou "Agence France Presse" d'où est dérivé notre corpus de test. Les requêtes sont créées selon des sujets d'intérêts variés et sont sélectionnées seulement si elles ont un nombre raisonnable de documents pertinents pour un tel sujet. Ces requêtes, comme les corpus de documents, sont distribués aux participants. Ces derniers utilisent leurs systèmes pour retourner des listes triées de documents relatives à chaque requête. A la fin, ces documents retrouvés seront évalués pour la pertinence par un ensemble d'évaluateurs [77]. Notons aussi que le CLEF (Cross Language Evaluation Forum), un forum européen équivalent à TREC, a été créé depuis 2000 et met à la disposition des chercheurs diverses ressources dans presque toutes les langues européennes pour évaluer les performances de la RI, comme la RI

translinguistique et multilingue. NTCIR est le pendant en Asie organisé au Japon, qui se concentre sur les langues asiatiques.

2.4 Métriques

Les deux principales métriques d'évaluation en recherche d'information sont la précision et le rappel. Ces métriques reflètent la comparaison des réponses d'un système pour l'ensemble des requêtes avec les réponses idéales (liste de références). Plus précisément, ces métriques sont définies comme suit :

Précision : La précision mesure le pourcentage des documents pertinents retrouvés parmi tous les documents retrouvés par le système.

$$\text{Précision} = \frac{\text{nombre de documents pertinents retrouvés}}{\text{nombre de documents retrouvés}} \tag{2.1}$$

Rappel : Le rappel mesure le pourcentage des documents pertinents retrouvés parmi tous les documents pertinents dans la base.

$$\text{Rappel} = \frac{\text{nombre de documents pertinents retrouvés}}{\text{nombre de documents pertinents dans la base}} \tag{2.2}$$

L'idéal pour un système de RI est d'avoir de bons taux de précision et de rappel en même temps. Les deux métriques ne sont pas indépendantes, il y a une forte relation entre elles : quand l'une augmente l'autre diminue. Ainsi, pour un système, on a une courbe de précision-rappel qui a en général la forme de la figure 2.1.

Un système qui retourne tous les documents de la base aura un rappel de 100% mais la précision sera très faible. D'un autre côté, un système retrouvant peu de documents aura sûrement une précision élevée, mais le rappel souffrira. Il faut

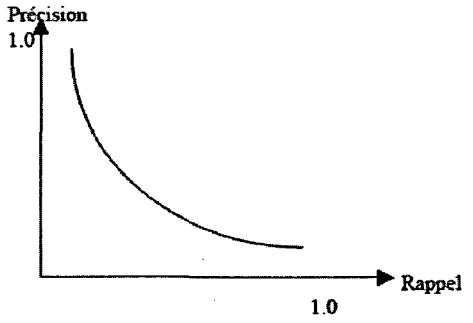

Figure 2.1: La courbe de précision-rappel

donc utiliser les deux métriques ensemble.

En pratique, la précision évolue en fonction du rappel et vice versa. Il est donc difficile de comparer deux systèmes avec un seul point de précision-rappel pour chaque système. Ainsi, deux autres métriques de moyenne sont largement utilisées pour l'évaluation des systèmes de recherche d'information. Ces deux métriques sont : la précision moyenne interpolée et non interpolée, exprimant la précision moyenne sur l'ensemble des points de rappel.

Précision moyenne sur 11 points : La précision moyenne sur 11 points consiste simplement à faire la moyenne des 11 précisions interpolées obtenues pour les points de rappels fixes, de 0 %, à 100 % par pas de 10 %. La règle d'interpolation est la suivante : la valeur interpolée de la précision pour un niveau de rappel i est la précision maximale obtenue pour tous les rappels supérieurs ou égaux à i.

Mean Average Precision (MAP) : La MAP caractérise la qualité du classement d'un système. Un système de RI calcule un score de pertinence pour l'ensemble des documents qui constituent la base de test, et les classe par ordre décroissant de pertinence à la manière des moteurs de recherche sur le web. En parcourant

cette liste, la précision est calculée pour chaque document pertinent. La MAP est obtenue en faisant la moyenne de ces différentes précisions. Pour un ensemble de requêtes, la MAP est calculée comme suit :

$$MAP = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{j}{rang(d_{ij})} \tag{2.3}$$

où d_{ij} est le j^e document pertinent pour la requête i, $rang(d_{ij})$ est le rang de ce document dans la liste de réponses du système, n_i est le nombre de documents pertinents pour la requête i, et N est le nombre de requêtes.

2.5 Mots outils

Les premiers modèles de RI utilisaient des approches naïves pour l'indexation résumant la recherche en une simple comparaison de chaînes de caractères entre les termes constituant la requête avec ceux des documents. En effet tous les mots de la requête étaient pris en considération mais ceci ne correspond guère au contexte de la RI où les documents sont considérés comme des sacs de mots. Intuitivement, on peut faire un certain nombre d'améliorations pour rendre la recherche plus utile. Le filtrage des mots outils et la lemmatisation sont les principales techniques d'amélioration du score de recherche et sont partagés par tous les modèles de RI que nous décrivons dans la suite de ce chapitre.

"Stop words" ou les mots outils sont des mots fonctionnels dans une langue dont la présence ne donne aucune spécificité ou informativité pour un document. Par conséquent, on ne veut pas les garder comme index parce qu'ils sont vides de sens pour les besoins de la RI. Ces mots apparaissent souvent dans tous les documents comme "le", "de", "à", etc. En pratique, l'élimination de ces mots des documents augmente généralement la précision de recherche de documents pertinents, d'où l'idée de créer une table contenant tous ces mots, appelée souvent "stoplist". Cette table renferme généralement des prépositions, des pronoms, certains adjectifs et

adverbes propres à chaque langue.

Pour mieux illustrer cette étape de RI, notre corpus spécimen dépourvu de ces mots fonctionnels devient :

1. google moteur recherche information connu Web.
2. gouvernement alloue budget recherche scientifique.
3. Web devenu source information populaire.

- Requête : recherche information Web.

2.6 Lemmatisation/troncature

Il y a une autre façon pour rendre les termes d'un document plus utiles pour une recherche efficace. Cette voie consiste en l'élimination des différences morphologiques non significatives sémantiquement. L'idée est de fusionner les termes ayant un sens similaire avec de petites différences sur la forme morphologique. On peut remarquer facilement que beaucoup de mots ont des formes légèrement différentes, mais leur sens reste le même ou très similaire. C'est notamment le cas des mots conjugués ou dérivés. Par exemple, les mots suivants ont des sens très similaires : informer, informés, informent, information, informateur. Si tous ces mots sont traités séparément, le rappel de recherche souffrira parce que cette différence de forme empêche le système de retrouver un texte dans lequel un mot similaire apparaît. Ainsi, l'idéal est d'éliminer toutes ces différences non significatives et ramener tous ces mots à une forme identique qu'on appellera le lemme (stem) ou la racine. L'idée de lemmatisation/troncature est d'éliminer ces indices de forme ou terminaisons à partir des termes et de ne garder que la racine ou le lemme. Il y a plusieurs méthodes de lemmatisation/troncature des mots. Nous présentons ici quelques unes :

1. **Méthode de troncature de Porter :** Cette méthode consiste à examiner seulement la forme de mot, et selon la forme, on déduit la racine. Cet algorithme élimine les terminaisons d'un mot anglais en cinq étapes : la première étape essaie de transformer le pluriel en singulier. Les étapes suivantes éliminent au fur et à mesure les dérivations comme par exemple le suffixe (ness) qu'on ajoute derrière certains adjectifs comme "happiness" [60]. Ainsi, "happiness" et "happy" seront tous transformés à "happi".

2. **Utilisation des dictionnaires :** Pour savoir si une séquence de lettres à la fin d'un mot correspond à une terminaison, il suffit de faire une élimination ou une transformation, et de voir si la forme obtenue existe dans le dictionnaire. Dans le cas contraire, d'autres possibilités sont ensuite envisagées. L'utilisation d'une telle méthode requiert un dictionnaire électronique et des tables renfermant tous les affixes possibles d'une langue. Cette approche a été utilisée pour le français dans [66]. Nos lemmatiseurs pour l'arabe que nous avons développés (chapitre 3) utilisent le même principe.

3. **Lemmatisation basée sur des étiqueteurs (taggeurs) :** L'idée de cette technique part du principe que pour trouver correctement le lemme d'un mot, il faut reconnaître sa catégorie grammaticale. Pour ce faire, il faut intégrer un étiqueteur automatique (analyseur de catégorie) dans un processus de lemmatisation. Avec ce mécanisme de reconnaissance de catégorie, on peut se permettre de transformer une forme de mot en une forme standard (dite aussi forme de citation).

Si on applique la première méthode de troncature (pour le français) sur notre corpus, il devient comme suit :

1. googl moteur recherch inform connaîtr Web.

2. gouvern allou budget recherch scientif.

3. Web deven sourc inform populair.

- Requête : recherch inform Web.

2.7 Les modèles de recherche d'information

Dans cette partie, nous allons discuter des quatre approches principales de modélisation de la recherche d'information, à savoir le modèle booléen, le modèle vectoriel de Salton, le modèle probabiliste, et l'approche utilisant les modèles de langue. Ces quatre classes constituent le fondement théorique de la matière.

2.7.1 Le modèle Booléen

Dans ce modèle, un document est représenté comme une conjonction logique de termes non pondérés i.e. les composants d'un document sont la présence ou l'absence des termes. Chaque terme du corpus est représenté par une variable logique. Ainsi un document d peut être vu comme ceci :

$$d = t_1 \wedge t_2 \wedge ... \wedge t_n \tag{2.4}$$

où t_i est un littéral formé avec le i^e terme du corpus. D'un point de vue booléen chaque document est une conjonction des termes le constituant. Les termes qui n'apparaissent pas dans un document sont représentés par leur négation comme par exemple dans notre corpus de trois documents :

1. $\neg$ allou $\wedge \neg$ budget $\wedge$ connaîtr $\wedge \neg$ deven $\wedge$ moteur $\wedge$ googl $\wedge \neg$ gouvern $\wedge$ inform $\wedge \neg$ populair $\wedge$ recherch $\wedge \neg$ scientif $\wedge \neg$ sourc $\wedge$ Web

2. allou $\wedge$ budget $\wedge \neg$ connaîtr $\wedge \neg$ deven $\wedge \neg$ moteur $\wedge \neg$ googl $\wedge$ gouvern $\wedge \neg$ inform $\wedge \neg$ populair $\wedge$ recherch $\wedge$ scientif $\wedge \neg$ sourc $\wedge \neg$ Web

3. $\neg$allou $\wedge \neg$ budget $\wedge \neg$ connaîtr $\wedge$ deven $\wedge \neg$ moteur $\wedge \neg$ googl $\wedge \neg$ gouvern $\wedge$ inform $\wedge$ populair $\wedge \neg$ recherch $\wedge \neg$ scientif $\wedge$ sourc $\wedge$ Web

D'un autre côté, la requête peut être vue comme une expression logique quelconque incluant les opérateurs ($\wedge$, $\vee$, $\neg$) comme $q = (t_1 \wedge t_2) \vee t_3$. L'objectif est de trouver

l'ensemble des documents qui impliquent la requête. La relation de pertinence $R(d, q)$ entre une requête q et un document d est déterminée par les formules suivantes :

$$R(d, q_i) = \begin{cases} 1 & \text{si } q_i \in d \ ; \quad q_i \text{ est un terme de } q \\ 0 & \text{sinon} \end{cases} \tag{2.5}$$

$$R(d, q_1 \wedge q_2) = \begin{cases} 1 & \text{si } R(d, q_1) = 1 \text{ et } R(d, q_2) = 1 \\ 0 & \text{sinon} \end{cases} \tag{2.6}$$

$$R(d, q_1 \vee q_2) = \begin{cases} 1 & \text{si } R(d, q_1) = 1 \text{ ou } R(d, q_2) = 1 \\ 0 & \text{sinon} \end{cases} \tag{2.7}$$

$$R(d, \neg q_i) = \begin{cases} 1 & \text{si } R(d, q_i) = 0 \\ 0 & \text{sinon} \end{cases} \tag{2.8}$$

La relation de pertinence dans ce modèle est binaire. Elle est vraie ou fausse. En effet le système retourne un ensemble de documents non ordonnés comme réponse à une requête. Si cette liste est longue, l'utilisateur doit encore fouiller dans cette liste non ordonnée pour identifier les documents qui sont vraiment pertinents à ses yeux. Cette façon de faire devient non fonctionnelle pour les utilisateurs d'un système de recherche booléen dans le cas où le système retourne beaucoup de documents, ou tout simplement il ne retourne rien.

D'autre part, le nombre d'occurrences d'un terme dans un document n'est pas pris en compte dans ce modèle ($t_i \wedge t_i = t_i$). Un terme présent dans un document 10 fois est vu de la même façon qu'un document où ce même terme n'apparaissant qu'une seule fois. Ceci ne répond guère aux besoins d'une recherche réelle.

Néanmoins le modèle de recherche booléen est assez puissant pour des usagers capables de formuler leurs besoins d'information d'une façon concise et précise : le langage d'interrogation est une expression quelconque de la logique de propositions, offrant une grande flexibilité aux usagers d'exprimer leurs besoins. Mais malheureusement, ceci n'est pas toujours le cas dans la pratique : un usager ordinaire manipule difficilement sinon mal les opérateurs logiques et par conséquent n'exprime pas bien son besoin d'information. En pratique, les systèmes basés sur cette approche naïve présentent généralement des performances faibles. Pour contourner ses lacunes et améliorer ses performances, des extensions sont proposées comme le modèle booléen basé sur les ensembles flous, qui tient compte des fréquences (où une mesure dérivée comme tf.idf - voir section 2.7.2) de mots dans le document. Nous ne décrivons pas ces extensions ici.

2.7.2 Le modèle vectoriel [63] [64]

Un modèle incontournable en recherche d'information est le modèle vectoriel. Il a été introduit par Salton en 1975 et développé par la suite par d'autres chercheurs. Il représente les documents et les requêtes comme des vecteurs de poids des termes. Le poids dans un vecteur peut être le nombre d'occurrences d'un terme correspondant dans ce document ou dans la requête, ou bien une mesure dérivée. Ces vecteurs prennent leur signification dans un espace vectoriel qui est défini par l'ensemble des termes que le système a rencontré durant l'indexation. Ainsi ce modèle représente une collection de documents par une matrice documents-termes où l'élément [i,j] dans la matrice indique l'association entre le i^e document et le j^e terme. Dans notre exemple, le petit corpus de trois documents ainsi que la requête sont représentés par le modèle vectoriel comme suit :

$$D = \begin{bmatrix} \text{allou} & \text{budget} & \text{connaîtr} & \text{deven} & \text{moteur} & \text{googl} & \text{gouvern} & \text{inform} & \text{populair} & \text{recherch} & \text{scientif} & \text{sourc} & \text{Web} \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix} \tag{2.9}$$

$$q = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \tag{2.10}$$

Ici la pertinence est représentée par la similarité entre le vecteur d'un document et le vecteur d'une requête suivant l'information présente dans l'espace. L'exemple le plus simple de cette similarité est le produit scalaire i.e. le document le plus proche à une requête est celui qui a le produit scalaire le plus élevé avec une requête.

$$Sim1(d, q) = d \,.q^t \tag{2.11}$$

Pour notre corpus spécimen, en utilisant la similarité du produit scalaire, cette mesure est calculée comme suit :

$$D \,.\, q^t = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix} \tag{2.12}$$

En observant ce schéma et contrairement au modèle booléen, le procédé de recherche produit présentement une liste triée. Ainsi le document 1 est classé le premier suivi du 3^e et enfin le 2^e à la fin de la liste.

Notons aussi que hormis la similarité du produit scalaire, il y a d'autres mesures de calcul dont la similarité cosinus qui est la plus connue et la plus utilisée. La

similarité cosinus est le produit scalaire d'un document avec une requête normalisée par la norme L-2 des vecteurs document et requête. Géométriquement, le document à la similarité cosinus la plus élevée avec la requête représente le document au plus petit angle avec la requête [72] :

$$Sim2(d,q) = \frac{d \,.q^t}{\| d \|_2 \,.\, \| q \|_2} \tag{2.13}$$

Où $\| d \|_2 = \sqrt{d^t \,.d} = \sqrt{\sum_{i=1}^{n} t_i^2}$ et $\| q \|_2 = \sqrt{q^t \,.q} = \sqrt{\sum_{i=1}^{n} q_i^2}$.

Dans cette partie, nous avons décrit le modèle vectoriel dans sa forme de base dans laquelle le poids d'un terme dans un document n'est que la fréquence de ce terme dans le document tf (term frequency). Des méthodes utilisant la distribution des termes dans la collection peuvent être utilisées. Plus précisément, ces méthodes introduisent la notion de l'IDF (inverse document frequency) qui calcule la valeur de discrimination d'un terme dans une collection. L'IDF pour un terme t_i de la collection est calculé comme suit :

$$idf_i = \log \frac{N}{n_i} \tag{2.14}$$

où N est le nombre de documents dans la collection et n_i est le nombre de documents dans la collection contenant le terme t_i. IDF tente d'attribuer des poids plus élevés aux termes qui sont rares dans la collection, et des poids plus faibles aux termes courants. Une fois ce facteur de discrimination introduit, la pondération d'un terme devient $tf.idf$:

$$w(t_i, d_j) = \log(tf_{ij} + 1).\ \log \frac{N}{n_i} \tag{2.15}$$

où tf_{ij} est la fréquence du terme t_i dans le document d_j.

2.7.3 Le modèle probabiliste [74]

Dans la RI, le modèle probabiliste classe les documents dans un ordre décroissant selon leur probabilité de pertinence pour une requête exprimant le besoin d'information d'un usager. Les concepteurs de ce modèle ont bâti leur approche sur la théorie de probabilités ainsi que celle des statistiques pour évaluer les probabilités de pertinence.

Etant donné deux ensembles R et NR représentant respectivement l'ensemble des documents pertinents et l'ensemble des documents non pertinents, le principe de base du modèle probabiliste est d'estimer les probabilités $P(R \mid D)$ et $P(NR \mid D)$ pour une requête donnée. Ces deux probabilités peuvent être interprétées de la façon suivante : si on retrouve le document D, quelle est la probabilité qu'on obtient l'information pertinente (R) et non pertinente (NR), c'est-à-dire $P(R \mid D)$ et $P(NR \mid D)$.

C'est autour de cette question que tournent toutes les initiatives pour la formalisation de la notion de pertinence dans le modèle probabiliste. Soit D un document et q une requête, le classement de ce document par rapport à cette requête est déterminé par la fonction O (odd) qui est égale à :

$$O(D) = \frac{P(R \mid D)}{P(NR \mid D)} \tag{2.16}$$

C'est la quantification de cette fonction qui permet de classer les documents. Plus $O(D)$ est élevée pour un document, mieux ce document est classé. Les deux probabilités exprimées dans la fonction $O(D)$ sont transformées par le théorème de Bayes :

$$P(R \mid D) = \frac{P(D \mid R)\ P(R)}{P(D)} \tag{2.17}$$

$$P(NR \mid D) = \frac{P(D \mid NR)\ P(NR)}{P(D)} \tag{2.18}$$

où $P(D \mid R)$ est la probabilité que D fait partie de l'ensemble des documents pertinents. $P(R)$ est la probabilité de pertinence, c'est-à-dire, si on choisit un document au hasard dans le corpus, quelle est la chance de tomber sur un document pertinent. $P(D)$ est la probabilité que le document soit choisi (si on prend au hasard un document dans le corpus, la chance de tomber sur D).

Comme $P(R)$ et $P(NR)$ sont des constantes pour une requête donnée, on peut les ignorer pour le but d'ordonner les documents. De plus, $P(D)$ est aussi considéré comme une constante (dans la première génération des modèles probabilistes). Ainsi, $O(D)$ peut être simplifiée seulement par la proportion de ces deux probabilités :

$$O(D) \simeq \frac{P(D \mid R)}{P(D \mid NR)} \tag{2.19}$$

Un document, comme une requête, est représenté par un ensemble de termes qui représentent des "événements". La présence ou l'absence de ces termes sont les seules caractéristiques observables dans les documents comme dans les requêtes. Ainsi un événement dénote soit la présence soit l'absence d'un terme dans un document. Pour un document contenant seulement deux termes t_1 et t_2, $P(D \mid R)$ et $P(D \mid NR)$ sont évaluées comme suit :

$$P(D \mid R) = P(t_1 = x_1, t_2 = x_2 \mid R) \tag{2.20}$$

$$P(D \mid NR) = P(t_1 = x_1, t_2 = x_2 \mid NR) \tag{2.21}$$

où $t_i = x_i$ exprime la présence ($x_i = 1$) ou l'absence ($x_i = 0$) du terme t_i dans le document D.

Dans la théorie des probabilités, la probabilité de la combinaison de plusieurs événements pris ensemble doit tenir compte des dépendances entre les événements, représentées dans la formule qui suit par des probabilités conditionnelles. Il est clair que dans le contexte de la RI, les présences et les absences des termes sont dépendantes. Ainsi, nous avons les formules suivantes :

$$P(t_1 = x_1, t_2 = x_2 \mid R) = P(t_1 = x_1 \mid R)\ P(t_2 = x_2 \mid t_1 = x_1, R) \tag{2.22}$$

$$P(t_1 = x_1, t_2 = x_2 \mid R) = P(t_2 = x_2 \mid R)\ P(t_1 = x_1 \mid t_2 = x_2, R) \tag{2.23}$$

Cependant, si on doit tenir compte de toutes les dépendances entre les termes dans un document, le calcul de $P(D \mid R)$ et de $P(D \mid NR)$ nous amène à un processus très complexe et très onéreux. En effet, c'est ce problème de complexité qui a conduit à l'hypothèse d'indépendance où on suppose que les événements liés à différents termes sont indépendants. Ainsi, le calcul de $P(D \mid R)$ et $P(D \mid NR)$ est réduit au produit des probabilités de chaque terme pris indépendamment des autres termes.

$$P(D \mid R) = \prod_{(t_i = x_i) \in D} P(t_i = x_i \mid R) = \prod_{(t_i = x_i) \in D} P(t_i = 1 \mid R)^{x_i}\ P(t_i = 0 \mid R)^{(1 - x_i)} \tag{2.24}$$

$$P(D \mid NR) = \prod_{(t_i = x_i) \in D} P(t_i = x_i \mid NR) = \prod_{(t_i = x_i) \in D} P(t_i = 1 \mid NR)^{x_i}\ P(t_i = 0 \mid NR)^{(1 - x_i)} \tag{2.25}$$

Maintenant l'estimation de ces deux probabilités revient à l'évaluation de $P(t_i = 1 \mid R)$ et $P(t_i = 1 \mid NR)$. Pour ce faire, on doit disposer d'un ensemble d'échantillons de documents dont le jugement de pertinence est déjà fait par un humain et tout ceci pour une requête fixée. Avec ces échantillons, on considère les paramètres suivants pour calculer $P(t_i = 1 \mid R)$ et $P(t_i = 1 \mid NR)$:

- N : le nombre d'échantillons,
- n : le nombre de documents pertinents,
- Ri : le nombre de documents contenant t_i,
- r_i : le nombre de documents pertinents contenant t_i.

Les probabilités sont calculées comme suit :

$$p_i = P(t_i = 1 \mid R) = \frac{r_i}{n} \tag{2.26}$$

$$(1 - p_i) = P(t_i = 0 \mid R) = \frac{n - r_i}{n} \tag{2.27}$$

$$q_i = P(t_i = 1 \mid NR) = \frac{R_i - r_i}{N - n} \tag{2.28}$$

$$(1 - q_i) = P(t_i = 0 \mid NR) = \frac{N - R_i - n + r_i}{N - n} \tag{2.29}$$

Après remplacement dans $O(D) = \frac{P(D|R)}{P(D|NR)}$, la fonction O devient :

$$O(D) = \frac{\prod_{t_i} p_i^{x_i} \, (1 - p_i)^{1 - x_i}}{\prod_{t_i} q_i^{x_i} \, (1 - q_i)^{1 - x_i}} \tag{2.30}$$

En introduisant la fonction *log* sur $O(D)$, on obtient :

$$\log O(D) = \sum_{t_i} x_i [\log \frac{p_i}{(1-p_i)} - \log \frac{q_i}{(1-q_i)}] + \sum_{t_i} \log \frac{(1-p_i)}{(1-q_i)} \tag{2.31}$$

Le deuxième terme de cette équation ne dépend pas du document (x_i). C'est une constante pour tous les documents et on peut donc l'ignorer pour classer les documents.

$$\log O(D) = \sum_{t_i} x_i \log \frac{p_i(1-q_i)}{q_i(1-p_i)} + c \simeq \sum_{t_i} x_i \; w_i \tag{2.32}$$

avec $w_i = \log \frac{p_i(1-q_i)}{q_i(1-p_i)}$ est le poids du terme t_i.

Finalement pour généraliser l'évaluation d'un document pour une nouvelle requête Q, on utilise la valeur de recherche (score du document) (Retrieval Status Value - RSV) suivante pour classer les documents:

$$RSV(D,Q) = \sum_{t_i} x_i \; y_i \; w_i \tag{2.33}$$

où y_i est le poids du terme t_i dans la requête. Il existe d'autres adoucissements de cette formule de pertinence du modèle probabiliste, dont la plus connue est la formule de Robertson-Sparck-Jones [61]. Le lecteur peut se référer au livre de van Rijsbergen [74] pour de plus amples détails sur le calcul et la simplification de ces probabilités.

2.7.4 Modèles de langue pour la RI

L'approche probabiliste traditionnelle tourne autour de la modélisation du concept de la pertinence. Pour une requête Q exprimant le besoin d'information d'un usager, le modèle probabiliste tente de déterminer $P(R|Q,D)$ la probabilité de per-

tinence d'un document D par rapport à la requête Q. Les approches utilisant les modèles de langue sont différentes. La modélisation de la notion de pertinence est contournée. Une de ces approches construit un modèle de langue probabiliste M_D à partir de chaque document D, et ordonne les documents selon la probabilité que la requête puisse être générée par le modèle de langue du document M_D. Cette approche est connue comme un modèle de vraisemblance de la requête (Query likelihood). Une autre approche construit deux modèles de langue : un pour le document et l'autre pour la requête et compare la similarité des deux modèles.

Le modèle de vraisemblance de la requête (Query likelihood)

Le premier modèle de la RI basé sur les modèles de langue est le modèle de vraisemblance de la requête [50]. Dans cette approche, on construit à partir de chaque document D de la collection C un modèle de langue M_D. Les documents sont ordonnés par $P(D|Q)$. Cette probabilité est interprétée par la vraisemblance que ce document soit pertinent pour la requête Q. En utilisant la règle de Bayes, $P(D|Q)$ est formulée par :

$$P(D \mid Q) = \frac{P(Q \mid D)P(D)}{P(Q)} \tag{2.34}$$

$P(Q)$ est la même pour tous les documents. Elle peut être donc ignorée car elle n'influence pas le classement des documents. La probabilité à priori d'un document $P(D)$ est souvent considérée comme uniforme pour tous les documents D et par conséquent elle peut être aussi ignorée. Avec ces simplifications, les documents seront ordonnées par seulement $P(Q|D)$, la probabilité pour que la requête Q soit générée par le modèle de langue M_D dérivée de D.

L'intuition derrière cette formulation est qu'au moment de formuler une requête, l'usager a une idée sur le document pertinent prototype et génère une requête basée

sur les mots qui peuvent apparaître dans ce document. C'est donc une génération de la requête à partir du document. Ainsi, le modèle tente de reproduire cette génération à partir d'un document de la collection.

Pour une requête Q constituée d'un ensemble de mots $Q = t_1 t_2 ... t_n$ et supposant que les mots qui apparaissent dans la requête sont indépendants, la probabilité de génération de requête est estimée comme suit :

$$P(Q \mid D) = P(t_1\ t_2\ ...\ t_n \mid \theta_D) = \prod_{t \in Q} P(t \mid \theta_D)^{c(t;Q)} \tag{2.35}$$

où $c(t;Q)$ est la fréquence du mot t dans la requête Q, et θ_D est le modèle de langue du document D. En utilisant l'estimation de maximum de vraisemblance (MLE), cette probabilité est estimée par :

$$P(Q \mid D) = \prod_{t \in Q} P_{MLE}(t \mid \theta_D) = \prod_{t \in Q} \frac{tf(t,D)}{\mid D \mid} \tag{2.36}$$

où $tf(t,D)$ est la fréquence du terme t dans le document D et $\mid D \mid$ est le nombre de mots (tokens) dans le document D. Le modèle du document considéré ici est un modèle unigramme.

Le problème classique avec ce type d'estimation est le cas où un terme de la requête est absent du document D : $P(t \mid \theta_D) = 0$ conduisant à $P(Q \mid D) = 0$. Or, un document dans lequel un terme de la requête est absent peut aussi être pertinent. Pour résoudre ce problème, il est essentiel d'introduire les techniques de lissage de probabilité. Une des stratégies qui fonctionne bien en pratique est l'utilisation d'une interpolation entre le modèle de document et le modèle de collection, plus connue sous le nom de lissage de Jelinek-Mercer :

$$P(t \mid \theta_D) = (1 - \lambda) P_{MLE}(t \mid \theta_D) + \lambda P_{MLE}(t \mid \theta_C) \tag{2.37}$$

où $P_{MLE}(t \mid \theta_C) = \frac{tf(t,C)}{|C|}$ est l'estimé du maximum de vraisemblance du modèle de langue unigramme basé sur la collection des documents C. λ $(0 \leq \lambda \leq 1)$ est un paramètre qui contrôle l'influence de chacun des deux modèles. Notons aussi que ce type de lissage modélise bien la spécificité des termes de la requête. Zhai *et al.* ont montré que l'introduction du facteur $P_{MLE}(t \mid \theta_C)$ produit un effet similaire à IDF du modèle vectoriel [81]. D'autres techniques de lissage peuvent être aussi utilisées comme "Dirichlet priors" et "Absolute discounting". Le lecteur peut se référer à [81] pour une étude exhaustive sur les techniques de lissage pour les modèles de langue appliqués à la RI. Finalement la formule générale pour classer les documents dans ce modèle de vraisemblance de la requête est :

$$P(Q \mid D) \simeq \prod_{t \in Q} (1 - \lambda) P_{MLE}(t \mid \theta_D) + \lambda P_{MLE}(t \mid \theta_C) \tag{2.38}$$

La comparaison des modèles (KL-divergence)

La similarité entre un document et une requête peut également être exprimée par la mesure de divergence de Kullback-Leibler (KL-divergence) [44]. L'idée de cette approche est de construire d'abord deux modèles de langue - un pour le document et l'autre pour la requête, ensuite de mesurer la divergence entre les deux modèles. D'un point de vue pratique, la KL-divergence exprime la distance entre deux distributions de probabilités. En théorie de l'information, ceci est interprété comme une différence entre deux entropies croisées ou bien le coût supplémentaire nécessaire pour encoder la requête dans le modèle de document. Ainsi, plus les deux modèles se divergent, moins le document doit être considéré comme une réponse appropriée à la requête.

Soient θ_Q le modèle de langue de la requête Q et θ_D le modèle de langue du

document D. La mesure de similarité entre un document et une requête par la KL-divergence est présentée comme suit :

$$Score(Q, D) = -KL(\theta_Q \parallel \theta_D) = \sum_{t \in V} P(t \mid \theta_Q) \log \frac{P(t \mid \theta_D)}{P(t \mid \theta_Q)} \tag{2.39}$$

$$Score(Q, D) = \sum_{t \in V} P(t \mid \theta_Q) \log P(t \mid \theta_D) - \sum_{t \in V} P(t \mid \theta_Q) \log P(t \mid \theta_Q) \tag{2.40}$$

Où V est le vocabulaire du langage. Le deuxième terme de l'équation 2.40 ne dépend que de la requête. Par conséquent, il n'influence pas le classement des documents. L'équation d'ordonnancement des documents peut être simplifiée à :

$$Score(Q, D) \simeq \sum_{t \in V} P(t \mid \theta_Q) \log P(t \mid \theta_D) \tag{2.41}$$

Dans la pratique courante, $P(t \mid \theta_Q)$ est estimé par le maximum de vraisemblance. Ainsi, la sommation sur $t \in V$ peut être simplifiée à $t \in Q$, car pour les termes non présents dans Q, $P(t \mid \theta_Q) = 0$.

$$Score(Q, D) \simeq \sum_{t \in Q} P(t \mid \theta_Q) \log P(t \mid \theta_D) \tag{2.42}$$

Lafferty et Zhai [44] ont démontré une relation d'équivalence entre les deux méthodes d'ordonnancement des documents : KL-divergence et le modèle de vraisemblance de la requête sont équivalents quant à l'ordre de documents quand une estimation de maximum de vraisemblance est utilisée pour le modèle de la requête. Pour montrer cette équivalence, il suffit de multiplier l'équation 2.42 par la taille de la requête pour obtenir la fonction d'ordonnancement des documents par la vraisemblance de la requête (équation 2.35) :

$$Score(Q, D) \simeq \sum_{t \in Q} c(t; Q) \log P(t \mid \theta_D) \tag{2.43}$$

$$Score(Q, D) \simeq \log \prod_{t \in Q} P(t \mid \theta_D)^{c(t;Q)} = \log P(Q \mid \theta_D) \simeq P(Q \mid D) \tag{2.44}$$

2.8 Conclusion

Ce chapitre nous a servi à présenter l'état de l'art du domaine de la recherche d'information qui est modélisé autour de deux concepts, à savoir la structure d'indexation des documents et le formalisme de recherche. Nous avons défini d'abord les notions fondamentales de la discipline comme la pertinence des documents par rapport à une requête, l'évaluation des systèmes et les principales métriques utilisées pour mesurer les performances d'une technique par rapport à l'autre. D'autre part, les approches fondamentales de la modélisation du procédé de recherche ont été énumérés et expliqués : le modèle booléen, l'approche vectorielle, le modèle probabiliste, ainsi que les approches basées sur les modèles de langue. Dans notre étude, nous allons utiliser les modèles de langue parce que ces derniers ont prouvé leur performance dans des expérimentations récentes et leur supériorité par rapport aux autres approches [59].

CHAPITRE 3

RECHERCHE D'INFORMATION TRANSLINGUISTIQUE

3.1 Introduction

Sur le Web, les documents sont écrits dans différentes langues. Si un usager exprime sa requête en anglais, il ne peut pas trouver l'information pertinente dans des documents écrits dans une langue autre que celle de la requête. C'est pour traiter ce problème qu'est née la Recherche d'Information Translinguistique (RIT); aussi connu sous le nom de "Cross Language Information Retrieval" (CLIR). Son but est de trouver des documents pertinents écrits dans une langue différente de la langue dans laquelle est exprimée la requête.

La principale différence entre la RIT et la recherche d'information monolingue est la différence de langue entre la requête et les documents. Le problème revient donc à trouver des méthodes qui traduisent avec succès les requêtes ou les documents dans une autre langue en vue de trouver des documents pertinents à une requête. Le problème supplémentaire dans la RIT est la traduction, soit de la requête soit des documents, pour que la requête devienne comparable avec les documents. La RIT fait partie d'une classe plus générale de la RI qui est la RI multilingue. En effet nous distinguons trois types de RI :

- RI monolingue : Recherche d'information utilisant une seule langue (les documents et les requêtes sont écrits dans la même langue).
- RI translinguistique (RIT) : retrouver des documents dans une autre langue.
- RI multilingue : retrouver des documents dans plusieurs langues.

Si une simple traduction des documents ou de la requête répond à la deuxième catégorie, dans la RI multilingue, en plus du module de traduction, il faudrait

un autre module de combinaison des résultats dans les différentes langues pour aboutir à une seule liste fusionnée (la façon traditionnelle de présenter les résultats de recherche). C'est ce problème de fusion qui augmente la complexité de la RI multilingue, en plus bien sûr de la difficulté de la traduction.

La RIT tire son profit et son importance dans les cas où on ne peut pas satisfaire des besoins d'information par des systèmes de RI monolingues. Les exemples qui suivent fournissent la motivation principale de la RIT :

- Pour une collection contenant des documents écrits en plusieurs langues, il est peu pratique de formuler une requête dans chaque langue pour la recherche.
- Un même document peut être écrit dans plus qu'une langue. A titre d'exemple : des documents dans lesquels des passages en anglais apparaissent mélangés avec le texte de récit dans une autre langue.
- Un utilisateur qui ne maîtrise pas suffisamment la langue de la collection pour exprimer sa requête dans cette langue, mais est capable de se servir des documents qui sont identifiés, sera intéressé à utiliser la RIT. C'est le cas d'un utilisateur qui est capable de lire, mais incapable de bien écrire dans la langue de la collection.

Dans notre travail, on s'intéresse plutôt à la RIT, et plus exactement la RIT avec l'arabe, c'est-à-dire on formule des requêtes en anglais et la recherche portera sur des documents écrits en arabe. Pour bien situer notre travail, nous présentons d'abord les principales approches utilisées pour la RIT.

Oard classe toutes ces approches en deux grandes catégories [56] : les techniques basées sur des bases de connaissances employant les ressources de langage pour franchir la barrière linguistique, et les techniques basées sur les textes parallèles. Toutefois, la première catégorie renferme elle-même trois techniques; la Traduction Automatique (TA), les dictionnaires bilingues et les vocabulaires contrôlés ou les

thésaurus. Nous nous attardons dans ce qui suit en détail sur chacune de ces méthodes.

3.2 Traduction Automatique (TA)

Ces techniques emploient un système de traduction automatique conventionnel pour traduire la requête ou le corpus de documents de sorte que tous les deux soient dans le même langage. On emploie alors un système de recherche monolingue conventionnel pour rechercher les documents. Il existe deux approches de base dans la traduction automatique : traduction des documents ou des requêtes.

Traduction de corpus des documents : Cette approche propose des aspects très intéressants, en particulier l'espoir que l'ambiguïté sera moins prononcée dans les documents relativement longs. En effet, la traduction automatique pourrait bénéficier de l'information contextuelle. Un autre avantage de cette approche est que les utilisateurs peuvent immédiatement recevoir les documents en leur langue préférée, ce qui leur permet de les consulter directement. Cependant, le volume à traduire est potentiellement très important, et il est difficile de prédire dans quelles langues il faut traduire.

Traduction des requêtes : Les pourvoyeurs de cette technique préfèrent la traduction des requêtes au lieu des documents. La traduction d'une requête de taille d'une phrase est considérablement moins coûteuse que la traduction des documents. Par contre, Yamabana *et al.* observent que les techniques adoptées avec succès par la communauté de la TA sont peu adaptées à traduire des requêtes, puisque les requêtes sont rarement des phrases complètes et plus souvent juste une séquence de mots [80]. Ils constatent aussi que le problème de résoudre l'ambiguïté pour les mots qui ont plus d'un sens dans des systèmes de traduction automatique reste toujours un défi majeur dans ce domaine.

Un autre inconvénient de la TA est que les systèmes opérationnels existants de la TA ne couvrent qu'un nombre réduit de paires de langues. Beaucoup de langues ne bénéficient pas de tels systèmes de TA. C'est notamment le cas de l'arabe.

3.3 Les dictionnaires bilingues

De nos jours, les dictionnaires bilingues conçus sous une forme utilisable par une machine sont plus largement disponibles et utilisés [7]. Ces dictionnaires sont souvent les analogues électroniques des dictionnaires imprimés.

L'approche de base est de prendre chaque terme dans la requête et de le remplacer par une liste de toutes ses traductions possibles, ceci par l'entremise d'une recherche des termes de la requête dans un dictionnaire compréhensible par une machine (Machine-Readable Dictionary - MRD). Cette utilisation du dictionnaire est simple à implanter. Néanmoins, il y a un certain nombre de facteurs qui limitent la performance de cette approche. Citons à titre d'exemple que cette façon de traduire les termes de la requête produit habituellement une expansion importante de la requête qui devient bruitée. En effet, les termes ont beaucoup de traductions possibles. Parmi ces traductions, non seulement on retrouve des synonymes, mais également des termes correspondant à des sens différents. Malheureusement, il est souvent difficile de désambiguïser les traductions. Par exemple, en utilisant un dictionnaire bilingue, le mot anglais "drug" est traduit en arabe par : "دوَاء" (médicament), "عقَّار" (drogue), "عقَاقير" (drogues), "خدّر" (droguer). Nous remarquons que ces traductions ont des sens différents et sans le contexte, on ne peut pas savoir quelle traduction est appropriée pour le mot "drug".

Un autre défi pour ces approches basées sur les dictionnaires bilingues est le fait que le dictionnaire est potentiellement incomplet i.e. il pourrait manquer la couverture de quelques termes. Ce cas de figure apparaît clairement pour une requête contenant des termes techniques pour lesquels un dictionnaire général ne peut répondre favorablement. Ces types de dictionnaires ne contiennent normalement pas de vocabulaire spécialisé. Les utilisateurs peuvent aussi entrer des abréviations, des acronymes ou de nouveaux termes dans leurs requêtes qui ne sont pas inclus dans les dictionnaires.

D'autre part, la traduction basée sur des mots ne réussit pas toujours à traduire correctement des mots composés ou des expressions qui contiennent plus d'un mot. Ce sont les dictionnaires idiomatiques ou terminologiques qui peuvent résoudre ce problème. Malheureusement, il est difficile de trouver ce genre de dictionnaires à large couverture. Beaucoup de paires de langues telles que anglais-arabe ne bénéficient pas de ce type de dictionnaires.

Les études faites par Ballesteros [7] et Grefenstette [27] montrent que la traduction des requêtes par les dictionnaires bilingues peut mener à une baisse de 40 à 60 % de performance de la RIT par rapport à la performance monolingue. Ils attribuent ceci aux trois facteurs introduits dans cette section : le manque de vocabulaire spécialisé dans les dictionnaires, l'introduction de l'ambiguïté par l'effet de polysémie, et l'impossibilité de traduire des concepts exprimés par des termes composés.

3.4 Le vocabulaire contrôlé (Thésaurus)

Un thésaurus est un dictionnaire hiérarchisé de concepts. Dans un thésaurus, les mots sont classés dans des catégories appelées concepts et les concepts sont reliés selon leurs relations sémantiques. Pour chaque concept, sont indiqués ses

synonymes (synset). Les relations entre les concepts sont de type hypernymie, hyponymie, etc. Dans un vocabulaire contrôlé (thésaurus), chaque concept est étiqueté par un terme descriptif unique dans le sens où l'usager peut facilement spécifier les concepts appropriés dans sa requête.

L'utilisation du vocabulaire contrôlé suppose qu'il existe un ensemble défini de concepts qu'on peut utiliser pour l'indexation et la recherche. Un usager émet un terme dans son langage pour trouver l'identificateur du concept correspondant au terme dans l'objectif de repérer les documents pertinents. La traduction des concepts se fait par une simple consultation d'un thésaurus incluant pour chaque concept les termes correspondants dans le langage cible. Ainsi, les techniques de traduction basées sur les thésaurus emploient les termes dans un thésaurus pour indexer les documents et les requêtes. Ensuite, un document est comparé à une requête en se basant sur les termes du thésaurus qu'ils ont en commun.

Un thésaurus multilingue peut être construit manuellement. L'exemple typique de ces théasaurus est celui du domaine juridique en Suisse pour les langues allemande, française et italienne. Une autre méthode pour construire un thésaurus bilingue est la traduction d'un thésaurus monolingue déjà existant. Pour le cas multilingue la construction d'un thésaurus multilingue se fait par la traduction d'un thésaurus monolingue vers plusieurs langages. Ainsi le principe qui découle de cette traduction est de donner la traduction correspondante pour chaque terme du thésaurus vers le langage cible.

On peut citer un des travaux les plus récents dans cette optique, le système CINDOR réalisé par une équipe de TextWise basé sur un thésaurus central qui est le WordNet [20]. Ce thésaurus contient hiérarchiquement différents termes anglais structurés en ensembles de synonymes (synsets). Chacun de ces ensembles a un seul identificateur. Chaque synset représente un concept dans la hiérarchie. Afin d'effectuer la RIT, chaque synset (ayant une signification et un numéro

d'identification unique) est approvisionné avec des termes du nouveau langage. Ceci produit un thésaurus parallèle dans le nouveau langage. On peut citer aussi le thésaurus EuroWordnet qui est structuré de la même manière que WordNet pour l'anglais. EuroWordnet couvre plusieurs langues européennes (Français, Espagnol, Italien, Allemand et Néerlandais) [78].

Un des avantages de l'approche des thésaurus est le contrôle des synonymes et de la polysémie par l'utilisation d'informations syntaxiques et sémantiques. Dans un thésaurus, l'information sémantique (appelée "scope note") est souvent fournie pour aider les utilisateurs à choisir manuellement les termes corrects. Un système de recherche basé sur les concepts peut appliquer cette idée par l'étiquetage automatique des mots avec leurs catégories grammaticales (Part-Of-Speech - POS), puis la sélection des traductions appropriées à ces catégories grammaticales [56]. En contrôlant les synonymes, cette approche permet de réduire l'ambiguïté et par conséquent augmenter la précision. Mais l'inconvénient de cette approche est que les applications translinguistiques se sont toujours heurtées à la complexité de développer des thésaurus multilingues.

Un autre problème possible est le manque de correspondance entre un mot dans le langage source et un mot dans le langage cible. Dans d'autres cas, un terme dans la langue source signifiant un concept ne peut être exprimé par un seul terme dans la langue cible; sa traduction nécessite une expression ou un syntagme regroupant plus d'un terme. De plus, beaucoup de termes peuvent manquer dans un thésaurus.

3.5 Les corpus parallèles

Les approches à base de corpus parallèles tentent de surmonter les limitations des techniques basées sur les connaissances établies manuellement et surtout la difficulté de construire des dictionnaires bilingues ou des thésaurus sophistiqués pour des applications diversifiées. Ces approches analysent de grandes collections de

textes et extraient automatiquement les informations nécessaires pour construire des techniques de traduction statistique. Les collections analysées sont des textes parallèles (un texte avec son équivalent traduit). Le tableau 3.1 présente un exemple de textes parallèles anglais-arabe. Les traductions d'une requête sont obtenues par le remplacement de chacun des termes de la requête dans sa langue source avec des termes qui ont une haute probabilité de traduction dans la langue cible.

Swiss Judge Demands Pinochet's Extradition to Switzerland. The Prosecutor-General of the Canton of Geneva, Bernard Bertosa, announced today Monday, that a Swiss judge decided to issue a temporary arrest warrant against Augusto Pinochet in preparation for his extradition to Switzerland. The Prosecutor-General explained, in a statement received by Agence France Presse, that the investigation judge's request was sent today, Monday, to the Federal Police Authority, which has the responsibility of deciding whether or not it will forward this request to the British authorities.
قَاض سويسري يطَالب بتسليم بينوشيه الَى سويسرَا . اعلن المدعي العَام في كَانتون جنيف برنَار بيرتوسَا اليوم الَاثنين ان قَاضيَا سويسريَا قرر إِصدَار مذكرة توقيف موءقتة في حق اوغوستو بينوشيه تمهيدَا لتسليمه الَى سويسرَا . وَاوضح المدعي العَام في بيَان تلقته وكَالة فرَانس برس ان طلب قَاضي التحقيق رفع اليوم الَاثنين الَى الهيئة الَاتحَادية للشرطة التي تعود اليهَا مسوءولية تقرير مَا اذَا ستحول هذَا الطلب الَى السلطَات البريطَانية .

Tableau 3.1: Un exemple de textes parallèles anglais-arabe

Plus précisément, ces techniques emploient des corpus écrits en deux langages différents pour produire un modèle probabiliste. On tente de relier les termes d'un langage à leurs plus proches traductions dans l'autre langage. Pour que cette méthode réussisse, les corpus doivent être parallèles. En général, les méthodes procèdent en alignant les phrases des corpus phrase par phrase. Ensuite, le système crée une représentation globale permettant de traduire un terme en un ensemble de termes probables selon plusieurs paramètres tels que la position des mots dans les phrases. Cette tâche est l'objet des modèles IBM de traduction statistique [11]. Ces modèles

tentent de calculer la probabilité conditionnelle $p(f_j \mid e_i)$ entre les mots e_i et f_j qu'on appelle la probabilité de traduction. Le principe de calcul de ces probabilités est que plus une paire de mots apparaît souvent dans des phrases parallèles, plus ces deux mots sont la traduction probable l'un de l'autre. L'algorithme de maximisation d'estimation EM est la base d'estimation des paramètres des modèles IBM [11]. Cet algorithme itératif EM estime les probabilités de traduction $p(f_j \mid e_i)$ de manière à maximiser la vraisemblance d'un corpus parallèle d'entraînement. Les détails de calcul de ces probabilités sont présentés au cinquième chapitre.

Les textes parallèles ont été utilisés par Nie pour traduire des requêtes de l'anglais vers le chinois et le français et ont révélé de bonnes performances [54] [52]. L'utilisation des corpus parallèles en recherche d'information multilingue a fait l'objet d'autres études pour les langues européennes. Nous pouvons citer les travaux effectués par Sheridan et Ballerini sur l'allemand et l'italien ainsi que sur d'autres paires de langues [69].

L'application de cette approche pour la traduction des requêtes nécessite la disponibilité de corpus parallèles. Cependant, l'acquisition de tels corpus est très coûteuse. A l'heure actuelle, il n'y a pas de corpus parallèles significatifs pour beaucoup de paires de langues (y compris l'arabe) à l'image de la collection Hansard du parlement canadien regroupant des textes anglais et français. D'autre part, la qualité de traduction qu'offre cette approche dépend aussi de la taille, du temps de la création, de la culture et de la thématique des corpus parallèles.

3.6 Approches combinées

Comme chacune des méthodes de traduction a ses limites, il est naturel de penser à leur combinaison pour bénéficier des avantages qu'offre chacune d'elles. Par la combinaison des techniques ou des ressources de traduction, on s'attend à améliorer la qualité de traduction des requêtes et par conséquent avoir une bonne

performance de la RIT. On peut dégager trois cas où la combinaison des ressources peut apporter un plus pour la RIT :

- Augmenter la couverture d'une ressource : Une ressource de traduction prise séparément peut ne pas couvrir tous les termes d'une requête. Ce cas de figure arrive souvent avec des mots représentant des entités nommées où les dictionnaires bilingues ne couvrent pas toujours ce type de mots. Dans un tel cas, il est primordial de combiner cette ressource de traduction avec d'autres pour compenser les termes manqués.

- Raffinement des traductions et des probabilités de traduction : Une ressource de traduction ne couvre pas tous les mots avec le même degré de confiance. Pour certains mots, les traductions peuvent être exactes, alors que pour d'autres, elles sont inappropriées. De même pour les poids de traduction. Il arrive parfois que différentes ressources proposent la même traduction mais avec des probabilités différentes. Ainsi, une combinaison judicieuse des ressources permet de raffiner les traductions correctes avec des probabilités raisonnables.

- Expansion de requête : Comme différentes ressources peuvent suggérer différentes traductions, il vaut mieux combiner ces ressources afin d'obtenir autant de traductions correctes que possible. En conséquence, quand plusieurs traductions correctes sont fournies pour une requête, ceci conduit naturellement à l'effet d'expansion de requête, qui est souhaitable en recherche d'information.

En recherche d'information translinguistique, la plupart des expériences ont combiné les techniques basées sur des bases de connaissances avec les techniques basées sur les textes parallèles [53] [79]. Dans d'autres cas, on a employé une simple combinaison des techniques basées sur les dictionnaires bilingues et les lexiques de termes bilingues pour couvrir la traduction des entités nommées [47]. Dans ces

études, une combinaison linéaire de diverses ressources de traduction est généralement appliquée par l'attribution d'un poids de confiance à chacune des ressources de traduction.

Dans cette thèse, nous introduisons une nouvelle technique de combinaison des ressources en RIT. Cette technique, utilisée initialement en reconnaissance de la parole, est basée sur l'estimation de confiance des traductions. Dans la combinaison linéaire, on se limite à un simple regroupement des traductions proposées par des ressources différentes et non homogènes. Dans la nouvelle méthode de combinaison, en introduisant des attributs additionnels, les traductions candidates proposées par les différentes ressources sont réévaluées et filtrées selon des poids plus fiables.

3.7 Recherche d'information translinguistique avec l'arabe

Dans cette section, nous décrivons brièvement les approches principales pour la RIT avec l'arabe. Les premières expériences en recherche d'information translinguistique avec l'arabe ont commencé en 2001 et 2002 dans la conférence TREC (Text REtrieval Conference). Dans cette conférence, plusieurs travaux ont étudié la problématique de recherche de documents pertinents dans une large collection de documents en arabe en utilisant des requêtes en anglais [57]. La plupart des systèmes qui ont étudié cette problématique, ont utilisé une approche de traduction des requêtes basée sur des dictionnaires bilingues, où chaque terme de requête est remplacé par plusieurs de ses traductions fournies par la ressource de traduction pour créer une requête représentée en langage des documents [57] [47] [26]. D'autres systèmes ont exploité un modèle de traduction statistique entraîné sur le corpus parallèles des nations unies [22] [18] [17]. La troisième catégorie de ces systèmes de RIT arabe a opéré une combinaison d'un modèle de traduction statistique avec des dictionnaires bilingues. Cette combinaison a donné de meilleures performances par rapport aux autres approches utilisant des ressources individuelles [19] [13] [79]. La raison principale est qu'une telle combinaison peut tirer profit de

plusieurs ressources de traduction. Cependant, cette combinaison a généralement utilisé des méthodes simples i.e. on a combiné linéairement diverses traductions pour un même terme de requête par l'attribution d'un poids de confiance à chaque ressource de traduction. Un exemple typique est le travail de Xu *et al.*, qui ont utilisé une combinaison d'un lexique bilingue avec un corpus parallèle comme suit [79] :

$$p(e \mid a) = 0.8p_{UN}(e \mid a) + 0.2p_{Lexicon}(e \mid a) \tag{3.1}$$

où e est un mot anglais, a est un mot arabe, p_{UN} et $p_{Lexicon}$ sont les probabilités de traduction extraites respectivement du corpus parallèle des nations unies et d'un lexique bilingue. Des poids fixes (0.8 et 0.2) sont attribués aux deux ressources de traduction représentant le degré de confiance donné à chacune des ressources.

3.8 Discussion

Dans ce chapitre, nous avons énuméré les principales méthodes utilisées pour la traduction des requêtes dans les systèmes de la RIT : la traduction automatique, les dictionnaires bilingues, les thésaurus, les corpus parallèles et la combinaison de plusieurs méthodes. Toutes ces approches cherchent une traduction pour une requête, mais elles diffèrent dans ce qu'elles traduisent et comment elles le font. Nous avons aussi vu que chaque méthode présente certains avantages, mais son application exclusive en dépit des autres a ses inconvénients. A la fin de ce chapitre, nous avons présenté l'état de l'art de la RIT avec l'arabe.

Au moment où une traduction automatique de bonne qualité est toujours un problème non résolu, le choix d'une méthode de traduction appropriée dépend des ressources disponibles et du rôle de la traduction dans la RIT. En effet, compte tenu des avantages et des inconvénients de chaque méthode et la question de disponibilité des ressources nécessaires pour une paire de langues telle que anglais-arabe, le

problème revient d'une part, à collecter différentes ressources de traduction et d'autre part, à proposer des méthodes efficaces de combinaison de ces différentes ressources de traduction afin d'améliorer la performance de recherche d'information translinguistique. Mais avant l'étape de traduction, un traitement approprié sur les textes arabes est aussi un pré-requis. Ainsi, dans le chapitre suivant, nous allons traiter le problème de RI monolingue en arabe, avant de traiter le problème de RIT dans le chapitre 5.

CHAPITRE 4

RECHERCHE D'INFORMATION MONOLINGUE ARABE

4.1 Introduction

Dans ce chapitre, on étudie la première partie de notre projet, à savoir la recherche des documents pertinents dans une collection en langue arabe en utilisant des requêtes en arabe (le cas monolingue). Plus particulièrement, nous cherchons à identifier la meilleure technique de lemmatisation des mots arabes. Une fois ce problème résolu, nous aborderons le cas translinguistique.

Rappelons que la lemmatisation est utile en recherche d'information (voir section 2.6 du chapitre 2). La plupart des études faites dans le contexte de la lemmatisation concluent que l'utilisation des termes obtenus à partir d'une analyse morphologique est plus efficace que l'utilisation des mots sans transformation [60] [66] [30]. Pour déterminer les meilleurs termes d'index, nous avons réalisé une série d'expérimentations sur une large collection de textes arabes associée avec un ensemble de requêtes, et bien sûr les jugements de pertinence des documents pour chaque requête.

Dans la suite de ce chapitre, nous nous attardons d'une part, sur les propriétés morphologiques de l'arabe et leurs prétraitements nécessaires, et d'autre part sur les techniques de lemmatisation pour choisir les termes d'index les plus efficaces pour l'indexation et la recherche. Enfin nous concluons notre étude par la présentation de résultats des différentes expérimentations commentés par des analyses et évaluations.

4.2 Propriétés morphologiques de l'arabe

L'arabe est d'essence génétique différent des langues européennes. Il comprend 28 lettres. Il s'écrit cursivement de droite à gauche et sa représentation morphologique est assez complexe de par sa variation orthographique et son phénomène d'agglutination [31]. Ses lettres changent de forme en fonction de leur position dans le mot (début, milieu, fin, isolé) (Tableau 4.1). Pour certaines lettres, malgré que leurs formes changent, leurs encodages ne changent pas et par conséquent leur traitement ne pose pas de problème. Mais pour certaines d'autres comme la lettre *hamza* (أ), leur encodage change d'un mot à un autre. Pour contourner ce problème, un traitement de normalisation orthographique est nécessaire. Dans les paragraphes qui suivent, nous dressons quelques particularités de cette langue tout en mettant en exergue les principaux problèmes relatifs à son traitement automatique pour le besoin de la recherche d'information :

Début	Milieu	Fin	Séparé
غراب	بغداد	تبغ	بلاغ
(corbeau)	(Baghdad)	(tabac)	(communiqué)

Tableau 4.1: Représentation du caractère "غ" (gh) au début, au milieu, à la fin et séparé (isolé) dans un mot

- Les mots arabes sont divisés en trois catégories : noms, verbes et particules. Les particules représentent les connections entre les mots à l'instar des prépositions et des pronoms. Notons ici que ces particules feront l'objet de la table des mots vides de sens (stop words) pour les besoins de la RI.

- La majorité des noms et des verbes sont dérivés d'un nombre réduit (environ

10 000) de racines verbales de type "فل". Ces racines sont des unités linguistiques porteuses d'un sens et la plupart de ces racines consistent en seulement trois consonnes, rarement quatre ou cinq consonnes.

- A partir des racines, on peut générer des lemmes ou des dérivés aussi bien nominaux que verbaux par l'application des patrons (règles morphologiques). On peut générer jusqu'à 30 lemmes à partir d'une racine à trois consonnes (trigramme) [32]. Voyons un exemple de la racine trigramme كتب (ktb) (écrire) où on peut produire plusieurs lemmes (Tableau 4.2) :

Ecrire	Livre	Ecrivain	Ecrit	Petit livre
كتب	كتاب	كاتب	مكتوب	كتيب

Tableau 4.2: Dérivation de plusieurs mots à partir de la racine "كتب" (écrire)

De cet exemple, nous remarquons qu'à partir d'une racine trigramme "كتب", on peut générer plusieurs mots. Dans tous ces mots générés, les trois lettres (ك, ت, ب) qui forment la racine sont toujours présentes, mais d'autres lettres représentant les patrons sont insérées au début, au milieu ou à la fin. Ainsi, pour générer les mots "كاتب" (écrivain) et "كتيب" (petit livre), les patrons (lettres) "ا" et "ي" sont ajoutés au milieu de la racine "كتب".

- Dans l'arabe écrit, les voyelles (diacritiques) sont normalement omises et comme résultat de la non voyellation, les mots tendent à avoir un haut niveau d'ambiguïté [31]. De ce fait, un mot arabe peut avoir plusieurs significations. Prenons à titre d'exemple le mot arabe non voyellé "جزر" qui a au moins

deux significations :" جُزُر" (îles) et " جَزَر"[1] (carottes). Pour une requête contenant ce mot non voyellé, la précision sera pénalisée dans le sens où un système de recherche d'information retournera tous les documents contenant le mot "îles" et ceux contenant le mot "carotte". En résumé, l'absence des voyelles dans l'arabe écrit génère une ambiguïté sur certains mots et cette ambiguïté peut pénaliser la performance de la recherche d'information arabe.

- Dans les langues européennes, les mots sont séparés par les séparateurs habituels comme l'espace et les autres signes de ponctuation. En arabe, ces mêmes séparateurs sont utilisés mais l'existence du phénomène d'agglutination rend parfois difficile la séparation entre les mots. Ce phénomène se manifeste quand certaines prépositions précédant des mots se collent avec ces derniers, et/ou certains pronoms sont rattachés aux mots vers la fin. Si la préposition ne se colle pas au mot comme dans l'exemple : في أمان (en sécurité), ceci ne pose pas de problème pour la tokenisation. Mais dans le cas où une préposition et/ou un pronom se collent à un mot, il est très difficile de détecter le lemme à partir de ce mot. Ce problème peut être illustré dans le mot arabe كعادتها (comme son habitude). La préposition ك (comme) se joint au mot عادة (habitude) et le pronom ها (sa) est rattaché vers la fin pour former un seul mot كعادتها (ك/عادة/ها). L'opération d'extraction d'un lemme comme عادة à partir d'un mot comme كعادتها n'est pas toujours facile à faire sans risque d'erreur, parce qu'une préposition comme ك (comme) peut aussi faire partie des lettres constituant le lemme. D'ailleurs le jeu de cette troncature fera l'objet de notre problématique sur la lemmatisation des mots arabes. Dans l'exemple précédent, nous remarquons aussi que certaines lettres changent de forme selon leur position dans le mot, ajoutant une autre difficulté. Pour les lettres ك et ها, malgré que leurs formes changent, leurs encodages ne changent pas. Mais pour la lettre ة, l'encodage change d'une

[1]Les marques qui sont ajoutées au dessus de ces deux mots arabes représentent les voyelles.

forme عادة (la lettre ة est à la fin du mot) à une autre كعادتها (la lettre ة est au milieu du mot). Ainsi, une normalisation est nécessaire pour traiter ce problème.

- En arabe, les mots sont constitués à partir d'une racine ou un lemme linguistique concaténés à des affixes. Ces affixes collés au début et à la fin des mots sont les antéfixes, les préfixes, les suffixes et les postfixes. L'exemple précédent كعادتها présente un lemme linguistique عادة concaténé à un antéfixe ك et un postfixe ها. L'exemple du tableau 4.3 présente le cas général où les quatre catégories d'affixes sont présents dans le mot ليفاوضونهم (pour qu'ils négocient avec eux).

Antéfixe	Préfixe	Noyau	Suffixe	Postfixe
ل	ي	فاوض	ون	هم
Préposition signifiant (pour que)	Une lettre signifiant le temps et la personne de conjugaison	négocier	Terminaison de conjugaison	Pronom signifiant (eux)

Tableau 4.3: Forme agglutinée d'un mot arabe signifiant "pour qu'ils négocient avec eux"

Dans ce mot, on note la présence d'une part, d'un antéfixe ل (pour que) et un préfixe ي (une lettre signifiant le temps et la personne de conjugaison) au début du lemme linguistique فاوض[2], et d'autre part un suffixe ون (deux

[2]L'origine de ce lemme est le verbe arabe تفاوض signifiant "négocier".

lettres exprimant la terminaison de conjugaison) et un postfixe هم (eux) à la fin du lemme. En arabe, ces affixes sont catégorisés selon leur rôle syntaxique. Les antéfixes sont généralement des prépositions. Les préfixes représentés par une seule lettre indiquent la personne de la conjugaison des verbes au présent. Les suffixes sont les terminaisons de conjugaison des verbes et les signes du duel, du pluriel et du féminin pour les noms. Enfin, les postfixes représentent des pronoms. L'ensemble de tous ces affixes feront l'objet du jeu de troncature durant l'étape de lemmatisation de la RI.

- En plus du phénomène d'ambiguïté, il y a un autre problème de la forme pluriel de certains noms irréguliers dit aussi le pluriel cassé. Ceci relève du fait qu'un nom au pluriel prend une autre forme orthographique que sa forme initiale au singulier. Il est difficile d'écrire un algorithme à base de règles pour réduire ce genre de pluriel au singulier.

- Evidemment, l'arabe est également très différent des langues européennes au niveau syntaxique. Mais ceci est au delà de notre intérêt dans cette thèse parce que l'aspect syntaxique n'est généralement pas pris en compte dans l'état de l'art en RI. Ainsi, nous nous limitons au traitement morphologique dans cette thèse.

4.3 Prétraitements nécessaires

4.3.1 Encodage

L'arabe est encodé sur le Web suivant plusieurs formats d'encodage comme Unicode, ISO-8859-6 et CP1256. Les textes recherchés et les requêtes peuvent être encodés différemment, les rendant incomparables. Par exemple, notre corpus de test provient de la collection TREC 2001. Les documents sont représentés en Unicode (UTF-8) et les requêtes, en ISO-8859-6. Un autre encodage (Windows CP1256) est utilisé sur le Web pour représenter les textes arabes. Afin d'apparier les documents avec les requêtes, nous devons réutiliser des outils de conversion

entre différents encodages en utilisant des tables de l'alphabet arabe[3]. Ainsi, tout a été transformé en format Unicode dans notre cas.

4.3.2 Tokenisation

La tokenisation consiste à identifier les mots dans une séquence de lettres. Pour la tokenisation des textes arabes, en plus des mêmes ponctuations présentes dans les textes anglais, nous avons ajouté d'autres signes de ponctuation arabe (encodés en arabe) comme la virgule, le point virgule et le point d'interrogation et nous les avons considérés comme des séparateurs. Evidemment, il y a des cas où ces signes ne séparent pas de mots dans les langues européennes comme "aujourd'hui". Mais ce phénomène ne se présente pas en arabe. Ainsi, tous ces signes agissent comme séparateurs de mots arabes.

4.3.3 Normalisation orthographique

Dans l'arabe écrit, les voyelles sont souvent omises dans les textes et un lecteur familier avec ce langage ne trouvera pas vraiment de difficulté pour lire correctement un texte sans voyelles. Néanmoins, on peut parfois trouver quelques voyelles présentes avec les mots dans les textes. Ainsi, l'élimination de ces voyelles est nécessaire pour fin de normalisation. D'autre part, certaines lettres subissent une légère modification dans l'écriture qui n'influe pas considérablement sur le sens du mot. Mais l'encodage de ces lettres change d'un mot à un autre. On peut citer à titre d'exemple la lettre *hamza* "أ". Au début des mots, elle peut être représentée par "أ" comme dans أكل (manger), "إ" dans إبل (chameaux) ou encore "آ" dans آبار (puits). Une autre raison pour ce prétraitement est qu'on a tendance fréquemment à mal écrire ces différentes formes de *hamza*. Ce genre d'erreurs est très répandu

[3]native2ascii Linux command

dans les textes arabes. Par exemple, le mot أكل est généralement écrit اكل. Il y a aussi la lettre "ة" à la fin des mots qui peut être écrite de deux façons : ة ou ه. Les deux mots arabes عادة et عاده signifient le même mot (habitude) malgré que leur dernière lettre est représentée différemment. Afin de tenir compte de toutes ces spécificités de ce langage et de pallier au problème de variation de représentation des caractères arabes dans les textes comme dans les requêtes, il est nécessaire d'adopter et d'appliquer quelques méthodes de normalisation sur le corpus avant l'indexation :

- Remplacer les *hamzas* (أ , إ , آ) par alifbar "ا" (A).
- Remplacer "ى" par "ي" (Y) à la fin des mots.
- Remplacer "ة" par "ه" (H) toujours à la fin des mots.
- Remplacer la séquence "ىء" par "يئ".
- Eliminer le caractère "tatweel" "kashida" (–) utilisé pour l'esthétique dans les textes arabes.
- Eliminer les diacritiques (voyelles) et la "chedda"[4].

Dans la section de résultats d'expérimentations, nous remarquons clairement une amélioration de performance de la recherche quand une normalisation est prise en considération.

4.3.4 Construction de Stoplist (Liste des mots outils)

Comme pour d'autres langues, l'arabe contient aussi des mots fonctionnels (ou mots outils) qui ne véhiculent pas un sens particulier utile pour la recherche d'information. Nous éliminons ainsi ces mots outils avant l'indexation des documents. Pour ce faire, une table de mots outils a été établie pour l'arabe. Elle contient aux environs de 350 mots et peut être enrichie. Cette table est présentée

[4]La chedda est une voyelle particulière en arabe. Son effet est de renforcer (doubler) la consonne sur laquelle elle est posée.

en annexe V. Cette table de mots outils renferme presque toutes les prépositions et les particules de l'arabe et les traductions de quelques mots outils (stopwords) anglais qu'on peut trouver dans les requêtes comme وثيقة (document). Cette table est d'une taille comparable à celle de l'anglais. D'autres tables de mots outils ont été conçues pour l'arabe. La plus répandue, et largement reprise par d'autres travaux dans ce domaine, est celle de Khoja renfermant 168 entités [37]. La raison pour laquelle la taille de notre table est plus large par rapport à celle de Khoja est la prise en compte du fait de l'agglutination que nous avons expliqué dans la section 4.2. Le fait qu'un pronom est souvent collé à une préposition en arabe génère une multitude de formes pour un même mot outil. Ainsi, nous pouvons trouver dans notre table plusieurs formes pour une même particule comme on le voit dans l'exemple suivant : قبل (avant), قبله (avant lui), قبلها (avant elle), قبلهم (avant eux), قبلهن (avant eux), قبلهما (avant eux), قبلك (avant toi). Notons aussi que Chen a conçu une autre table de façon automatique et d'une taille beaucoup plus importante que la nôtre [16]. A partir de la collection de documents TREC, Chen a extrait tous les mots uniques de cette collection et les a traduit en anglais. Après, il a filtré tous ceux dont la traduction est un mot outil en anglais. Avec ce procédé, il a pu établir une table d'environ 2942 entités. Dans notre cas, nous avons analysé directement les mots en arabe et les mots outils sont déterminés selon leur sens et leurs rôles syntaxiques.

4.4 Lemmatisation

Un des traitements les plus importants pour la langue arabe, en vue de la recherche d'information, est la lemmatisation des mots. L'objectif de la lemmati-

sation est de trouver la forme représentative d'index d'un mot à partir de sa forme représentée dans le document par l'application de la troncature des affixes. Comme nous l'avons introduit précédemment dans la section 4.2, la forme des mots arabes peut avoir quatre catégories d'affixes : les antéfixes, les préfixes, les suffixes et les postfixes. Ainsi un mot arabe peut avoir une forme plus compliquée si tous ces affixes sont attachés à sa forme standard. De telles situations se présentent fréquemment dans la morphologie arabe.

La question qui se pose pour la lemmatisation est la suivante : quel lemme linguistique doit-on choisir à un mot pour fin de recherche d'information ? Pour l'exemple du tableau 4.3, il ne sera pas suffisant pour la RI de tronquer seulement un préfixe et seulement un suffixe de ce mot. La forme résultante يفاوضون (négocient) ne sera pas commune à d'autres mots sémantiquement similaires. Par exemple si on appliquait la même stratégie, un mot très similaire ليفاوضهم (pour qu'il négocie avec eux) serait lemmatisé à une forme différente يفاوض (négocie). Nous voyons clairement que même si les deux mots sont sémantiquement similaires, leurs lemmes ainsi déterminés sont différents. En effet, dans cet exemple, nous pouvons constater que, si nous tronquons tous les affixes de ces deux mots, nous obtenons une forme d'index identique (فاوض). Cette forme représente le noyau sémantique du mot. Ainsi, une intuition serait d'éliminer tous les affixes.

Dans ce travail, on tente de répondre à la question suscitée en proposant et comparant deux méthodes de lemmatisation, l'une motivée linguistiquement et l'autre assouplie. Avant de décrire les deux méthodes, on dresse un bilan des

travaux qui ont été réalisés dans ce domaine.

4.4.1 Difficultés de la lemmatisation des mots arabes

La langue arabe soulève plusieurs défis au traitement automatique des langages naturels, en grande partie, dus à sa morphologie très riche et variable. La composition de ses mots est régie par de multiples et complexes règles morphologiques. Dans cette langue, le traitement morphologique devient particulièrement important pour la recherche d'information, parce que la RI vise à déterminer une forme appropriée d'index aux mots. Cependant, ce langage renferme un certain nombre de complexités (voir section 4.2) qui rend son traitement difficile :

- D'abord, l'aspect de non voyellation génère de l'ambiguïté. Les articles de journaux sur lesquels le corpus de test est construit ne sont pas voyellés. Par conséquent, certains mots non voyellés dans les textes sont ambigus et peuvent être confondus avec d'autres mots ayant la même forme dans les requêtes ou dans les dictionnaires.

- D'autre part, l'arabe est une langue hautement flexionnelle. Son phénomène qu'une racine de trois consonnes comme فعل peut engendrer jusqu'à (voire plus de) 30 dérivés entre nominaux et verbaux (Tableau 4.2), entraîne une multitude de formes pour un mot donné. Si on arrive à réduire certaines formes d'un mot donné en une seule forme, cela peut être bénéfique pour la recherche d'information. Mais cette tâche n'est pas toujours aisée à accomplir.

- Nous avons vu dans la section 4.2 que plusieurs prépositions, particules, ou préfixes peuvent être attachés au début d'un mot. De même, plusieurs

pronoms ou suffixes sont généralement attachés à la fin d'un mot. A un niveau plus profond, parfois des lettres représentant les patrons de dérivation sont insérées à l'intérieur d'un mot pour produire certaines formes (Tableau 4.2). Dans d'autres cas, au milieu du mot, certaines lettres sont supprimées ou modifiées par d'autres lettres pour engendrer d'autres formes (le cas des verbes irréguliers).

- Enfin, certaines formes les plus proches l'une de l'autre comme les formes du singulier et du pluriel pour les noms, sont irrégulières. Elles ne sont pas reliées par de simples inflexions i.e. ajouts de préfixes ou suffixes. Par exemple, la forme pluriel du mot قافله (caravane) est قوافل (caravanes). Sans utiliser un lexique pour ces types de mots, il est difficile d'écrire un algorithme à base de règles pour réduire ce genre de pluriel au singulier.

En résumé, pour la recherche d'information, ce phénomène de multitude de formes morphologiques d'un mot arabe, rend difficile la correspondance entre la forme d'un mot dans une requête avec ses autres formes qu'on peut trouver dans les documents pertinents pour cette requête.

4.4.2 Travaux reliés

La lemmatisation des mots arabes a été une problématique majeure dans plusieurs travaux dans la recherche d'information arabe. Une première approche intuitive consiste à utiliser les racines trigrammes comme index. Le lemmatiseur de Khoja [37], le plus cité dans la littérature, applique cette méthode. Il essaye de trouver

des racines pour les mots arabes qui sont plus abstraites que des lemmes[5]. D'abord, il élimine les préfixes et les suffixes, ensuite il essaye d'extraire la racine à partir de la forme réduite. Pour ce faire, il procède à une analyse morphologique qui vérifie une liste de patrons et de racines pour déterminer si la forme dépourvue d'affixes peut être obtenue en appliquant un certain patron sur une racine connue.

Pour le besoin de la RI, si on tente d'indexer les mots par ces racines trigrammes, les résultats de recherche des documents pertinents ne seraient pas satisfaisants du fait que ces racines sont très abstraites et ne sont pas spécifiques aux sens que peuvent représenter les mots originaux. Par exemple si nous indexons les mots " كاتب" (écrivain) et "كتاب" (livre) par leur racine "كتب", cet index pourrait représenter plusieurs mots de sens différents et par conséquent il ne serait pas discriminatoire. Ce lemmatiseur a été testé et ses performances étaient modestes devant les lemmatiseurs assouplis [48] [17] que nous introduirons plus tard dans cette section.

McNamee [51] utilise les n-grammes de longueurs multiples (3, 4, 5, 6) pour indexer les mots. Un mot est indexé par plusieurs de ses n-grammes possibles. Par exemple, si on tente d'indexer le mot بنتهم (leur fille) par ses 3-grammes, le résultat serait : بنت, نته, تهم. Avec cette méthode (matching n-grams), malgré qu'on arrive toujours à identifier les bons index, d'autres index erronés peuvent s'infiltrer. Dans l'exemple du mot précédent بنتهم, le bon index بنت (fille) est identifié, mais un autre index erroné تهم (accusations) est introduit. Le résultat de cet index erroné est qu'un système de RI appliquant une lemmatisation n-grammes retourne beaucoup de documents qui ne sont pas forcément pertinents et la précision sera

[5]Dans cette thèse, nous distinguons entre une racine trigramme et un lemme pour les mots arabes. Contrairement à la racine, un lemme n'est obtenu que par troncature sur les deux extrémités du mot sans modification interne sur le mot.

pénalisée. L'autre inconvénient de cette méthode est la production d'un fichier d'index de taille exorbitante. Par exemple, pour indexer la collection TREC de 383 872 textes, on crée un fichier d'index d'une capacité de 441 MB en utilisant les 3-grammes et de 1 691 MB en utilisant les 6-grammes [51].

Une autre approche dite "lemmatisation assouplie" plus connue sous le nom "Light stemming", est adoptée pour la RI arabe. Cette dernière approche est inspirée par le processus de lemmatisation de l'anglais. Plusieurs lemmatiseurs de ce type ont été développés par Larkey [48], Darwish [18] et Chen [14]. Ces lemmatiseurs opèrent une légère troncature sur le début et la fin des mots. Pour ce faire, des listes de préfixes et de suffixes à une lettre, à deux lettres et à trois lettres sont établies. Le choix de ces listes est déterminé généralement selon des statistiques de corpus. Ces statistiques analysent les fréquences d'occurrence des préfixes et des suffixes sur les mots d'un grand corpus de textes. La décision de tronquer un préfixe ou un suffixe d'un mot est faite selon de simples règles comme la longueur de mots. Par exemple, on ne peut pas tronquer un préfixe à trois lettres d'un mot de longueur quatre. Plus de détails sur cette approche sont présentés dans la section 4.4.4. Puisqu'elle donne de bonnes performances, cette approche est largement utilisée maintenant en RI.

Malgré ces études, il est encore peu clair quel type de lemmatisation est approprié pour la RI arabe. D'une part, une lemmatisation assouplie peut empêcher de grouper deux mots différents; mais elle court également le risque de ne pas grouper deux mots sémantiquement semblables, menant à un rappel plus faible. Nous avons vu ce cas de figure dans les deux mots de la section précédente : (ليفاوضونهم) (pour qu'ils négocient avec eux), ليفاوضهم (pour qu'il négocie avec eux). Si on appliquait une lemmatisation assouplie sur les deux mots, leurs lemmes résultants seraient

يفاوضون (négocient) et يفاوض (négocie) respectivement. Nous voyons clairement que même si les deux mots sont sémantiquement similaires, leurs lemmes déterminés par une lemmatisation assouplie sont différents. D'autre part, une lemmatisation plus sévère peut grouper incorrectement des mots sémantiquement non similaires dans un même index, menant à une précision plus faible. Par conséquent, plus d'investigations sur les effets de la lemmatisation sur la performance de la RI sont nécessaires.

Dans le premier volet de notre étude, nous proposons de comparer deux méthodes de lemmatisation. Une première méthode assouplie, classique, opère quelques troncatures sur un mot aux deux extrémités. Cette approche est semblable à celles proposées par Larkey, Darwish et Chen. L'autre méthode, nouvelle et linguistiquement motivée, essaye de déterminer le noyau d'un mot. Cette deuxième approche s'inspire de la composition des mots en arabe : Les mots arabes sont habituellement formés d'une séquence de {antéfixe, préfixe, noyau, suffixe, postfixe}. Nous croyons qu'une bonne stratégie de lemmatisation doit indexer les mots par leur noyau[6], et ces index vont encoder la sémantique de base dans la langue arabe.

4.4.3 Lemmatisation à base linguistique

Cette méthode est motivée par la composition des mots en arabe. Nous avons mentionné que les mots arabes sont habituellement formés d'une séquence de {antéfixe, préfixe, noyau, suffixe, postfixe} [33]. Le tableau 4.4 présente les affixes les plus utilisés en arabe.

Intuitivement, une méthode directe serait de tronquer ces éventuels affixes selon le tableau 4.4. Cependant, cette approche intuitive nous mène à plusieurs cas d'ambiguïté : une séquence particulière de lettres peut ou non jouer un rôle d'affixe, selon le mot. A titre d'exemple nous citons la séquence ان dans deux mots différents : طفلان (deux enfants) et بستان (jardin). Dans le premier mot طفلان, elle joue le

[6]Le noyau est un lemme linguistique. Il n'est pas toujours la racine trigramme. La différence entre un lemme et une racine trigramme est perçue surtout pour les mots irréguliers. Un lemme n'est obtenu que par troncature sur les deux extrémités du mot sans modification interne sur le mot.

Antéfixes		Préfixes		Suffixes		Postfixes	
وبال	et avec le	ا	Lettres	تما	Terminaisons	كما	votre(s)
وال	et le	ن	représentant	يون	de	هما	leur(s)
بال	avec le	ي	la	تين	conjugaison	كن	votre(s)
فال	ensuite le	ت	personne	تان	pour les	هن	leur(s)
كال	comme le		de	ات	verbes et	تي	mon(ma)(mes)
ولـل	et pour le		conjugaison	ان	marqueurs	ها	son(sa)(ses)
ال	le (la)		des	ون	de	نا	notre(s)
وب	et avec		verbes	ين	duel/pluriel/	هم	leurs
ول	et pour		au	وا	féminin	كم	vôtres
لـل	pour le		présent	تا	pour	ك	votre(ton)(ta)
فس	ensuite -			تم	les noms	ه	son(sa)(ses)
فب	ensuite avec			تن		ي	mon(ma)(mes)
فل	ensuite pour			نا			
وس	et -			ت			
ك	comme			ن			
ف	ensuite			ا			
و	et			ي			
ب	avec			و			
ل	pour						

Tableau 4.4: Les affixes de l'arabe

rôle d'un suffixe pour désigner le duel du mot طفل (enfant). Tandis que dans le deuxième mot بستان, cette séquence fait partie intégrante du mot et son éventuelle troncature produirait un lemme erroné.

Nous remarquons qu'en arabe les racines seules sont aussi des mots. Si une racine est souvent utilisée comme mot dans un corpus de textes, alors nous pouvons juger que cette racine est fréquente en arabe. Ainsi, afin de résoudre le problème d'ambiguïté, nous proposons de tirer profit des statistiques du corpus[7] : nous appliquons des règles qui génèrent un ensemble de lemmes candidats pour un mot; ensuite nous choisissons le lemme le plus fréquent selon les statistiques de corpus - le lemme le plus utilisé est choisi. Le tableau 4.5 présente quelques exemples de règles de troncature pour générer les lemmes.

- Si une séquence de lettres au début d'un mot représente un préfixe alors cette séquence est tronquée du mot et la partie restante du mot est considérée comme un lemme.
- Si une séquence de lettres à la fin d'un mot représente un suffixe alors cette séquence est tronquée du mot et le reste est considéré comme un lemme.
- Si une séquence de lettres au début d'un mot représente un préfixe et une autre séquence à la fin de ce mot représente un suffixe alors les deux séquences sont tronquées du mot et le reste est considéré comme un lemme.

Tableau 4.5: Exemples de règles de troncature pour générer les lemmes

Les statistiques de corpus sont compilées sur les 523 359 mots différents de la collection TREC. Une première approche intuitive serait de tenter de tronquer un mot de différentes façons, de comparer les résultats avec les mots de la collection, et de garder seulement le lemme le plus fréquent dans la collection. Cette méthode choisit généralement des formes de mots non lemmatisées parce que ces formes sont plus fréquentes que les lemmes ou les racines dans la collection.

[7]La collection TREC arabe. Elle contient 523 359 mots différents. http://trec.nist.gov/

Une deuxième approche plus efficace serait de construire d'abord un lexique de tous les lemmes possibles pour tous les mots de la collection. Pour ce faire, chaque mot de la collection subit différentes décompositions pour obtenir tous les lemmes possibles pour ce mot. En faisant ainsi pour tous les mots, nous construisons un corpus de lemmes avec leurs fréquences d'occurrence dans la collection. Quand un mot est soumis à la lemmatisation, nous générons un ensemble de lemmes candidats pour ce mot; ensuite nous choisissons le lemme le plus fréquent dans la collection. Nous notons que cette approche est raisonnable dans la mesure où le lemme est également un mot qui apparaît dans les textes. Les statistiques de corpus peuvent ainsi révéler les lemmes les plus utilisés généralement en arabe.

La sélection de cette forme commune de lemme peut enlever plusieurs cas d'ambiguïté. Nous avons testé cet aspect sur l'ensemble des mots distincts présentant une ambiguïté des deux ensembles de requêtes TREC (voir section 4.7.1) et nous sommes arrivés à résoudre 50.72 % de cas d'ambiguïté. Pour le reste de cas d'ambiguïté présents dans les requêtes et non résolus, malgré que cette méthode de lemmatisation génère de mauvais lemmes pour ces mots, certains cas ne posent pas de problème pour la recherche d'information, parce que ces mêmes mots présents dans les textes, sont lemmatisés de la même façon et par conséquent leurs lemmes identifiés sont identiques aux lemmes obtenus pour ces mots dans les requêtes.

4.4.4 Lemmatisation assouplie (Light stemming)

Cette approche est similaire aux lemmatiseurs généralement utilisés sous l'appellation "light stemming". Elle tronque un mot aux deux extrémités. Le choix des affixes de mots à tronquer est fait selon des statistiques de corpus ainsi que leur rôle syntaxique [32]. Nous avons groupé tous les affixes dans deux classes : préfixes et suffixes. Puis nous avons dressé une table de statistiques basée sur les fréquences d'occurrence de ces affixes sur les 523 359 mots différents de la collection TREC arabe. Le tableau 4.6 résume ces statistiques.

Préfixe	Fréquence	Suffixe	Fréquence
و (et)	111512	ا (terminaison de conjugaison)	89591
ا (un patron de dérivation des noms ou une marque du présent)	90048	ه son(sa)(ses)	71586
ال (le)	51157	ن (terminaison de conjugaison)	68634
ب (avec)	47874	ي (mon)(ma)(mes)	55175
ل (pour)	47006	ت (terminaison de conjugaison)	32056
ت (marque du présent)	24722	ها (sa)	25392
س (-)	23072	ين (marque du duel/pluriel masculin)	24235
ف (ensuite)	21671	و (marque du pluriel masculin)	21150
ك (comme)	19524	ان (marque du duel)	18834
وال (et le)	18506	ات (marque du pluriel féminin)	17332
ي (marque du présent)	18125	ون (marque du pluriel masculin)	15477
ن (marque du présent)	10550	هم (leurs)	13147
لل (pour le)	10442	نا (notre(s))	11495
بال (avec le)	8781	ك (votre)(ton)(ta)	9700
وب (et avec)	8675	وا (terminaison de conjugaison)	7860
ول (et le)	6403	هما (leur(s))	6182
فال (ensuite le)	2199	تي (mon)(ma)(mes)	4509
كال (comme le)	1733	هن (leur(s))	1926
ولل (et pour le)	1370	كم (vôtres)	1899
فل (ensuite pour)	1308	تم (terminaison de conjugaison)	645
وبال (et avec le)	1148	تن (terminaison de conjugaison)	536
فب (ensuite avec)	372	كن (votre(s))	385
		كما (votre(s))	107

Tableau 4.6: Fréquences d'occurrence des affixes sur les mots de la collection TREC

Finalement, nous avons établi deux listes d'affixes les plus fréquents pour les tronquer des mots : une pour les préfixes et l'autre pour les suffixes. Ces choix sont aussi guidés par le rôle syntaxique que jouent ces affixes dans les textes arabes. Les préfixes sont généralement des particules collés aux débuts des mots comme "و" exprimant la conjonction "et", "ل" la préposition "pour", "ال" l'article de définition "le"/"la" ou parfois plusieurs prépositions attachées aux mots comme "وب" exprimant "et avec", "وبال" "et avec le". La liste de ces préfixes à enlever que nous avons établie est la suivante :
(فب | وبال | فل | ولل | كال | فال | ول | وب | بال | لل | وال | ل | ب | ال | ا | و).

Quant aux suffixes que nous avons jugés nécessaires de tronquer sont ceux qui sont les plus fréquents et représentent généralement des pronoms attachés à la fin des mots exprimant le nombre ou le genre des noms arabes [33]:
(تي | هما | وا | ك | نا | هم | ون | ات | ان | و | ين | ها | ت | ي | ن | ه | ا).

Notons également que notre méthode partage plusieurs préfixes et suffixes à tronquer avec les "light stemmers" développés par Larkey, Darwish et Chen. Néanmoins, nous remarquons que la différence entre notre lemmatiseur et les autres lemmatiseurs réside dans le fait qu'une méthode est plus ou moins agressive qu'une autre. Le lemmatiseur "Al-stem" de Darwish est le plus agressif de ces méthodes. Quant à celui de Larkey, il est très assoupli. Il n'établit que deux petites listes pour les préfixes et les suffixes à tronquer. Par contre, plusieurs préfixes comme لل, ولل, وبال et plusieurs suffixes comme تي, هم, وا ne sont pas tronqués. La méthode de Chen est aussi agressive et introduit de nouveaux préfixes, du moins non connus

dans les autres lemmatiseurs, comme مال et سال.

En comparaison, notre méthode assouplie partage le même principe de troncature que les autres lemmatiseurs assouplis (light stemmers), et donne des performances comparables à ces derniers.

Ce que nous ressortissons dans notre travail est la classification de tous les affixes en quatre classes avec leur rôle syntaxique quand ils sont attachés aux mots arabes. D'autre part, nous avons donné une argumentation linguistique et statistique pour choisir les préfixes et les suffixes potentiels à éliminer. Nous avons choisi les préfixes qui sont généralement des prépositions attachées aux débuts des mots, et les suffixes qui sont des pronoms collés à la fin des mots. Nous présenterons une comparaison de performances entre les deux lemmatiseurs que nous avons décrits dans la partie expérimentation et évaluation.

4.5 Modèle de recherche

Le modèle de recherche utilisé est basé sur les modèles de langue unigrammes et une fonction de score basée sur la divergence de Kullback-Leibler décrits dans la section 2.7.4 du chapitre 2. Pour faciliter la lecture, nous le rappelons ici.

Etant donnés une requête Q et un document D, nous calculons la pertinence de ce document par rapport à la requête selon la divergence négative entre le modèle de langage de requête θ_Q et le modèle de langage du document θ_D [82]:

$$Score(Q, D) = \sum_{t \in Q} P(t \mid \theta_Q) \log P(t \mid \theta_D) \simeq -KL(\theta_Q \parallel \theta_D) \qquad (4.1)$$

Le modèle de langue de la requête θ_Q est généralement estimé par la fréquence relative des mots-clefs qui la composent, c'est-à-dire : $P(t \mid \theta_Q) = \frac{tf(t,Q)}{|Q|}$. Pour contourner le problème d'attribuer des probabilités nulles aux termes de la requête

non présents dans le document D : $P(t \mid \theta_D) = 0$ conduisant à $Score(Q, D) = -\infty$, la technique de lissage Jelinek-Mercer est utilisée pour lisser $P(t \mid \theta_D)$ [81] :

$$P(t \mid \theta_D) = (1 - \lambda)P_{MLE}(t \mid \theta_D) + \lambda P_{MLE}(t \mid \theta_C) \tag{4.2}$$

Où $P_{MLE}(t \mid \theta_D) = \frac{tf(t,D)}{|D|}$ et $p_{MLE}(t \mid \theta_C) = \frac{tf(t,C)}{|C|}$ sont les estimés du maximum de vraisemblance des modèles de langue unigrammes basés respectivement sur le document D et la collection des documents C; λ est un paramètre qui contrôle l'influence de chacun des deux modèles. Dans nos expérimentations, λ est fixé à 0.5.

Nous avons opté pour le choix d'une approche basée sur les modèles de langue parce que ces derniers ont prouvé leurs performances dans des expérimentations récentes sur les autres approches [59].

4.6 Rétroaction de pertinence

Une technique d'amélioration de performance de la RI est la rétroaction de pertinence (relevance feedback). Le principe de cette technique est d'étendre la requête en utilisant l'information issue des documents pertinents. Dans le contexte des modèles de langue, on utilise une interpolation linéaire pour combiner la requête originale et l'information pertinente issue de la rétroaction. Soient θ_Q le modèle de langue de la requête originale et θ_F une estimation du modèle de la rétroaction de pertinence basé sur les documents pertinents. Le nouveau modèle de la requête étendue θ'_Q est estimé comme suit :

$$\theta'_Q = (1 - \alpha)\ \theta_Q + \alpha\ \theta_F \tag{4.3}$$

Où α est un paramètre qui contrôle l'influence du modèle de rétroaction de per-

tinence. Remarquons que si $\alpha = 0$, le modèle de la requête étendue est réduit au modèle de la requête originale.

Dans la pratique de RI, les jugements de pertinence sont généralement non-disponibles. Ainsi, on utilise souvent la pseudo-rétroaction de pertinence, en supposant que les n premiers documents trouvés par le système sont "pertinents" : $F = (d_1, d_2, ..., d_n)$.

Une méthode naturelle pour estimer le modèle de rétroaction de pertinence θ_F est de supposer que les documents de la rétroaction sont générés par un modèle probabiliste $P(F \mid \theta)$ unigramme qui génère chaque terme dans F indépendamment selon θ [82]:

$$P(F \mid \theta) = \prod_i \prod_w P(w \mid \theta_F)^{c(w;d_i)} \tag{4.4}$$

Où $c(w; d_i)$ est le nombre de termes w dans le document d_i. Notant que ce modèle ne contient pas seulement l'information pertinente. Les documents de rétroaction peuvent contenir aussi des informations bruitées ou des informations de base sur la collection des documents. Un modèle plus raisonnable serait un modèle plus spécifique extrait des documents de rétroaction. Pour ce faire, on considère qu'un document de rétroaction est généré par la combinaison du modèle spécifique de rétroaction $P(w \mid \theta_F)$ avec le modèle de langue de la collection $P(w \mid \theta_C)$ [82]. Le modèle de rétroaction de pertinence θ_F sera extrait de ce modèle mixte en utilisant l'algorithme EM (Expectation Maximization) [82] de telle manière que la log-vraisemblance des documents de rétroaction F sera maximisée :

$$\theta_F = arg\ max_{\theta_F} \sum_{d_i \in F} \sum_{w \in d_i} c(w; d_i)\ \log[(1-\lambda)P(w \mid \theta_F) + \lambda P(w \mid \theta_C)] \tag{4.5}$$

La valeur de λ est fixée (0.5 dans [82]). L'algorithme EM calcule itérativement les deux paramètres suivants :

$$t^n(w) = \frac{(1-\lambda)P_\lambda^n(w \mid \theta_F)}{(1-\lambda)P_\lambda^n(w \mid \theta_F) + \lambda P(w \mid \theta_C)} \qquad (E-Etape) \qquad (4.6)$$

$$P_\lambda^{n+1}(w \mid \theta_F) = \frac{\sum_{d_j \in F} c(w; d_j) t^n(w)}{\sum_{w_i \in F} \sum_{d_j \in F} c(w_i; d_j) t^n(w_i)} \qquad (M-Etape) \qquad (4.7)$$

Intuitivement, quand on estime le modèle de rétroaction θ_F, on tente de purifier les documents par l'élimination du bruit présent dans ces documents. Ainsi, le modèle de rétroaction estimé sera concentré sur les termes communs dans l'ensemble des documents de rétroaction, mais peu observés dans la collection. Ceci a le même effet que la plupart des méthodes traditionnelles de rétroaction de pertinence comme la pseudo rétroaction de pertinence de Rocchio [62].

Finalement pour évaluer un document d utilisant le modèle estimé de la rétroaction θ_F, il faut d'abord l'interpoler avec le modèle original de la requête θ_Q pour obtenir un nouveau modèle de la requête étendue $\theta_Q^{'}$. Ensuite on calcule la KL-divergence entre le modèle de langue de la requête étendue et le modèle de langue du document comme dans l'équation 4.1.

4.7 Expérimentation et évaluation

Le but de nos expérimentations est d'évaluer l'impact des traitements morphologiques ainsi que les différentes méthodes de lemmatisation sur la performance de recherche d'information arabe monolingue. Rappelons ici que la collection des documents ainsi que les requêtes sont toutes en arabe. Avant de présenter les résultats, nous décrivons notre collection de test sur laquelle on a fait nos expérimentations.

4.7.1 Description du corpus de test

Toutes nos expérimentations ont été faites sur la collection LDC2001T55 de LDC (Linguistic Data Consortium)[8], qui a été utilisée pour les tests dans TREC 2001 et 2002 (Text REtrieval Conference)[9] dans la piste (track) de recherche d'information translinguistique. Notons que cette collection TREC incluant les documents, les requêtes et les jugements de pertinence est la plus grande collection en arabe actuellement disponible. Elle contient 383 872 articles provenant de Arabic Newswire de l'AFP (Agence France Presse). La collection représente un volume de 884 MOctets. Ce sont des articles de journaux arabes couvrant la période de mai 1994 jusqu'à décembre 2000. La structure d'un document de cette collection est présentée en annexe I. Le tableau 4.7 présente un exemple de documents.

Les requêtes que nous avons évaluées proviennent aussi du TREC. Elles ont été développées par NIST (National Institute of Standards and Technology). La structure d'une requête (Tableau 4.8) inclut un champ (tagged field) "title" identifiant le sujet de la requête, un champ "description" qui consiste généralement en une phrase simple décrivant la requête, et un champ "narrative" explicitant comment un document peut être jugé pertinent par un expert humain. Le tableau 4.9 présente la version anglaise de la requête arabe du tableau 4.8. Nous utilisons le titre et la description dans toutes nos expérimentations i.e. des requêtes de type TD.

Nous disposons de deux ensembles de requêtes : le premier utilisé par TREC2001 contient 25 requêtes et le deuxième - TREC2002 englobe 50 requêtes, faisant un total de 75 requêtes disponibles. La liste des requêtes de la collection TREC2001 est présentée en annexe II). Le tableau 4.10 présente quelques caractéristiques de la collection TREC.

[8]http://www.ldc.upenn.edu/
[9]http://trec.nist.gov/

<DOC>
<DOCNO>19940513_AFP_ARB.0034 </DOCNO>
<HEADER>اراا3110 4 ع 1110 قبرص / افب-تصذ56 اوكرانيا/نووي</HEADER>
<BODY>
<HEADLINE> كرافتشوك يعلن تعطيل الصواريخ النووية الاوكرانية ال 64 العابرة للقارات
</HEADLINE>
<TEXT>
<P>
موسكو 31-5 (اف ب) - افادت وكالة «انترفاكس» للانباء ان الرءيس الاوكراني
ليونيد كرافتشوك اعلن اليوم الجمعة تعطيل الصواريخ النووية ال64
العابرة للقارات من طراز "اس اس 42" التي تملكها اوكرانيا
</P>
<P>
وكان اتفاق ثلاثي وقعه كل من كرافتشوك و نظيراه الروسي بوريس يلتسين والامريكي
بيل كلينتون في 14 كانون الثاني / يناير في موسكو نص على ازالة هذه الصواريخ
وعلى جعل اوكرانيا دولة خالية تماما من السلاح النووي
</P>
<P>
وقد ورثت كييف عن الاتحاد السوفياتي السابق 671 صاروخا من بينها
64 صاروخا نوويا عابرا للقارات يفوق مداها عشرة آلاف كيلومتر
</P>
</TEXT>
<FOOTER> كاذ/ن ا/فن مواا482 افب</FOOTER>
</BODY>
<TRAILER> 639131 49 جمت ماي </TRAILER>
</DOC>

Tableau 4.7: Exemple d'un document arabe dans la collection TREC

<top>
<num> Number: AR25
<title> الدور الاوروبي و الامريكي في عملية السلام في الشرق الاوسط
<desc> Description:
ما هي أدوار الدول الاوربية و أمريكا في عملية السلام في الشرق الاوسط ؟
<narr> Narrative:
يتعلق بهذا الموضوع كل مقال يخص التدخل الاوروبي و الامريكي
في القرارات العربية لتوجيه عملية السلام في الشرق الاوسط
و ما لا يرتبط بهذا الموضوع هو التدخل الاوروبي و الامريكي في
الشوءون الداخلية لدول الشرق الاوسط
</top>

Tableau 4.8: Exemple d'une requête de la collection TREC

<top>
<num> Number: AR25
<title> European and American roles in Middle East peace process
<desc> Description:
What are the roles of the European countries and America
in the peace process in the Middle East?
<narr> Narrative:
Relevant articles are about the involvement of Europe and the US in
directing the peace process and stopping violence in the Middle East.
Articles are about European and US involvement in the internal
affairs of Middle Eastern countries are irrelevant.
</top>

Tableau 4.9: La version anglaise de la requête arabe précédente

	TREC2001	TREC2002
Langue du corpus des documents	arabe	arabe
Nombre de documents	383 872	383 872
Capacité du corpus (MB)	884	884
Nombre total de mots (tokens)	76 millions	76 millions
Nombre de mots différents	666 094	666 094
Taille moyenne des documents (mots)	150	150
Langues des requêtes	arabe et anglais	arabe et anglais
Nombre de requêtes	25	50
Taille moyenne des requêtes (mots)	12.6 (arabe) 12.1 (anglais)	11.2 (arabe) 11 (anglais)
Nombre moyen de documents pertinents par requête	164	118.18

Tableau 4.10: Caractéristiques de la collection TREC arabe

Le fichier de jugement de pertinence est développé aussi par NIST. Le nombre de documents pertinents pour toutes les requêtes de la collection TREC2001 est de l'ordre de 4100 documents et de 5909 pour la collection TREC2002. Le format de ce fichier est présenté en annexe III.

4.7.2 Impact des prétraitements morphologiques

Une série d'expérimentations a été menée sur les deux collections de requêtes pour montrer l'effet de chaque aspect des prétraitements sur la performance de la recherche. Les documents sont classés par rapport aux requêtes selon la formule 4.1 dans la section 4.5. Dans nos expériences, nous utilisons la mesure classique de recherche d'information : précision moyenne non interpolée (Mean Average Precision : MAP). Le rappel est également employé comme une deuxième mesure. Le tableau 4.11 présente les résultats des expériences en considérant les aspects de normalisation des lettres et l'élimination des mots outils. Les nombres entre parenthèses dans le tableau représentent les pourcentages d'amélioration en précision moyenne de chaque aspect de prétraitement par rapport à la base - quand aucun

prétraitement n'est effectué.

Collection de requêtes	Stoplist (Elimination des mots outils)	Normalisation des lettres	MAP	Rappel
TREC2001	-	-	0.1674	1648/4122
	+	-	0.1820 (8.72%)	1787/4122
	-	+	0.2059 (22.99%)	1785/4122
	+	+	0.2210 (32.02%)	1965/4122
TREC2002	-	-	0.2268	4088/5909
	+	-	0.2298 (1.32%)	4107/5909
	-	+	0.2324 (2.46%)	4104/5909
	+	+	0.2355 (3.83%)	4130/5909
TREC2001-2002	-	-	0.2070	5736/10031
	+	-	0.2138 (3.28%)	5894/10031
	-	+	0.2236 (8.01%)	5889/10031
	+	+	0.2306 (11.40%)	6095/10031

Tableau 4.11: L'impact de la normalisation des lettres et la suppression des mots outils sur la RI monolingue arabe

Nous remarquons que si nous appliquons une recherche naïve où aucun prétraitement n'est pris en considération, le score de recherche est plus faible, la précision moyenne n'est que 16.74% pour la collection TREC2001 et 22.68% pour la collection TREC2002. L'introduction du facteur de stoplist pour éliminer les mots outils des documents augmente légèrement le score de recherche. On note un gain de 3.28% en précision moyenne sur les deux collections fusionnées. L'effet de la

normalisation orthographique de certains caractères arabes est aussi important. La précision moyenne gagne 8.01% sur les deux collections fusionnées, atteignant 22.36%. Le rappel est aussi augmenté quand on utilise la normalisation des lettres et on élimine les mots outils. Enfin, la prise en compte des deux prétraitements (normalisation orthographique et élimination des mots outils) améliore de 11.40% la précision moyenne sur les deux collections. En conclusion, l'élimination des mots outils et la normalisation orthographique de certains caractères arabes sont utiles pour la RI arabe.

On remarque aussi que les impacts de ces deux facteurs (élimination des mots outils et normalisation orthographique) sont différents sur les deux collections. La raison principale est que les requêtes de la collection TREC2002 ne comptent pas parmi elles beaucoup de mots outils et de mots susceptibles d'être normalisés. Statistiquement parlant, la moyenne par requête des fréquences relatives des mots outils par rapport à tous les mots est de 31.54% pour la collection TREC2001 et de 23.51% pour la collection TREC2002. De même, la moyenne par requête des fréquences relatives des mots susceptibles d'être normalisés par rapport à tous les mots est de 28.16% pour la collection TREC2001 et de 21.70% pour la collection TREC2002.

4.7.3 Impact de lemmatisation

Dans cette section, nous présentons l'apport de la lemmatisation sur la RI et entre autres nous comparons les deux méthodes de lemmatisation que nous avons proposées. Avant l'indexation, les documents ainsi que le requêtes sont normalisés et les mots outils sont supprimés. Le tableau 4.12 présente les performances de la RI monolingue arabe selon les deux stratégies de lemmatisation avec les deux collections de requêtes. La figure 4.1 dresse une comparaison entre les deux méthodes de lemmatisation sur la collection TREC2001-2002 en fonction de leurs courbes rappel-précision.

Collection de requêtes	Métrique	Sans lemmatisation	Lemmatisation à base linguistique	Lemmatisation assouplie
TREC2001	MAP	0.2210	0.3326 (50.49%)	0.3220 (45.70%)
	Rappel/4122	1965	2704	2664
TREC2002	MAP	0.2355	0.2828 (20.08%)	0.2671 (13.41%)
	Rappel/5909	4130	4301	4443
TREC 2001-2002	MAP	0.2306	0.3107 (34.73%)	0.2868 (24.37%)
	Rappel/10 031	6095	7181	7121

Tableau 4.12: Les performances de la RI monolingue arabe selon les deux méthodes de lemmatisation

L'introduction du facteur de lemmatisation est capital pour la performance de RI. Les deux méthodes avec lemmatisation se comportent nettement mieux que lorsque la lemmatisation n'est pas faite. En comparant à une recherche qui normalise les lettres et supprime les mots outils mais n'effectue aucune lemmatisation (la troisième colonne du tableau 4.12), on observe une nette amélioration en précision moyenne de 24.37% pour une recherche utilisant la lemmatisation assouplie et de 34.73% pour une recherche utilisant la lemmatisation à base linguistique sur la collection TREC2001-2002 (Tableau 4.12). Ceci est dû au fait que la lemmatisation permet de fusionner les termes ayant un sens similaire avec de petites différences sur la forme morphologique en un seul index, et par conséquent elle permet d'améliorer la qualité de recherche.

Sur les deux collections de requêtes, les résultats montrent que la technique de lemmatisation à base linguistique est uniformément plus efficace que la technique de lemmatisation assouplie sur tous les points de rappel. On peut observer ce comportement des deux méthodes dans la figure 4.1 : la courbe de lemmatisation à base linguistique représentant la précision de recherche en fonction des points de rappel est toujours au dessus de la courbe de la lemmatisation assouplie. Sur l'ensemble des 75 requêtes, nous avons obtenu 31.07% de précision moyenne (MAP) avec la méthode de lemmatisation à base linguistique contre 28.68% pour la technique as-

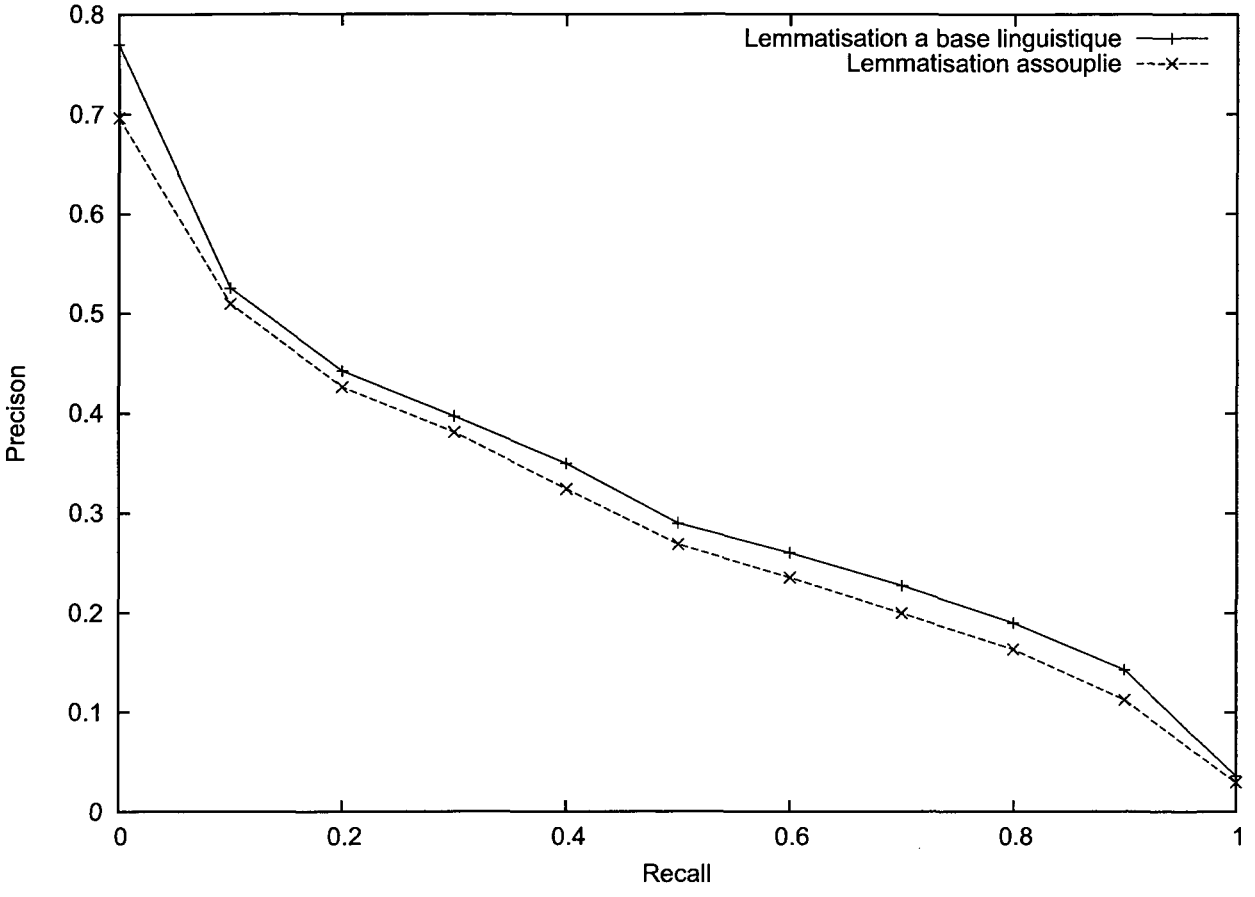

Figure 4.1: Les courbes rappel-précision des deux méthodes de lemmatisation (TREC2001-2002)

souplie (Tableau 4.12). Ces résultats prouvent qu'une lemmatisation assouplie pour les mots arabes n'est pas la meilleure approche pour la RI arabe. Cette stratégie ne réussit pas à grouper beaucoup de mots sémantiquement similaires dans le même index. En revanche, la méthode à base linguistique peut mieux déterminer le noyau sémantique d'un mot. Plus les index groupent les mots sémantiquement similaires, meilleure est la performance de la RI.

Nous donnons quelques exemples pour établir une comparaison entre les deux méthodes de lemmatisation (Tableau 4.13).

Le tableau 4.13 montre les résultats de lemmatisation de quelques mots arabes selon les deux techniques. La méthode de lemmatisation à base linguistique produit des lemmes plus appropriés que ceux de la méthode assouplie. Pour les exemples suivants : عراقيين (Irakiens), البوسنيه (La bosniaque), السياسيون (Les politiciens), la première méthode élimine tous les affixes et détermine correctement les lemmes. En revanche, la lemmatisation assouplie ne réussit pas à déterminer le suffixe à une lettre (ي) et produit en conséquence un lemme (عراقي) qui n'est pas représentatif pour beaucoup de mots sémantiquement similaires. Par exemple avec la même méthode, si on essaye de lemmatiser le mot عراقي (Irakien), très similaire au premier mot عراقيين, le lemme produit serait عراق, qui est différent du premier lemme عراقي. Pour le troisième exemple dans le tableau 4.13 مهرجان (Festival), les deux lettres ان à la fin du mot ne représentent pas un suffixe pour ce mot. Quoique ce ne soit pas un suffixe, la stratégie assouplie le tronque du mot et produit un lemme

Mot	Lemmatisation à base linguistique	Lemmatisation assouplie
عراقيين (Irakiens)	عراق (Irak)	عراقي (Irakien)
البوسنيه (La bosniaque)	بوسن (Bosni)	بوسني (Bosniaqu)
مهرجان (Festival)	مهرجان (Festival)	مهرج (clown)
السياسيون (Les politiciens)	سياس (Polit)	سياسي (Politicien ou politique)
العربيه (L'arabe)	عرب (Arab)	عربي (Arabe)
الأرضيه (La terrestre)	أرض (Terre)	أرضي (Terrestr)
قانون (Loi)	قانون (Loi)	قان (-)
بناء (Construction)	بناء (Construction)	ناء (-)
لقاح (Vaccin)	لقاح (Vaccin)	قاح (-)
مياه (Eaux)	مياه (Eaux)	ميا (-)

Tableau 4.13: Résultats de lemmatisation de quelques mots selon les deux méthodes

erroné مهرج signifiant "clown".

La lemmatisation à base linguistique procède différemment. Elle applique différentes décompositions sur le mot original, produit en conséquence un ensemble de lemmes candidats, et en utilisant les statistiques de corpus, choisit le lemme le plus utilisé dans le corpus. Nous pouvons aussi remarquer cet effet sur les mots : بناء (Construction) et لقاح (Vaccin). La lemmatisation à base linguistique, grâce aux statistiques de corpus, réussit à trouver les lemmes appropriés بناء et لقاح. Par contre la lemmatisation assouplie, en tronquant les préfixes ب et ل, produit des lemmes erronés parce que ces préfixes sont considérés comme des prépositions attachées aux mots. Or ce n'est pas le cas dans ces mots. Les lettres tronquées, considérées comme des préfixes, font partie intégrante des mots.

En analysant les performances individuelles de chaque requête des deux collections TREC, on observe une amélioration pour plusieurs requêtes de la lemmatisation à base linguistique par rapport à la lemmatisation assouplie mais peu d'amélioration pour certaines d'autres. Deux raisons sont derrière cette amélioration:

- La première raison est que la méthode à base linguistique, en cas de présence d'affixes multiples dans un mot, réussit à éliminer tous les affixes et détermine correctement les lemmes. Cependant, la lemmatisation assouplie, opérant une légère troncature sur les mots, ne réussit pas toujours à déterminer tous les

affixes présents dans un mot, et par conséquent elle produit des lemmes non représentatifs pour beaucoup de mots sémantiquement similaires.

- La deuxième raison est l'apport des statistiques de corpus. Les requêtes, où on a observé une amélioration, comptaient parmi leurs termes des mots ambigus[10]. Pour ces mots, la méthode assouplie tronque naïvement des lettres qui ne sont pas des affixes malgré leurs apparences. En revanche, la méthode à base linguistique ne fait pas une troncature aveugle. Elle produit d'abord un ensemble de lemmes candidats et, grâce aux statistiques de corpus, elle choisit le lemme le plus utilisé dans le corpus. Les quatre derniers mots du tableau 4.13 : قانون (loi), بناء (construction), لقاح (vaccin) et مياه (eaux) sont des exemples de ces mots ambigus. La lemmatisation à base linguistique réussit à trouver les lemmes appropriés pour ces mots. Par contre la lemmatisation assouplie, en tronquant des séquences de lettres de ces mots, produit des lemmes erronés parce que ces séquences de lettres font partie intégrante de ces mots et ne représentent pas des affixes.

Pour les requêtes où on n'a pas observé d'amélioration, c'est à cause de la présence de ces mots ambigus. Malgré l'application des statistiques de corpus, la lemmatisation à base linguistique n'a pas réussi à sélectionner les lemmes appropriés pour ces mots ambigus. Une raison pour laquelle certains mots ambigus n'ont pas été lemmatisés correctement est que parfois, un lemme obtenu après décomposition d'un mot ambigu, représente un autre mot plus fréquent dans la collection. Comme exemple, le mot suivant الفن (l'art) possède deux lemmes possibles : فن (art) الف (mille). L'application des statistiques du corpus n'a pas sélectionné le lemme

[10]Une séquence particulière de lettres peut ou non jouer un rôle d'affixe, selon le mot

approprié فن (art) parce que le deuxième lemme الف (mille), représentant un autre mot, est plus fréquent que le premier dans la collection.

En résumé, comme nous l'avons déjà mentionné, la méthode à base linguistique n'est pas parfaite et ne réussit pas à identifier les lemmes corrects pour certains mots ambigus. C'est d'ailleurs dans cet aspect que notre méthode doit être améliorée. Plus de traitement au niveau des statistiques de corpus doit être fait pour choisir le lemme correct quand différents lemmes candidats sont proposés pour un mot.

4.7.4 Impact de la rétroaction de pertinence

Le tableau 4.14 présente les performances de la recherche d'information monolingue arabe avec la rétroaction de pertinence selon les deux méthodes de lemmatisation. Les valeurs des paramètres utilisés en rétroaction de pertinence sont : 10 pour le nombre de documents et 50 pour le nombre de termes c'est-à-dire que le modèle de feedback est extrait des 10 premiers documents retrouvés, et les 50 termes les plus forts de ce modèle sont utilisés pour étendre la requête initiale. Le paramètre qui contrôle l'influence du modèle de la collection sur le modèle de la rétroaction est fixé à 0.5.

Collection de requêtes	Métrique	Lemmatisation à base linguistique	Lemmatisation assouplie
TREC 2001	MAP	0.3900	0.3784
	Rappel/4122	2893	2865
TREC 2002	MAP	0.3536	0.3133
	Rappel/5909	4763	4836
TREC 2001-2002	MAP	0.3657	0.3349
	Rappel/10 031	7654	7704

Tableau 4.14: Les performances de la RI monolingue arabe avec la rétroaction de pertinence selon les deux méthodes de lemmatisation

La prise en compte de la rétroaction de pertinence améliore nettement la per-

formance de recherche. Ce résultat est cohérent à d'autres expérimentations dans les langues européennes. Son introduction ici pour le cas de la RI arabe a donné les mêmes améliorations que celles observées sur les autres langues. A noter aussi que ces résultats sont comparables aux meilleurs résultats des systèmes présentés en TREC 2002 (Tableau 4.15).

Système	Précision moyenne (MAP)
Al-Stem (Darwish)	0.3482
n-grams stemmer (McNamee)	0.3410
Deeper light stemmer (Chowdhury)	0.3419
Light stemmer (Chen)	0.3666
Light stemmer (Savoy) [68]	0.3712

Tableau 4.15: Résultats des meilleurs systèmes présentées en TREC 2002

Ce que nous dégageons de ces expérimentations est que tous les traitements que nous avons introduits sont nécessaires pour la recherche d'information sur les documents arabes, bien sûr avec des degrés d'importance différents. L'importance de la technique de lemmatisation est clairement observée. Ceci est dû au choix des termes d'indexation. Plus ces index regroupent des mots sémantiquement similaires, meilleures sont les performances de recherche des documents pertinents.

4.8 Récapitulatif

En recherche d'information arabe, le traitement morphologique et plus particulièrement la lemmatisation joue un rôle important. Malgré que la lemmatisation a fait l'objet de plusieurs travaux, cet aspect n'est pas encore largement et systématiquement étudié. Il est encore peu clair quel type de lemmatisation est approprié pour la RI arabe.

Dans cette optique, nous avons défini une méthodologie pour résoudre le problème de la performance de recherche des documents pertinents arabes. Le prob-

lème central dans cette méthodologie était comment identifier les meilleurs termes d'index pour avoir des performances raisonnables en recherche monolingue. Plus particulièrement, nous avons proposé une nouvelle méthode de lemmatisation qui essaye de déterminer le noyau d'un mot selon des règles linguistiques appuyées par des statistiques de corpus. Cette méthode est comparée à une technique de lemmatisation assouplie largement utilisée dans la littérature. La nouvelle méthode présente une meilleure performance de recherche que la technique assouplie. Cette dernière ne réussit pas à grouper beaucoup de mots sémantiquement similaires dans le même index. Au contraire, la méthode à base linguistique peut mieux déterminer le noyau sémantique d'un mot.

Cependant, la nouvelle méthode peut également entraîner des erreurs. Ce cas de figure apparaît quand un affixe est tronqué d'un mot mais il ne devait pas l'être parce qu'il fait partie des lettres du lemme. Pour surmonter ce genre d'ambiguïté, nous estimons que cette méthode peut être améliorée au niveau des statistiques de corpus. Plus de traitement doit être fait pour choisir le lemme correct quand différents lemmes candidats sont proposés pour un mot. Une première réflexion serait de choisir le lemme le plus utilisé dans un document ou une requête au lieu de favoriser le lemme le plus fréquent dans toute la collection des documents.
Dans le prochain chapitre nous investiguons les techniques de traduction de requêtes de l'anglais vers l'arabe et la façon de les combiner dans le modèle de recherche pour retrouver le plus de documents pertinents possibles.

CHAPITRE 5

RECHERCHE D'INFORMATION TRANSLINGUISTIQUE ANGLAIS-ARABE

5.1 Introduction

Dans ce chapitre, on s'intéresse à la deuxième phase de notre travail, à savoir la RI translinguistique où la requête est représentée dans un langage source (anglais) et la collection de documents dans un autre langage cible (arabe). La tâche centrale dans la RIT est de permettre aux documents de se comparer à une requête, typique via une traduction.

Une des approches de l'unification des langues de la requête et des documents est la traduction de la requête vers la langue des documents. Cette approche est réalisée avec un coût moindre du fait que la taille de la requête est généralement réduite à quelques termes. C'est aussi une approche plus flexible parce que l'utilisateur peut choisir la ou les langue(s) qui l'intéressent. Nous avons présenté plus en détail dans le troisième chapitre les façons d'implanter cette approche. Le choix d'une méthode par rapport à une autre dépend de la disponibilité des ressources et la qualité de traduction qu'elle produit. Rappelons quelques éléments décrits dans le chapitre 3.

La traduction automatique de bonne qualité est toujours un problème non résolu surtout pour une paire de langues (anglais-arabe) en stade d'exploration. Le coût élevé en temps pour implémenter un système de traduction automatique fiable nous a aussi poussé à écarter cette piste.

D'autre part, les requêtes sont constituées généralement des mots clés et ne respectent pas souvent une structure syntaxique. Dans ce contexte, l'utilisation des

dictionnaires bilingues et les corpus parallèles deviennent des alternatifs intéressants, surtout que la traduction d'une requête vise à suggérer de bons termes pour trouver des documents et non à produire une phrase compréhensible par l'humain. La traduction des termes de requête peut aussi s'étendre à des termes reliés.

Cependant les ressources reliées à ces deux méthodes ne sont pas disponibles pour l'arabe. En l'absence de corpus parallèles comme le Hansard, l'exploitation du Web est une piste privilégiée pour construire un corpus de textes parallèles pour la paire anglais-arabe. Sur la base de ce corpus, un Modèle de Traduction Statistique (MTS) est entraîné spécifiquement pour la RIT anglais-arabe [32].

Toutefois, l'utilisation d'une seule ressource de traduction est souvent insuffisante du fait du contexte général des thèmes de la requête. Autrement dit, Il arrive parfois que des termes d'une requête ne sont pas couverts par une ressource de traduction. En effet, pour consolider la traduction des requêtes, l'exploration d'autres techniques de traduction basées sur les Dictionnaires Bilingues (DB) ainsi que les textes parallèles est nécessaire. Ainsi, un des objectifs de notre étude est de voir comment différentes ressources linguistiques disponibles peuvent être combinées dans la tâche de la traduction de requête.

Pour notre étude, nous allons exploiter notamment les ressources suivantes :

- Un modèle de traduction statistique entraîné sur les pages Web parallèles.
- Un autre modèle de traduction statistique entraîné sur le corpus parallèle des nations unies.
- Deux dictionnaires bilingues extraits du Web : Ajeeb et Almisbar.

Une fois les ressources de traduction identifiées, la question de la combinaison des ressources s'impose naturellement. La combinaison de plusieurs modèles se résout traditionnellement par l'attribution des poids à chacun des modèles ou

des ressources. Ces poids mesurent le degré de confiance accordé à chacune des ressources. Au lieu d'attribuer une mesure de confiance à une ressource globalement, nous allons étudier une autre méthode de combinaison qui utilise des facteurs de confiance associés à chaque traduction. Cette méthode fournit une ré-estimation des différentes traductions en examinant des critères supplémentaires. Nous allons décrire ceci plus en détail.

Dans la section suivante, on présente les travaux reliés à cette problématique. Ensuite, nous développerons successivement différentes approches de traduction des requêtes conjointement avec les ressources qu'on a pu identifier pour la paire de langues anglais-arabe. L'intégration des probabilités de traduction dans un modèle de recherche basé sur les modèles de langue sera étudiée avec les différentes techniques de combinaison des traductions. On s'attardera aussi sur l'apprentissage des facteurs de confiance. A la fin de ce chapitre, on présentera nos expérimentations suivies d'analyses des différentes méthodes de traduction de requêtes.

5.2 Travaux reliés

Dans la plupart des travaux sur la RIT, le problème majeur est la traduction des requêtes. Les principales approches utilisées pour la traduction des requêtes sont la traduction automatique, les dictionnaires bilingues et les corpus parallèles [56]. Malgré le fait que la traduction automatique peut être utilisée pour la traduction de requêtes, elle n'est pas disponible pour beaucoup de paires de langues. Ainsi, on doit utiliser d'autres alternatives. De plus, la traduction automatique n'est pas une approche parfaitement adaptée à traduire des requêtes : Les requêtes sont rarement des phrases et plus souvent juste une séquence de mots sans structure syntaxique. Ceci pose un grand problème aux systèmes de traduction automatique, qui sont généralement conçus pour des phrases complètes. Un autre inconvénient de la traduction automatique est que les systèmes de traduction automatique fournissent une seule traduction pour un mot source, alors qu'il est généralement préférable

qu'un terme de la requête soit traduit par plusieurs mots dans le but de produire l'effet d'expansion de requête, fortement souhaité en RI. De plus, le choix d'une seule traduction pour un mot source par un système de traduction automatique peut être erroné.

Les dictionnaires bilingues sont largement utilisés en RIT [7] [27], grâce, d'une part, à leur large disponibilité, et d'autre part, à la facilité de les intégrer dans une approche de traduction. Une approche simple d'utilisation d'un dictionnaire consiste à remplacer chaque terme de requête par une liste de ses traductions possibles. La multitude de choix de traductions qu'ils offrent, a aussi favorisé leur préférence d'utilisation en RIT par rapport à la traduction automatique. Cependant, comme chaque méthode a ses limites, les dictionnaires bilingues souffrent de manque de couverture de traduction pour certains termes de requêtes tels que les noms propres. L'autre problème des dictionnaires bilingues est l'effet de polysémie qui engendre parfois des traductions ambigües. Ajoutons aussi à ces inconvénients la difficulté d'identifier de telles ressources pour la paire de langues anglais-arabe.

La troisième approche exploite des corpus parallèles. Les corpus parallèles contiennent des informations utiles pour la traduction des mots dans des domaines particuliers. Ils peuvent aussi surmonter quelques limitations des techniques basées sur les dictionnaires bilingues. On peut utiliser de tels corpus pour entraîner des modèles de traduction statistique, qui peuvent ensuite être utilisés pour traduire les termes d'une requête. Cette approche présente l'avantage que peu d'interventions manuelles sont nécessaires pour produire un modèle de traduction statistique. L'autre avantage des techniques basées sur les corpus parallèles est que chaque mot source peut être traduit par plusieurs mots cibles et que ces derniers sont pondérés [54]. Cette technique a été utilisée en RIT pour traduire des requêtes de l'anglais vers le chinois et le français et a révélée de bonnes performances [52] [53] [41]. Cependant, le problème avec l'utilisation des corpus parallèles est que ces derniers ne sont pas disponibles et leur acquisition est coûteuse. D'une part,

il est difficile de trouver des traductions déjà existantes de bonne qualité pour des documents et d'autre part, les versions traduites sont coûteuses à créer. Le corpus Hansard anglais-français du parlement canadien est parmi les rares corpus parallèles existants non seulement par sa disponibilité mais aussi par sa taille considérable et sa qualité de traduction. Malheureusement, on ne peut pas se procurer des corpus bilingues semblables à Hansard pour la paire de langues anglais-arabe.

Une autre catégorie d'approches pour la traduction des requêtes, combine plusieurs ressources de traduction notamment les corpus parallèles et les dictionnaires bilingues [53]. Cette combinaison repose sur deux principes. Le premier s'annonce comme suit : Comme différentes ressources peuvent suggérer différentes traductions, il vaut mieux combiner ces ressources afin d'obtenir autant de traductions correctes que possible. Le deuxième principe est celui des évidences multiples : lorsque plusieurs traductions sont fournies pour une requête, celle proposée par plusieurs ressources aura une chance plus élevée d'être correcte.

L'attribution de poids appropriés aux termes de la requête est un autre problème crucial pour la RIT. Si les modèles de traduction statistiques sont utilisés pour la traduction des requêtes, les probabilités de traduction peuvent être utilisées comme poids associés aux termes de la requête [41]. Mais, les dictionnaires bilingues ne fournissent pas de probabilités aux traductions. Ainsi, dans les études antérieures, les poids sont souvent déterminés selon une distribution uniforme ou selon les occurrences et les cooccurrences des traductions dans un corpus [75]. Quand plusieurs outils ou ressources de traduction sont utilisés, la question qui se manifeste est comment combiner correctement toutes les traductions candidates? Dans les études précédentes, de simples méthodes ont souvent été employées, i.e. on combine linéairement diverses traductions pour un même terme de requête par l'attribution d'un poids de confiance à la ressource de traduction [79] [19]. Dans d'autres études, ces poids de confiance accordés aux ressources de traduction ont été optimisés automatiquement sur des corpus de validation indépendants des pre-

mières ressources utilisées pour obtenir les traductions et les probabilités [53].

On remarque que l'utilisation d'une combinaison linéaire des ressources attribue un poids unique pour chaque ressource de traduction. Même si on peut combiner ces poids avec les probabilités initiales de traduction (si elles sont disponibles), les scores résultants peuvent seulement faire la différence entre les traductions proposées par les différentes ressources. Ces scores ne modifient pas l'importance relative des traductions de la même ressource [34]. En pratique, quand de nouveaux critères sont considérés, une traduction suggérée par une ressource avec un poids faible, peut s'avérer être une meilleure traduction. Dans un tel cas, il est nécessaire de modifier l'importance relative de cette traduction dans l'ensemble de traductions. Par exemple, dans une requête TREC 2001, le mot anglais "develop" dans "to develop tourism in Cairo" (لتطوير السياحه العربيه) est traduit en arabe par un modèle de traduction statistique avec l'ensemble des traductions suivantes :

{تنميه 0.48 (développement), نامي 0.13 (développé), إنماء 0.08 (développement), تطور 0.06 (évolution), تطوير 0.04 (développement)}.

Nous observons que la traduction la plus commune "تطوير" (développement) prend seulement la cinquième position avec une probabilité beaucoup plus faible que celle de la traduction en première position "تنميه"[1]. Si une combinaison linéaire est employée pour combiner ce modèle de traduction avec une autre ressource (comme un dictionnaire bilingue), il est peu probable que la bonne traduction "تطوير" puisse gagner un poids plus grand que "تنميه". Il est alors important de reconsidérer chaque traduction candidate selon des critères additionnels afin

[1]"تنميه est aussi une traduction pour le mot "develop" mais la traduction "تطوير" est plus appropriée. D'autant plus que dans la version arabe de la requête, on utilise la traduction "تطوير" au lieu de la traduction "تنميه pour le mot "develop".

d'en produire de nouveaux poids. En faisant ceci, l'ordre initial des traductions candidates peut être changé. Plus précisément, en utilisant la méthode des facteurs de confiance, nous pouvons réordonner les traductions candidates comme suit:

{تطوير 0.51, تنميه 0.29}.

Dans cette nouvelle liste, le poids de la meilleure traduction "تطوير" est considérablement augmenté tel que souhaité.

La technique des mesures de confiance peut ajuster le poids des traductions d'un terme de requête selon des attributs informatifs additionnels. Cet estimé de confiance mesure le degré de certitude que la traduction soit correcte. Il nous fournit un moyen pour réévaluer les traductions candidates provenant de différentes ressources d'une façon homogène. Les avantages de cette approche sont donc doubles. D'une part, la mesure de confiance nous permet d'ajuster le poids original des traductions et de filtrer les meilleurs termes de traduction. D'autre part, ces mesures de confiance nous fournissent également des poids comparables pour les traductions candidates à travers des ressources de traduction différentes. En conséquence, la mesure de confiance peut être vue comme un mécanisme général pour combiner, d'une manière efficace, différentes ressources de traduction.

A la fin de ce chapitre, nous présentons des expériences qui prouvent que cette méthode surclasse la méthode de combinaison linéaire sur deux collections de test. Dans la section 5.9, nous présentons les étapes d'apprentissage de ces estimés de confiance.

5.3 Un modèle de traduction basé sur les pages Web parallèles

Nous avons mentionné qu'une des approches souvent utilisées consiste à exploiter une large collection de textes parallèles pour entraîner un modèle de traduction statistique. Toutefois la mise en application d'un tel modèle performant nécessite un très grand nombre de textes parallèles. Or, pour plusieurs paires de langues, on ne peut pas obtenir suffisamment de textes parallèles. Afin de surmonter cet obstacle, Nie a eu l'idée de fouiller le Web pour chercher automatiquement des pages Web parallèles à l'aide d'un système dit PTMiner [52] [53]. Dans ce contexte, on a repris la même expérience avec PTMiner pour construire un corpus bilingue pour la paire de langues anglais-arabe. Sur la base de ce corpus, un modèle de traduction statistique est entraîné [32]. Dans les sections suivantes, on décrit les différentes étapes du système PTMiner.

5.3.1 PTMiner

Le système PTMiner [16] vise à fouiller le Web et à construire un corpus parallèle anglais-arabe. Le principe de PTMiner est que sur le Web, il y a beaucoup de textes parallèles (un exemple est présenté dans le tableau 3.1 du chapitre 3). De plus, ces textes parallèles sont très souvent organisés selon certains schémas ou règles communes comme :

- Les textes parallèles utilisent des textes d'ancrage (anchor text) comme " النسخه العربيه" (Arabic version), "عربي" (Arabic), "English version" ou "English" pour pointer d'une langue à une autre.

- Les textes parallèles ont généralement des URLs similaires. La différence

entre elles réside seulement et souvent dans les préfixes ou suffixes indiquant le langage dans lequel est écrit le document. On peut citer à titre d'exemple, les segments ".en", ".e", "english-" pour les documents anglais et leurs segments équivalents pour les documents arabes; ".ar", ".a", "arabic-".

Ainsi, en explorant le Web, il est possible de reconnaître des textes parallèles automatiquement. L'application de ce processus a permis de constituer des corpus parallèles pour plusieurs paires de langues. Le tableau 5.1 montre le nombre de textes parallèles avec leurs capacités pour la paire de langues anglais-arabe ainsi que pour quelques autres paires de langues étudiées antérieurement [53].

Paire	anglais-arabe	anglais-français	anglais-allemand	anglais-italien
Nombre de paires	2816	18807	10195	8499
Capacité (en MB)	37	300	115	85

Tableau 5.1: Les quatre corpus parallèles constitués à l'aide de PTMiner. Les corpus autres que anglais-arabe sont collectés dans [53].

Pour la paire anglais-arabe, un corpus modeste en taille (2 816 paires de pages) est constitué. Ce petit nombre de pages parallèles est expliqué à notre avis par la rareté des sites Web pour cette paire de langues. Nous notons que ce petit nombre (2 816) de textes parallèles anglais-arabe n'est pas suffisant pour construire un modèle de traduction statistique raisonnable. En plus de la taille du corpus parallèle, la qualité de traduction qu'offre un modèle de traduction statistique dépend aussi de la qualité typographique du corpus parallèle. D'autres facteurs comme le temps de la création du corpus parallèle et la variation linguistique et culturelle d'un pays à l'autre jouent un rôle important dans la qualité des traductions.

5.3.2 Lemmatisation

Les textes parallèles collectionnés à partir du Web, après un premier nettoyage, ont subi deux procédures de lemmatisation avant leur passage pour l'entraînement

avec les modèles de traduction. Les documents arabes sont lemmatisés par la méthode de troncature à base linguistique [33] que nous avons décrit précédemment au chapitre quatre. Quant aux documents anglais, la technique de Porter [60] a été appliquée pour trouver les lemmes des mots anglais. Notant aussi que les mots outils ont été éliminés des textes. Pour l'arabe, nous avons utilisé la liste des mots outils arabes que nous avons conçue (Chapitre 4). La table des stopwords anglais a été procurée du package Snowball de Porter [2].

5.3.3 Alignement

L'entraînement des modèles de traduction statistiques nécessite l'existence des bitextes (textes alignés en phrases). On peut obtenir des bitextes à partir d'un corpus parallèle en alignant ce corpus au niveau des phrases. L'alignement est fait de différentes façons, Langlais dresse une comparaison de plusieurs algorithmes d'alignements de phrases [46]. Pour ce faire, deux types d'informations sont exploités dans les algorithmes d'alignement :

Informations métriques : Ces critères utilisent la position et la longueur des phrases dans les textes : Les phrases parallèles dans deux textes parallèles ont généralement des positions similaires. Les longueurs des phrases alignées sont aussi similaires. Church et Gale ont montré qu'il existe un rapport de proportionnalité entre la longueur d'une phrase source et la longueur de sa traduction [23]. Ainsi, en essayant de faire correspondre les phrases de longueurs similaires dans l'ordre, en utilisant la programmation dynamique, on peut déterminer les meilleurs alignements de phrases selon ces critères.

Informations linguistiques : Simard *et al.* utilisent d'autres unités linguistiques qu'ils appellent les cognats pour aligner des bitextes [70]. Un cognat désigne des mots semblables dans deux langues comme "système" en français et "system" en anglais. L'idée derrière l'exploitation de cognats est que deux phrases en relation de traduction partagent souvent des mots communs ou proches comme les noms

propres, les symboles, les chiffres ou tout simplement partagent une forme identique dans les deux langues. Ainsi, deux phrases contenant ces cognats ont plus de chance d'être parallèles. Naturellement, l'approche de cognat ne peut s'appliquer pour des langues si différentes comme l'anglais et l'arabe mais elle peut être étendue en utilisant un dictionnaire bilingue, i.e. en considérant une traduction stockée dans un dictionnaire comme l'équivalent d'un cognat.

Pour nos expérimentations, on a utilisé le système d'alignement SFIAL basé sur les positions et les longueurs des phrases ainsi que les cognats [70]. Ainsi, avant l'entraînement des modèles de traduction statistique, on a aligné le corpus parallèle anglais-arabe recueilli à partir du Web à l'aide du système SFIAL pour avoir des bitextes qui seront les entrées des modèles de traduction IBM.

5.3.4 Modèles de traduction probabiliste IBM

Quand on arrive à aligner les textes parallèles phrase par phrase, c'est au tour des mots d'être alignés. Ceci revient à estimer les relations de traduction de mots. Cette tâche est l'objet des modèles de traduction statistique.

Brown *et al.* [11] proposent cinq modèles de traduction 1, 2, 3, 4 et 5. Chaque modèle a sa propre prescription pour calculer la probabilité conditionnelle $p(f_j \mid e_i)$ entre les mots e_i et f_j qu'on appelle la probabilité de traduction. Intuitivement, plus une paire de mots apparaît dans des phrases parallèles, plus grande est la chance que les deux mots de cette paire soient la traduction l'un de l'autre. Plus précisément, on utilise généralement le processus de maximisation d'estimation (EM) afin de maximiser la vraisemblance de traduction entre les phrases parallèles du corpus. L'algorithme EM est la base d'estimation des paramètres des modèles IBM. Cet algorithme est décrit dans [11]. A la fin, ce processus de maximisation EM produit une fonction de probabilité $p(f_j \mid e_i)$ exprimant la probabilité que le mot f_j soit la traduction du mot e_i. Avec cette fonction, on peut déterminer un ensemble de traductions probables dans la langue cible pour chaque mot de la

langue source.

Formellement, ces modèles cherchent à modéliser $P(F = f \mid E = e)$ où E est l'ensemble de phrases sources, F est l'ensemble de phrases cibles. $e = e_1, ..., e_l$ et $f = f_1, ..., f_m$ sont deux phrases particulières de E et F. Soit $A(e, f)$ l'ensemble des alignements liant une phrase source donnée à une phrase cible. On note par $P(F = f, A = a \mid E = e)$ la probabilité jointe de f et d'un alignement particulier a étant donné e. Si on somme sur tous les alignements possibles entre e et f, on peut estimer $P(F = f \mid E = e)$ par :

$$P(F = f \mid E = e) = \sum_{a \in A(e,f)} P(F = f, A = a \mid E = e) = \sum_{a} P(f, a \mid e) \qquad (5.1)$$

Le dernier élément dans l'équation 5.1 est juste une écriture abrégée. $a = (a_1, ..., a_m)$ avec $a_i \in [0, l] \forall i \in [1, m]$.

Dans le modèle IBM 1, tous les alignements sont considérés équiprobables et indépendants de la position du mot dans la phrase cible. Chaque mot cible f_j possède donc $(l + 1)$ positions possibles ($+1$ car le mot "NULL" e_0 est considéré)[2] [45]. Sachant qu'il y a $(l + 1)^m$ alignements possibles entre une phrase source e de longueur l et une phrase cible f de longueur m, $P(f, a \mid e)$ est estimé comme suit :

$$P(f, a \mid e) = P(m \mid e) \prod_{j=1}^{m} \frac{t(f_j \mid e_{a_j})}{(l + 1)} = \frac{\epsilon}{(l + 1)^m} \prod_{j=1}^{m} t(f_j \mid e_{a_j}) \qquad (5.2)$$

où ϵ est un facteur qui exprime la probabilité de générer une phrase cible f de longueur m sachant la phrase source e. $t(f_j \mid e_{a_j})$ est la probabilité de transfert ou la probabilité que le mot e_{a_j} génère le mot f_j. Pour expliquer cette notation,

[2] la phrase e est étendue par le mot e_0 auquel seront associés les mots de f qui n'ont pas de correspondant direct dans la langue source.

on peut imaginer un alignement comme $a = (1, 2, 4, 3)$, où a_j est une position d'un mot dans la phrase source e. e_{a_j} est le mot qui est responsable de la génération du j^e mot du f. Si l'on somme sur tous les alignements possibles alors la probabilité d'une phrase f sachant une phrase e :

$$P(f \mid e) = \sum_a P(f, a \mid e) = \sum_a \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} t(f_j \mid e_{a_j}) \tag{5.3}$$

$$= \frac{\epsilon}{(l+1)^m} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j \mid e_{a_j}) \tag{5.4}$$

$$= \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j \mid e_i) \tag{5.5}$$

Cette dernière équation signifie que chaque mot cible de la phrase f est généré en consultant les probabilités de transfert de chaque mot source de la phrase e vers ce mot cible.

Le modèle IBM 1 possède la propriété que l'on peut calculer $P(f \mid e)$ de manière exacte et efficace. Dans cette équation, on cherche à estimer les probabilités de transfert $t(f_j \mid e_i)$ de manière à augmenter le logarithme de vraisemblance d'un corpus d'entraînement $LL(C)$ sous les contraintes stochastiques de l'équation 5.7:

$$LL(C) = \sum_{s=1}^{S} \log P(f^{(s)} \mid e^{(s)}) \tag{5.6}$$

$$\sum_{f_j} t(f_j \mid e_i) = 1 \ , \ \forall e_i \tag{5.7}$$

où $(f^{(1)}, e^{(1)}), (f^{(2)}, e^{(2)}), ..., (f^{(S)}, e^{(S)})$ est l'ensemble des paires de phrases du corpus. f_j désigne n'importe quel mot de la langue F et e_i désigne n'importe quel mot de la langue E. Les probabilités $t(f_j \mid e_i)$ sont optimisées itérativement par

un processus de maximisation d'estimation (EM) :

$$c(f_j \mid e_i; e, f) = \frac{t(f_j \mid e_i)}{\sum_{i=0}^{l} t(f_j \mid e_i)} \; nb(e_i, e) \; nb(f_j, f) \qquad (Etape - E) \qquad (5.8)$$

$$t(f_j \mid e_i) = \lambda_e \sum_{s=1}^{S} c(f_j \mid e_i; f^{(s)}, e^{(s)}) \qquad (Etape - M) \qquad (5.9)$$

où $nb(e_i, e)$ est le nombre de mot e_i dans la phrase e et $nb(f_j, f)$ est le nombre de mot f_j dans la phrase f. λ_e est un facteur de normalisation qui assure les contraintes stochastiques de l'équation 5.7.

Rappelant que le modèle IBM 1 estime la valeur de la probabilité de transfert en supposant que toutes les positions d'alignement d'un mot source avec un mot cible sont équiprobables. Le modèle IBM 2 remédie à cette simplification en introduisant des probabilités d'alignement selon les positions des mots sources et cibles. Son argumentation est le fait qu'un mot source dans une position donnée se connecte généralement à un mot cible dans la même position de la phrase cible. Ce nouveau paramètre calculé par le modèle IBM 2 est dit probabilité de distorsion.

Le modèle IBM 3 introduit un nouveau paramètre exprimant la probabilité de fertilité qui représente la distribution du nombre de mots cibles générés par un mot source particulier. Tous les détails de calcul de ces paramètres pour les modèles IBM sont décrits par Brown dans [11]. On peut aussi se référer au document de Knight qui présente ces modèles d'une façon plus accessible [38].

Pour l'entraînement de ces modèles, on a utilisé le package GIZA++ [58], une extension de GIZA [6], qui implémente les modèles IBM de Brown. On appelle la sortie d'un modèle de traduction, les probabilités de transfert ou encore un dictionnaire bilingue probabilisé. Comme exemple, on peut voir dans le tableau 5.2 la

sortie du modèle IBM 1 pour la traduction du mot anglais "develop" vers l'arabe. Chaque mot source a plusieurs traductions dans la langue cible et à chaque traduction est associé un poids représentant la probabilité de transfert. Dans nos expérimentations, nous avons utilisé juste les probabilités du modèle IBM 1. Ce modèle ne prend pas en considération l'ordre des mots, ce qui correspond bien au contexte de recherche d'information dans la mesure où l'ordre des mots est aussi ignoré dans la plupart des modèles actuels de recherche d'information.

(تنميه, 0.48) (développement)
(نامي, 0.13) (développé)
(إنماء, 0.08) (développement)
(تطور, 0.06) (évolution)
(تطوير, 0.04) (développement)

Tableau 5.2: Exemple de la sortie d'un modèle de traduction pour le mot source "develop"

La couverture de ce modèle de traduction statistique entraîné sur les pages Web est de 11.30 % des mots uniques de la collection de documents TREC sur laquelle nous avons fait nos expérimentations. Quant à la couverture des termes de requêtes, ce modèle couvre 95.37 % des mots uniques des deux collections TREC. Ces taux montrent que la majorité des termes de requêtes peuvent être traduits par ce modèle mais les traductions proposées ont moins de chance d'être trouvées parmi les termes de la collection de documents.

5.4 Un autre modèle de traduction statistique basé sur le corpus parallèle des Nations Unies

Cette ressource ressemble au modèle de traduction statistique construit à partir des textes parallèles extraits du Web. C'est une table de traduction probabiliste de l'anglais vers l'arabe. Cette table est bâtie par l'équipe BBN [22] à partir d'une large collection de textes bilingues alignés anglais-arabe. Le corpus provient des

archives des Nations Unies (NU), ses documents représentent les discours officiels et sont de bonne qualité de traduction. Il renferme 38 000 paires de documents. Ce modèle de traduction a été entraîné avec GIZA++ [58]. La couverture de la table de traduction de l'arabe vers l'anglais est de 29 % des mots uniques de la collection de documents TREC sur laquelle nous avons fait nos expérimentations. Ce taux est supérieur à celui du modèle précédent. Pour les termes de requêtes en anglais, la table de traduction de l'anglais vers l'arabe couvre 97.83 % des mots uniques des deux collections de requêtes TREC. Malgré que le taux de couverture des termes de requêtes du modèle précédent est comparable au taux de ce modèle, les traductions fournies par le modèle des NU ont plus de chance d'être trouvées dans la collection des documents que celles proposées par le modèle entraîné sur les pages Web.

Malheureusement, nous n'avons pas obtenu les textes parallèles des NU pour entraîner nous-mêmes un modèle de traduction. Ainsi, nous allons utiliser le modèle entraîné par BBN, ce qui pose quelques difficultés additionnelles parce que BBN n'a pas utilisé la même méthode de lemmatisation que nous. Ainsi, une relemmatisation sur les mots dans ce modèle sera nécessaire. Nous avons créé un ensemble de règles pour transformer les lemmes de BBN en les nôtres.

5.5 Dictionnaires bilingues

Les autres sources de traduction qu'on a exploitées pour la traduction des requêtes sont des dictionnaires bilingues. Parmi une variété disponible sur le Web, on a sélectionné deux sites : Almisbar [4] et Ajeeb [3] qui offrent une meilleure qualité de traduction. Pour bâtir ces deux ressources, on a implémenté un script qui extrait en ligne sur le Web les traductions d'un ensemble de mots anglais [3]. Ajeeb inclut 20 K paires d'entrées et Almisbar 11 K de paires.

[3]Ces mots sont collectés de deux corpus de LDC : LDC2004T18 et LDC2004T17, plus les termes des deux collections de requêtes TREC

5.6 Combinaison des ressources

L'utilisation de plusieurs ressources donne généralement une meilleure qualité de traduction de requête. D'une part, la combinaison des ressources permet de contourner le problème de couverture d'une ressource pour certains termes de requêtes. D'autre part, elle nous permet de bénéficier du principe d'évidence multiple ainsi que l'effet d'expansion de requête, très souhaité en recherche d'information.

Plus précisément, on a identifié et conçu quatre modèles de traduction différents entre l'anglais et l'arabe provenant de quatre ressources différentes : deux modèles de traduction statistique (Web et NU) et deux dictionnaires bilingues (Almisbar et Ajeeb). La question qui se pose pertinemment à ce stade est comment combiner ces ressources potentielles ensemble d'une façon raisonnable pour augmenter la qualité de traduction de requête et par conséquent améliorer la performance de recherche des documents pertinents dans la collection des documents?

Pour répondre à cette question, deux méthodes de combinaison sont étudiées: Une méthode traditionnelle, combinant linéairement les quatre ressources par l'attribution d'un poids de confiance à chaque ressource. L'autre méthode est plus sophistiquée. Elle reconsidère toutes les traductions candidates proposées par les différentes ressources et, en introduisant des attributs additionnels, elle les réévalue plus radicalement que dans la combinaison linéaire. Cette deuxième méthode fournit également de nouveaux poids comparables pour les traductions candidates à travers des ressources de traduction différentes. Ainsi, la question de combinaison des ressources est dépendante de la problématique de la façon d'associer des poids d'importance à des traductions différentes proposées par une ou plusieurs ressources pour un même mot.

Dans les sections suivantes, nous présentons le cadre général d'un modèle de recherche d'information permettant d'intégrer des poids avec les traductions ainsi

que le procédé de calcul de ces poids.

5.7 Intégration des probabilités de traduction dans le processus de recherche

Dans ce paragraphe on décrit le cadre général d'un modèle de recherche d'information permettant d'intégrer des poids avec les différentes traductions d'un terme de la requête. Pour ceci, on utilise un modèle de recherche basé sur les modèles de langue. Etant donné une requête Q_E écrite dans un langage source E et un document D_A représenté dans un autre langage cible A, on peut calculer la pertinence de ce document par rapport à la requête avec la divergence négative entre le modèle de langage de requête θ_{Q_E} et le modèle de langage du document θ_{D_A} [82].

$$R(Q_E, D_A) \simeq \sum_{t_A \in V_A} p(t_A \mid \theta_{Q_E}) \log p(t_A \mid \theta_{D_A}) \tag{5.10}$$

Où V_A = le vocabulaire de A. Comme nous l'avons décrit dans le chapitre 2 (section 2.7.4), pour contourner le problème d'attribuer des probabilités zéro aux termes de la requête non présents dans le document D_A, les techniques de lissage sont utilisées pour estimer $p(t_A \mid \theta_{D_A})$. On peut utiliser la technique de lissage Jelinek-Mercer qui est une méthode d'interpolation entre les modèles de langage du document θ_{D_A} et de celui de la collection θ_{C_A} [81]. Le lissage de $p(t_A \mid \theta_{D_A})$ est calculé comme suit:

$$p(t_A \mid \theta_{D_A}) = (1 - \lambda) p_{MLE}(t_A \mid \theta_{D_A}) + \lambda p_{MLE}(t_A \mid \theta_{C_A}) \tag{5.11}$$

Où $p_{MLE}(t_A \mid \theta_{D_A}) = \frac{tf(t_A, D_A)}{|D_A|}$ et $p_{MLE}(t_A \mid \theta_{C_A}) = \frac{tf(t_A, C_A)}{|C_A|}$ sont les estimés du maximum de vraisemblance des modèles de langue unigrammes basés respectivement sur le document D_A et la collection des documents C_A. λ est un paramètre qui contrôle l'influence de chacun des deux modèles.

Le terme $p(t_A \mid \theta_{Q_E})$ dans l'équation 5.10 représentant le modèle de la requête peut être estimé dans la langue source avec :

$$p(t_A \mid \theta_{Q_E}) = \sum_{q_E \in V_E} p(t_A, q_E \mid \theta_{Q_E}) = \sum_{q_E \in V_E} p(t_A \mid q_E, \theta_{Q_E}) p(q_E \mid \theta_{Q_E}) \qquad (5.12)$$

On suppose que $p(t_A \mid q_E, \theta_{Q_E}) \simeq p(t_A \mid q_E)$. C'est-à-dire que le terme de traduction t_A, est essentiellement déterminé par le terme source q_E, et indépendamment de la requête θ_{Q_E}. Evidemment, ceci est une simplification. Ainsi :

$$p(t_A \mid \theta_{Q_E}) \simeq \sum_{q_E \in V_E} p(t_A \mid q_E) p_{MLE}(q_E \mid \theta_{Q_E}) \qquad (5.13)$$

Où V_E = le vocabulaire de E, q_E est un terme en langage source, $p_{MLE}(q_E \mid \theta_{Q_E})$ est l'estimé du maximum de vraisemblance du modèle de langue unigramme basé sur la requête Q_E : $p_{MLE}(q_E \mid \theta_{Q_E}) = \frac{tf(q_E, Q_E)}{|Q_E|}$ et $p(t_A \mid q_E)$ est le modèle de traduction. En remplaçant la formule 5.13 dans 5.10, nous obtenons la formule générale de classement des documents pertinents [41]:

$$R(Q_E, D_A) \simeq \sum_{t_A} \sum_{q_E} p(t_A \mid q_E) p_{MLE}(q_E \mid \theta_{Q_E}) \log p(t_A \mid \theta_{D_A}) \qquad (5.14)$$

Notre travail maintenant se concentre sur l'estimation du modèle de traduction $p(t_A \mid q_E)$. Quand une seule ressource est utilisée, on utilise la probabilité de traduction comme poids associé à la traduction. Pour les modèles de traduction statistique, le modèle de traduction $p(t_A \mid q_E)$ est estimé avec le modèle IBM 1 [11]. Pour les dictionnaires bilingues, on utilise un corpus bilingue pour calculer des probabilités de traduction pour chaque paire de traduction (voir la section 5.9.4).

Quand plusieurs ressources de traduction sont utilisées, le modèle de traduction $p(t_A \mid q_E)$ est estimé de deux façons :

1. Par une combinaison linéaire :

$$p(t_A \mid q_E) = z_{q_E} \sum_i \lambda_i p_i(t_A \mid q_E) \tag{5.15}$$

où λ_i est le paramètre relié à la ressource de traduction i et z_{q_E} est un facteur de normalisation de sorte que $\sum_{t_A} p(t_A \mid q_E) = 1$. $p_i(t_A \mid q_E)$ est la probabilité de traduire le mot source q_E par le mot cible t_A utilisant la ressource i.

2. par les facteurs de confiance : Etant donné une traduction candidate t_A pour un mot source q_E et F un ensemble d'attributs, $p(t_A \mid q_E)$ est calculé avec la somme des estimés de confiance portés sur cette traduction utilisant différentes ressources, i.e.:

$$p(t_A \mid q_E) = z_{q_E} \sum_i p_i(C = 1 \mid t_A, q_E, F) \tag{5.16}$$

où $p_i(C = 1 \mid t_A, q_E, F)$ est la probabilité d'exactitude de t_A pour traduire q_E. Cette probabilité est normalisée telle que : $\sum_{t_A} p(C = 1 \mid t_A, q_E, F) = 1$.

Dans les deux sections qui suivent, nous décrivons plus en détail comment estimer les paramètres de combinaison linéaire ainsi que les facteurs de confiance.

5.8 Combinaison linéaire

Pour déterminer les paramètres appropriés pour chacune des ressources de traduction, l'algorithme EM est utilisé pour trouver les valeurs qui maximisent le logarithme de vraisemblance LL d'un ensemble de données C selon le modèle combiné [36]. L'ensemble de validation C est un corpus de phrases alignées anglais-arabe.

$$LL(C) = \sum_{(a,e)} p(a,e) \sum_{j=1}^{|a|} \log \sum_{k=1}^{r} \sum_{i=1}^{|e|} \lambda_k t_k(a_j \mid e_i) p(e_i) \tag{5.17}$$

Où $p(a,e) = \frac{\#(a,e)}{|C|}$ est la probabilité à priori de la paire de phrases (a,e) dans le corpus C, $|a|$ est le nombre de mots de la phrase cible a et $|e|$ est le nombre de mots de la phrase source e. r est le nombre de ressources combinées. λ_k est le coefficient relié à la ressource k que nous voulons optimiser. $t_k(a_j \mid e_i)$ est la probabilité de traduire le mot source e_i par le mot cible a_j utilisant la ressource k. $p(e_i)$ est la probabilité à priori du mot source e_i dans le corpus C. Ainsi pour estimer les valeurs des λ_k, l'algorithme EM calcule itérativement les deux paramètres suivants jusqu'à la convergence:

$$C_k^n(a_j \mid e) = \frac{\sum_{i=1}^{|e|} \lambda_k^n \; t_k(a_j \mid e_i) \; p(e_i)}{\sum_{k=1}^{r} \lambda_k^n \; \sum_{i=1}^{|e|} t_k(a_j \mid e_i) \; p(e_i)} \qquad (Etape - E) \tag{5.18}$$

$$\lambda_k^{n+1} = \frac{\sum_{(a,e)} p(a,e) \sum_{j=1}^{|a|} C_k^n(a_j \mid e)}{\sum_{(a,e)} p(a,e) \;\; \mid a \mid} \qquad (Etape - M) \tag{5.19}$$

5.9 Facteurs de confiance

L'estimation de confiance a été à l'origine utilisée en reconnaissance de la parole [28]. Lorsque les erreurs se produisent fréquemment en reconnaissance de la parole, une mesure de confiance précise peut aider à déterminer si le résultat de reconnaissance est correct. L'estimation de confiance a été appliquée pour améliorer la reconnaissance en incorporant des informations supplémentaires dans le processus de reconnaissance. L'introduction de ces facteurs de confiance a contribué substantiellement à la réduction du taux d'erreur de reconnaissance. Cette performance est due au fait que lorsqu'une faible confiance est attribuée à un mot hypothèse, ce dernier est souvent un mot erronément reconnu selon l'information précédente, mais selon des informations supplémentaires, il peut être rejeté.

Gandrabur *et al.* (2003) ont introduit l'estimation de confiance dans une tâche de prédiction de traduction. Ils ont utilisé les réseaux de neurones pour estimer la probabilité conditionnelle d'exactitude $p(C = 1 \mid w_m, h, s)$ pour une prédiction w_m qui suit l'historique h dans la traduction d'une phrase source s. Ici aussi, on a observé un gain significatif quand on a utilisé l'estimation de confiance avec les modèles de traduction [24]. Récemment, plusieurs études ont été l'objet d'intégration des facteurs de confiance en traduction automatique [73] [10] [25].

Dans la RIT, on observe le même problème que dans la reconnaissance de la parole. En reconnaissance de parole, afin de gagner plus de robustesse, on utilise plusieurs systèmes de reconnaissance et on intègre les facteurs de confiance comme une mesure d'évaluation de reconnaissance à postériori. De même dans la RIT, la traduction de requête est effectuée avec plusieurs ressources. Nous avons alors le même problème que dans la reconnaissance de la parole : déterminer les traductions correctes des termes de requêtes parmi toutes les traductions candidates. De plus, les traductions suggérées par différentes ressources sont assignées des poids différents et souvent incompatibles.

Les modèles utilisés dans le processus de traduction sont incompatibles, pour plusieurs raisons.

D'abord, les modèles basés sur les dictionnaires bilingues diffèrent des modèles de traduction statistique entraînés sur des corpus parallèles.

Ensuite, des modèles de traduction statistique différents sont entraînés de différentes manières et probablement sur différents corpus. Par conséquent, ils produisent souvent différentes alternatives de traduction et différentes probabilités. Pour ces raisons, une combinaison naïve ne sera pas appropriée pour la RIT si nous nous fions à toutes les "hypothèses" de traduction d'un terme de la requête

sans une réévaluation et un filtrage supplémentaires. Dans ce contexte, les mesures de confiance peuvent être employées pour apprendre comment ajuster les scores originaux de traduction en observant leur performance sur de nouveaux textes. Ces estimés de confiance seront utilisés dans cette étude comme une mesure uniforme sur les traductions au lieu des probabilités originales. Concrètement, étant donné une traduction produite par n'importe quelle ressource, un modèle de traduction statistique ou un dictionnaire bilingue, on vise à mesurer la confiance que cette traduction soit correcte, selon des attributs informatifs.

Enfin, l'estimation de confiance peut être vue comme une technique de "rescoring" sur les traductions initiales. Elle nous aidera aussi à mieux filtrer les traductions appropriées et à ajuster les probabilités originales de traduction.

5.9.1 Définition

La confiance pour une traduction est définie par la probabilité postérieure que cette traduction est correcte $p(C = 1 \mid X)$, étant donné $X-$ le mot source, une traduction et un ensemble d'attributs associés. Pour obtenir la mesure de confiance pour chaque traduction candidate, on implémente un classificateur binaire qui prend en entrée un mot source, sa traduction et les attributs associés, et retourne en sortie un score estimant la probabilité d'exactitude de la traduction selon l'ensemble des attributs.

5.9.2 Apprentissage des facteurs de confiance

Dans la plupart des études précédentes sur la reconnaissance de la parole et la traduction automatique, les réseaux neuronaux ont été couramment utilisés pour produire les mesures de confiance. Même si d'autres techniques comme "naïve Bayes" ont été testées, elles n'ont pas été performantes comme les réseaux de neurones [10]. Les réseaux de neurones ont la capacité de manipuler des données d'entrée de différentes natures et sont bien adaptés pour des tâches de classification.

Notre tâche de classification d'une traduction sur deux classes (correct/incorrect) est similaire à la reconnaissance de parole ou la traduction automatique. Ainsi, on a opté pour l'utilisation d'un perceptron multicouche pour estimer la probabilité d'exactitude d'une traduction $p(C = 1 \mid X)$.

Nos données d'entraînement peuvent être vues comme un ensemble de paires (X, C), où X, un vecteur d'attributs relatif à une traduction[4], est utilisé comme entrée du réseau, et C est la sortie désirée du réseau (l'exactitude de la traduction, soit 0 ou 1). Le perceptron multicouche implémente une projection non linéaire des attributs d'entrée en combinant des couches de transformation linéaire et une fonction de transfert non linéaire [9]. Formellement, le perceptron multicouche implémente une fonction discriminante d'une entrée X de la forme :

$$g(X;\theta) = o(V \times h(W \times X)) \tag{5.20}$$

où $\theta = \{W, V\}$ est l'ensemble des poids optimisés durant l'étape d'apprentissage. W est une matrice de poids entre la couche d'entrée et la couche cachée et V est un vecteur de poids entre la couche cachée et la couche de sortie. h est une fonction d'activation non linéaire pour les unités cachées qui transforme la combinaison linéaire des entrées $W \times X$; o est également une fonction d'activation non linéaire pour l'unité de sortie, qui transforme la sortie du réseau multicouche à la probabilité d'exactitude $p(C = 1 \mid X)$. La fonction d'activation logistique $\frac{1}{(1+e^{-X})}$ est utilisée pour les unités cachées ainsi que pour l'unité de sortie. Sous ces conditions, notre réseau est entraîné pour minimiser une fonction objective du taux d'erreur (section 5.9.3). Pendant l'étape de test, la confiance d'une traduction X est estimée par la fonction discriminante $g(X;\theta)$ ci-dessus, i.e. $p(C = 1 \mid X) = g(X;\theta)$.

[4]On entend par la traduction, la paire: le mot source et sa traduction

5.9.3 La fonction objective à minimiser

Les données d'entraînement et de test sont des paires de phrases considérées comme des traductions mutuelles. La fonction objective vise à refléter la correspondance entre ces phrases. Une métrique traditionnelle pour évaluer ces estimés de probabilité est le logarithme de vraisemblance négatif (ou l'entropie croisée CE) attribué au corpus de test par le modèle et normalisé par le nombre d'exemples dans le corpus de test [10]. Cette métrique évalue les probabilités d'exactitude. Elle mesure l'entropie croisée entre la distribution empirique sur les deux classes (correct/incorrect) et la distribution du modèle de confiance à travers tous les exemples X^i dans le corpus. L'entropie croisée est définie comme suit :

$$CE = -\frac{1}{n}\sum_i \log p(C^i \mid X^i) \tag{5.21}$$

où C^i est égale à 1 si la traduction X^i est correcte et égale à 0 si elle est incorrecte. Pour enlever la dépendance à l'égard de la probabilité d'exactitude à priori, l'entropie croisée normalisée (NCE) [10] est utilisée :

$$NCE = \frac{(CE_b - CE)}{CE_b} \tag{5.22}$$

Le modèle de base CE_b est un modèle qui attribue des probabilités d'exactitude fixes basées sur les fréquences empiriques des deux classes :

$$CE_b = -\frac{n_0}{n}\log\frac{n_0}{n} - \frac{n_1}{n}\log\frac{n_1}{n} \tag{5.23}$$

où n_0 et n_1 sont les nombres de traductions correctes et incorrectes parmi les n exemples du corpus de test.

5.9.4 Attributs de confiance (features)

Le perceptron multicouche vise à extraire la relation entre l'exactitude de la traduction et les attributs, et sa performance dépend de l'identification d'attributs informatifs. Ces attributs sont utilisés ensemble pour estimer la valeur de la confiance. Dans notre travail, on a identifié intuitivement sept classes d'attributs qui sont présumés être informatifs sur l'exactitude d'une traduction.

L'index du modèle de traduction : L'index identifie la source de traduction. Dans notre cas, on utilise quatre modèles : un modèle de traduction statistique construit sur un ensemble de pages Web parallèles [32], un autre modèle de traduction statistique bâti sur le corpus anglais-arabe des nations unies [22], deux dictionnaire bilingues : Ajeeb [3] et Almisbar [4].

Les probabilités de traduction : La probabilité de traduire un mot source par un mot cible. Pour les modèles de traduction statistique, ces probabilités sont estimées avec le modèle IBM 1 [11] sur des corpus parallèles. Pour les dictionnaires bilingues, comme il n'y a aucune probabilité fournie avec ces dictionnaires, les poids sont souvent déterminés selon une distribution uniforme ou selon les occurrences et les cooccurrences des traductions dans un corpus. Plusieurs méthodes peuvent être utilisées pour attribuer une probabilité à chaque paire de traduction (e, a) dans le dictionnaire bilingue :

1. **Distribution uniforme :** Quand n alternatives de traduction sont connues pour un mot source, chacune des traductions est attribuée une probabilité de $1/n$.

2. **Utilisation d'un corpus monolingue :** Les traductions sont pondérées selon leur fréquence d'occurrence sur un corpus monolingue de grande taille. La collection de documents TREC sur laquelle on a fait nos expérimentations a été utilisée dans ce sens. Ainsi, à chaque traduction pour un mot source est attribuée une probabilité relative à sa fréquence d'occurrence sur la collection

TREC. Cette technique peut vraisemblablement introduire des erreurs dans le sens où les évènements rares seront mal estimés voire même assignés des probabilités nulles. Pour contourner ce dernier phénomène, les probabilités sont lissées par une technique de lissage comme celle de Laplace (i.e. la fréquence de chaque mot est augmentée de 1).

3. **Utilisation d'un corpus parallèle :** Les probabilités pour les paires de traduction du dictionnaire bilingue sont déterminées selon les cooccurrences des mots de ces paires dans un corpus parallèle aligné phrase par phrase. Un modèle d'alignement de mots comme le modèle IBM 1 de Brown [11] peut être entraîné. Bien sûr pour que la somme de probabilités de traductions d'un mot source dans le dictionnaire soit 1, une normalisation doit être faite. Les probabilités ainsi obtenues seront attribuées aux paires de traduction du dictionnaire bilingue.

4. **Le dictionnaire bilingue comme corpus parallèle :** Cette méthode considère le dictionnaire bilingue comme un corpus parallèle. Chaque entrée du dictionnaire bilingue est une paire de phrases alignées. Pour attribuer des probabilités aux paires de traduction du dictionnaire, il suffit d'entraîner un modèle standard d'alignement de mots sur le corpus. L'effet de cette estimation est d'attribuer une plus grande probabilité à une traduction qui apparaît plusieurs fois où un mot source apparaît.

Il est évident que les traductions proposées par une ressource pour un mot source n'ont pas le même poids d'importance. Ainsi, une méthode attribuant des probabilités selon une distribution uniforme n'est pas raisonnable. La méthode qui attribue des probabilités aux traductions selon leur fréquence d'occurrence sur un corpus monolingue n'est pas fiable dans la mesure où les traductions rares ou absentes dans le corpus sont mal estimées. La méthode qui utilise des corpus parallèles produit des probabilités plus raisonnables que les autres méthodes. Ceci dans le sens où l'attribution des probabilités aux traductions par cette méthode, est effectuée avec un modèle plus formel et plus puissant (le modèle IBM 1 [11]). Encore

plus, ces probabilités seront mieux estimées si on dispose d'un corpus parallèle de taille importante et de bonne qualité pour entraîner un modèle d'alignement de mots. Cette méthode est largement utilisée en traduction automatique [75].

En résumé, dans nos expérimentations, pour attribuer des probabilités aux traductions d'un dictionnaire bilingue, un modèle de traduction statistique IBM 1 est entraîné sur un corpus parallèle[5]. Ensuite, la probabilité produite par ce modèle de traduction statistique pour chaque paire de traduction (e, a) du dictionnaire bilingue est extraite. Enfin, cette probabilité est normalisée avec la méthode de lissage de Laplace [35]:

$$p_{DB}(a \mid e) = \frac{p_{MTS}(a \mid e) + 1}{\sum_{i=1}^{n}[p_{MTS}(a_i \mid e) + 1]} \tag{5.24}$$

Où n est le nombre de traductions proposées par le dictionnaire bilingue au mot e.

L'ordre de la traduction : Cette classe inclut deux attributs : Le rang de la traduction dans la liste des traductions proposées par chacune des quatre ressources et la différence de probabilité entre la traduction en question et la traduction dont la probabilité est la plus élevée. L'intuition derrière le deuxième attribut est que la traduction du premier rang (la traduction dont la probabilité est la plus élevée) est souvent correcte. Pour les autres traductions, plus la différence de probabilité avec la traduction du premier rang est petite, plus ces traductions sont correctes.

L'information sur la traduction inverse : Ceci inclut la probabilité de traduction d'un mot cible par un mot source et le rang du mot source dans la liste des traductions du mot cible. Ces attributs mesurent le degré de bijection entre le mot source et la traduction. C'est-à-dire, si y est une traduction de x et si on tente

[5]http://www.ldc.upenn.edu/
Arabic-English Parallel News Part 1 (LDC2004T18)
Arabic News Translation Text Part 1 (LDC2004T17)

de faire la traduction dans le sens inverse, à quel point y (le mot source) est une traduction spécifique de x. Un autre attribut identifie si le mot source est présent dans la liste des cinq meilleures traductions du mot cible dans la traduction inverse.

Le vote des ressources sur la traduction : Le but de cet attribut est de vérifier si la traduction est votée par plus d'une ressource. Plus une traduction est proposée par plusieurs ressources, plus il est probable qu'elle soit correcte.

Les attributs relatifs à la phrase source : Les attributs dans cette classe visent à extraire la relation de traduction entre les mots de la phrase source et la traduction en question. Un attribut mesure la fréquence du mot source dans la phrase source. Un autre attribut calcule le nombre de mots dans la phrase source qui sont reliés à la traduction en question. Un mot source est relié à un mot cible (traduction) si le mot source fait partie des meilleures traductions du mot cible dans la traduction inverse.

Les attributs de modèles de langue : On utilise les modèles de langue unigramme, bigramme et trigramme pour les mots sources et cibles sur les données d'entraînement.

Le tableau 5.3 dresse une liste exhaustive des attributs utilisés.

5.9.5 Expérimentation sur les facteurs de confiance

Corpus d'entraînement et annotation

La disponibilité d'une collection de données d'entraînement est nécessaire pour l'implémentation du modèle de confiance. Ces données doivent être différentes de celles employées pour entraîner les modèles de base. Le corpus LDC qui est un bitexte arabe-anglais de nouvelles de quelques agences de presse est un bon ensemble de données de référence pour l'entraînement du modèle de confiance. Le

Attribut	Description
index	l'index de la ressource de traduction
prob	la probabilité de traduction $p(a \mid e)$
rangTrad	le rang de la traduction dans la liste des traductions
diffProb	la différence de probabilité entre la traduction en question et la traduction dont la probabilité est la plus élevée
probTradInv	la probabilité de traduction d'un mot cible à un mot source $p(e \mid a)$
rangSourceInv	le rang du mot source dans la liste des traductions du mot cible dans la traduction inverse
sourcePresInv	présence/absence du mot source dans la liste des meilleures traductions du mot cible dans la traduction inverse
vote	le nombre de ressources qui proposent la traduction
freqSource	la fréquence du mot source dans la phrase source
nbSourceTrad	le nombre de mots dans la phrase source qui sont reliés à la traduction en question
source1gr	la probabilité du mot source dans le corpus $p(e_i)$
source2gr	la probabilité du bigramme source (le mot source et son prédécesseur dans la phrase source) dans le corpus $p(e_i \mid e_{i-1})$
source3gr	la probabilité du trigramme source (le mot source et ses deux prédécesseurs dans la phrase source) dans le corpus $p(e_i \mid e_{i-2}, e_{i-1})$
cible1gr	la probabilité de la traduction dans le corpus $p(a_i)$
cible2gr	la probabilité du bigramme cible (la traduction et le mot qui la précède) dans le corpus $p(a_i \mid a_{i-1})$
cible3gr	la probabilité du trigramme cible (la traduction et les deux mots qui la précèdent) dans le corpus $p(a_i \mid a_{i-2}, a_{i-1})$

Tableau 5.3: La liste complète des attributs utilisés

tableau 5.4 décrit quelques caractéristiques de ce corpus:

Langues du corpus	arabe et anglais
Nombre de mots anglais	3 040 304
Nombre de mots arabes	2 470 086
Nombre de paires de phrases	83 571
Source des documents	Xinhua, AFP, An Nahar, Ummah press service

Tableau 5.4: Caractéristiques du corpus parallèle anglais-arabe de LDC

Ce corpus consiste en 83 K paires de phrases alignées. Les phrases sources (en anglais) sont traduites mot par mot vers l'arabe en utilisant les modèles de traduction de base (deux MTSs et deux DBs). Chaque mot source est traduit par les traductions les plus probables[6] en utilisant les MTSs, et par les cinq meilleures traductions en utilisant les DBs. Ces traductions sont alors comparées à la phrase de référence pour construire un corpus étiqueté. Cependant, les relations de traduction exactes entre les mots dans les phrases correspondantes (source et cible) ne sont pas connues. Autrement dit, on ne sait pas exactement quel(s) mot(s) cible(s) traduit(traduisent) un mot source. En l'absence de cette information, deux approches sont proposées et testées pour la construction d'un corpus étiqueté de façon approximative :

1. **PER :** Cette approche est inspirée des métriques d'évaluation en traduction automatique. Une traduction d'un mot de la phrase source est considérée comme une traduction correcte si elle est présente dans la phrase de référence. L'ordre des mots est ignoré, mais le nombre d'occurrences est pris en considération. Cette métrique s'accommode bien avec notre contexte de recherche d'information [34] : Les modèles de recherche d'information sont basés sur le principe de "sac de mots" et l'ordre des mots n'est pas considéré.

2. **Alignement IBM 1 :** Comme son nom l'indique, cette approche utilise le résultat d'alignement de mots issu de l'entraînement du modèle IBM 1

[6]Les traductions dont la probabilité $p(a \mid e) \geq 0.1$

sur le corpus parallèle. Chaque traduction d'un mot de la phrase source est considérée comme une traduction correcte si elle est équivalente à la meilleure[7] traduction proposée par le modèle IBM 1 pour ce mot source.

Pour illustrer le fonctionnement des deux méthodes d'annotation, nous proposons les deux phrases alignées du corpus d'entraînement (tableau 5.5) :

Swiss Judge Demands Pinochet's Extradition to Switzerland.
قاض سويسري يطالب بتسليم بينوشيه الى سويسرا .

Tableau 5.5: Un exemple de phrases alignées anglais-arabe du corpus LDC

Chaque mot source de la phrase source est traduit en arabe en utilisant toutes les ressources. Par exemple, le dictionnaire bilingue Ajeeb propose les traductions {قاض (juge), حاكم (gouverneur)} pour le mot source "judge". En utilisant la méthode PER, chaque traduction est vérifiée si elle est présente dans la phrase cible. L'ordre de la traduction dans la phrase cible est ignoré mais son nombre d'occurrences est pris en considération. Ainsi, avec cette méthode la première traduction que propose Ajeeb est annotée correcte (1) parce qu'elle est présente dans la phrase cible. Mais, la deuxième traduction est annotée erronée (0) parce qu'elle est absente dans la phrase cible. La deuxième méthode (Alignement IBM 1) compare chaque traduction à la traduction avec la probabilité la plus élevée que propose le modèle IBM 1, entraîné sur le corpus d'entraînement dont la paire de phrase fait partie, pour le mot source "judge". Pour ce mot "judge", la traduction avec la probabilité la plus élevée proposée par le modèle IBM 1 est قاض (juge).

[7]la traduction avec la probabilité la plus élevée

Ainsi, la première traduction قاض proposée par Ajeeb est annotée correcte (1) mais la deuxième est annotée erronée (0).

Par l'application d'une des deux méthodes d'annotation, on s'attend à obtenir des données d'entraînement pour le réseau de neurones. La première approche ne cherche la validité d'une traduction que dans le contexte d'une paire de phrases, contrairement à la deuxième méthode, où la validité d'une traduction est perçue dans tout le contexte du corpus parallèle. C'est-à-dire, la deuxième approche considère une traduction correcte celle qui est la plus souvent alignée à un mot source dans tout le corpus des paires de phrases, tandis que dans la première approche, une traduction est considérée correcte si elle est "alignée" à un mot source juste dans la paire de phrases en question. Pour la deuxième méthode, d'une part, le fait de ne considérer que la meilleure traduction comme traduction de référence, avantage certaines traductions par rapport aux autres. D'autre part, si le nombre de traductions de référence est augmenté, on s'attend à plus de bruit. En résumé, chaque stratégie a ses avantages et ses inconvénients. Après leur implémentation sur le corpus parallèle LDC, les résultats d'étiquetage de traductions des deux méthodes se sont avérés raisonnables et plus ou moins similaires.

Impact des couches cachées

Habituellement, le nombre d'unités cachées dans un perceptron multicouche a un impact sur sa performance. Dans ce sens, une série d'expérimentations est effectuée en utilisant différents nombres d'unités cachées. Le tableau 5.6 montre les performances de ces diverses architectures mesurées par la métrique d'évaluation (NCE) sur le corpus de test. L'entropie croisée normalisée (NCE) (section 5.9.3) mesure la baisse relative dans le logarithme de vraisemblance négatif comparé au modèle de base qui dépend des probabilités d'exactitude à priori. Plus l'entropie croisée normalisée est élevée, meilleure est la performance. Toutes ces expérimen-

tations ont été menées avec la librairie d'apprentissage machine Plearn[8].

Nombre d'unités cachées	NCE
5	62.58
10	62.56
20	62.58
50	62.64
100	62.41

Tableau 5.6: Résultats de différents perceptrons multicouches

Le tableau 5.6 montre l'amélioration dans l'entropie croisée comparée au modèle de base décrit dans la section 5.9.3. On rappelle que plus NCE est élevée, plus le modèle entraîné est considéré performant. Ainsi, les résultats de ce tableau révèlent une nette amélioration de tous les perceptrons par rapport au modèle de base mais la différence entre les différents perceptrons n'est pas très significative. Le perceptron multicouche avec 50 unités cachées présente néanmoins la meilleure performance. C'est ce nombre que nous allons utiliser dans la suite de nos expérimentations.

Impact des attributs

Pour mesurer la performance individuelle des différents attributs sur le corpus de test, deux manières peuvent être utilisées : prendre en considération tous les attributs sauf l'attribut en question ou utiliser exclusivement chacun des attributs. Dans nos expérimentations, chaque classe d'attributs est utilisée exclusivement et son comportement est évalué en fonction de l'entropie croisée normalisée. Le tableau 5.7 montre les résultats:

A partir de ce tableau, on peut facilement remarquer que le meilleur attribut est le vote des ressources sur la traduction. Ceci démontre qu'il est très utile et bénéfique d'utiliser des évidences multiples dans la traduction des requêtes : plus

[8]http://plearn.berlios.de/

Attributs	NCE
Attributs relatifs à la phrase source	32.20
Index du modèle de traduction	33.03
Information sur la traduction inverse	36.93
Ordre de la traduction	38.28
Probabilités de traduction	44.46
Attributs de modèles de langue	50.89
Vote des ressources sur la traduction	57.63
Tous les attributs	62.56

Tableau 5.7: Les performances des différents attributs

des ressources ou évidences votent en faveur d'un candidat, plus ce candidat aura la chance d'être correcte. Les autres attributs très utiles sont les attributs de modèles de langue et les probabilités de traduction. Les modèles de langue sont informatifs dans la mesure où une traduction dans un bigramme ou un trigramme très fréquent dans un corpus aura plus de chance d'être correcte qu'une traduction faisant partie d'un bigramme ou un trigramme rare dans le corpus. Les probabilités de traduction, calculées par le modèle formel (IBM 1) d'une façon efficace et exacte, donnent aussi une information significative sur les traductions. Par contre, l'ordre de la traduction, l'information sur la traduction inverse, l'index du modèle de traduction et les attributs relatifs à la phrase source fournissent quelques informations marginalement utiles.

5.10 Expérimentation et évaluation en RIT

5.10.1 Description du corpus de test

Afin d'évaluer les différentes méthodes de traduction de requêtes: modèles de traduction statistique, dictionnaires bilingues, combinaison linéaire et facteurs de confiance, on utilise des requêtes en anglais pour rechercher des documents en arabe. Pour la collection des documents, la même collection arabe de TREC que nous avons évaluée pour nos expériences en monolingue arabe est employée. Pour

les requêtes, deux collections sont évaluées : la première collection TREC2001 contient 25 requêtes et la deuxième (TREC2002) englobe 50 requêtes. Les deux collections sont ensuite fusionnées pour constituer une collection plus large (75 requêtes). On rappelle que les caractéristiques de ces collections (documents et requêtes) sont déjà présentées au chapitre précédent. Chaque requête est constituée de trois parties : le titre, la description et le récit (narrative). Dans nos expériences, nous avons utilisé seulement les deux parties (titre et description) i.e. des requêtes de type TD. Le tableau 5.8 présente un exemple de requêtes. La liste des requêtes de la collection TREC2001 est présentée en annexe IV.

<top>
<num> Number: AR1
<title> Performing Arts and Islamic Institutions in the Arab World.
<desc> Description:
What is the impact of Islamic Institutions on the performing arts such as dance and music in the Arab World?
<narr> Narrative:
Articles concerning sports or spatial arts, performance arts outside of the Arab World, religious behavior not related to the performing arts, or debts and banking loans, are not relevant to this topic.
</top>

Tableau 5.8: Exemple d'une requête en anglais de la collection TREC

Les documents en arabe sont traités comme dans le cas de recherche monolingue (chapitre 4). Dans les sections suivantes, nous allons tester différentes manières d'attribuer des poids aux traductions candidates : les probabilités originales de traduction attribuées par chaque ressource, la combinaison linéaire et les mesures de confiance.

5.10.2 Utilisation des modèles séparés

Avant la combinaison des ressources, nous avons évalué chacune des quatre ressources séparément. Dans ce cas, nous assignons les probabilités de traduction du modèle IBM 1 à nos traductions comme poids. Pour chaque terme de la requête, nous prenons seulement les traductions dont la probabilité $p(a \mid e) \geq 0.1$ quand on utilise les modèles de traduction statistique et les cinq meilleures traductions quand on utilise les dictionnaires bilingues. Les raisons de ces choix sont les suivantes : Pour les modèles de traduction statistique, les traductions avec une probabilité $p(a \mid e) < 0.1$ correspondent souvent aux bruits. Pour les dictionnaires bilingues, on a sélectionné les cinq premières traductions parce qu'au delà de cinq traductions, on ne connaît pas un gain significatif en précision moyenne de recherche.

5.10.3 Combinaison linéaire

Dans cette combinaison linéaire, nous utilisons quatre ressources de traduction : Un modèle de traduction statistique entraîné sur les pages Web, un autre modèle de traduction statistique entraîné sur le corpus des nations unies, et deux dictionnaires bilingues (Ajeeb et Almisbar). Le corpus utilisé pour déterminer les meilleurs paramètres de la combinaison linéaire est le même corpus parallèle arabe-anglais de LDC qu'on a utilisé pour entraîner le modèle de confiance. Le processus d'optimisation des paramètres est décrit dans la section 5.8. Les paramètres optimisés assignés à chaque ressource de traduction sont décrits dans le tableau 5.9.

Les résultats obtenus dans le tableau 5.9 montrent que les MTSs sont attribués des coefficients plus élevés que ceux des dictionnaires bilingues. La raison principale de ces résultats est que l'algorithme EM pénalise les modèles qui assignent des probabilités nulles aux mots des textes cibles. Malheureusement à la différence des modèles de traduction statistique, les dictionnaires bilingues attribuent des probabilités nulles aux mots cibles qui ne sont pas forcément des traductions. Par conséquent cette combinaison va avantager les modèles de traduction statistique sur les dictionnaires bilingues bien que les traductions que proposent les modèles

Le modèle de traduction	Le paramètre optimisé accordé
MTS entraîné sur les pages Webs (MTS-Web)	0.3115
MTS entraîné sur le corpus des nations unies (MTS-UN)	0.3235
Dictionnaire bilingue Ajeeb	0.1436
Dictionnaire bilingue Almisbar	0.2213

Tableau 5.9: Les valeurs optimisées pour les paramètres reliés à chaque ressource de traduction

de traduction statistique ne sont pas toujours meilleures. Nous pouvons remarquer ce cas de figure dans l'exemple du tableau 5.11. La traduction تدابير (mesures) du mot anglais "measures", proposée par le modèle de traduction statistique entraîné sur le corpus des NU, prend un poids fort (0.61) avec la combinaison linéaire comparativement au poids (0.029) de la meilleure traduction إجراءات (mesures) (aussi correcte) proposée par le dictionnaire bilingue Almisbar.

Nous rappelons que le poids associé à chaque traduction en utilisant la combinaison linéaire est calculé avec la formule 5.15 (section 5.7). A noter aussi qu'après la combinaison, si un terme de requête est traduit avec plusieurs alternatives de traduction, nous gardons au plus quatre d'entre elles. Cette sélection nous a donnée la meilleure performance.

Le tableau 5.10 et la figure 5.1 montrent les performance de la RIT en termes de précision moyenne (Mean Average Precision : MAP) utilisant les quatre ressources séparément et combiné linéairement. Les nombres entre parenthèses dans le tableau 5.10 représentent les pourcentages de la performance de la RIT par rapport à la précision moyenne (MAP) de la RI monolingue. Nous observons que la performance est tout à fait différente d'un modèle à l'autre. Le score faible enregistré par le MTS entraîné sur les pages Web est dû à la petite taille des données sur lequel le modèle est entraîné. Le MTS-Web utilise seulement 2 816

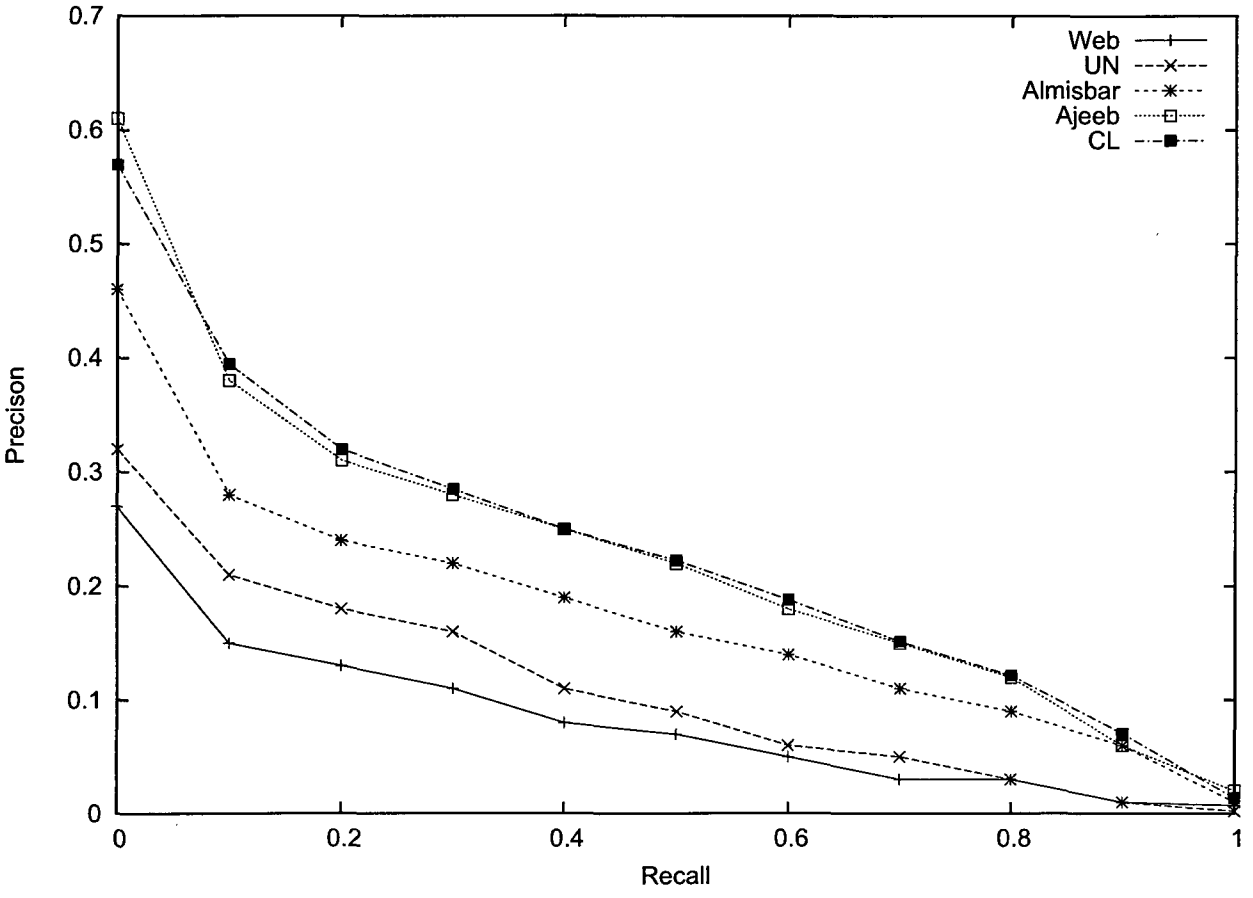

Figure 5.1: Les courbes rappel-précision de chaque modèle séparé et de la combinaison linéaire (TREC2001-2002)

Le modèle de traduction	TREC2001	TREC2002	TREC2001-2002
Recherche d'information monolingue	0.33	0.28	0.31
MTS-Web	0.14 (42%)	0.04 (17%)	0.07 (25%)
MTS-UN	0.11 (33%)	0.09 (34%)	0.10 (33%)
DB Ajeeb	0.27 (81%)	0.19 (70%)	0.22 (70%)
DB Almisbar	0.17 (51%)	0.16 (58%)	0.16 (54%)
Combinaison linéaire	0.24 (72%)	0.20 (71%)	0.21 (67%)

Tableau 5.10: Les performances (MAP) de chaque modèle séparé et de la combinaison linéaire

paires de documents anglais-arabe, ce qui n'est pas suffisant pour construire un MTS raisonnable. L'autre MTS entraîné sur un corpus parallèle plus volumineux produit des résultats légèrement meilleurs. Cependant, les documents du corpus des NU ne sont pas dans le même domaine que les documents recherchés. Ainsi, les traductions peuvent être non-appropriées. Ceci ne fait que confirmer que la qualité d'un modèle de traduction statistique dépend aussi de la période, de la thématique, de la source linguistique du corpus parallèle sur lequel le modèle de traduction a été entraîné.

Ici, les dictionnaires bilingues présentent des performances meilleures que celles des MTSs parce qu'ils fournissent multiples et "bonnes" traductions à chaque terme de la requête. Cependant, les résultats d'Almisbar ne sont pas aussi bons que ceux d'Ajeeb parce que beaucoup de termes de requêtes ne sont pas couverts par ce dictionnaire.

Lors de la combinaison de ces ressources, la performance est censée être meilleure qu'avec n'importe quelle ressource individuelle parce que tous les termes des requêtes peuvent être traduits correctement par au moins une ressource. Les résultats dans le tableau 5.10 ne confirment pas cette hypothèse, plus particulièrement pour la collection TREC2001, où une ressource individuelle (le dictionnaire bilingue Ajeeb) se comporte mieux que la combinaison linéaire. Ceci est dû au faible co-

efficient de confiance attribué à cette ressource lors de la combinaison selon notre méthode d'estimation de paramètre. Voici un exemple des requêtes anglaises de la collection TREC2001 : "What measures are being taken to develop tourism in Cairo?". La traduction arabe manuelle fournie par TREC au mot "measures" est : "إجراءات" (mesures). Le tableau 5.11 montre les différentes traductions fournies par les quatre méthodes à ce mot. Cette traduction est incluse dans celles du dictionnaire bilingue Almisbar et du MTS-UN.

Le modèle de traduction	Traduction(s) du mot "measures"
DB Ajeeb	تدبير 0.05 (mesure), عيار 0.05 (calibre), قياس 0.05 (mesure), مقياس 0.05 (mesure), معيار 0.05 (critère), مكيال 0.05 (mesure), ميزان 0.05 (balance)
DB Almisbar	إجراءات 0.05 (mesures), مقياس 0.03, قدر 0.03 (mesure), مقدار 0.03 (montant)
MTS-Web	تدابير 0.69 (mesures)
MTS-NU	إجراءات 0.09 (cette traduction n'est pas prise en considération parce que $p(a \mid e) < 0.1$)
Combinaison linéaire	تدابير 0.61, مقياس 0.037, إجراءات 0.029, قياس 0.020

Tableau 5.11: Traduction du mot anglais "measures" avec les différentes méthodes

Nous voyons clairement que les traductions avec les différentes ressources ne sont pas identiques. Quelques ressources proposent des traductions inappropriées telles que "ميزان" (balance) ou "مكيال" (mesure). Même si deux ressources suggèrent les mêmes traductions, les poids sont différents. Pour cette requête, la combinaison linéaire produit des traductions meilleures que n'importe quelle ressource prise séparément : Les traductions les plus probables sont filtrées à partir de la liste combinée. Cependant, cette méthode ne peut pas attribuer un poids approprié à

la meilleure traduction "إجراءات"; elle est sélectionnée mais classée juste en troisième position avec un poids faible. Cet exemple montre les limites de la combinaison linéaire des ressources : Même si elle est capable de sélectionner les traductions les plus probables elle ne peut pas attribuer des poids appropriés à ces traductions.

On peut aussi se questionner sur la façon que nous avons utilisée pour déterminer les paramètres : on tente de maximiser la log-vraisemblance des données alignées. Cette fonction objective peut diverger de la qualité de traduction : une meilleure traduction ne produit pas toujours une meilleure log-vraisemblance et vice versa. Ainsi, il est possible qu'avec une autre façon de déterminer les paramètres, on pourrait obtenir une meilleure performance avec la combinaison que les ressources séparées. Nous laissons ceci à une étude future. Malgré ce fait, les performances que nous avons obtenues sont bien représentatives de la combinaison linéaire.

5.10.4 Facteurs de confiance

Dans ces expériences, nous utilisons les valeurs estimées pour les facteurs de confiance comme poids pour les traductions au lieu des probabilités originales. Ainsi, le poids associé à chaque traduction est recalculé avec la formule 5.16 de la section 5.7. Selon ces mesures de confiance, nous choisissons les quatre traductions avec les meilleurs estimés de confiance pour chaque terme de requête (comme dans la combinaison linéaire). Le tableau 5.12 montre les résultats :

Collection	TREC2001	TREC2002	TREC2001-2002
MAP de Combinaison Linéaire (CL)	0.2426	0.2032	0.2163
MAP des Mesures de Confiance (MC)	0.2775	0.2052	0.2290
% d'amélioration des MC par rapport à la CL	14.35 %	1.0 %	5.17 %

Tableau 5.12: Comparaison de performance de la RIT entre la CL et les MC

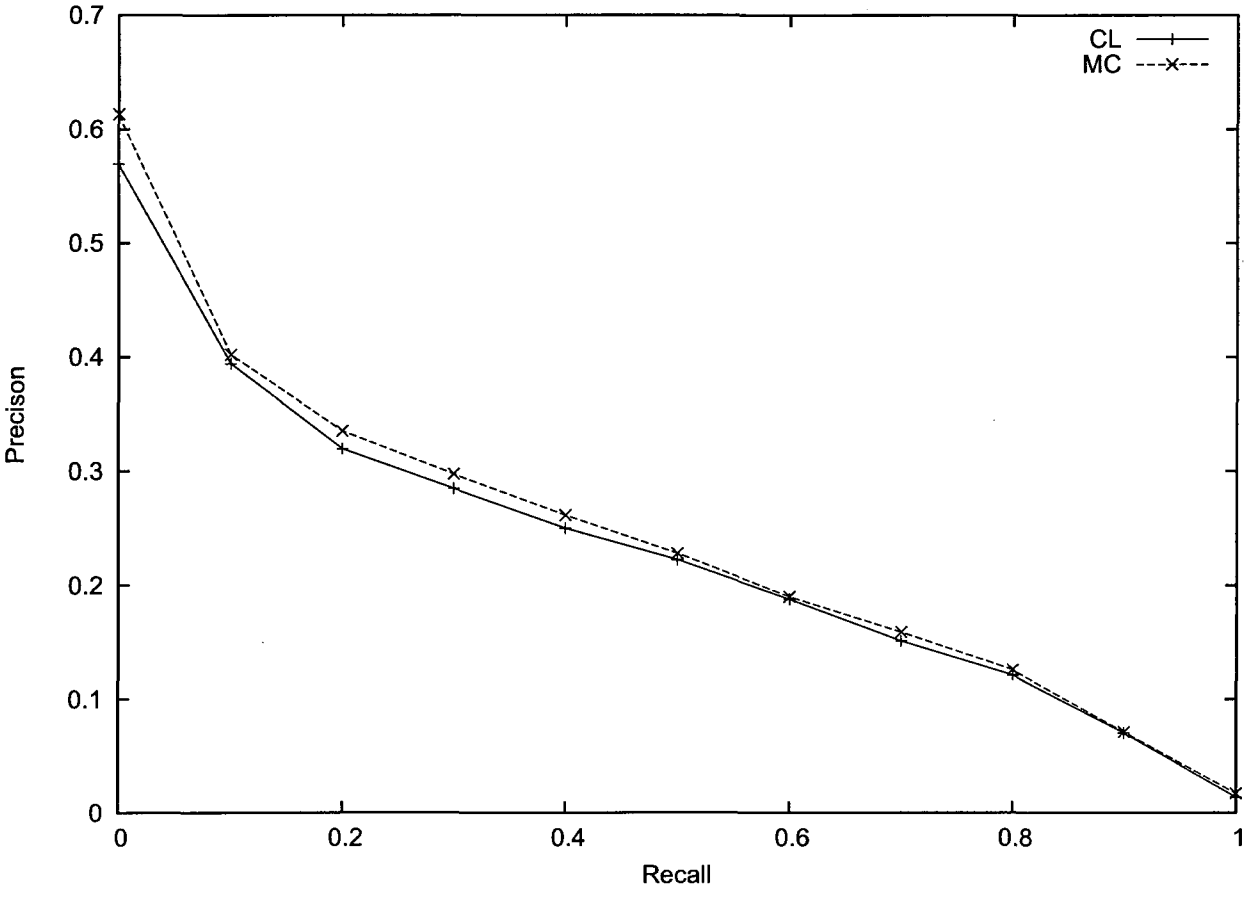

Figure 5.2: Les courbes rappel-précision de la combinaison linéaire et des mesures de confiance (TREC2001-2002)

Précision	TREC2001		TREC2002		TREC2001-2002	
	CL	MC	CL	MC	CL	MC
à 5 Docs	0.47	0.51	0.32	0.33	0.37	0.39
à 10 Docs	0.45	0.48	0.31	0.31	0.35	0.36
à 15 Docs	0.40	0.47	0.27	0.29	0.32	0.34
à 20 Docs	0.41	0.45	0.26	0.28	0.31	0.33
à 30 Docs	0.38	0.41	0.25	0.26	0.29	0.31
à 100 Docs	0.30	0.33	0.20	0.21	0.24	0.25
à 200 Docs	0.25	0.26	0.17	0.18	0.19	0.20
à 500 Docs	0.16	0.15	0.11	0.11	0.13	0.12
à 1000 Docs	0.10	0.09	0.07	0.07	0.08	0.08

Tableau 5.13: La précision à n documents retrouvés avec la combinaison linéaire et les mesures de confiance

5.10.5 Analyse des résultats

En terme de précision moyenne (MAP), nous voyons que les résultats de la méthode employant les mesures de confiance sont meilleurs que ceux obtenus avec la combinaison linéaire sur les deux collections particulièrement aux bas niveaux du rappel (tableau 5.13) (figure 5.2). Sur la collection fusionnée TREC2001-2002, le t-test prouve que l'amélioration apportée par les mesures de confiance par rapport à la combinaison linéaire est statistiquement significative au niveau $P < 0.05$. Cette amélioration de la performance de la RIT est attribuée à la capacité de la méthode des mesures de confiance de réévaluer correctement le poids de chaque traduction candidate. La liste des traductions sélectionnées avec leurs probabilités est plus raisonnable que celle de la combinaison linéaire. A titre d'exemple, nous avons obtenu une grande amélioration en précision moyenne pour la requête de la collection TREC2001 "What measures are being taken to develop tourism in Cairo? ", quand elle est traduite avec la méthode utilisant les mesures de confiance. Le terme de la requête "measures" est traduit par les deux méthodes dans le tableau 5.14.

Dans cet exemple, la mesure de confiance a pu augmenter le poids de la tra-

Le modèle de traduction	Traduction(s) du mot "measures"
Combinaison linéaire	0.020 قياس, 0.029 إجراءات, 0.037 مقياس, 0.61 تدابير
Mesures de confiance	0.06 قياس, 0.10 قدر, 0.51 إجراءات

Tableau 5.14: Traduction du mot anglais "measures" avec la combinaison linéaire et les mesures de confiance

duction correcte "إجراءات" à un niveau plus haut que les autres traductions incorrectes ou inappropriées. Cet exemple montre l'avantage potentiel des mesures de confiance sur la combinaison linéaire : La combinaison linéaire suppose que toutes les traductions candidates proposées par les différentes ressources sont correctes et les regroupe simplement ensemble sans réévaluation à postériori. Au contraire, le modèle de confiance ne fait pas confiance aveugle à toutes les traductions. Il examine leur validité selon de nouvelles informations. Ainsi, les traductions candidates sont réévaluées et filtrées selon des informations supplémentaires.

Dans quelques exemples, même si la méthode des mesures de confiance propose les mêmes traductions que la combinaison linéaire, les poids sont réajustés. Dans le tableau 5.15, nous montrons un exemple où les deux méthodes proposent la bonne traduction mais avec des poids différents. L'exemple présente la traduction du mot "refugee" dans la requête "Refugees in Zaire" de la collection TREC2002.

Le modèle de traduction	Traduction(s) du mot "refugee"
Combinaison linéaire	لاجىء 0.43 (réfugié), لاجا 0.36 (réfugié mal écrit), اجا 0.052 (-), مشرد 0.037 (sans abri)
Mesures de confiance	لاجىء 0.95, شريد 0.015 (vagabond), مهجر 0.011 (exilé,déporté), مشرد 0.005

Tableau 5.15: Traduction du mot anglais "refugee" avec la combinaison linéaire et les mesures de confiance

Pour cette requête, les deux méthodes réussissent à identifier la meilleure tra-

duction "لاجىء" (réfugié) au premier rang, mais avec des probabilités différentes. La méthode des mesures de confiance donne un poids nettement plus élevé à la meilleure traduction comparativement à la combinaison linéaire. Cette méthode est aussi capable d'une part, d'éliminer des traductions bruitées comme "لاجا" (réfugié mal écrit) et d'autres erronées comme "اجا", et d'autre part, de proposer d'autres traductions proches dans le sens comme "شريد" (vagabond) et "مهجر" (exilé, déporté) mais avec des poids faibles.

Nous ne prétendons pas que les mesures de confiance peuvent attribuer des poids d'importance précis pour toutes les traductions, mais les traductions les plus vraisemblables sont validées et valorisées avec des poids plus élevés. Ceci influence fortement sur l'efficacité de la RIT. Par conséquent, la mesure de confiance propose un mécanisme prometteur pour filtrer les traductions les plus appropriées d'une requête.

La fonction objective avec laquelle on a optimisé les facteurs de confiance peut aussi être discutée. Rappelons que pour estimer la probabilité d'exactitude d'une traduction, on a tenté de minimiser le logarithme de vraisemblance négatif (l'entropie croisée) d'un corpus de test. Cette fonction objective est plus appropriée pour des tâches de classification qu'à des tâches d'estimation de vraies probabilités comme la nôtre. Ainsi, il est possible qu'avec d'autres fonctions d'optimisation des paramètres, on pourrait obtenir une meilleure estimation des probabilités d'exactitude de traductions.

Un autre problème qui limite la performance de la traduction des requêtes est le manque de couverture des ressources de traduction pour certains termes de requêtes. Ces termes représentent généralement des entités nommées. La traduction des entités nommées n'a pas été étudiée dans cette thèse. Malheureusement, si on tente de traduire de telles entités nommées par une traduction générique basée

sur les dictionnaires bilingues ou des modèles statistiques, la qualité de traduction souffrira. Nous pouvons remarquer ce cas dans une des requêtes TREC "Caspian Beluga Conservation". Même si une ressource propose une traduction pour l'entité nommée "Beluga", cette traduction "بلوجا" (beluga mal translitéré) est erronée. Ainsi, une méthode de traduction spécifique pour ces entités doit être développée.

5.11 Conclusion

Dans ce chapitre, nous avons étudié la question de traduction des requêtes pour répondre à la problématique de la RIT anglais-arabe. Le problème majeur dans cette partie était l'identification et la collection des ressources de traduction pour cette paire de langues : une tâche non triviale. Après la collection des ressources, nous avons défini une méthodologie pour répondre à la problématique de combinaison des ressources de traduction basée sur les facteurs de confiance.

L'idée derrière la traduction des requêtes avec des ressources multiples est de renforcer la traduction. Cette combinaison des ressources est basée sur le principe d'évidence multiple. Une telle combinaison produit toujours de meilleurs résultats si la combinaison est faite correctement. Il y a eu peu d'études sur la manière d'effectuer cette combinaison. La plupart des études précédentes dans la RIT ont utilisé une combinaison linéaire des ressources. Cette méthode ne peut pas combiner correctement des ressources non homogènes telles que les dictionnaires bilingues et les modèles de traduction statistique. Cette combinaison naïve peut maintenir du bruit engendré par les différentes ressources, ou attribuer des poids

incorrects aux traductions.

Dans ce chapitre, nous avons proposé d'utiliser les mesures de confiance pour combiner les ressources pour la tâche de la traduction des requêtes [35]. Cette méthode réévalue chaque traduction candidate proposée par les différentes ressources en introduisant des attributs additionnels. Elle est capable de revaloriser une traduction candidate plus radicalement que dans la combinaison linéaire. Nos expérimentations ont montré des résultats très encourageants. Nous avons obtenu une amélioration moyenne de 5.87 % comparativement à la combinaison linéaire.

Cependant, cette approche peut être encore améliorée sur plusieurs aspects. Par exemple, nous pouvons optimiser cette technique en identifiant d'autres attributs informatifs. Evidemment le réflexe nous conduira vers les attributs sémantiques qui peuvent apporter des informations nouvelles. En l'absence des ressources comme WordNet[9] pour l'arabe, à l'heure actuelle, nous n'avons pas pu introduire de telles informations lors du calcul d'estimés de confiance. Rappelons qu'un thésaurus bilingue peut être construit par la traduction d'un thésaurus monolingue déjà existant comme WordNet. Dans ce type de thésaurus, chaque concept est traduit par les termes correspondants dans le langage cible, comme dans EuroWordnet.

Plus particulièrement, l'extraction d'informations sémantiques dans le cadre de la RIT signifie aussi la quantification du degré de similarité sémantique entre un mot source et sa traduction. Pour ce faire, la similarité sémantique peut être calculée avec des mesures comme la distance WordNet. Dans notre contexte de recherche d'information où un mot peut être traduit avec des synonymes ou des mots reliés, plusieurs techniques de calcul de distance WordNet peuvent être testées [12]. On peut citer la mesure de Hirst-St-Onge où la similarité est calculée entre mots par le poids du chemin le plus court qui mène d'un terme à un autre [29]. L'idée est

[9]http://wordnet.princeton.edu/

que deux concepts sont proches sémantiquement si leurs synsets sont connectés par un chemin qui n'est pas très long et qui ne change pas souvent de direction. Nous pouvons aussi utiliser la méthode conçue par Banerjee et Pedersen où la similarité entre deux mots est représentée par le degré de recouvrement entre les mots utilisés dans leurs définitions (glosses) [8]. L'intuition derrière cette mesure est que deux mots ayant un même sens sont définis en utilisant les mêmes mots.

Nous considérons que l'identification de telles informations sémantiques peut ajouter des informations nouvelles aux traductions et leur association aux autres attributs permet de mieux calculer les poids de confiance associés aux traductions. Enfin, La méthode des facteurs de confiance pour combiner les ressources de traduction peut être appliquée à d'autres paires de langues et révéler de bonnes performances comme pour le cas de l'anglais-français [36].

CHAPITRE 6

CONCLUSION

Dans cette thèse, nous nous sommes intéressés à la recherche d'information monolingue et translinguistique sur les documents en arabe. La RIT s'intéresse à la recherche des documents écrits dans une langue différente de celle des requêtes. Cette recherche est de plus en plus en demande sur le Web. Dans notre cas, la collection des documents est en arabe et les requêtes sont en anglais. Avant de traiter la RIT avec l'arabe, nous avons étudié la RI arabe monolingue où les documents et les requêtes sont écrits en arabe. L'idée sous-jaccente est d'opérer des traitements morphologiques efficaces spécifiques à l'arabe relatifs aux choix des termes d'index à la RI monolingue, avant de franchir la barrière linguistique. Force est de noter que durant cette étude, le problème majeur rencontré est le manque de ressources linguistiques pour cette langue. En effet, notre premier défi était la création de certaines ressources relatives et nécessaires pour les traitements de la RI arabe monolingue et de la RIT anglais-arabe.

6.1 RI arabe monolingue

La langue arabe possède une morphologie très riche et complexe qu'elle a suscité de nombreux défis au traitement automatique des langages naturels. Le défi principal soulevé en recherche d'information est le traitement morphologique qui vise à déterminer une forme appropriée d'index aux mots. Dans ce contexte, la problématique de lemmatisation des mots arabes a fait l'objet d'étude pour plusieurs travaux sur la recherche d'information arabe. Dans ces études, on distingue trois approches majeures. Une première approche utilise les racines trigrammes pour indexer les mots arabes. Une deuxième approche de lemmatisation indexe un mot arabe par plusieurs de ses n-grammes possibles. La dernière approche, inspirée par le processus de lemmatisation de l'anglais, exerce une troncature assouplie sur les

mots arabes.

Dans cette étude, nous avons proposé deux méthodes de lemmatisation pour les mots arabes [33]. Une première méthode assouplie, opère quelques légères troncatures sur un mot aux deux extrémités. Le choix des affixes de mots à tronquer est fait selon des statistiques de corpus ainsi que leur rôle syntaxique. Cette première méthode correspond à celles qu'on trouve dans les études antérieures. L'autre méthode, nouvelle et linguistiquement motivée, essaye de déterminer le noyau d'un mot selon des règles linguistiques appuyées par des statistiques de corpus. Cette deuxième approche opère plusieurs décompositions possibles sur un mot, produit d'abord un ensemble de lemmes candidats, et ensuite, grâce aux statistiques de corpus, elle choisit le lemme le plus utilisé dans le corpus.

Les deux méthodes sont testées et comparées sur une collection de volume important (TREC) de documents arabes associée avec deux ensembles de requêtes. Les résultats des expérimentations ont montré que la nouvelle méthode à base linguistique présente une meilleure performance de recherche que la technique assouplie.

6.2 RIT avec l'arabe

En recherche d'information translinguistique, le défi majeur est la traduction de la requête vers la langue des documents. Les principales approches de la traduction des requêtes utilisent respectivement la traduction automatique, les dictionnaires bilingues et les corpus parallèles. Nous avons vu dans le chapitre 5 que la traduction automatique a des faiblesses face à la traduction des requêtes parce que les requêtes sont rarement des phrases et plus souvent juste une séquence de mots sans règles grammaticales. Les techniques basées sur les dictionnaires bilingues et les corpus parallèles sont souvent utilisées comme solution de rechange pour la traduction des requêtes et sont largement utilisées en RIT [7] [54]. L'introduction des méthodes

basées sur les dictionnaires en RIT est motivée par leur facilité d'utilisation et la disponibilité des dictionnaires électroniques. Les corpus parallèles, de leur côté, contiennent des informations utiles pour la traduction des mots dans des domaines particuliers. Sur la base de ces corpus, des modèles de traduction statistique peuvent être entraînés et utilisés par la suite pour traduire les termes d'une requête. Cette approche présente l'avantage que peu d'interventions manuelles sont nécessaires pour produire un modèle de traduction statistique.

Par rapport à la traduction automatique, les dictionnaires et les corpus parallèles peuvent suggérer des traductions multiples pour un mot source. Pour la RI, ceci équivaut à une expansion de requête, qui est désirable afin de couvrir plus de documents pertinents. L'autre avantage des techniques basées sur les corpus parallèles est qu'un mot source peut aussi être traduit par des mots cibles reliés. Ceci renforce encore plus l'effet d'expansion de requête. Cependant les ressources linguistiques nécessaires pour ces deux méthodes, i.e. les dictionnaires bilingues et les corpus parallèles, ne sont pas toujours disponibles pour des paires de langues telles que anglais-arabe et même si elles existent, elles ne sont pas suffisantes pour fournir une traduction de qualité qui répond au besoin de la RIT.

En l'absence des corpus parallèles comme le Hansard, nous avons exploité le Web pour construire un corpus de textes parallèles pour la paire de langues anglais-arabe. Ceci a eu un succès limité à cause du peu de sites bilingues anglais-arabe sur le Web. Sur la base de ce corpus, un modèle de traduction statistique est entraîné spécifiquement pour la RIT anglais-arabe [32]. Pour consolider la traduction des requêtes de l'anglais vers l'arabe, d'autres ressources de traduction sont identifiées. Parmi ces ressources, un autre modèle de traduction statistique entraîné sur le corpus parallèle des Nations Unies est utilisé. Nous avons aussi tiré profit de deux dictionnaires bilingues extraits du Web. Ces deux ressources sont construites par l'extraction des traductions fournies par deux dictionnaires pour chaque entrée d'un corpus de mots anglais [35].

Quand plusieurs ressources de traduction sont identifiées, la question de combinaison des ressources s'impose naturellement. L'objectif de la combinaison est de tirer profit des avantages qu'offre chaque méthode de traduction et par conséquent augmenter la performance de la RIT. Dans cette thèse, deux techniques de combinaison sont étudiées [36]. La première méthode est traditionnelle : elle effectue une combinaison linéaire des ressources permettant de regrouper différentes traductions suggérées par différentes ressources pour un même mot en attribuant un poids de confiance pour chaque ressource. Les coefficients de confiance assignés à chacune des ressources de traduction sont optimisés sur un ensemble de validation de phrases alignées anglais-arabe. Nous avons proposé une deuxième méthode de combinaison, qui utilise des facteurs de confiance associés à chaque traduction. C'est la première fois que cette méthode est utilisée dans la traduction de requête en RIT. Dans cette méthode, un facteur de confiance mesure la probabilité que la traduction soit correcte. Cette nouvelle méthode de combinaison de ressources reconsidère toutes les traductions candidates proposées par les différentes ressources et, en introduisant des attributs additionnels, elle les réévalue plus radicalement que dans la combinaison linéaire.

Pour mettre à l'épreuve ces deux méthodes de combinaison, nous avons réalisé une série d'expérimentations sur différentes collections de RIT [36]. Les résultats de ces expérimentations ont montré que la méthode des facteurs de confiance est plus raisonnable et plus performante que la méthode traditionnelle pour combiner des ressources de traduction différentes et non homogènes telles que les dictionnaires bilingues et les modèles de traduction statistique.

6.3 Perspectives et améliorations possibles

Nous avons abordé certains problèmes dans la RI monolingue et translinguistique pour l'arabe. Mais ces problèmes sont loin d'être résolus complètement.

Notamment, nous observons qu'il y a encore un grand écart de performance entre la RIT et la RI monolingue. Des améliorations sont possibles sur plusieurs aspects. Dans cette section, nous discutons de quelques améliorations possibles.

6.3.1 Lemmatisation des mots arabes

Nous avons expérimenté deux méthodes de lemmatisation sur une collection de documents arabes de volume important et les résultats ont montré que l'aspect de lemmatisation est capital pour la RI arabe. Les résultats ont aussi montré que la technique de lemmatisation à base linguistique est plus efficace que la technique de lemmatisation assouplie. Mais, l'analyse des performances individuelles des requêtes a révélé que la méthode à base linguistique a ses limites et entraîne parfois des erreurs de lemmatisation. Ces erreurs sont principalement liées aux mots ambigus. Pour ces mots, un affixe est tronqué mais normalement il ne devrait pas l'être parce qu'il fait partie des lettres de ces mots. Nous estimons que la lemmatisation à base linguistique peut être améliorée à ce niveau.

Plus de traitement au niveau des statistiques de corpus doit être fait pour choisir le lemme correct quand différents lemmes candidats sont proposés pour un mot. Par exemple, on choisit le lemme le plus utilisé dans un document ou une requête au lieu de sélectionner le lemme le plus fréquent dans toute la collection des documents. Une autre approche qui peut être envisagée pour le traitement de ces mots ambigus est d'utiliser les techniques de désambiguïsation (word sense disambiguation) appliquées à la recherche d'information [42] [76]. Par exemple, pour les mots qui admettent plusieurs décompositions (lemmes), on peut utiliser un dictionnaire et exploiter les informations qui se trouvent dans la définition d'un mot pour déterminer le lemme approprié comme le fait Lesk pour la désambiguïsation des mots à plusieurs sens [49]. Pour ce faire, on utilise les informations du contexte i.e. les mots qui co-occurrent avec le mot ambigu. On compare les mots présents dans la définition des mots du contexte avec ceux des lemmes possibles pour le mot ambigu. Ensuite, on sélectionne le lemme qui a plus d'affinités (mots

en commun) avec les mots du contexte.

Enfin, la lemmatisation peut être améliorée par l'utilisation de certains lexiques comme le lexique des formes singulière et plurielle des mots irréguliers. Nous avons vu que, pour ces mots, les formes du singulier et du pluriel ne sont pas reliées par de simples inflexions. Sans utiliser un lexique pour ces types de mots, il est difficile d'écrire un algorithme à base de règles pour réduire ce genre de pluriel au singulier.

La lemmatisation peut aussi provoquer des erreurs pour certains mots comme les entités nommées. Par exemple, le mot السودان (Le Soudan) peut facilement être lemmatisé par سود (noirs). De tels mots devraient être normalement exemptés de la lemmatisation. Ainsi, un lexique pour certaines entités nommées comme les noms propres et les noms de lieux géographiques peut épargner la lemmatisation de certaines erreurs.

6.3.2 Traduction des entités nommées

Dans le deuxième volet de cette étude, notre défi était la traduction des requêtes pour résoudre la problématique de la RIT anglais-arabe. Le premier problème rencontré dans ce volet était l'identification des ressources de traduction pour cette paire de langues. Un autre problème qui n'a pas été traité dans cette thèse est la question de la traduction des entités nommées. Il s'agit particulièrement des noms propres, des noms d'organisations, des termes techniques, etc. Actuellement, de nouveaux mots et de nouvelles phrases sont introduits de plus en plus dans les documents. Ces entités de par leur spécificité posent un problème majeur lors de

leur traduction. Parce que la majorité des termes constituant une requête sont des mots génériques, leur traduction est faite en consultant des ressources communes comme les dictionnaires bilingues ou des modèles de traduction statistique. Cependant, une requête peut aussi contenir des entités nommées. Si on tente de traduire de telles entités nommées par une traduction générique basée sur les dictionnaires bilingues ou des modèles statistiques, la qualité de traduction souffrira. Ceci est notamment dû aux limitations et insuffisances suivantes :

- Les dictionnaires sont statiques et ne sont pas toujours mis à jour pour contenir les traductions des entités nommées. Ils ne couvrent pas la majorité des noms propres, des noms d'organisations, des noms de lieux, etc.
- La qualité de traduction des mots par les modèles statistiques dépend du domaine des textes parallèles sur lesquels ces modèles sont entraînés. En plus, la couverture de ces modèles est limitée pour les entités nommées.

A cause de ces limitations, les entités nommées ne seront pas trouvées dans les dictionnaires ou dans les modèles de traduction statistique et leur traduction échoue automatiquement. Ainsi, une méthode de traduction spécifique pour ces entités doit être développée.

Une telle approche consiste en deux étapes : identification des entités ensuite leur traduction. L'identification peut être réalisé automatiquement par des outils d'extraction d'entités nommées comme "Annie" du package "Gate" [1]. La traduction peut se faire par l'exploitation des dictionnaires bilingues spécialisés ou des corpus parallèles d'un domaine spécifique. Pour le cas de la traduction anglais-arabe, la translitération peut aussi résoudre la traduction de certaines entités comme les noms propres et les noms de lieux. La traduction par la translitération est utilisée notamment pour les langues ayant des alphabets et des sons différents [39]. La translitération est typiquement basée sur la projection du son dans la langue cible.

Par exemple, "Knight" est translitéré en arabe à "نايت" (nayt). Ce type de traduction basé sur le son est déjà appliqué en RIT pour la traduction des noms propres en chinois [15]. Pour l'arabe, on peut citer le travail de Al-onaizan [5].

6.3.3 Combinaison des ressources de traduction

Quand une ressource de traduction ne peut pas couvrir tous les termes de requêtes, on utilise plusieurs ressources pour renforcer la traduction. Dans cette thèse, nous avons étudié deux approches de combinaison de ressources : une combinaison linéaire et une combinaison par facteurs de confiance. Les deux méthodes ont été expérimentées sur deux collections de RIT et les résultats ont montré que la méthode des facteurs de confiance est plus efficace pour combiner des ressources de traductions non homogènes comme les dictionnaires bilingues et les modèles de traduction statistique [34] [35]. Néanmoins, l'approche des facteurs de confiance peut encore être optimisée en identifiant d'autres attributs informatifs. Nous envisageons notamment de définir des attributs sémantiques pour les traductions. La similarité sémantique entre mots est une piste privilégiée pour la définition de ces attributs. Pour quantifier cette similarité, on peut s'inspirer des méthodes du calcul de la distance WordNet [12]. Enfin, nous estimons que la définition de ce type d'attributs peut apporter des informations nouvelles et utiles pour l'estimation des facteurs de confiance associés aux traductions.

En conclusion, dans cette thèse, nous avons abordé les problèmes de la RI monolingue et translinguistique en arabe. Mais il reste beaucoup de problèmes à résoudre avant qu'on obtienne un système de RIT avec l'arabe plus performant.

BIBLIOGRAPHIE

[1] http://gate.ac.uk/ie/annie.html.

[2] http://snowball.tartarus.org/.

[3] http://www.ajeeb.com.

[4] http://www.almisbar.com.

[5] Y. Al-Onaizan. *Named entity translation: A statistical approach.* PhD thesis, University of southern California, 2002.

[6] Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, D. Melamed, F-J Och andd D. Purdy, N. A. Smith, and D. Yarowsky. Statistical Machine Translation. Technical report, Center for Speech and Language Processing, Johns Hopkins University, Baltimore, MD, 1999. Summer Workshop on Language Engineering.

[7] L. Ballesteros and B. Croft. Dictionary methods for cross-lingual information retrieval. In *Proceedings of the 7th International Conference on Database and Expert Systems Applications, DEXA'96*, pages 791–801, Zurich, Switzerland, 1996. Springer-Verlag Berlin, Germany.

[8] S. Banerjee and T. Pedersen. An adapted lesk algorithm for word sense disambiguation using WordNet. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, pages 136–145, London, UK, 2002. Springer-Verlag.

[9] C. M. Bishop. *Neural networks for pattern recognition.* Oxford University Press, 1995.

[10] J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. Confidence estimation for machine translation.

In *Proceedings of the 20th international conference on Computational Linguistics*, page 315, Geneva, Switzerland, 2004.

[11] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[12] A. Busdanistky and G. Hirst. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Proceedings of the workshop on WordNet and other lexical resources, second meeting of the north American chapter of the association for computational linguistics, Pittsburgh, PA, USA*, pages 29–34, 2001.

[13] A. Chen and F. Gey. Translation term weighting and combining translation resources in cross-language retrieval. In *Proceedings of the Text REtrieval Conference (TREC-10)*, pages 529–533, 2001.

[14] A. Chen and F. Gey. Building an Arabic stemmer for information retrieval. In *Proceedings of the Text REtrieval Conference (TREC-11)*, pages 631–639, 2002.

[15] H. H. Chen, S. J. Huang, Y. W. Ding, and S. C. Tsai. Proper Name Translation in Cross-Language Information Retrieval. In *Proceedings of the conference on computational linguistics, COLING-ACL*, pages 232–236, Montréal, Canada, 1998.

[16] J. Chen and J.Y. Nie. Parrallel Web text mining for cross-language. In *Proceedings de la conférence sur la Recherche d'Information Assistée par Ordinateur (RIAO)*, pages 62–77, Paris, 2000.

[17] A. Chowdhury, M. Aljlayl, E. Jensen, S. Beitzel, D. Grossman, and O. Frieder. IIT at TREC 2002 : Linear combination based on document structure and varied stemming for Arabic retrieval. In *Proceedings of the Text REtrieval Conference (TREC-11)*, pages 299–310, 2002.

[18] K. Darwish and D. W. Oard. CLIR experiments at Maryland for TREC-2002: Evidence combination for Arabic-English retrieval. In *Proceedings of the Text REtrieval Conference (TREC-11)*, pages 703–710, 2002.

[19] K. Darwish and D. W. Oard. Probabilistic structured query methods. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 338–344, 2003.

[20] A. Diekema, F. Oroumchian, P. Sheridan, and E. D. Liddy. TREC-7 Evaluation of Conceptual Interlingua Document Retrieval (CINDOR) in English and French. In *Proceedings of the Text REtrieval Conference (TREC-7)*, pages 169–180, 1998.

[21] W. Frakes and R. Baeza-Yates. *Information retrieval: data structures and algorithms.* 1992.

[22] A. Fraser, J. Xu, and R. Weischedel. TREC 2002 Cross-lingual Retrieval at BBN. In *Proceedings of the Text REtrieval Conference (TREC-11)*, pages 102–106, 2002.

[23] W. A. Gale and K. W. Church. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102, 1993.

[24] S. Gandrabur and G. Foster. Confidence Estimation for Text Prediction. In *Proceedings of the Conference on Natural Language Learning (CoNLL 2003)*, pages 95–102, 2003.

[25] S. Gandrabur, G. Foster, and G. Lapalme. Confidence Estimation for NLP Applications. *Transactions on Speech and Language Processing*, 3(3):1–29, 2006.

[26] F. C. Gey and D. W. Oard. The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic using English, French, or Arabic Queries. In *Proceedings of the Text REtrieval Conference (TREC-10)*, pages 16–25, NIST, Gaithersburg, MD., 2001.

[27] G. Grefenstette. The Problem of Cross-Language Information Retrieval. In *Cross Language Information retrieval. G. Grefenstette (Ed)*, pages 1–10. Norwell, Kulwer Academic Publishers, 1998.

[28] T. J. Hazen, T. Burianek, J. Polifroni, and S. Seneff. Recognition confidence scoring for use in speech understanding systems. *Computer Speech and Language*, (16):49–67, 2002.

[29] G. Hirst and D. St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. WordNet-An Electronic Lexical Database. *The MIT Press, C. Fellbaum (Ed)*, pages 305–332, 1998.

[30] D. A. Hull. Stemming algorithms - a case study for detailed evaluation. *Journal of the American Society for Information Science (JASIS)*, 47(1):70–84, 1996.

[31] Y. Kadri and A. Benyamina. Un système d'analyse syntaxico-sémantique du langage arabe non voyellé. Mémoire d'ingénieur, Université d'oran, 1992.

[32] Y. Kadri and J. Y. Nie. Traduction des requêtes pour la recherche d'information translinguistique Anglais-Arabe. In *Proceedings de la conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, pages 291–296, 2004.

[33] Y. Kadri and J. Y. Nie. Effective stemming for Arabic information retrieval. In *The challenge of Arabic for NLP/MT Conference*, pages 68–74. Proceedings of The British Computer Society. London, UK, 2006.

[34] Y. Kadri and J. Y. Nie. Improving Query Translation with Confidence Estimation for Cross Language Information Retrieval. In *Conference on Information and Knowledge Management*, pages 818–819. CIKM-2006, Arligton VA, USA, 2006.

[35] Y. Kadri and J. Y. Nie. Combining resources with Confidence Measures for Cross Language Information Retrieval. In *Proceedings of the Ph.D. Workshop*

on Information and Knowledge Management (PIKM), pages 131–138. Lisbon, Portugal, 2007.

[36] Y. Kadri and J. Y. Nie. A Comparative Study for Query Translation using Linear Combination and Confidence Measure. In *The Third International Joint Conference on Natural Language Processing*, pages 181–188. IJCNLP-2008, Hyderabad, India, 2008.

[37] S. Khoja and S. Garside. Stemming Arabic Text. Technical report, Computing department, Lancaster University, Lancaster, U.K., 1999. http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps.

[38] K. Knight. A Statistical Machine Translation Tutorial Workbook. Technical report, 1999. Unpublished, http://www.clsp.jhu.edu/ws/projects/mt/wkbk.rtf.

[39] K. Knight and J. Graehl. Machine Transliteration. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 128–135, 1997.

[40] W. Kraaij. TNO at CLEF-2001: Comparing translation resources. In *Workshop of the Cross-Language Evaluation Forum (CLEF), Darmstadt*, pages 29–40, 2001.

[41] W. Kraaij, J. Y. Nie, and M. Simard. Embedding Web-based statistical translation models in CLIR. *Computational Linguistics*, 29(3):381–419, 2003.

[42] R. Krovetz and W. B. Croft. Lexical ambiguity in information retrieval. *ACM Transactions on Information Systems*, 10(2):115–141, April 1992.

[43] K. L. Kwok. English-Chinese cross-language retrieval based on a translation package. In *Workshop of Machine Translation for Cross Language Information Retrieval, Machine Translation Summit VII, Singapore*, pages 8–13, 1999.

[44] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 111–119, 2001.

[45] P. Langlais. Traitement statistique des langues naturelles. Support de cours de graduation. Université de Montréal.

[46] P. Langlais, M. Simard, and J. Véronis. Methods and Practical Issues in Evaluating Alignment Techniques. In *Proceedings of the conference on computational linguistics, COLING-ACL*, pages 711–717, Montréal, Québec, 1998.

[47] L. S. Larkey, L. Ballesteros, and M. 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*, pages 275–282, ACM, NY, 2002.

[48] L. S. Larkey and M. E. Connell. Arabic information retrieval at UMass in TREC-10. In *Proceedings of the Text REtrieval Conference (TREC-10)*, pages 562–570, Gaithersburg, Maryland, 2001.

[49] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, pages 24–26, Toronto, Canada, 1986.

[50] C. D. Manning, P. Raghavan, and H. Schütze. *An introduction to Information Retrieval.* Cambridge University Press, 2007.

[51] P. McNamee, C. Piatko, and J. Mayfield. JHU/APL at TREC 2002: Experiments in Filtering and Arabic Retrieval. In *Proceedings of the Text REtrieval Conference (TREC-11)*, pages 358–363, Gaithersburg, Maryland, 2002.

[52] J. Y. Nie. CLIR using a Probabilistic Translation Model based on Web Documents. In *Proceedings of the Text REtrieval Conference (TREC-8)*, page 607, 1999.

[53] J. Y. Nie, M. Simard, and G. Foster. Multilingual information retrieval based on parallel texts from the Web. In C. Peters, editor, *Proceedings of the Cross-Language Evaluation Forum (CLEF)*, pages 188–201, 2000.

[54] J.Y. Nie. TREC-7 CLIR using a Probabilistic Translation Model. In *Proceedings of the Text REtrieval Conference (TREC-7)*, pages 547–553, 1998.

[55] J.Y. Nie and J. Chen. Exploiting the Web as Parallel Corpora for Cross-Language Information Retrieval. *Web Intelligence, eds. N. Zhong, J. Liu, Y. Yao, Springer*, pages 218–239, 2002.

[56] D. W. Oard and A. Diekema. Cross-Language Information Retrieval. In *Annual review of Information science. M. Williams (Ed)*, volume 33, 1998.

[57] D. W. Oard and F. C. Gey. The TREC-2002 Arabic/English CLIR Track. In *Proceedings of the Text REtrieval Conference (TREC-11)*, pages 17–26, NIST, Gaithersburg, MD., 2002.

[58] F. J. Och and H. Ney. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 440–447, 2000.

[59] J. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281, 1998.

[60] M. Porter. An algorithm for suffix stripping. *Automated Library and Information Systems*, 14(03):130–137, 1980.

[61] S. E. Robertson and K. Sparck Jones. Weighting of search terms. *Journal of the American Society of Information Science*, 27:129–146, 1976.

[62] J. Rocchio. *Relevance feedback in information retrieval.* 1971.

[63] G. Salton and al. A vector space model for automatic indexing. *Communications of the Association for Computing Machinery (CACM)*, 18:613–620, 1975.

[64] G. Salton and M. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill (Ed), New-York, 1983.

[65] T. Saracevic. The concept of "relevance" in information science: A historical review. *In Saracevic, T. (Ed.), Introduction to Information Science. New York: R.R. Bowker*, pages 111–151, 1970.

[66] J. Savoy. Stemming of French Words Based on Grammatical Category. *Journal of the American Society for Information Science*, 44(1):1–9, 1993.

[67] J. Savoy. Recherche multilingue d'information. *Information-Interaction-Intelligence*, 2(2):9–36, 2002.

[68] J. Savoy and Y. Rasolofo. Report on the TREC-11 Experiment: Arabic, Named Page and Topic Distillation Searches. In *Proceedings of the Text REtrieval Conference (TREC-11)*, pages 765–774. NIST, 2002.

[69] P. Sheridan and J. P. Ballerini. Experiments in multilingual information retrieval using the SPIDER system. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 58–65, 1996. http://www.acm.org/pubs/articles/proceedings/ir/243199/p58-sheridan/p58-sheridan.pdf.

[70] M. Simard, G. Foster, and P. Isabelle. Using Cognates to Align Sentences in Bilingual Corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 67–81, Montréal, Québec, 1992.

[71] K. Sparck-Jones and P. Willett. *Readings in information retrieval.* 1993.

[72] J. B. Teevan. *Improving IR with textual analysis : Bayesian models and beyond.* PhD thesis, 2001.

[73] N. Ueffing, K. Machery, and H. Ney. Confidence measures for statistical machine translation. In *Proceedings of the Machine Translation Summit IX*, pages 394–401, 2003.

[74] K. Van-Rijsbergen. *Information retrieval.* 1979.

[75] S. Vogel and C. Monson. Augmenting Manual Dictionaries for Statistical Machine Translation Systems. In *Proceedings of the Fourth Conference on Language Resources and Evaluation (LREC)*, pages 1593–1596, 2004.

[76] E. Voorhees. Using WordNet to Disambiguate Word Senses for Text Retrieval. In *Proceedings of the 16th Annual International ACM SIGIR conference of Research and Development in Information Retrieval*, pages 171–180, 1993.

[77] E. Voorhees and D. Harman. Overview of the Text REtrieval Conference (TREC-7). In *Proceedings of the Text REtrieval Conference (TREC-7)*, page 1. NIST, 1999.

[78] P. Vossen. EuroWordNet: a multilingual database for information retrieval. In *Proceedings of the DELOS workshop on Cross-language Information Retrieval, Zurich*, pages 5–7, 1997.

[79] J. Xu and R. Weischedel. Empirical studies on the impact of lexical resources on CLIR performance. *Information processing and management*, 41(3):475–487, 2005.

[80] K. Yamabana, K. Murald, S. Doi, and S. Kamei. A Language Conversion Front-End for Cross Linguistic Information Retrieval. In *Proceedings of the SIGIR Workshop on Cross-linguistic Information Retrieval*, pages 34–39, 1996.

[81] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the Annual International ACM SIGIR conference of Research and Development in Information Retrieval*, pages 334–342, 2001.

[82] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Conference on Information and Knowledge Management (CIKM)*, pages 403–410, 2001.

Annexe I

Structure d'un document arabe dans la collection TREC

<DOC>
<DOCNO>19940513_AFP_ARB.0022</DOCNO>
<HEADER>
ار1010 4 ع 3320 قبرص /اف ب-تصغ38 توقعات
</HEADER>
<BODY>
<HEADLINE>
&HT; الاحداث المتوقعة للاسابيع المقبلة
</HEADLINE>
<TEXT>
<P>
نيقوسيا 13-5 (اف ب) - في ما يلي ابرز الاحداث المتوقعة للاسابيع المقبلة: 14/5 طهران: رئيس وزراء كزاخستان يقوم بزيارة رسمية الى ايران
</P>
<P>
15 - 16/5 دمشق : زيارة وزير الخارجية الاميركي وارن كريستوفر
</P>
<P>
16/5 نيروبي : استئناف مفاوضات السلام حول جنوب السودان
</P>
<P>
17 - 18/5 القدس: زيارة وزير الخارجية الاميركي
</P>
<P>
16 - 19/5 بيروت: زيارة وفد تجاري بريطاني
</P>
<P>
18/5 غزة (مبدئيا) انتهاء الانسحاب العسكري الاسرائيلي وانتقال السلطة الى الفلسطينيين
</P>
<P>
18/5 مكة (السعودية): تظاهرة ايرانية بمناسبة موسم الحج
</P>
<P>
20/5 مكة (السعودية): صعود الحجاج الى جبل عرفات
</P>
<P>
21/5 احتفال المسلمين بعيد الاضحى
</P>
<P>
22 - 26/5 القدس : وفد من ارباب العمل الفرنسيين يزور اسرائيل والاراضي المحتلة
</P>
<P>
23/5 بيروت: زيارة وفد تجاري من مقاطعة كيبيك 25/5 (مبدئيا) غزة/اريحا : تنصيب السلطة الفلسطينية الجديدة
</P>
<P>
28 - 31/5 القاهرة : مؤتمر وزراء خارجية الدول الاعضاء في حركة عدم الانحياز
</P>
<P>
30/5 الكويت : صدور الحكم في دعوى استئناف حكم الاعدام الصادر ضد عشرة فلسطينيين بتهمة التعاون مع العراق
</P>
<P>
30/5 طرابلس : الموعد المقرر لانتهاء الانسحاب الليبي من شريط اوزو
</P>
<P>
4/6 الكويت : صدور الحكم في قضية العراقيين المتهمين بمحاولة اغتيال الرئيس الاميركي السابق جورج بوش
</P>
<P>
4/6 طهران: احياء الذكرى الخامسة لوفاة الامام الخميني
</P>
<P>
14/6 ابو ظبي : صدور الحكم في قضية 31 موظفا سابقا في بنك الاعتماد والتجارة الدولي
</P>
</TEXT>
<FOOTER>
لبا/فن/هدى نيف اف ب _____
</FOOTER>
</BODY>
<TRAILER>
407131 جمت ماي 49
</TRAILER>
</DOC>

Annexe II

La liste des requêtes de la collection TREC 2001 en arabe

<top>
<num> Number: AR1
<title> فنون العرض و المؤسسات الاسلامية في العالم العربي
<desc> Description:
ما هو اثر المؤسسات الاسلامية على فنون العرض مثل الرقص و الموسيقى في العالم العربي؟
<narr> Narrative:
المقالات المتعلقة بالفنون الرياضية او التشكيلية او بفنون العرض خارج العالم العربي او بالسلوكيات الدينية خارج اطار فنون العرض او بالديون و القروض المالية لا علاقة لها بالموضوع
</top>

<top>
<num> Number: AR2
<title> استهلاك العرب للسينما العربية و الغربية
<desc> Description:
ما هو اثر مهرجانات كان و هوليوود مثلا للسينما على انتاج و استهلاك العرب للأفلام العربية؟
<narr> Narrative:
لا يدخل في هذا الموضوع مقالات عن الجوائز التي حصل عليها الممثلون الغربيون في المهرجانات الغربية سواء في امريكا او في اوربا او اخبار عن حياة أو موت أي فنان عربي و من المفضل لتوسيع الموضوع أخبار عن مشاركة العرب في المهرجانات الدولية للسينما و مقالات عن مشاركة العرب في الأعمال السينمائية الدولية داخل الوطن العربي و أخبار عن مدى اهتمام الجمهور العربي بالسينما العربية و احصاءات لأنواع الافلام المستوردة من الغرب و نوعية الافلام التي يقبل عليها الشباب في الوطن العربي
</top>

<top>
<num> Number: AR3
<title> الفنون بعد حرب الخليج في العالم العربي
<desc> Description:
كيف اثرت حرب الخليج في الانتاج الفني عامة في العالم العربي ؟
<narr> Narrative:
يدخل في هذا الموضوع كل ما يتعلق بالعلاقات السياسية بين الشرق الأوسط و الدول المشاركة في حرب الخليج و التبادلات الثقافية فيما بين العرب أو بين العرب و الدول الاروبية و انتاجات العرب الفنية داخل و خارج الشرق الأوسط و القوانين الصادرة في حق الفنون و الأنشطة الفنية في العالم العربي
</top>

<top>
<num> Number: AR4
<title> الرسوم الفطرية و الغربيون في الوطن العربي
<desc> Description:
لماذا يشجع تجار الفن الغربيون من اروبا و الامريكتين الرسوم الفطرية في بعض الدول العربية؟
<narr> Narrative:
الفنون الفطرية هي كل الفنون التي ينجزها فنانون أميون لم يتلقوا أي تكوين فني أو تقني يتعلق بفن الرسم أو النحت لا يدخل في هذا الموضوع اي خبر عن الفنانين الفطريين في الغرب و يتعلق بهذا الموضوع أخبار الفنانين العرب تساهم في توسيع الرؤية حول وضع الفنون الفطرية في العالم العربي
</top>

<top>
<num> Number: AR5
<title> الصناع التقليديون في مواجهة التكنولوجيا
<desc> Description:
كيف يواجه الصناع التقليديون كلا من التحديات التكنولوجية و غياب سياسة اقتصادية حقيقية في العالم العربي؟
<narr> Narrative:
يتعلق بالموضوع كل من المقالات و الأخبار الخاصة عن المشاكل التي يواجهها الصناع التقليديون في العالم العربي بسبب الغزو التكنولوجي لميادين تخصصاتهم
</top>

<top>
<num> Number: AR6
<title> كبريات المدن العربية و الطرق الحديثة للاشهار
<desc> Description:
ما هي التدابير المتخذة في المشاريع البلدية لتكييف البيئة في المدن الكبرى مع المعطيات التكنولوجية الجديدة الخاصة بفنون الاشهار مثل اضواء الليزر، اللافتات، الملصقات، السيارات المتجولة
<narr> Narrative:
الانتخابات البلدية و التشريعية لا تدخل في هذا الموضوع العلاقات الخارجية و التبادلات بين البلديات العربية و الغربية هي من صميم الموضوع من المفضل أن تضاف مقالات عن المنجزات البلدية في هذا الاطار
</top>

<top>
<num> Number: AR7
<title> النقد و الشعر السياسي في العالم العربي
<desc> Description:
كيف يعبر النقاد العرب عن مواقفهم تجاه الشعر السياسي سواء كان مع او ضد النظام السياسي في بلدهم؟
<narr> Narrative:
يمكن ارفاق الاخبار المتعلقة بالمهرجانات الشعرية في العالم العربي بالموضوع و لكن الاخبار السياسية و الاخبار الفنية الخارجة عن ميدان النقد الادبي لا علاقة لها بالموضوع
</top>

<top>
<num> Number: AR8
<title> الطفل العربي و الفنون في المدارس الابتدائية و الثانوية
<desc> Description:
ما هي الفنون التي يدرسها الطفل العربي في المدارس الابتدائية و الثانوية؟ أي تدابير بيداغوجية تتخذ من أجل ذلك؟
<narr> Narrative:
تدخل في هذا الموضوع كل المقالات التي تتحدث عن التربية الفنية في العالم العربي و القرارات الحكومية المتخذة لتدبير الشؤون التربوية و التبادلات الدولية بين العرب و الغرب لتطوير تعليم الفنون في المدارس الابتدائية و الثانويات في العالم العربي
</top>

<top>
<num> Number: AR9
<title> ضحايا الحرب بين الجراحة التجميلية و الاسلام
<desc> Description:
كيف توفق المؤسسات الصحية بين حاجة ضحايا الحرب نساء و رجالا الى التعديلات التجميلية و بين الاوامر و النواهي الاسلامية في بلدان العالم العربي ؟
<narr> Narrative:
لايدخل في هذا الموضوع كل ما يتحدث عن سياسة الحرب أو أخبارها أو نتائجها الأيكولوجية اخبار الجديد في ميدان التجميل أيضا لا علاقة لها بالموضوع
</top>

<top>
<num> Number: AR10
<title>
القضاء على شلل الأطفال في الشرق الأوسط
<desc> Description:
هل تحقق أي تقدم في القضاء على شلل الأطفال في الشرق الأوسط؟
<narr> Narrative:
ان المقالات التي تتحدث عن القضاء على شلل الاطفال عالميا او في مناطق اخرى في العالم من غير ذكر الشرق الاوسط ليست لها علاقة بالموضوع و يشمل الشرق الاوسط هنا الصومال و افغانستان و باكستان
</top>

<top>
<num> Number: AR11
<title> حملات التلقيح ضد الحصبة في الشرق الأوسط
<desc> Description:
هل نفذت حملات التلقيح أو التطعيم ضد مرض الحصبة في الشرق الأوسط؟
<narr> Narrative:
ان المقالات التي تتحدث عن مرض الحصبة او التلقيح ضد الحصبة بشكل عام او في مناطق اخرى للعالم من غير ذكر الشرق الاوسط ليست لها علاقة بالموضوع و يشمل الشرق الاوسط هنا الصومال و افغانستان و باكستان
</top>

<top>
<num> Number: AR12
<title> منع البلهارسيا في مصر
<desc> Description:
ماذا قالت الصحف عن برامج منع البلهارسيا في مصر؟
<narr> Narrative:
تتعلق بالموضوع المقالات عن تجارب لقاح ضد البلهارسيا في مصر ولكن المقالات عن اللقاح بشكل عام لا علاقة لها بالموضوع و تتعلق بالموضوع كذلك كل أخبار الاجتماعات العلمية المنعقدة في موضوع البلهارسيا في مصر
</top>

<top>
<num> Number: AR13
<title> المسرح في مصر
<desc> Description:
ماذا قالت الصحف عن المسرح في مصر؟
<narr> Narrative:
يتعلق بالموضوع كل المقالات التي تتحدث عن الساحة المسرحية في مصر في الماضي و في الحاضر
</top>

<top>
<num> Number: AR14
<title> السياحة الاسرائيلية في الأردن
<desc> Description:
ماذا قالت الصحف عن السياحة الاسرائيلية في الأردن بعد اتفاقية السلام
<narr> Narrative:
تتعلق بالموضوع المقالات التي تتحدث عن الجهود المشتركة بين الأردن و اسرائيل لتشجيع السياحة
</top>

<top>
<num> Number: AR15
<title> القنوات الفضائية في المناطق الريفية
<desc> Description:
هل تتوفر القنوات الفضائية في المناطق الريفية العربية؟
<narr> Narrative:
ان المقالات التي تتحدث عن اطلاق قنوات فضائية جديدة في البلدان العربية لا علاقة لها بالموضوع الا اذا تحدثت عن توفر هذه القنوات في المناطق الريفية
</top>

<top>
<num> Number: AR16
<title> قوانين حماية البيئة في مصر
<desc> Description:
هل تنفذ القوانين لحماية البيئة في مصر؟
<narr> Narrative:
يتعلق بهذا الموضوع كل المقالات التي تتحدث عن تنفيذ القوانين لحماية البيئة المصرية و مواردها الطبيعية
</top>

<top>
<num> Number: AR17
<title> الكفاءة النووية لدى اسرائيل
<desc> Description:
ماذا قالت الصحف عن الكفاءة النووية لدى اسرائيل ؟
<narr> Narrative:

يتعلق بالموضوع كل المقالات التي تتحدث عن الكفاءة النووية لدى اسرائيل مهما كانت عسكرية او مدنية
</top>

<top>
<num> Number: AR18
<title> العلاقات بين مصر و ليبيا أثناء التسعينات
<desc> Description:
كيف صورت الصحف العلاقات بين مصر و ليبيا أثناء التسعينات؟
<narr> Narrative:
المقالات التي تشمل الاقوال البسيطة المتعلقة بالدولتين مثل نتائج المباريات الرياضية لا علاقة لها بالموضوع
</top>

<top>
<num> Number: AR19
<title> مكتبات الاسكندرية
<desc> Description:
هل هناك مكتبات جديدة بنيت في الاسكندرية؟
<narr> Narrative:
يرتبط بهذا الموضوع كل ما يتعلق بتاريخ مكتبات الاسكندرية القديمة والمكتبات الحديثة ان
المقالات المتعلقة بموقع ومناخ الا سكندرية لا علاقة لها بالموضوع
</top>

<top>
<num> Number: AR20
<title> السياحة فى القاهرة
<desc> Description:
ما هى الاجراءات التى اتخذت لتطوير السياحة فى القاهرة ؟
<narr> Narrative:
يدخل في هذا الموضوع كل المقالات التي تتعلق بالا رهاب الذى يهدد السياحة فى مصر المقالات المتعلقة
بالقضايا البلدية مثل الازدحام المرورى في القاهرة لا علاقة لها بالموضوع
</top>

<top>
<num> Number: AR21
<title> إكتشافات أثرية هامة في منطقة البحر الميت
<desc> Description:
هل هناك أي إكتشافات أثرية مهمة فى منطقة البحر الميت مؤخرا؟
<narr> Narrative:
يدخل فى هذا الموضوع كل المقالات المتعلقة بالاكتشافات الاثرية الحديثة فى منطقة البحر الميت
ان المقالات الخاصة عن وصف طبيعة البحر الميت لا تتعلق بالموضوع
</top>

<top>
<num> Number: AR22
<title> الصحف المحلية و قانون الصحافة الجديد فى الأردن
<desc> Description:
هل أغلقت الحكومة الأردنية أي صحف محلية بسبب قانون الصحافة الجديد؟
<narr> Narrative:
كل المقالات التي تتحدث عن قانون الصحافة فى الأردن وتأثيره على الصحف المحلية و على الرأي العام
لها علاقة بالموضوع لا يتعلق بهذا الموضوع اي خبر عن معاناة الصحفيين الشخصية
</top>

<top>
<num> Number: AR23
<title> عصر المعلومات و العالم العربى
<desc> Description:
كيف تأثر العالم العربي بدخوله إلى عصر المعلومات؟
<narr> Narrative:
تتعلق بالموضوع كل المقالات المتعلقة بتأثير تكنولوجيا المعلومات على احداث تغييرات فى العالم العربى و رد فعل العامة
إزا تلك المتغيرات ان المقالات التي تتحدث عن السوق الاقتصادية المشتركة بين الدول العربية لا تدخل فى اطار هذا الموضوع
</top>

<top>
<num> Number: AR24
<title> مشكلات موارد المياه فى حوض وادى النيل
<desc> Description:
ما هى المشكلات الأساسية في المياه فى دول حوض النيل؟
<narr> Narrative:
مناقشة بناء السدود والا جتماعات المستمرة والتفاهم بين دول حوض وادى النيل هى من الموضوعات
المرتبطة اما مناقشة التبادل التجارى بين دول حوض النيل هى من الموضوعات غير المرتبطة
</top>

<top>
<num> Number: AR25
<title> الدور الاوروبى والامريكى فى عملية السلام فى الشرق الأوسط
<desc> Description:
ما هى أدوار الدول الأوربية وامريكا في عملية السلام فى الشرق الأسط؟
<narr> Narrative:
يتعلق بهذا الموضوع كل مقال يخص التدخل الاوروبى والامريكى فى القرارات العربية لتوجيه عملية
السلام في الشرق الاوسط وما لا يرتبط بهذا الموضوع هو التدخل الاوروبى و الامريكى فى الشؤون الداخلية لدول الشرق الاوسط
</top>

Annexe III

Structure du fichier du jugement de pertinence

1	0	19940515AFPARB.	0095	0
1	0	19940519AFPARB.	0085	0
1	0	19940526AFPARB.	0068	1
1	0	19940526AFPARB.	0080	1

Annexe IV

La liste des requêtes de la collection TREC 2001 en anglais

<top>
<num> Number: AR1
<title> Performing Arts and Islamic Institutions in the Arab World.
<desc> Description:
What is the impact of Islamic Institutions on the performing arts such as dance and music in the Arab World?
<narr> Narrative:
Articles concerning sports or spatial arts, performance arts outside of the Arab World, religious behavior not related to the performing arts, or debts and banking loans, are not relevant to this topic.
</top>

<top>
<num> Number: AR2
<title> Arab consumption of Arab and western Cinema.
<desc> Description:
What is the impact of Western cinema festivals (Cannes, Hollywood, others) on the Arab production and consumption of Arab films?
<narr> Narrative:
Items about prizes won by Western artists in Western festivals held in the West, or death announcements or other news about individual Arab artists are not relevant to this topic. *News about Arab participation in international cinema festivals, about Arab participation* in international cinema in the Arab world, about the interest that the Arab public has toward Arab cinema and the kinds of films popular with youth in the Arab world are very relevant to the topic, as are statistics about the kinds of films imported from Western countries.
</top>

<top>
<num> Number: AR3
<title> Arts in the Arab World after the Gulf War.
<desc> Description:
How did the Gulf War affect Arab arts in general?
<narr> Narrative:
Documents concerning political relations between the Middle East and Western nations that participated in the Gulf War, and about cultural exchanges between Arab nations and also between Arab and European nations, are relevant to this topic, as are items discussing Arab artistic products realized both within and outside of the Middle East and laws and rules concerning the arts or artistic activities in the Arab world.
</top>

<top>
<num> Number: AR4
<title> Naive painting and Westerners in the Arab world
<desc> Description:
Why do Western art dealers, from both Europe and the Americas, encourage "naive painting" in some Arab countries?
<narr> Narrative:
Naive arts are the arts that illiterate artists realize without any artistic or technical training concerning paintings or sculpture. This topic does not include articles about naive artists in the West. Stories about Arab Artists could expand understanding of the status of naive arts in the Arab world and are therefore relevant to this topic.
</top>

<top>
<num> Number: AR5
<title> Traditional craftsmen facing technology
<desc> Description:
How are traditional craftsmen facing both the technological challenges and the lack of real economic policies in the Arab world?
<narr> Narrative:
This topic includes articles concerning the problems that traditional craftsmen are facing in the Arab World because of technological developments in their fields.
</top>

<top>
<num> Number: AR6
<title> Major Arab cities and new ways of advertising
<desc> Description:
What measures are being taken by municipalities to regulate new urban advertising technologies (e.g. lasers, billboards, circulating vehicles, etc.)?
<narr> Narrative:
Legislative and municipal elections are not relevant to this topic International relations and exchanges between Arab and Western municipalities are the main part of this topic. It would be fruitful to add articles about what municipalities have accomplished in this regard.
</top>

<top>
<num> Number: AR7
<title> Criticism and political poetry in the Arab World
<desc> Description:
How have Arab critics expressed their attitude towards the new political poetry, whether for or against the political regime in their countries?

<narr> Narrative:
Stories about festivals of poetry in the Arab World may be related to this topic. Any article concerning politics or arts that does not mention literary criticism is not relevant to this topic.
</top>

<top>
<num> Number: AR8
<title> Arab children and Arts in primary and high schools
<desc> Description:
What arts are taught to Arab children in primary and high schools, and what pedagogical measures have been taken for this goal?
<narr> Narrative:
Articles relevant to this topic would include those concerning artistic education in the Arab World, governmental decisions in educational affairs, and international exchanges between Arab and Western countries to develop art instruction in elementary and high schools in the Arab world.
</top>

<top>
<num> Number: AR9
<title> War victims, plastic surgery and Islam
<desc> Description:
How do health institutions reach a compromise between the need of both male and female war victims for plastic surgery and Islamic commands and interdictions in the countries of the Arab world?
<narr> Narrative:
Articles concerning military policy or the ecological effects of war are not relevant to this topic, nor are articles regarding recent developments in plastic surgery.
</top>

<top>
<num> Number: AR10
<title> Polio eradication in the Middle East
<desc> Description:
What progress has been made toward polio eradication in the Middle East?
<narr> Narrative:
Articles which discuss global polio eradication, or polio eradication in other regions of the world, without mentioning the Middle East specifically, are not relevant. Pakistan, Afghanistan and Somalia are included here as parts of the Middle East.
</top>

<top>
<num> Number: AR11
<title> Measles immunization campaigns in the Middle East.
<desc> Description:
Have measles immunization campaigns been conducted in Middle Eastern countries?
<narr> Narrative:
Articles which discuss measles and measles immunization in general,
without mentioning the Middle East specifically, are not relevant.
Pakistan, Afghanistan and Somalia are included here as parts of the Middle East.
</top>

<top>
<num> Number: AR12
<title> Bilharzia/Schistosomaisis prevention in Egypt.
<desc> Description:
Find documents that discuss Bilharzia/Schistosomaisis prevention programs in Egypt
<narr> Narrative:
Articles about testing a vaccine for bilharzia in Egypt are relevant, but articles about the vaccine in general are not. Articles about scientific meetings about bilharzia held in Egypt are also relevant.
</top>

<top>
<num> Number: AR13
<title> Theater in Egypt
<desc> Description:
Find documents that discuss the theater in Egypt
<narr> Narrative:
Any articles discussing the Egyptian theater scene, past or present, are relevant.
</top>

<top>
<num> Number: AR14
<title> Israeli tourism in Jordan
<desc> Description:
Find documents that discuss Israeli tourism in Jordan after the peace treaty.
<narr> Narrative:
Articles which discuss joint efforts by Jordan and Israel to encourage tourism are relevant.
</top>

<top>
<num> Number: AR15
<title> Satellite TV in rural areas
<desc> Description:

What is the availability of satellite TV in rural areas in Arab countries?
<narr> Narrative:
Articles which discuss the launching of new satellite TV channels in Arab countries are not relevant unless they also discuss the availability of these channels in the rural areas.
</top>

<top>
<num> Number: AR16
<title> Environmental protection laws in Egypt
<desc> Description:
Are environmental protection laws being enforced in Egypt?
<narr> Narrative:
Any articles discussing enforcement of laws protecting Egypt's environment and natural resources are relevant.
</top>

<top>
<num> Number: AR17
<title> Israel's nuclear capability
<desc> Description:
What has the press reported about Israel's nuclear capability?
<narr> Narrative:
Any articles dealing with Israel's nuclear capability, military or civilian, are relevant.
</top>

<top>
<num> Number: AR18
<title> Egyptian-Libyan relations during the 1990s
<desc> Description:
Find documents that characterize Egyptian-Libyan relations during the 1990s.
<narr> Narrative:
Articles containing nothing but simple statements of facts about the two countries, such as sports scores, are not relevant.
</top>

<top>
<num> Number: AR19
<title> Alexandrian libraries
<desc> Description:
Are there any new libraries being built in Alexandria?
<narr> Narrative:
Articles about the history of the old and the new libraries in Alexandria are relevant to this topic. Articles about the location and climate of Alexandria are irrelevant to this topic.
</top>

<top>
<num> Number: AR20
<title> Tourism in Cairo
<desc> Description:
What measures are being taken to develop tourism in Cairo?
<narr> Narrative:
Articles about government steps to abolish terrorism that threatens tourism and tourist services are relevant to this topic. Articles about municipal issues such as traffic jams in Cairo are irrelevant to this topic.
</top>

<top>
<num> Number: AR21
<title> Significant archaeological finds in the Dead Sea area
<desc> Description:
Have there been any significant archaeological finds in the Dead Sea area recently?
<narr> Narrative:
Articles about any recent archaeological discoveries in the Dead Sea area are relevant to this topic. Articles about the nature of the Dead Sea are irrelevant.
</top>

<top>
<num> Number: AR22
<title>
Local newspapers and the new press law in Jordan
<desc> Description:
Has the Jordanian government closed down any local newspapers due to the new press law?
<narr> Narrative:
Any articles about the press law in Jordan and its effect on the local newspapers and the reaction of the public and journalists toward the new press law are relevant. The articles that deal with the personal suffering of the journalists are irrelevant.
</top>

<top>
<num> Number: AR23
<title> Information technology and the Arab World
<desc> Description:
How has the IT age affected the Arab World?

<narr> Narrative:
Any articles about the effect of information technology on the Arab World, changes which have occurred and the reaction of the public toward these changes are relevant. Articles about the economic common market between the Arab countries are irrelevant.
</top>

<top>
<num> Number: AR24
<title> Water resource problems in the Nile Valley
<desc> Description:
What are the main problems regarding water resources in the Nile valley countries?
<narr> Narrative:
Discussion of dam-building and ongoing meetings and attempts to reach understanding between the countries of the Nile Valley are relevant topics. Documents about commercial exchange between the Nile Valley countries are irrelevant.
</top>

<top>
<num> Number: AR25
<title> European and American roles in Middle East peace process
<desc> Description:
What are the roles of the European countries and America in the peace process in the Middle East?
<narr> Narrative:
Relevant articles are about the involvement of Europe and the US in directing the peace process and stopping violence in the Middle East. Articles are about European and US involvement in the internal affairs of Middle Eastern countries are irrelevant.
</top>

Annexe V

Liste de mots outils (stop words) arabes

ا , ب , ت , ث , ج , ح , خ , د , ذ , ر , ز , س , ش , ص , ض , ط , ظ ,
ع , غ , ف , ق , ك , ل , م , ن , و , ه , ي , ؤ , ة , ئ , ء , أ , ى ,
إ , آ , من , منه , منها , منهم , منهما , لمن , فمن , فمنه , فمنها , فمنهم ,
فمنهما , ومن , ومنه , ومنها , ومنهم , ومنهما , عن , عنه , عنها , عنهم , عنهما ,
الذي , للذي , فالذي , كالذي , التي , للتي , فالتي , كالتي , اللتان , اللتين ,
اللاتي , اللاتى , اللواتي , اللواتى , الذين , اللذان , إلى , الي , اليه ,
اليها , اليهم , اليهما , الى , كم , كما , كيف , أين , اين , ما , إن , ان ,
انه , انها , انهم , انهما , أن , لم , لما , فلم , فلما , قد , وقد , تحت , حيث ,
ثم , أنا , انا , نحن , أنت , انت , أنتما , انتما , أنتم , انتم , أنتن , انتن ,
هو , هما , هم , هي , هن , حتى , حتي , يا , أيا , ايا , هيا , منذ , إذما ,
اذما , إذن , اذن , نعم , هلم , إليك , إلا , ألا , الا , عدا , ماعدا ,
مازال , في , وفي , لفي , فيه , فيهم , فيهما , في , اما , أما , لو , لولا ,
إما , لن , كأن , لكن , لكنه , لكنها , لكنهم , لكنهما , ليس , ليت , ليته ,
ليتها , ليتهم , كان , اهلا , أهلا , سهلا , اولاء , أولاء , هؤلاء , ذا ,
ذو , ذي , ذى , اولو , أولو , اولات , أولات , اياك , أياك , إياك , سوف , لا ,
حيثما , بعدما , قبلما , عندما , كذا , مم , مما , كل , كلا , كلما , بلا ,
فوق , مرة , مره , جل , ككل , وحيد , وحيدة , وحيده , إحدى , احدى , احده ,
احد , أحد , بل , بلا , ايضا , أيضا , دائما , وأبدا , أبدا , ابدا , آخر ,
اخر , الآخر , الاخر , اخرى , اخري , أخرى , الأخرى , الاخري , لأن , لان , أصبح ,
اصبح , غير , شيء , أشياء , اشياء , الخ , قليل , قلة , قله , اكثر , أكثر ,
إذا , اذا , على , علي , أو , او , بأن , بان , ذلك , فإن , فان , لأن ,
مابعد , ماقبل , بين , بينه , بينها , بينهم , بينهما , بينما , ربما , حيثما ,
ريثما , مع , ضد , بعض , بن , هل , هناك , اي , مثل , مثله , مثلها , مثلهم ,
مثلا , بعد , ماذا , متى , اثناء , لها , له , لهم , هذا , وهذا , فهذا , فلهذا ,
هذه , وهذه , لهذه , فهذه , ولهذه , فلهذه , هذان , هذين , هاتان , هاتين ,
عليها , عليه , عليهم , سواء , سواءا , فيما , بهذا , لهذا , ولهذا , بهذا ,
حول , يمكن , تمكن , ليست , هنا , كذلك , وكذلك , لذلك , ولذلك , فذلك , لدي ,
لديه , لديها , لديهم , مهما , كان , كانت , ازاء , تلك , وما , لما , ذات ,
ذاته , ذاتها , ذاتهم , مرارا , عده , اف , يوم , يوما , اليوم , وراء , وراءه ,
وراءها , وراءهم , وراءهما , احسن , احسنه , احسنها , احسنهم , احسنهما , افضل ,
افضله , افضلها , افضلهم , افضلهما , يقدر , تقدر , قادر , قادره , قادرا , متحقق ,
متحققه , متيقن , متيقنه , فعل , يفعل , تفعل , فاعل , فاعله , فعله , فعلها ,
بفعل , لفعل , وفعل , وبفعل , ولفعل , لفعله , لفعلها , لفعلهم , بفعله , بفعلها ,
بفعلهم , الفعل , بالفعل , وبالفعل , فعلا , ابان , قطع , قطعا , القطع , مؤثر ,
مؤثرا , مؤثره , المؤثر , المؤثره , ذوات , ذوي , ذوو , مقال , مقالا , المقال ,
المقالات , اوجد , وثائق , الوثائق , خاصه , يناقش , تناقش , خصوص , خصوصا , بخصوص ,
متعلق , المتعلق , متعلقه , المتعلقه , موضوع , الموضوع , يكون , سيكون , تكون ,
ستكون , لماذا , يتضمن , تتضمن , تضمن , تضمنت , اخرون , اي , اذ , تجاه , اتجاه ,