Désambiguïsation des Traductions des Requêtes dans la Recherche d'Information Translinguistique

Wiem Ben Romdhane

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université de la Manouba

Ecole Nationale des Sciences de l'Informatique

Thèse

Présentée en vue de l'obtention du diplôme de

Docteur en Informatique

Désambiguïsation des Traductions des Requêtes dans la Recherche d'Information Translinguistique

---

par :

Wiem Ben Romdhane

Soutenue le 13 /04 /2019 devant le jury composé de :

Président : Pr. Moncef Tagina, Professeur, ENSI, Univ. de Manouba
Rapporteur : Pr. Lamia Labed Jilani, Maître de Conférences, ISGT, Univ. de Tunis
Rapporteur : Pr. Kais Haddar, Professeur, FSS, Univ. de Sfax
Examinateur : Pr. Afef Kacem, Maître de Conférences, ENSIT, Univ. de Tunis
Directeur : Pr. Narjès Bellamine Ben Saoud, Professeur, ENSI, Univ. de Manouba
Encadrant : Dr. Bilel Elayeb, Maître Assistant Habilité, FSB, Univ. de Carthage

Je dédie cette thèse

À mes chers parents

À mon mari

À mon frère et mes sœurs

À toute ma famille

À mes amis

À tous ceux que j'aime.

Remerciements

Tout d'abord, je tiens à remercier les membres du jury qui m'ont honoré d'avoir accepté d'évaluer ce travail. En particulier, je remercie :

Professeur Moncef Tagina, d'avoir accepté de présider le jury de ma thèse.

Professeur Lamia Labed Jilani et Professeur Kais Haddar pour l'honneur qu'ils m'ont fait en acceptant d'être les rapporteurs de cette thèse.

Professeur Afef Kacem pour avoir accepté d'être l'examinatrice de ma thèse.

J'adresse mes plus vifs remerciements à Professeur Narjès Bellamine Ben Saoud, ma directrice de thèse, pour ses directives et ses multiples conseils. J'ai été extrêmement sensible à ses qualités humaines d'écoute et de compréhension tout au long de la réalisation de cette thèse. Ce fut un plaisir et un grand honneur de travailler sous sa direction.

J'adresse tout particulièrement ma reconnaissance à mon encadrant Dr. Bilel Elayeb. Merci pour m'avoir cadrée et formée durant les années de thèse. Je vous remercie pour tout ce que vous m'avez appris scientifiquement et humainement. Vous êtes l'exemple du « chercheur » pour moi et tu le resteras toujours. Merci pour ta grande disponibilité, ta gentillesse, ton exigence qui me fait râler.

Je tiens à remercier aussi tous les membres du laboratoire RIADI, en particulier les membres de l'équipe : Dr. Oussama Ben Khiroun, j'ai beaucoup appris de toi, Dr. Raja Ayed merci pour vos encouragements et pour le soutien moral que tu m'as apporté.

J'adresse aussi mes remerciements à Dr. Ibrahim Bounhas, Mme Nadia Soudani et Mme Wiem Lahbib, merci pour vos encouragements.

Mes remerciements vont également vers mes amies de toujours, Nour Selmi, Sana Gabsi, Rihab Loudhaief, Takwa Demni, Camiléa Merabet, Sabrin Hichri, Jihène Romdhane, Rihab Said, Manel Boujlel, Amel Jlassi, Anissa Larbi et Sarra Lasmer.

Résumé

Nous avons proposé dans cette thèse quatre approches de désambiguïsation des traductions des requêtes (TR), dans lesquelles nous avons essayé de franchir les difficultés retrouvées dans les approches existantes dans la littérature, à savoir : (i) le problème de la désambiguïsation des traductions, (ii) le manque de couverture des dictionnaires bilingues, et (iii) le manque de contexte fourni par les termes des requêtes sources. Les trois premières approches sont basées principalement sur les mesures des possibilités ainsi que la transformation probabilité-possibilité. Alors que, la quatrième approche est à base de proximité bilingue. Nos contributions sont résumées comme suite :

Premièrement, nous avons proposé une approche possibiliste de TR à base d'un dictionnaire bilingue Français-Anglais. Tout d'abord, nous avons enrichi la couverture du dictionnaire en générant les mots ainsi que leurs traductions automatiquement à partir d'un corpus parallèle. Le dictionnaire obtenu a été vérifié et enrichi via des correcteurs intelligents en ligne. En outre, nous avons gardé le maximum d'informations contextuelles existantes dans les requêtes sources pour garantir des documents pertinents. Puis, nous avons identifié, à partir de la requête source, les syntagmes nominaux afin de les traduire en unités en utilisant des patrons de traduction et un modèle de langue (trigramme).Les termes de requête source qui ne font pas partie des syntagmes nominaux sélectionnés sont traduits mot-à-mot en utilisant une approche possibiliste de traduction des mots simples.

Deuxièmement, nous avons proposé et testé une approche possibiliste discriminative de désambiguïsation de TR. Dans cette approche, la pertinence des traductions d'un terme est modélisée par deux mesures, à savoir : la pertinence possible et la pertinence nécessaire. La pertinence possible permet de rejeter les traductions non-pertinentes, alors que la pertinence nécessaire permet de renforcer les traductions non-éliminées par la possibilité.

Troisièmement, nous avons proposé et évalué une approche possibiliste hybride de TR. Cette approche présente une combinaison des deux premières méthodes. Nous

avons profité des points forts de l'approche à base de transformation probabilité-possibilité dans la traduction des syntagmes nominaux ainsi que les points forts de l'approche possibiliste discriminative dans la traduction des termes restants de la requête.

Quatrièmement, une approche de désambiguïsation de TR à base de proximité a été proposé et discuté. Dans cette approche nous avons combiné un dictionnaire bilingue Français-Anglais avec un corpus parallèle Français-Anglais (*Europarl*), pour construire un dictionnaire sémantique bilingue de contexte (*DSBC*). Cette nouvelle ressource linguistique assure un apprentissage automatique des requêtes source. Cette approche est basée sur des méthodes stochastiques, où les termes des requêtes ainsi que leurs traductions sont représentés sous forme des graphes sémantiques bilingues. Ce graphe est transformé en des chaînes de Markov, où les états sont les nœuds du graphe et les transitions sont ses arêtes. Nous avons utilisé le principe de PROX qui calcule un score de similarité sémantique (proximité) entre les termes polysémiques des requêtes sources afin d'identifier la traduction appropriée de chaque terme de la requête source.

Finalement, nous avons conçu et implémenté un outil de recherche d'information translinguistique, dénommé SPORT (Système POssibiliste de tRaduction de requêTes) afin d'évaluer nos contributions. Nous avons implémenté les interfaces de cet outil en langage Java et nous avons intégré la plateforme *Terrier* qui fournit des modèles d'indexation et d'appariement. Dans notre cas, nous avons évalué les résultats en utilisant le modèle d'appariement OKAPI BM25 ainsi que la collection de test CLEF-2003.

Mots-clés : Recherche d'information translinguistique, désambiguïsation de la traduction de requêtes, approche possibiliste à base de transformation, approche possibiliste discriminative, approche à base de proximité.

Abstract

We have proposed in this thesis four approaches for query translation (QT) disambiguation, in which we try to overcome some problems of existing approaches in the literature, namely : (i) the problem of translation disambiguation, (ii) the lack of coverage in bilingual dictionaries, and (iii) the lack of context provided by the source queries terms. The first three approaches are based on the measures of possibilities and the probability-to-possibility transformation. Whereas, the fourth approach is based on bilingual proximity. Our contributions are summarized as follows :

Firstly, we have proposed and tested a query translation approach based on probability-to-possibility transformation.This approach is based on a bilingual French-English dictionary. We start by enriching the dictionary by automatically generating words and their translations from a parallel corpus. The obtained dictionary was checked and enriched via an online intelligent proofreader. In addition, we have considered contextual information in the source query to guarantee relevant documents. Then, we have identified, from the source query, the name phrases (NPs) in order to translate them into units using translation patterns and a language model (trigram). Source query terms that are not part of the selected noun phrases are translated word-by-word using a possibilistic approach for single words translation.

Secondly, we have investigated a discriminative possibilistic approach for QT disambiguation. In this approach, the translation relevance of a term is modeled by two measures : possible relevance and necessary relevance. Possible relevance rejects irrelevant translations, while necessary relevance strengthens translations that are not eliminated by the possibility.

Thirdly, a hybrid possibilistic approach of QT disambiguation was proposed and tested. This approach is a combination of the first two methods. We have taken advantage of the strengths of the first approach based on probability-to-possibility transformation in the translation of noun phrases and the strengths of the discriminative possibilistic approach in the translation of the remaining terms of the

query.

Fourthly, a proximity-based disambiguation approach was suggested and assessed. In this approach, we have combined a bilingual French-English dictionary with a parallel French-English corpus (*Europarl*), to build a bilingual semantic dictionary of contexts (*BSDC*).This new linguistic resource ensures automatic learning of source queries. This approach is based on stochastic methods, where the query terms and their translations are represented as bilingual semantic graph. This graph is transformed into Markov chains, where the states are the nodes of the graph and the transitions are its edges. We used the PROX method which calculates a semantic similarity score (proximity) between the ambiguous terms of the source query in order to identify the suitable translation of each source query term.

Finally, we have designed and implemented a Cross-Language Information Retrieval (CLIR) tool, denoted SPORT in order to evaluate our contributions. We have implemented the SPORT interfaces using Java language and we have integrated the *Terrier* platform, which provides indexing and matching models. We have assessed the results using the OKAPI BM25 matching model and the CLEF-2003 test collection.

Keywords : Cross-language information retrieval, query translation disambiguation, possibilistic approach-based transformation, discriminative possibilistic approach, proximity-based approach.

Table des matières

Table des figures

Liste des tableaux

Liste des symboles

CLIR : Cross-language Information Retrieval

DPP : Degré de Pertinence Possibiliste

DSBC : Dictionnaire Sémantique Bilingue de Contextes

NP : Noun Phrase (groupe nominal)

RI : Recherche d'Information

RIT : Recherche d'Information Translinguistique

SPORT : Système POssibiliste de tRaduction de requêTes

SRI : Système de Recherche d'Information

SRIT : Système de Recherche d'Information Translinguistique

TR : Traduction de requêtes

WSD : Word Sense Disambiguation (désambiguïsation des sens de mots)

Introduction Générale

1. Contexte et problématique

La Recherche d'Information (RI) est un domaine de recherche complexe et multi-disciplinaire (Kadri, 2008) qui tire profit de plusieurs branches à savoir : l'informatique, le traitement automatique des langages naturels, la science cognitive, etc. Ce domaine n'est plus limité à la langue maternelle de l'utilisateur, il est devenu plus ouvert à d'autres langues, ce qui a soulevé le concept de la Recherche d'Information Translinguistique (RIT). Dans la RIT, nous étudions les principes d'identification des informations pertinentes exprimées dans une langue différente de celle de la requête à l'aide d'un Système de Recherche d'Information Translinguistique (SRIT) afin de répondre aux besoins d'un utilisateur.

Avec l'augmentation d'informations sur le Web et la disponibilité d'une grande quantité des documents non-anglais en ligne, les SRIT deviennent plus importants pour résoudre l'inadéquation entre la langue des requêtes et celle des documents. Cela peut être réalisé soit en traduisant les documents ou les requêtes, soit les deux vers une troisième langue pivot. En effet, nos travaux s'inscrivent spécifiquement dans le cadre de traduction des requêtes vers la langue des documents. Cette dernière est l'approche la plus populaire et la plus utilisée dans le domaine de RIT, grâce à son efficacité et sa facilité. Selon (Elayeb et Bounhas, 2016 ; Zhou et al., 2012) nous pouvons classifier les approches de Traduction de Requêtes (TR) en trois catégories principales : approches à base de dictionnaire, approches à base de traduction automatique et approches à base des corpus parallèles ou comparables. Dans cette thèse, nous nous sommes intéressés aux approches à base de dictionnaire grâce à leurs disponibilités pour de nombreuses langues.

Néanmoins, ces approches souffrent de nombreux problèmes tels que leurs couvertures limitées et la difficulté de choisir la traduction appropriée d'un terme donné d'une requête source. Notre principal défi est de chercher à identifier la meilleure

approche pour la traduction des requêtes. Pour le faire, nous proposons des approches de désambiguïsation de TR dans le domaine de RIT. Notre objectif serait de choisir la traduction adéquate pour améliorer la qualité des requêtes traduites, et en conséquence la qualité des documents retournés par un système performant de RIT dans le couple des langues Français-Anglais.

Par ailleurs, l'ambiguïté des traductions d'un terme d'une requête source est considérée comme un cas d'imprécision, alors que la théorie des possibilités s'applique naturellement à ce genre d'imperfection. À notre connaissance et à partir des approches de TR existantes dans la littérature, il n'existe pas une approche disponible basée sur la théorie des possibilités dans la RIT. C'est ce qui nous a motivé à pousser nos recherches dans cette direction et de proposer des nouvelles approches pour la désambiguïsation de TR dans la RIT en exploitant la théorie des possibilités, tout en profitant d'une nouvelle ressource linguistique. Cette dernière est dôtée d'une couverture qui tire profit d'une part, d'un dictionnaire traditionnel bilingue, et d'autre part, d'un corpus parallèle.

2. Contributions

La présente thèse est scindée en deux parties, dans lesquelles, nous proposons des approches de désambiguïsation de traduction de requêtes afin d'améliorer la performance d'un SRIT. Nous résumons nos approches comme suite :

Proposition des approches possibilistes de désambiguïsation de TR dans la RIT, notre contribution est triple :

Premièrement, nous proposons une approche à base d'une transformation probabilité-possibilité comme un moyen pour introduire plus de tolérance dans le processus de TR. En effet, ce type de transformation peut être approprié lorsqu'il est difficile de travailler avec des mesures probabilistes, ou lorsque nous sommes en présence de pauvres sources d'information. Tout d'abord, nous identifions à partir de la requête source les syntagmes nominaux qui gardent un maximum d'informations contextuelles. Puis, nous les traduisons en unités en utilisant des patrons de traduction et un modèle de langue. Ensuite, les termes de requête source qui ne font pas partie des syntagmes sélectionnés sont traduits mot-à-mot en utilisant une approche possibiliste de traduction des mots simples. En fait, nous avons pris en considération tous les termes de la requête et leurs traductions lorsque nous choisissons la traduction appropriée d'un mot donné. Nous partons de l'idée que les traductions appropriées des termes ont tendance de coexister dans les documents de la langue cible, au contraire de ceux qui ne conviennent pas. Afin d'augmenter la

couverture du dictionnaire bilingue utilisé, nous avons généré automatiquement ses termes et ses traductions à partir d'un corpus parallèle. Ce dictionnaire est enrichi aussi par des correcteurs intelligents en ligne (Elayeb et al., 2018).

Deuxièmement, nous proposons une approche possibiliste discriminative pour la désambiguïsation des traductions, basée sur un réseau possibiliste. Au niveau de traduction, nous avons opté à une autre technique différente de celle de la première approche, dans laquelle la pertinence de traduction d'un terme d'une requête source est modélisée par deux mesures : la pertinence possible et la pertinence nécessaire. La pertinence possible permet de rejeter les traductions non-pertinentes, alors que la pertinence nécessaire permet de renforcer les traductions non éliminées par la possibilité (Ben Romdhane et al., 2017).

Troisièmement, nous proposons une approche possibiliste hybride qui est une combinaison des deux premières approches, en profitant des avantages des deux précédentes pour surmonter certaines faiblesses des techniques existantes basées sur un dictionnaire bilingue. Les syntagmes nominaux sont traduits via l'approche à base de transformation probabilité-possibilité, alors que les termes restants de la requête source sont traduits via l'approche possibiliste discriminative (Ben Romdhane et al., 2019).

Proposition d'une approche basée sur la proximité bilingue pour la désambiguïsation de TR dans la RIT :

Dans n'importe quel processus de TR, le contexte est pertinent pour identifier la traduction appropriée pour chaque terme de la requête source. Cependant, et en dépit de leurs avantages, les dictionnaires bilingues traditionnels souffrent du manque d'informations précises utiles pour la TR. En outre, il existe un manque de corpus parallèles bilingues de haute-couverture sur lesquels les méthodes d'apprentissage pourraient être formées. Pour ces multiples raisons, il devient important d'utiliser un dictionnaire sémantique bilingue de contextes (*DSBC*) pour assurer un apprentissage dans une plate-forme sémantique de TR. Ainsi, nous proposons une approche de désambiguïsation de TR basée sur une proximité bilingue pour surmonter certaines lacunes des approches existantes (Ben Romdhane et al., 2018). Notre approche combine un dictionnaire bilingue traditionnel avec un corpus parallèle pour construire une nouvelle ressource linguistique. En outre, le *DSBC* utilise les similarités entre les mots à l'aide d'une version étendue d'une distance sémantique probabiliste existante afin de choisir la meilleure traduction correspondante à chaque terme de la requête source.

Pour valider nos approches, nous avons utilisé le corpus parallèle bilingue *Europarl* Français-Anglais et la collection de test des campagnes d'évaluation CLEF-2003 (Cross Language Evaluation Forum). Les améliorations de nos approches sont sta-

tistiquement significatives dans la majorité des cas. En effet, l'approche possibiliste à base de transformation est efficace dans la traduction des syntagmes nominaux (requêtes longues). Tandis que, l'approche discriminative semble être performante dans la traduction des mots simples (requêtes courtes). De plus, l'approche hybride semble plus confortable en utilisant les requêtes longues. Quant à l'approche à base de proximité bilingue, elle est efficace dans le cas des requêtes longues et de large collection.

3. Organisation de la thèse

Cette thèse est organisée en deux parties : La première est intitulée «*La Désambiguïsation des traductions dans La Recherche d'Information Monolingue et Translinguistique*», elle présente une synthèse de l'état de l'art et donne une vue d'ensemble sur le contexte de recherche et la problématique dans le domaine de RIT. La deuxième partie est intitulée «*Contributions*», présente nos approches proposées, les résultats, et l'évaluation expérimentale. Les deux parties comprennent six chapitres que nous synthétisons comme suit :

Le premier chapitre est intitulé «*Désambiguïsation de requêtes dans la Recherche d'Information Monolingue*», présente une courte revue sur la RI Monolingue, en décrivant les concepts de base, les modèles existants et le processus de fonctionnement d'un SRI. Nous présentons aussi la désambiguïsation des requêtes dans la RI Monolingue, en détaillant les approches existantes de désambiguïsation des sens de mots ainsi que les problèmes qu'elles confrontent.

Dans le deuxième chapitre, intitulé «*Désambiguïsation de requêtes dans La Recherche d'Information Translinguistique*» nous étudions, d'abord, les principales approches existantes de désambiguïsation de TR dans la RIT. Ensuite, nous discutons une revue de littérature des différents SRI Translinguistiques. Enfin, nous présentons une étude comparative des SRIT étudiés.

Le troisième chapitre, intitulé «*Concepts de base des théories des probabilités et des possibilités*», est consacré à définir les théories des probabilités, des possibilités et les transformations probabilité-possibilité. Nous détaillons les méthodes de calcul dans chaque transformation. Nous présentons aussi une étude comparative entre ces deux théories afin de profiter de leurs avantages dans nos contributions.

Dans le quatrième chapitre, intitulé «*Approches possibilistes proposées pour la désambiguïsation des traductions de requêtes*», nous proposons trois approches à base de la théorie des possibilités pour la désambiguïsation de TR. Nous commençons par la première approche à base de transformation probabilité-possibilité

pour améliorer les approches de TR à base des dictionnaires, en nous focalisant sur la traduction des syntagmes nominaux. La deuxième approche est une approche possibiliste discriminative de désambiguïsation de TR basée sur un réseau possibiliste. La troisième approche est une approche possibiliste hybride qui combine les deux premières approches afin de profiter des avantages des deux.

Le cinquième chapitre, intitulé «*Validation des approches possibilistes pour la désambiguïsation des traductions de requêtes*», récapitule les expérimentations effectuées pour tester et évaluer les approches proposées dans le chapitre 4. Nous introduisons aussi le cadre du travail en dérivant toutes les ressources exploitées. De plus, nous détaillons, dans ce chapitre, toutes les fonctionnalités fournies par notre SRIT proposé dans lequel nous avons intégré nos approches.

Le sixième et dernier chapitre, intitulé «*Proposition et validation d'une approche à base de proximité bilingue pour la désambiguïsation des traductions de requêtes*», est divisé en deux grandes parties : dans la première partie, nous proposons une approche automatique à base d'une proximité bilingue pour la désambiguïsation de TR. Dans la deuxième partie, nous présentons les tests et les résultats retenus de cette approche.

En conclusion, un bilan de nos travaux met en pages nos contributions. Nous terminons par la proposition des perspectives possibles à ces travaux de recherche.

Première partie

LA DÉSAMBIGUÏSATION DES TRADUCTIONS DANS LA RECHERCHE D'INFORMATION MONOLINGUE ET TRANSLINGUISTIQUE

Chapitre

1 Désambiguïsation de requêtes dans la Recherche d'Information Monolingue

Sommaire

Introduction

Le concept de Recherche d'Information est apparu suite à l'augmentation exponentielle du volume d'information électronique échangée et stockée dans le monde. Ces informations peuvent être sous forme d'un texte, une vidéo ou une image. La RI est un domaine de l'informatique qui fournit un ensemble de modèles et de systèmes appelés « Systèmes de Recherche d'Information (SRI) ». L'objectif est de retrouver les informations pertinentes dans les documents afin de répondre au besoin d'un utilisateur représenté sous forme d'une requête. En outre, avec l'avènement

du Web, la tâche de désambiguïsation est devenue une nécessité dans les applications automatiques de traitement de langues, à savoir la recherche d'information, la traduction automatique, le Web sémantique, etc.

Ce chapitre est organisé comme suit : nous présentons dans la section 1.1 une courte revue sur la RI classique, en décrivant les concepts de base du domaine de RI Monolingue, le processus de fonctionnement d'un SRI, les étapes d'évaluation d'un SRI, les différents modèles existants dans la RI et enfin les mesures d'évaluation de performance des SRI. Nous présentons aussi dans la section 1.2 un état de l'art sur la désambiguïsation dans la RI Monolingue, en détaillant les différentes approches de désambiguïsation des sens de mots

1.1 La Recherche d'Information Monolingue

1.1.1 Les Concepts de base

La recherche dans un SRI consiste à comparer les mots-clés d'une requête utilisateur avec le contenu d'une base de documents, pour trouver l'information pertinente qui répond au besoin de l'utilisateur. À partir de cette définition, les termes fondamentaux dans un SRI sont : *la requête*, *les documents* et *la notion de pertinence*, à savoir :

La requête : l'utilisateur exprime son besoin d'information sous forme d'une requête qui peut être sous forme de mots-clés ou d'expressions écrites en langage naturel. La soumission de cette requête déclenche un processus de recherche documentaire.

Les documents : une collection ou un corpus, sont un ensemble de documents qui contiennent des informations exploitables par l'utilisateur. Un document peut être sous forme d'un texte, vidéo, image, etc. Il peut être aussi structuré, semi-structuré ou non structuré.

La pertinence : la pertinence est une notion fondamentale et cruciale dans le domaine de la RI (Brini, 2005). La pertinence est classée en deux types : le premier est *la pertinence système* dans laquelle l'évaluation se fait automatiquement par un système, et le deuxième est *la pertinence utilisateur* qui évalue chaque document récupéré.

1.1.2 Représentation d'un SRI

Le SRI suit plusieurs étapes pour aboutir à la correspondance entre la requête (Module de requête) et le corpus documentaire (Module de documents), pour retrouver

l'information pertinente qui satisfait au mieux l'utilisateur (figure 1.1).

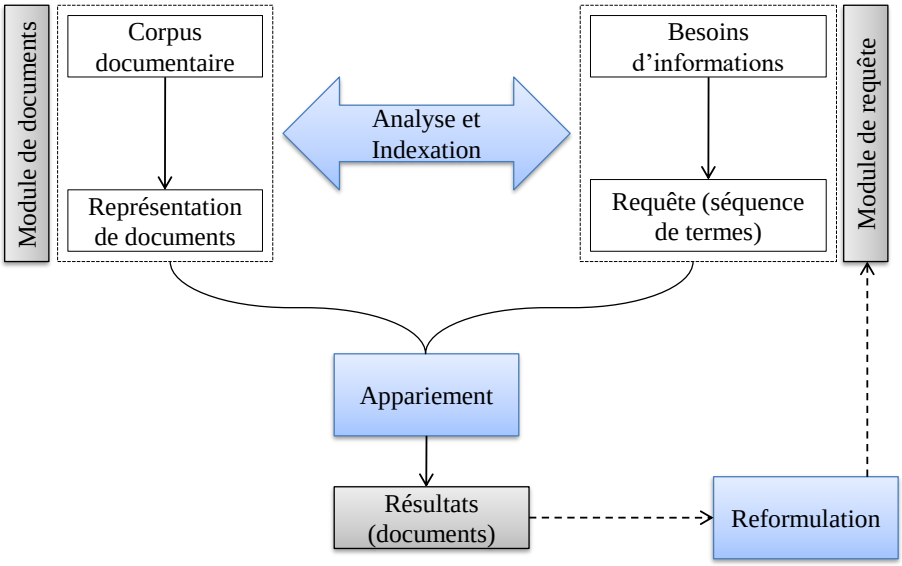

Figure 1.1 – Système de Recherche d'Information

Nous résumons le fonctionnement d'un SRI en trois étapes principales : l'indexation (cf. section 1.1.2.1), l'appariement (cf. section 1.1.2.2) et la reformulation (cf. section 1.1.2.3).

1.1.2.1 L'indexation

L'indexation facilite aux SRI l'accès rapide aux différents documents de la collection. L'objectif de cette étape est de fournir des représentations des documents et des requêtes facilement exploitables par le système dans la phase de recherche (Znaidi, 2016). Les représentations peuvent être sous forme d'une liste de mots pondérés les plus discriminants, appelés aussi des index. Le processus d'indexation peut être fait de trois manières, à savoir : manuelle, automatique ou semi-automatique.

- *L'indexation manuelle* est faite par un groupe de documentalistes et des spécialistes du domaine qui sont censés analyser les documents et de former une liste des termes jugés les plus significatifs.
- *L'indexation automatique* est la plus utilisée dans les SRI, dans cette technique, la détermination des descripteurs de documents est totalement informatisée.
- *L'indexation semi-automatique* est une indexation automatique qui nécessite une intervention humaine pour la validation finale des descripteurs.

Généralement, l'étape d'indexation est accompagnée par un ensemble d'analyses et de traitements nécessaires appliqués sur les documents et les requêtes, comme la lemmatisation, la suppression des mots vides et la pondération.

1.1.2.2 L'appariement

L'appariement est une étape indispensable qui évalue la pertinence des documents par rapport à la requête saisie. C'est une correspondance requête-document qui associe à chaque document un poids de pertinence qui est le degré de similarité entre la requête et le document. Par conséquent, le résultat est une liste de documents triés par ordre de valeur de correspondance du plus pertinent au moins pertinent.

1.1.2.3 La reformulation de requêtes

Cette étape a pour objectif d'enrichir la requête, pour générer une nouvelle qui exprime mieux le besoin de l'utilisateur. Il existe trois types de reformulation (Bouramoul, 2011) (figure 1.2), à savoir : (i) *la reformulation interactive* connue aussi par la réinjection de la pertinence ou « relevance feedback », dans cette technique l'utilisateur sélectionne, à partir des documents retournés jugés pertinents, les termes les plus expressifs à son besoin et les ajoute à sa requête initiale. (ii) *La reformulation automatique*, elle s'agit d'une expansion de requête qui s'appuie sur les termes des documents pertinents et les ajoute dans la requête initiale. (iii) *La combinaison des présentations de la requête* ; Lee (1998) a prouvé que la recherche est plus efficace en utilisant des représentations multiples des requêtes et en exploitant différents algorithmes de recherche et techniques de réinjection.

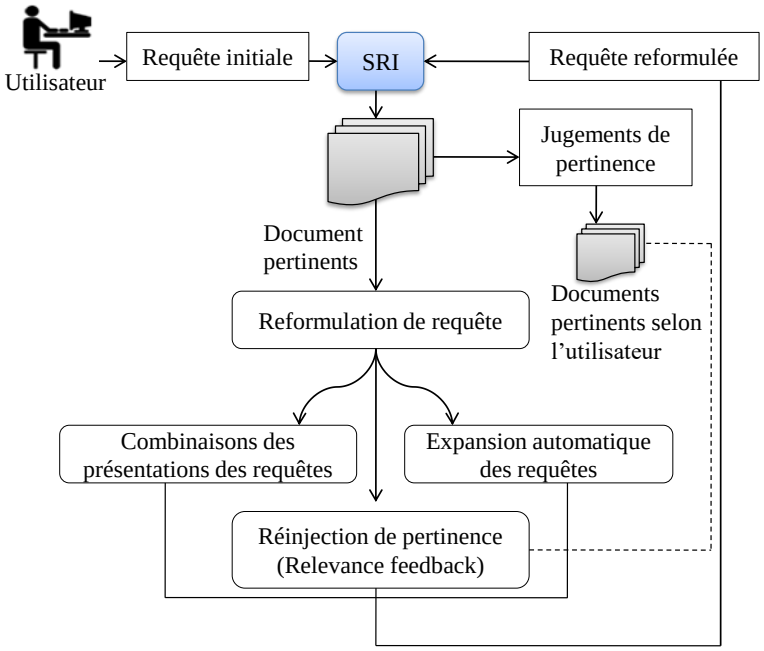

Figure 1.2 – Techniques d'améliorations des SRI par reformulation de requêtes (Bouramoul, 2011)

1.1.3 Les modèles de RI

Nous présentons, dans cette section, les différents modèles théoriques utilisés par les SRI afin de mettre une correspondance requête-documents pour retrouver l'ensemble des documents pertinents qui répondent aux besoins de l'utilisateur. Généralement, les fondements mathématiques sur lesquels se basent les modèles de la littérature reposent sur l'utilisation de l'algèbre, la théorie des ensembles, la théorie des probabilités, la théorie des possibilités et des statistiques.

Nous présentons dans la figure 1.3, une taxonomie des modèles existants en RI. Cette taxonomie est étendue de la classification proposée par (Baeza-Yates et Ribeiro-Neto, 2011).

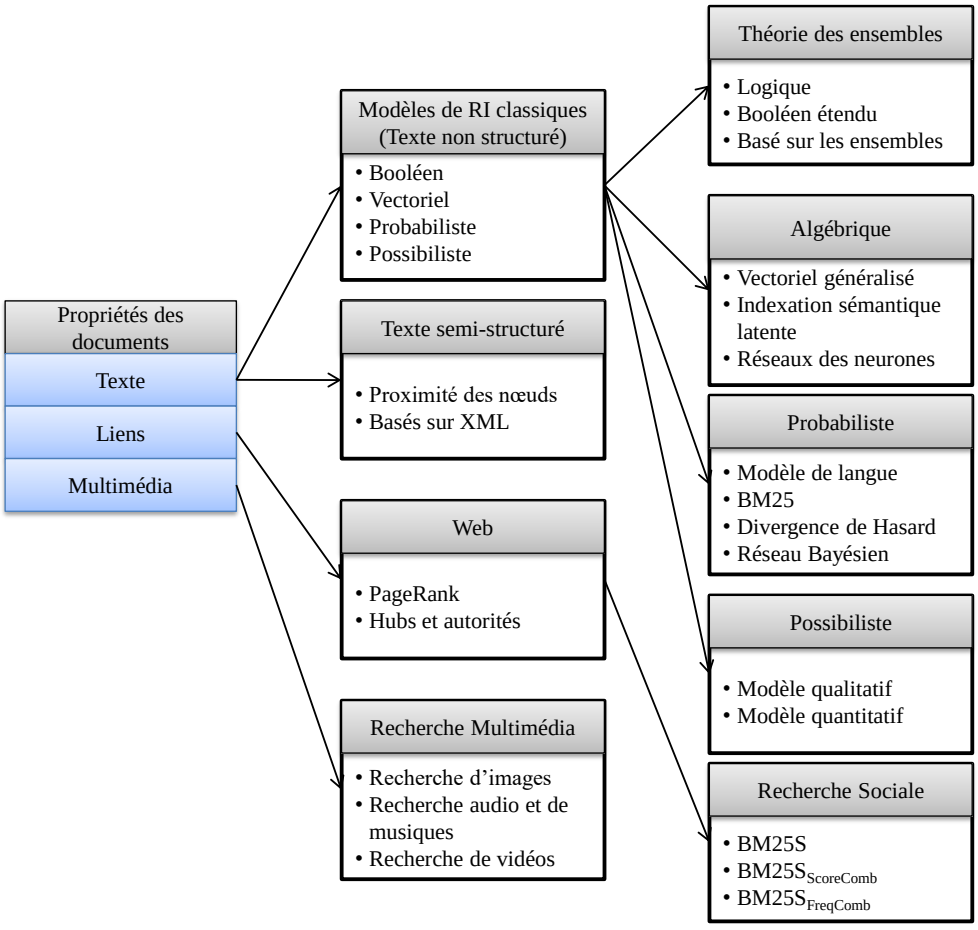

Figure 1.3 – Taxonomie des modèles de RI

Nous détaillons dans ce qui suit, les principaux modèles apparus :

Le modèle booléen (Salton et al., 1983) est le premier modèle utilisé, il est fondé sur la théorie des ensembles et l'algèbre de Boole. Dans ce modèle, les documents et les requêtes sont représentés comme des termes reliés par des opérateurs booléens

≪AND≫, ≪OR≫ et ≪NOT≫. La pertinence dans ce modèle est binaire, soit vraie ou fausse.

Le modèle vectoriel a été proposé par (Salton, 1971) dans le cadre du système SMART [1]. Dans ce modèle, les documents et les requêtes sont représentés par des vecteurs de poids des termes. Le poids dans un vecteur peut être un nombre d'occurrences d'un terme dans une requête ou un document.

Le modèle probabiliste (Fuhr, 1992) est adopté pour calculer les scores de pertinence des documents en utilisant la théorie des probabilités. Il existe des extensions du modèle probabiliste, tels que : le modèle probabiliste général, les modèles basés sur les réseaux Bayésiens (Fung et Del Favero, 1995), le modèle de langue, le modèle inférentiel (Turtle et Croft, 1989), le modèle de pondération BM25 (Robertson et al., 1995), INQUERY(Callan et al., 1992), etc.

Le modèle possibiliste est basé sur la théorie des possibilités (Dubois et Prade, 1992). Cette théorie offre deux cadres de travail : un cadre qualitatif ou ordinal et un cadre quantitatif. Le modèle de (Brini, 2005) s'inscrit dans le cadre quantitatif, alors que (Elayeb, 2009) propose un modèle possibiliste mixte qui tient en compte l'aspect quantitatif et l'a étendu vers un cadre qualitatif.

Le modèle de recherche d'information sociale est proposé par (Bouhini et al., 2014). Ce modèle intègre le contexte social des utilisateurs, construit à partir des annotations des réseaux sociaux. Bouhini et son équipe ont proposé trois modèles basés sur $BM25$, ils sont définis comme suit : (i) le modèle $BM25S$ retourne une liste classée de documents pertinents, en utilisant seulement le profil social de l'utilisateur. (ii) Le modèle $BM25S_{ScoreComb}$ retourne une liste classée des documents pertinents pour un utilisateur, en considérant sa requête initiale combinée au niveau de score avec son profil social. Et (iii) le modèle $BM25S_{FreqComb}$ retourne une liste classée de documents pertinents pour un utilisateur, en combinant sa requête au niveau des fréquences de termes avec son profil comme recommandé par (Robertson et al., 2004).

Le modèle à base des réseaux de neurones est un modèle d'appariement où les documents et les requêtes sont présentés à l'aide d'un réseau de neurones. (Huang et al., 2013) ont essayé de développer une série de modèles sémantiques structurés, formés de manière discriminante en maximisant la vraisemblance conditionnelle des documents. Dans le même contexte, (Severyn et Moschitti, 2015) ont présenté une architecture du réseau neuronal convolutionnel pour classer les paires de textes courts, où ils apprennent la représentation optimale de paires de textes et une fonction de similarité pour les relier de manière supervisée. Au-delà, (Nguyen et al., 2017)

1. SMART (System for the Mechanical Analysis and Retrieval of Text) est un système d'analyse mécanique et de recherche d'information développé à l'Université Cornell dans les années 1960.

ont conçu un modèle neuronal pour la RI ad-hoc qui exploite des représentations sémantiques latentes des documents et des requêtes.

(Brini, 2005) a répertorié les modèles étudiés en trois catégories selon leur définition de la pertinence, à savoir : (i) la pertinence est vue comme la similarité entre la requête et le document ; (ii) la pertinence est modélisée par une variable binaire et des modèles probabilistes ; (iii) l'incertitude sur la pertinence est produite par l'inférence de la requête à partir des documents ou inversement.

Par ailleurs, (Bouhini et al.,2014) ont montré que l'interaction de l'utilisateur avec le système est une composante essentielle pour améliorer la qualité de RI. Cela conduit à une recherche d'informations sociales qui tente d'étendre la RI classique en prenant en compte le profil de l'utilisateur. Par exemple lors de saisie de la requête ou lors de l'indexation et du classement des documents. Dans ce dernier cas, le classement d'un document dépend non seulement de la correspondance entre le document et la requête, mais également de la correspondance entre l'intérêt de l'utilisateur et le document.

1.1.4 Evaluation des performances en RI

L'évaluation est une étape primordiale dans le processus de RI. C'est une stratégie qui permet d'étudier les SRI et d'identifier l'impact des méthodes et des techniques employées dans les approches de recherche (Znaidi, 2016). Nous présentons, dans ce qui suit, les ressources exploitées et les métriques principales d'évaluation.

1.1.4.1 La collection de test

La collection (ou standard) de test est composé(e) de trois éléments :

Un ensemble de documents (corpus) : contient généralement un nombre important de documents dans différentes langues et dans des sujets généraux ou spécialisés.

Un ensemble de requêtes : pour avoir une évaluation objective, le système doit prendre en considération les réponses du système pour toutes les requêtes. Selon (Buckley et Voorhees, 2000), le nombre de requêtes nécessaire pour une bonne expérience est d'au moins 25 requêtes et 50 requêtes pour garantir une meilleure qualité.

Les jugements de pertinence : c'est l'identification des documents pertinents de la collection de test pour chaque requête par des utilisateurs experts. Ces jugements sont généralement binaires ; soit le document est pertinent, soit non.

1.1.4.2 Les campagnes d'évaluation

Les campagnes d'évaluation les plus connues dans ce domaine sont TREC, CLEF et NTCIR à savoir :

TREC [2](Text REtrieval Conference), ce projet a été lancé en 1992 dans le cadre du programme TIPSTER [3]. Son but est de soutenir la recherche au sein de la communauté de RI en fournissant l'infrastructure nécessaire pour l'évaluation des SRI sur différentes collections.

CLEF [4](Cross Language Evaluation Forum), ce forum a été initié en 2000, il permet de promouvoir la performance de la RI multilingue et translinguistique dans les langues européennes. CLEF est aussi devenu un forum pour plusieurs évaluations spécialisées.

NTCIR [5](NII Testbeds and Community for Information access Research), cet atelier a eu lieu depuis 1997 au Japon. C'est une série d'ateliers d'évaluation conçue pour améliorer la recherche dans les langues asiatiques pour les technologies de RI, les systèmes de question-réponse, extraction d'information et les systèmes de résumé automatique de documents.

1.1.4.3 Les métriques d'évaluation

Il existe dans la littérature des métriques standards pour l'évaluation des SRI. Nous focalisons notre attention sur les six principales mesures que nous avons utilisées dans nos expérimentations qui sont : la précision, le rappel, la précision moyenne MAP (Mean average Precision), la précision exacte (R-Précision), les mesures de précision aux top documents ($P@5, P@10, P@15, \ldots, P@1000$) et le pourcentage d'amélioration.

Le rappel et la précision L'évaluation suppose qu'il existe un ensemble idéal que le système est censé chercher. Le rappel et la précision sont définis comme suit :

Le rappel est le pourcentage de documents pertinents retrouvés par le système parmi tous les documents pertinents dans la base :

$$Rappel = \frac{Le\ nombre\ de\ documents\ pertinents\ retrouvés}{Le\ nombre\ de\ documents\ pertinents\ dans\ la\ base}$$

2. http://trec.nist.gov/
3. un projet sponsorisé par la DARPA (Defense Advanced Research Projects Agency), son but est de développer les systèmes de recherche documentaire et l'extraction d'informations (Harman, 1992).
4. http://www.clef-campaign.org/
5. http://research.nii.ac.jp/ntcir/index-en.html

La précision est le pourcentage de documents pertinents retrouvés par le système parmi tous les documents retrouvés par le système :

$$Précision = \frac{Le\ nombre\ de\ documents\ pertinents\ retrouvés}{Le\ nombre\ de\ documents\ retrouvés}$$

Le nombre total des documents retrouvés est limité aux 1000 premiers, c'est une valeur fixée par les programmes d'évaluation utilisés (Baccini et al., 2010). Nous avons utilisé la précision (axe des Y) et le rappel (axe des X) sur 11 points $(0.0, 0.1, \ldots, 1.0)$ pour dessiner *les courbes de rappel-précision*.

Les deux mesures ne sont pas indépendantes, elles varient en sens inverse, quand l'une augmente l'autre diminue. Un système idéal qui restitue tous les documents pertinents a pour rappel égal à 100 et tous les documents récupérés sont pertinents (précision =100%).

La précision moyenne (MAP) Nous avons également utilisé la MAP (Mean Average Precision), qui est la moyenne des scores de précision moyenne sur un ensemble de 11 points de rappel. Elle est considérée comme une mesure de performance sur un ensemble de n requêtes, et elle se focalise principalement sur les documents pertinents classés dans les premiers rangs. La formule de calcul du MAP est la suivante (Kompaoré, 2008 ; Manning et al., 2008) :

$$MAP = \frac{1}{n} \sum_{Q_i} \frac{1}{|R_i|} \sum_{D_j \in R_i} \frac{i}{r_{ij}} \tag{1.1}$$

avec :

r_{ij} : correspond au rang du $j - ième$ document pertinent pour la requête Q_i ;

$|R_i|$: indique le nombre de documents pertinents pour la requête Q_i ;

n : correspond au nombre de requêtes de test.

Nous présentons dans le tableau 1.1, un simple exemple illustratif pour le calcul de MAP.

Q_1	Q_2	Commentaires
1	4	Rang du premier document pertinent
5	8	Rang du deuxième document pertinent
10		Rang du troisième document pertinent

Table 1.1 – Listes des documents restitués par un système en réponse aux requêtes Q_1 et Q_2 (Kompaoré, 2008)

La MAP vaut alors :

$MAP = \frac{1}{2} * [\frac{1}{3}(\frac{1}{1} + \frac{2}{5} + \frac{3}{10}) + \frac{1}{2}(\frac{1}{4} + \frac{2}{8})] = 0,4083.$

la précision exacte (R-Précision) La précision exacte (R-Précision) est la précision au rang R ; où R est le nombre total des documents pertinents pour la requête considérée. La formule de calcul du R-Précision est la suivante :

$$R - Précision = \frac{1}{n}\sum_{j=1}^{n} Precision\left(\{D_{kj}\}\right) \tag{1.2}$$

Où :

$Precision\left(\{D_{kj}\}\right)$ correspond à la précision des $k - premiers$ documents retournés pour la requête Q_j ;

k est le nombre des documents pertinents retournés pour la requête Q_j.

Les mesures de précision aux top documents P@X Les mesures de précision aux top documents permettent d'examiner les listes restituées par le système à différents niveaux de coupe (Moulahi, 2015). La précision est calculée en fonction de la valeur de X premiers documents, cette valeur peut être fixée à $5, 10, 15$ ou 100 documents. Cette mesure évalue le système pour retrouver un maximum de documents pertinents pour un faible nombre de documents récupérés. Par exemple, pour une requête Q, une précision à 5 documents notée $P@5$, un système qui restitue 3 documents pertinents aura une valeur de $P@5$ égale à 3/5.

Le pourcentage d'amélioration Parfois, nous sommes appelés à mesurer le pourcentage d'amélioration entre deux variations de variables du modèle. Ce pourcentage est généralement obtenu pour deux variables A et B mesurant le pourcentage de C

$$\%C = \frac{B - A}{A} * 100 \tag{1.3}$$

1.2 La désambiguïsation en RI Monolingue

Les SRI souffrent de nombreux défis, surtout liés aux requêtes des utilisateurs. En fait, ces derniers expriment principalement leurs besoins via des courtes requêtes, qui peuvent également contenir des termes ambigus. Par conséquent, les résultats des SRI peuvent inclure plusieurs documents non-pertinents (bruit), en raison du contexte limité fourni des requêtes. Ce bruit diminue l'efficacité de la recherche et ouvre les portes aux deux problèmes à résoudre, à savoir : l'expansion sémantique et la désambiguïsation sémantique des requêtes. Nous nous focalisons, dans cette étude, sur la désambiguïsation des requêtes.

Les langues naturelles contiennent plusieurs mots ambigus, chaque mot a différentes significations et plusieurs sens. L'identification du bon sens dépend purement du contexte dans lequel le mot apparaît. Donc, il est crucial d'éliminer l'ambiguïté des mots, en choisissant le sens le plus approprié pour chaque mot dans un texte donné.

La désambiguïsation des sens de mots (WSD : Word Sense Disambiguation) est considérée comme une étape intermédiaire, qui ne constitue pas une fin en soi, mais qui est plutôt nécessaire à un niveau ou à un autre pour accomplir la plupart des tâches de traitement des langues (Billami, 2015).

La résolution de l'ambiguïté améliore la qualité des diverses applications de traitement automatique de langue (TAL), telle que la recherche d'information, la traduction automatique, la recherche d'information translinguistique, la catégorisation et l'extraction de connaissances, la classification des entités nommées, etc.

Différentes approches de désambiguïsation ont été proposées dans la littérature. Selon (Audibert, 2003a ; Elayeb et al., 2015 ; Ide et Véronis, 1998 ; Navigli, 2009 ; Ranjan Pal et Saha, 2015 ; Sarmah et Sarma, 2016), les approches de WSD peuvent être classées en quatre catégories, à savoir : les approches de modélisation des connaissances ou du raisonnement, les approches à base de dictionnaire, les approches à base de corpus et les approches hybrides. Nous présentons, dans ce qui suit, un survol sur ces différentes catégories.

1.2.1 Approches à base de modélisation des connaissances ou du raisonnement

Ce type d'approches est basé sur la modélisation et la compréhension du langage humain à travers des modèles connexionnistes ou symboliques bien connus en Intelligence Artificielle. Les approches à base des connaissances reposent sur différentes sources de connaissance, telles que les dictionnaires lisibles par machine, les inventaires de sens, les thésaurus, etc.

(Masterman, 1961) a utilisé les «*réseaux sémantiques*» pour la modélisation des concepts de langage. Les distinctions de sens sont implicitement faites en choisissant des représentations qui reflètent des groupes de nœuds étroitement liés dans le réseau. Ce réseau est construit à partir des définitions d'un dictionnaire, puis est enrichi manuellement. Étant donné deux mots présentés au réseau, si plusieurs concepts sont rattachés à un même mot, la désambiguïsation se fait en choisissant le concept intervenant dans le plus court chemin reliant ces deux mots (Audibert, 2003a).

(Banerjee et Pedersen, 2002) ont adapté *l'algorithme de Lesk* (Lesk, 1986) pour la WSD. Au lieu d'utiliser un dictionnaire standard comme une source glossaire,

les auteurs ont employé la base de données lexicale *WordNet*[6]. Cette méthode est évaluée en utilisant les données de l'échantillon lexical anglais de l'exercice Senseval-2 (Kilgarriff, 1998). Le résultat a donné une précision globale de 32%, cela représente une amélioration significative par rapport aux 16% et 23% de précision obtenue par les variations de l'algorithme de Lesk, qui est utilisé comme un repère pour les systèmes de WSD, au cours de l'exercice comparatif Senseval-2.

Pour l'acquisition de connaissances, (Ponzetto et Navigli, 2010) ont présenté une méthodologie pour étendre automatiquement la base de données lexicale *WordNet* avec des grandes quantités de relations sémantiques. Les pages sont extraites à partir de la ressource encyclopédique *Wikipédia*. Les pages de *Wikipédia* sont automatiquement associées au sens de *WordNet* et les relations topiques et associatives sémantiques de *Wikipédia* sont transférées à *WordNet*, en produisant une ressource lexicale beaucoup plus riche. Plus tard (Navigli et Ponzetto, 2012) ont présenté une approche multilingue pour la WSD. Cette approche exploite *BabelNet*[7], pour effectuer une désambiguïsation à base des graphes dans différentes langues.

(Zhong et Ng, 2010) ont présenté *IMS* (It Makes Sense) : un système supervisé de WSD pour tous les mots anglais. Grâce à sa flexibilité, *IMS* permet aux utilisateurs d'intégrer différents outils de prétraitement, des fonctionnalités supplémentaires et des différents classifieurs. Par défaut, les auteurs ont utilisé des machines vectorielles de support linéaire, comme un classifieur avec plusieurs fonctionnalités basées sur la connaissance.

(Agirre et al., 2014) ont présenté un algorithme à base des promenades aléatoires sur les relations de grandes bases de connaissances lexicales. L'algorithme utilise efficacement le graphe complet de *WordNet*. Les résultats ont montré que cet algorithme est plus performant que *le PageRank* ou *les sous-graphes* (Navigli et Lapata, 2010 ; Ponzetto et Navigli, 2010).

(Chaplot et Salakhutdinov, 2018) ont proposé un algorithme de WSD à base de connaissances. Cet algorithme est capable d'utiliser un document en entier comme un contexte pour la désambiguïsation d'un mot. Pour la modélisation d'un document, les auteurs ont tiré parti du formalisme des modèles de sujet, en particulier de LDA (Latent Dirichlet Allocation) (Blei et al., 2003). Ils ont extrait aussi les connaissances à partir de *WordNet*. Les évaluations sont faites sur les campagnes d'évaluation Senseval-2, Senseval-3, SemEval-2007, SemEval-2013 et SemEval-2015. Les résultats ont montré que ce système surpasse les systèmes étudiés pour la désambiguïsation à base de connaissances.

6. https://wordnet.princeton.edu/

7. C'est une très grande base de connaissances multilingues, contenant entre 13 et 14 millions d'entrées dans 271 langues. Elle est considérée comme un réseau sémantique multilingue et une ontologie lexicalisée

1.2.2 Approches à base de dictionnaire

Le dictionnaire informatisé est un élément de recherche en désambiguïsation lexicale. L'idée principale de ce type d'approche est : lorsque plusieurs mots coexistent, le sens le plus probable pour chacun de ces mots est celui qui maximise sa similarité avec les sens des mots co-occurrents.

Lesk (1986) est parmi les premiers qui a construit un système de désambiguïsation à base de dictionnaire lisible par machine (MRD : Machine Readable Dictionnary). Il a créé une base de connaissances qui associe à chaque sens, une liste de mots qui apparaissent dans sa définition. L'algorithme a été évalué sur un échantillon des paires de mots ambiguës, une précision de $50-70\%$ a été observée en utilisant des exemples courts du roman «*Pride et Prejudice*» et des articles de presse «*Associated Press*».

(Veronis et Ide, 1990) ont étendu la méthode de Lesk pour construire automatiquement un grand réseau de neurones à partir des définitions de dictionnaire anglais *Collins (CED : Collins English Dictionary*) et aussi pour démontrer l'utilité de ses réseaux pour la WSD. Ces réseaux relient les mots à ses sens et eux-mêmes reliés aux mots apparaissant dans leur définition. La méthode a obtenu un taux de précision de 71.1%.

(Wilks et al., 1990) ont utilisé la méthode de Leks en utilisant le dictionnaire *LDOCE*[8] (*Longman Dictionary of Contemporary English*) afin de résoudre le problème des définitions courtes de Lesk (Boubekeur, 2008), en enrichissant manuellement toutes les définitions et en calculant la fréquence de co-occurrence des mots dans ces définitions.

(Ellman et al., 2000) ont décrit une approche simple de WSD en utilisant le filtrage et l'extraction des sens. L'approche exploite et étend les informations disponibles dans le dictionnaire anglais *HECTOR* de *SENSEVAL* pour éliminer les acceptions inappropriées au contexte. L'algorithme applique plusieurs filtres pour élaguer l'ensemble des sens de mots candidats les plus fréquents.

(Brun et al., 2001) ont présenté un système de désambiguïsation lexicale sémantique. Au début, ce système a été conçu pour l'anglais, puis adapté au français. Le système a utilisé le dictionnaire OHFD[9] (*Oxford-Hachette French Dictionary*) comme un corpus sémantiquement étiqueté pour extraire une base de règles de désambiguïsation sémantique, pour relier chaque mot à son sens compte tenu de son contexte.

(Soudani et al., 2014a) ont proposé une approche générique de normalisation des

8. http://ldoce.longmandictionariesonline.com/main/Home.html
9. https://global.oup.com/academic/product/oxford-hachette-french-dictionary-9780198614227?cc=tn&lang=en&

dictionnaires Arabes et de construction d'une ontologie arabe. En première étape, les auteurs ont commencé par la transformation des dictionnaires arabes non structurés (Al-Nihaya fi Gharib Al-Hadith w Al-Athar), qui sont extraits de Hadith [10], en des dictionnaires normalisés en format LMF [11]. En deuxième étape, ils ont construit l'ontologie de Hadith pour désambiguïser les termes de la requête de l'utilisateur.

1.2.3 Approches à base de corpus

Les approches à base de corpus sont généralement de type statistique. Elles utilisent des corpus sémantiquement annotés pour former des algorithmes d'apprentissage automatique. Les corpus utilisés sont de grandes tailles pour pouvoir construire une large connaissance pour la désambiguïsation. Ces approches se divisent en des approches supervisées qui s'appuient sur des corpus manuellement étiquetés et des approches non supervisées qui s'affranchissent de cette limitation (Boubekeur, 2008).

(Schutze, 1992) a décrit un ensemble d'approches non supervisées basées sur la notion de regroupement (*clustering*) contextuel. Chaque occurrence d'un mot donné dans un corpus est représentée comme un vecteur de contexte. Les vecteurs sont ensuite regroupés en clusters, et chaque cluster définit un sens du mot.

(Reymond, 2001) a proposé la construction d'un ≪*dictionnaire distributionnel*≫ de façon incrémentale à partir d'un corpus de référence. Ce dictionnaire comporte au début 60 vocables polysémiques : de 20 noms communs, 20 verbes et 20 adjectifs, représentant environ 60000 occurrences de ces 60 termes dans un corpus du projet *SyntSem* [12].

Parmi les difficultés de désambiguïsation : l'inadéquation des dictionnaires classiques, le manque des corpus lexicalement désambiguïsés utiles pour les expérimentations des approches d'apprentissage supervisé, et aussi l'absence d'un corpus dans la langue française. Pour ces multiples raisons, (Audibert, 2003a) a étudié les critères de désambiguïsation basés sur les co-occurrences. Cette étude est effectuée, dans le cadre du projet *SyntSem,* sur un corpus français étiqueté sémantiquement (Audibert, 2003b).

(Véronis, 2003) a proposé un algorithme *HyperLex* basé sur les propriétés des graphes constitués par les co-occurrences de mots. L'algorithme permet à partir d'une grande base textuelle de détecter automatiquement les composants de haute

10. http://shamela.ws/index.php/book/23691
11. http://www.lexicalmarkupframework.org/
12. SyntSem est composé de texte de genre varié comportant 6 millions de mots qui ont été étiquetés manuellement (Reymond, 2002). Le projet SyntSem vise à produire un corpus français d'amorçage étiqueté au niveau morphosyntaxique, lemmatisé et comportant un étiquetage syntaxique peu profond (shallow parsing en Anglais) ainsi qu'un étiquetage lexicale de 60 mots-cibles.

densité et permet d'isoler les usages (sens) de fréquence très faible. Un graphe de co-occurrence est construit de sorte que les nœuds sont les mots apparaissant dans les paragraphes d'un corpus de texte, dans lequel un mot donné se produit. Une arête, entre un couple de mots, est ajoutée au graphe si ces deux mots existent dans le même paragraphe. Un poids est attribué à chaque arête en fonction de la fréquence de co-occurrence relative des deux mots reliés par cette arête.

(Ben Khiroun et al., 2014) ont proposé une approche possibiliste pour étudier l'impact de la désambiguïsation des sens de mots sur l'expansion des requêtes. L'approche a été appliquée sur la langue française, et elle est également applicable à d'autres langues. Les auteurs ont opté pour la désambigüisation des graphes de co-occurrence, extraits à partir de la collection de test *ROMANSEVAL*[13], afin de modéliser des liens de similarité et de contexte.

(Basile et al., 2014) ont décrit un algorithme non supervisé pour la WSD qui étend la méthode de Lesk (Lesk, 1986). *L'algorithme de Lesk* exploite l'idée du nombre maximal des mots partagés (un maximum de chevauchement) entre le contexte d'un mot et chaque définition de ses sens (glose). La contribution principale de l'approche de Basile et son équipe, repose sur l'utilisation d'une fonction de similarité distributionnelle des mots par rapport aux contextes. Les mots sont définis sur un espace sémantique distributionnel, pour calculer le chevauchement glose-contexte, en utilisant les réseaux sémantiques *BabelNet* et *MultiWordNet*[14]. Les expérimentations sont faites sur deux langues : anglaise et italienne. Les espaces sémantiques des deux langues sont construites en utilisant un code propriétaire reposant sur deux index *Lucene*[15] qui contiennent des documents de «*British National Corpus*» (BNC) pour l'anglais et de *Wikipédia* pour l'italien.

Les modèles basés sur le regroupement contextuel traditionnel, nécessitent généralement un réglage minutieux des paramètres, car ils sont généralement moins performants pour les sens des mots peu fréquents. À cet effet, (Chen et al., 2015) ont présenté une approche qui aborde ces limitations, en incorporant la composition des glossaires *WordNet* et du modèle basé sur le regroupement contextuel, pour apprendre des représentations distribuées des sens de mots. Les représentations obtenues sont évaluées sur deux tâches : la première tâche est la similarité des mots et la deuxième est le raisonnement analogique fourni par *WordRep* (Gao et al., 2014).

13. http://www.lpl.univ-aix.fr/projects/romanseval
14. http://multiwordnet.fbk.eu/english/home.php
15. https://lucene.apache.org/core/3_0_3/fileformats.html

1.2.4 Approches hybrides

Les approches hybrides combinent des connaissances extraites de ressources lexicales avec des informations contextuelles/distributives apprises à partir des corpus. Ces approches ont atteint des taux de réussite encourageants en termes de précision de désambiguïsation.

(Mihalcea et Moldovan, 1998) ont présenté une approche basée sur l'idée de densité sémantique. La désambiguïsation se fait dans le contexte de *WordNet*. L'Internet est utilisé comme un corpus brut pour fournir des informations statistiques sur les conjonctions de mots. Une métrique est introduite et utilisée pour mesurer la densité sémantique et pour classer toutes les combinaisons possibles des sens de deux mots.

(Stevenson et Wilks, 2001) ont présenté un système combinant plusieurs sources d'information : un filtrage morphosyntaxique basé sur l'étiquetage, des collocations générées à partir d'un corpus, un chevauchement avec les définitions de dictionnaire *LDOCE*, des catégories de sujets et des restrictions de sélection. Les auteurs ont achevé des résultats impressionnants, avec une précision moyenne de 65.24%, une précision de désambiguïsation de 95% au niveau des homographes de *LDOCE* et de 90% au niveau des sens.

(Barathi et Valli, 2010) ont étudié l'expansion de requêtes à base d'ontologie, pour améliorer le rappel et la précision des résultats de la RI sémantique, en se concentrant sur le contexte des concepts. Ce travail est basé sur la désambiguïsation, le thésaurus *WordNet* et une ontologie du domaine médical, pour extraire des concepts basés sur la mesure de similarité et les relations sémantiques entre eux. La requête de l'utilisateur est enrichie avec le concept sélectionné et transmise au moteur de recherche pour une recherche efficace des pages Web pertinentes.

(Merhbene et al., 2014) ont proposé une approche basée sur les arbres sémantiques pour la désambiguïsation semi-supervisée de la langue arabe. La partie supervisée de cette approche utilise un corpus de 123.854.642 mots arabes et le dictionnaire «*Alwassit*[16]» comme ressources pour classifier les contextes des mots ambigus selon le sens. Les résultats ont montré que cette approche a obtenu un taux de rappel et de précision d'environ 83%.

Récemment (Elayeb et al., 2015) ont présenté une approche hybride qui combine des dictionnaires traditionnels avec des corpus étiquetés pour construire un *DSC* «Dictionnaire Sémantique de Contexte» et identifier le sens d'un mot donné en

16. C'est un dictionnaire contemporain de la langue arabe. Ce dictionnaire contient les définitions des milliers de termes et de mots scientifiques, techniques, des nouveaux mots de la langue arabe.

utilisant un modèle d'appariement possibiliste. Ils ont proposé aussi une deuxième approche probabiliste automatique pour la WSD. Cette dernière utilise et étend une distance sémantique probabiliste existante pour calculer les similarités entre les mots, en exploitant le *DSC* et un graphe sémantique d'un dictionnaire traditionnel. Les auteurs ont évalué et comparé ces deux approches avec 5 autres systèmes monolingue de WSD (Segond, 2000). Ces systèmes sont similaires, ils s'intéressent à la langue française et utilisent la même collection de tests *ROMANSEVAL*. Les expérimentations sont faites sur 3 catégories grammaticales (nom, verbe, adjectif), en utilisant la métrique *kappa* (Cohen, 1968) pour calculer le taux d'exactitude. Les résultats ont montré l'efficacité de l'approche possibiliste par rapport à toutes les autres approches. Mais l'approche probabiliste semble mieux que la possibiliste et les 5 systèmes pour la catégorie des noms.

1.2.5 Synthèse

Selon l'état de l'art que nous avons effectué sur la WSD, nous présentons à travers les tableaux 1.2 et 1.3 une synthèse des différentes approches de WSD en RI Monolingue.

En fait, les approches supervisées ont donné des bons résultats, à condition que toutes les ressources nécessaires sont disponibles. Les approches non supervisées sont moins performantes à celles non-supervisées et à base des connaissances. La performance des approches à base de dictionnaire dépend de la couverture et la qualité des dictionnaires utilisés. Cette étude nous conduit à conclure qu'il est recommandé de combiner plusieurs ressources lors de la tâche de WSD.

Conclusion

Nous nous sommes focalisés dans notre premier chapitre sur les notions et les concepts de base du domaine de RI Monolingue. Nous avons présenté brièvement les étapes principales de fonctionnement d'un SRI, et les modèles utilisés par ce dernier dans l'étape d'appariement. Nous avons présenté aussi la démarche d'évaluation d'un SRI en décrivant les campagnes et les mesures d'évaluation.

Ce chapitre nous a servi aussi à présenter la désambiguïsation des requêtes dans la RI Monolingue. Nous avons présenté les approches existantes dans la littérature qui

Type d'approche	**Description**	**Référence**	**Technique/Ressource exploitée**
Approches à base de dictionnaire	• Les dictionnaires sont conçus et organisés autour de la notion de sens et de mot, ils représentent un outil parfaitement adapté pour la sélection du sens contextuel d'un mot. Et, le découpage des entrées polysémiques en différents sens implique une grande précision de cette information (Brun et al., 2001). • Les performances dépendent des définitions des dictionnaires. • Les algorithmes de Lesk fournissent une précision de désambiguïsation raisonnable lorsque la seule ressource disponible est un ensemble des définitions de dictionnaire (Agirre et Edmonds, 2007).	(Lesk, 1986)	Algorithme de Lesk/Oxford Advanced Learner de l'Anglais courant
		(Veronis et Ide, 1990)	Méthode de Lesk /dictionnaire anglais Collins *CED*
		(Wilks et al., 1990)	Méthode de Lesk /*LDOCE*
		(Ellman et al., 2000)	Filtrage de sens/ *HECTOR* corpus
		(Brun et al., 2001)	Oxford-Hachette French Dictionary (*OHFD*)
		(Soudani et al., 2014b)	Al-Nihaya fi Gharib Al-Hadith w Al-Athar
Approches à base de corpus (supervisées et non supervisées)	• Les approches supervisées sont meilleures par rapport aux approches étudiées (Navigli, 2009 ; Ranjan Pal et Saha, 2015). • Les approches non supervisées ne dépendent pas des sources de connaissances externes (*WordNet*) ou d'inventaires de sens, de dictionnaires électroniques ou de corpus sémantiquement annotés (Ranjan Pal et Saha, 2015). • Les approches supervisées ne donnent pas des résultats satisfaisants pour les langues qui manquent de ressources. • Les approches non supervisées sont difficiles à mettre en œuvre et ses performances sont toujours inférieures à celles non supervisées et à base de connaissances.	(Schutze, 1992)	Non supervisé/regroupement contextuel
		(Reymond, 2001)	Dictionnaire distributionnel à partir d'un corpus de référence du projet SyntSem
		(Audibert, 2003b, 2003a)	Cooccurrence des mots/ corpus français du projet *SyntSem*
		(Véronis, 2003)	Graphe de co-occurrence/ *ROMANSEVAL*
		(Ben Khiroun et al., 2014)	Graphe de co-occurrence/ *ROMANSEVAL*, Le monde 94
		(Basile et al., 2014)	Similarité distributionnelle /*BabelNet + MultiWordNet*
		(Chen et al., 2015)	Représentations distribuées des sens de mots /*WordNet*

Table 1.2 – Synthèse des approches à base de dictionnaire et à base de corpus (supervisées et non supervisées)

Type d'approche	**Description**	**Référence**	**Technique/Ressource exploitée**
Approches à base de modélisation des connaissances ou du raisonnement	• Les approches basées sur la connaissance exploitent principalement les informations contenues dans les ressources lexicales à couverture étendue. • *WordNet* est couramment utilisée comme référentiel de sens dans les systèmes de WSD. • Ces approches donnent une bonne précision. • Les approches basées sur la connaissance ont un avantage important sur les approches basées sur les corpus : c'est le fait qu'elles ne sont pas adaptées à la disponibilité des corpus annotés par les sens, et qu'elles peuvent donc être facilement transférées vers d'autres langues ou domaines pour lesquels un inventaire des sens est disponible.	(Masterman, 1961)	Réseaux sémantiques construits à partir des définitions d'un dictionnaire
		(Banerjee et Pedersen, 2002)	Adaptation d'algorithme de Lesk/ *WordNet*
		(Ponzetto et Navigli, 2010)	Enrichir *WordNet* par des relations sémantiques extraits de *Wikipédia*/ *WordNet*, *Wikipédia*
		(Zhong et Ng, 2010)	Intégration des classifieurs/ *SEMCOR* +DSO corpus + 6 corpus parallèles
		(Navigli et Ponzetto, 2012)	Désambiguïsation à base de graphes dans plusieurs langues/ *BabelNet*
		(Agirre et al., 2014)	Promenades aléatoires sur les relations des ressources lexicales/ *WordNet*
		(Chaplot et Salakhutdinov, 2018)	Exploitation d'un document en entier comme un contexte pour la désambiguïsation d'un mot /*WordNet*, *LDA* (Latent Dirichlet Allocation)
Approches hybrides	• Les approches hybrides essaient d'éviter les limites des approches supervisées et non supervisées en les combinant (Elayeb et al., 2015). • Les approches hybrides ne sont pas encore bien étudiées. Pourtant, ce type d'approche a tendance à donner des bons résultats.	(Mihalcea et Moldovan, 1998)	Densité sémantique/*WordNet* + l'Internet comme un corpus brut pour fournir des informations statistiques
		(Stevenson et Wilks, 2001)	Combinaison des ressources/ *LDOCE*+ corpus construit à partir du Corpus Brown et du Wall Street Journal Corpus
		(Barathi et Valli, 2010)	RI sémantique à base de contexte des concepts/ Ontologie+*WordNet*
		(Merhbene et al., 2014)	Arbres sémantiques/ corpus de 123,8554,642 mots Arabes +≪ *Alwassit* ≫
		(Elayeb et al., 2015)	Distance sémantique et graphe sémantique/ *ROMANSEVAL*+Dictionnaire Sémantique de Contexte

Table 1.3 – Synthèse des approches à base de modélisation des connaissances ou du raisonnement et des approches hybrides

ont permis de clarifier les principaux problèmes de la désambiguïsation sémantique. Vu que l'ambiguïté des sens interagit avec l'ambiguïté des traductions, la WSD a une grande importance dans la RIT. Le chapitre qui suit est consacré à un état de l'art détaillé sur la RIT.

Chapitre

2 Désambiguïsation de requêtes dans la Recherche d'Information Translinguistique

Sommaire

Introduction

Il existe trois classes de RI, à savoir : la RI Monolingue, la RI Multilingue et la RI Translinguistique (RIT) :

- *La RI Monolingue* : c'est la recherche classique, où la requête de l'utilisateur et les documents retournés sont écrits dans une seule langue.
- *La RI Multilingue* : les documents retournés à l'utilisateur sont écrits dans plusieurs langues.
- *La RI Translinguistique* : les documents retournés à l'utilisateur sont écrits dans une seule langue qui est différente de celle de la requête.

Une simple recherche monolingue n'est plus efficace, à cause, de la diversité linguistique des documents disponibles sur le Web. Un document écrit dans une langue différente de celle de la requête peut être plus pertinent qu'un document écrit dans la langue de la requête. Dans ce cas, une RI multilingue est la solution. Cette dernière nécessite un module de traduction et un module de combinaison : pour la traduction, il faut traduire la requête ou les documents, et pour la combinaison, il faut fusionner les résultats traduits de différentes langues dans une seule liste, ce qui augmente la complexité de la RI multilingue (Kadri, 2008). Par contre, la RIT n'a pas besoin d'un module de combinaison dans son processus.

La RIT est une intersection entre la recherche d'information et la traduction. Les questions qui se posent au niveau de la traduction sont : quelle approche faut-il choisir pour la traduction ? Est-ce qu'il faut traduire la requête ou les documents ou les deux ? Et comment choisir la meilleure traduction parmi toutes les traductions possibles ?

Pour bien encadrer notre travail et répondre aux questions précédentes, nous présentons dans ce chapitre une vue d'ensemble sur la RIT. La section 2.1 présente le module de traduction dans la RIT. Dans la section 2.2, nous citons les principales approches utilisées dans la RIT. Dans la section 2.3, nous définissons la notion de désambiguïsation dans la RIT. Dans la section 2.4, nous détaillons les différents SRI existants en RIT.

2.1 La traduction des documents et des requêtes

Le module de traduction en RIT est divisé en trois catégories principales (Nie, 2010 ; Zhou et al., 2012) :

- *La traduction des documents vers la langue de requête* : pour ce type de traduction, il faut déterminer à l'avance à quelle langue chaque document doit

être traduit. Dans ce cas, toutes les versions traduites doivent être stockées. Pour le cas de la RI Multilingue, il faut traduire chaque document vers toutes les autres langues, ceci n'est pas du tout pratique, en raison de la multiplication des versions de documents et l'augmentation de la complexité et du besoin de stockage. Une fois le document est traduit vers la langue de la requête, l'utilisateur peut lire et comprendre la version traduite, sinon il sera obligé de passer par une deuxième traduction après récupération, pour rendre ces documents plus lisibles et compréhensibles.

- *La traduction de requête vers la langue des documents* : les recherches récentes dans la RIT sont faites sur la traduction de requêtes suite à l'avantage limité de la traduction de documents, prouvé par les expérimentations. La RIT est disponible dans plusieurs langues et a un faible coût de calcul, par rapport à l'effort de traduire des grands ensembles de documents. Cependant, la traduction de requête souffre souvent du problème d'ambiguïté de traduction, surtout dans les requêtes courtes, où le contexte est limité.
- *La traduction à travers une langue pivot* : les langues minoritaires en voie de disparition ont des ressources limitées, donc il est possible de ne pas trouver une traduction entre deux langues de ce type. Par ailleurs, il peut y exister des ressources entre ces langues avec une troisième langue qui peut être utilisée comme une langue pivot. Selon (Nie, 2010), il existe deux approches pour ce type de traduction : (i) soit de traduire le document et la requête vers la langue pivot, (ii) soit de traduire l'un des deux en premier vers la langue pivot, puis vers la langue cible.

La figure 2.1 résume les étapes principales de traduction des documents et des requêtes en RIT.

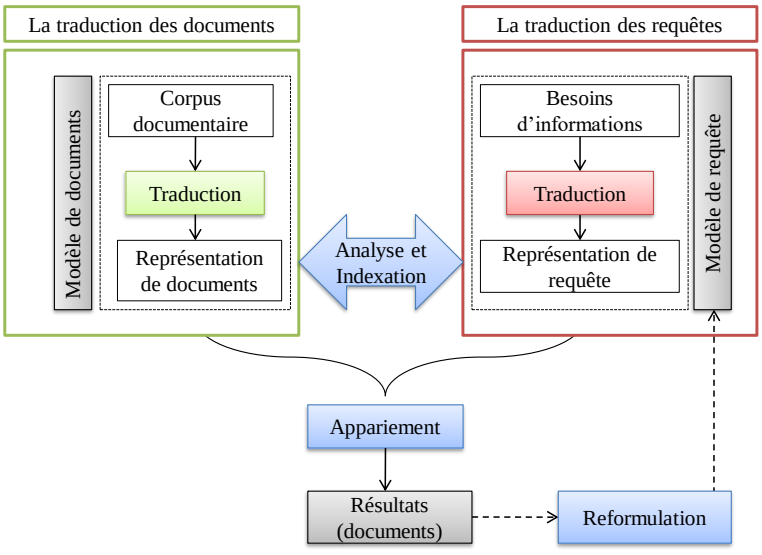

Figure 2.1 – Les techniques de traduction de documents et de requêtes en RIT

Nous nous focalisons dans cette thèse sur la traduction de requêtes (TR) en RIT. Nous détaillons dans ce qui suit les approches et les systèmes de TR en RIT.

2.2 Les approches de traduction de requêtes dans la RIT

Il existe trois catégories d'approches de TR (Elayeb et Bounhas, 2016 ; Zhou et al., 2012) : les approches basées sur les dictionnaires (section 2.2.1), les approches basées sur la traduction automatique (section 2.2.2), et les approches basées sur les corpus (section 2.2.3). Étant donné que chacune de ces approches souffre de certaines limites et bénéficie de certains avantages. Par ailleurs, les approches qui combinent les ressources de traduction (section 2.2.4) ont amélioré considérablement le processus de TR et en conséquence l'efficacité de la RIT.

2.2.1 Les approches à base de dictionnaire

Les dictionnaires bilingues sont de plus en plus disponibles pour de nombreuses langues, ce qui a supporté davantage l'utilité des approches basées sur les dictionnaires dans la TR (Aljlayl et Frieder, 2001 ; Hedlund et al., 2004 ; Levow et al., 2005). Ces approches deviennent plus populaires dans la RIT depuis le début des années 90, quand aucun système de traduction automatique de haute précision n'est disponible. Néanmoins, ces approches souffrent de deux faiblesses majeures :

- *L'ambiguïté* : les dictionnaires bilingues contiennent plusieurs traductions possibles pour chaque terme d'une requête source. Ainsi, choisir la traduction «*appropriée*» et «*correcte*» à partir d'une liste de termes candidats est une tâche importante et n'est pas facile.
- *Le manque de couverture* : certains types de termes comme les termes nouvellement inventés, les mots techniques, les noms propres, les acronymes, les abréviations, etc., ne sont pas tous représentés dans les dictionnaires bilingues. Ces termes hors vocabulaire (OOV : Out-Of-Vocabulary) peuvent considérablement réduire l'efficacité de la recherche dans un SRIT, surtout lorsque les requêtes sources sont très courtes.

Nous présentons dans ce qui suit les différentes approches proposées pour différentes paires de langues, la plupart d'elles sont Latines.

Les recherches et les évaluations effectuées, dans les SRIT dans l'Université de Tampere (UTA [1]) au laboratoire de RI, ont été développées par Pirkola et al. (Pirkola et al., 2001). Les chercheurs ont discuté des nombreux problèmes causés par

1. http://www2.uta.fi/en

les approches basées sur le dictionnaire, et ils ont suggéré «*le modèle des requêtes structurées de Pirkola*» (Pirkola et al., 2002). Les résultats expérimentaux ont été réalisés en utilisant quatre paires de langues différentes [2] pour bien évaluer l'efficacité de la structuration des requêtes dans la RIT.

(Pirkola et al., 2003) ont étudié la structuration des requêtes dans la RIT de diverses façons : premièrement, ils ont exploité un dictionnaire Anglais-Finnois dans leur processus de TR, afin de récupérer des documents d'une base de données d'un journal finnois de 55000 articles. Les différentes traductions finlandaises d'un terme anglais sont considérées comme des synonymes. Elles ont été combinées en utilisant l'opérateur *#syn()* [3] du système de recherche *InQuery* (Callan et al., 1992). Cette combinaison modifie implicitement la pondération des traductions, et permet de mieux équilibrer l'importance relative entre les parties de la requête. Les chercheurs ont confirmé que les requêtes structurées ont surpassé les requêtes non structurées : elles ont donné une efficacité de 77% par rapport à une RI Monolingue et 52% par rapport à une stratégie de traduction sans structuration. Deuxièmement, les auteurs ont étudié l'influence de la structuration à base des composants en utilisant un opérateur de proximité pour les composants des traductions, construits à partir des termes de la requête. Malheureusement, cette approche n'est pas utile dans les requêtes basées sur l'opérateur *#syn()*, car elle a réduit l'efficacité de la recherche. Troisièmement, ils ont proposé une technique de correspondance des chaînes de caractères basées sur le *n-gramme*, appelée la technique ciblée de correspondance *s-gramme* (Pirkola et al., 2002). Dans cette technique, les *n-grammes* sont classés en catégories, en se basant sur la contiguïté des caractères dans les mots. La technique a été comparée à la technique classique de *n-gramme*, en utilisant des caractères adjacents en tant que des *n-grammes*. Plusieurs types de mots et de paires de mots ont été étudiés : les mots-clés des requêtes en anglais, en allemand et en suédois ont été adaptés à leurs variantes morphologiques et orthographiques finnoises, en utilisant une liste de 119000 mots cibles en langue finnoise. Dans tous les tests translinguistiques effectués, la technique ciblée de correspondance *s-gramme* a surpassé la technique de correspondance de *n-gramme*.

Les noms propres et les termes techniques constituent un problème majeur dans la RIT, en raison de la couverture limitée des dictionnaires bilingues. Aussi, ces derniers ont de nombreuses variantes orthographiques, puisqu'ils partagent souvent la même origine grecque ou Latine. C'est pourquoi (Toivonen et al., 2005) ont suggéré et évalué, *en deux étapes*, une approche floue pour la traduction translinguistique

2. Finnois vers l'Anglais, Anglais vers le Finnois, Suédois vers l'Anglais, Allemand vers l'Anglais

3. dans opérateur #syn(), les termes de l'argument doivent être considérés comme des synonymes, dans le cas de Pirkola et al., les traductions d'un terme sont des synonymes.

des variantes orthographiques. Tout d'abord, des règles de transformation sont appliquées sur les termes des requêtes sources pour les rendre plus similaires à leurs traductions correspondantes dans la langue cible. Ensuite, les formes intermédiaires, générées à partir de la première étape, sont traduites vers la langue cible, en utilisant un modèle flou d'appariement. Enfin, les expérimentations sont effectuées dans cinq langues sources (finnois, français, allemand, espagnol et suédois) et l'anglais comme une langue cible. Les résultats ont montré que la technique «*en deux étapes*» a surpassé une simple technique floue d'appariement.

Par ailleurs, les méthodes probabilistes des requêtes structurées ont été suggérées par (Darwish et Oard, 2003). Ils ont exploité les estimations des probabilités de remplacement et le concept de «*fréquence de document*» du modèle des espaces vectoriels, pour calculer le poids de chaque terme de requête. Les résultats ont montré l'efficacité de ces méthodes dans la RIT, dans la paire de langues Arabe-Anglais et dans la recherche des documents arabes numérisés (scannés), basée sur la reconnaissance optique des caractères [4].

Plus tard, les probabilités de traduction bidirectionnelle [5] et la synonymie sont toutes les deux exploitées par (Wang et Oard, 2006) afin d'améliorer l'efficacité de la RIT. Les auteurs ont montré que la combinaison des connaissances de traduction bidirectionnelle avec la synonymie à base de similarité donne de bons résultats par rapport à certaines techniques existantes.

(Pourmahmoud et Shamsfard, 2008) ont proposé une approche à base d'un dictionnaire Anglais-Persan, pour retrouver des documents pertinents en anglais pour des requêtes persanes. Dans cette approche, les auteurs ont présenté une méthode d'identification et de traduction de phrases. Aussi, ils ont enrichi les requêtes avec des termes connexes, par une expansion avant et après traduction, afin d'améliorer la TR.

(Saralegi et De Lacalle, 2010) ont développé une approche de TR basée sur le dictionnaire qui aborde trois problèmes principaux dans la RIT Basque-Anglais : (i) la présence des mots hors vocabulaire OOV, (ii) la traduction des expressions multi-mots et (iii) le traitement des traductions ambiguës. Les auteurs ont proposé des méthodes basées sur les co-occurrences pour gérer la sélection de traduction, la détection du connexe (Knight et Graehl, 1998) pour traiter les termes OOV et un processus d'appariement naïf pour détecter les expressions multi-mots. De plus, les auteurs ont mesuré comment chaque problème affecte les performances de recherche dans la TR à base de dictionnaire et comment les méthodes proposées les traitent.

4. OCR (Optical Character Recognition) est la version des documents numérisés vers un texte éditable par les logiciels de traitement de texte.

5. traduire les requêtes et les documents, ou plus précisément avoir une traduction dans les deux sens.

Dans les expérimentations, les résultats changent en fonction de la longueur des requêtes, la diminution produite par la sélection de traduction et celle produite par les expressions multi-mots. Dans le cas de la sélection de traduction, on distingue deux cas : une mauvaise sélection dans le dictionnaire (baisse de $10-21\%$) et des traductions incorrectes dans le dictionnaire (baisse de $17-32\%$). Le traitement OOV (baisse de $4-12\%$) semble être le facteur le moins influent, probablement en raison de l'orthographe similaire des deux langues.

(Sharma et Mittal, 2018) ont construit un système de TR à base de dictionnaire dans lequel les requêtes sont segmentées et les termes multi-mots sont créés en utilisant la technique de *n-gramme*. Les termes hors vocabulaire (OOV) sont translittérés en utilisant la technique OOVTTM (OOV Terms Transliteration Mining). Les expérimentations ont montré que l'approche proposée a donné des meilleurs résultats.

2.2.2 Les approches à base de traduction automatique

L'objectif principal des approches basées sur la traduction automatique est de traduire automatiquement un texte ou un discours d'une langue source vers une langue cible différente. Ce processus consiste à analyser le contexte de la requête source avant de traduire ses termes vers une langue cible. L'analyses syntaxique et morphologique ont été utilisées pour améliorer les méthodes de traduction statistiques.

De nos jours, les récentes recherches ont progressivement amélioré les approches à base de traduction automatique dans la RIT. Par exemple, le traducteur en ligne *Google Translate*[6] a été exploité par (Wu et al., 2008) dans leur processus de TR. Cet outil fournit des traductions instantanées en 66 langues différentes (Yen, 2013). Les résultats ont montré l'efficacité de la traduction automatique pour les requêtes courtes et longues sur le Web, en utilisant une technique de réinjection de pertinence.

En outre, (Hefny et al., 2011) ont suggéré une méthode de sélection de requêtes afin d'améliorer l'efficacité de RI Translinguistique et Multilingue. Ils ont bénéficié du traducteur en ligne *Bing*[7] pour retrouver des documents anglais sur le Web, en utilisant des requêtes arabes. Cette méthode a eu un impact statistiquement significatif sur la RI Multilingue en montrant une amélioration par rapport à la RI Monolingue.

De plus, des nombreux SRIT (Udupa et al., 2009 ; Wu et al., 2008) ont exploité les approches de traduction automatique dans leurs tâches de TR. Les systèmes de

6. http://translate.google.com/

7. http://www.bing.com/translator/, Bing est crée par le groupe de recherche de Microsoft et il est capable de traduire 41 langes différentes.

traduction automatique sont formés à partir des corpus parallèles. Dans le cas de (Habash et Sadat, 2006), ils ont utilisé un corpus parallèle Arabe-Anglais fourni par le consortium des données linguistiques (LDC : Linguistic Data Consortium), d'environ 5 millions de mots. Au-delà, le programme GALE (The Global Autonomous Language Exploitation) [8] a impliqué un système de traduction automatique afin de supporter la diffusion des actualités multilingues (Olive et al., 2011).

Récemment, la traduction automatique est devenue statistique (SMT : Statistical Machine Translation), sa précision est relativement meilleure, ce qui améliore davantage les résultats de la RIT. Ceci a été confirmé dans l'évaluation des SRIT réalisés dans les campagnes CLEF. De plus, le nombre des traducteurs automatiques gratuits en ligne a augmenté, dans lesquels *Google* et *Bing* sont parmi les deux premiers les plus populaires aujourd'hui. L'évaluation de ces moteurs de recherche commerciaux sur le Web se fait à l'aide des requêtes réelles issues des fichiers de journaux des requêtes qui contiennent des informations, telles que : les expressions des requêtes, le nombre des résultats, le nombre des pages consultées, etc.

2.2.3 Les approches à base de corpus

Les approches basées sur les corpus (Ahmed et Nürnberger, 2008 ; McNamee et Mayfield, 2002 ; Yang et al., 2005) bénéficient d'un ensemble de listes des termes multilingues générés à partir des corpus comparables ou parallèles. Un corpus comparable renferme des documents similaires dans différentes langues, où les couples de documents sont conceptuellement similaires. Mais la collection des documents parallèles contient des couples ou un ensemble de documents identiques dans différentes langues. En fait, les corpus parallèles sont d'une importance capitale pour les applications multilingues de traitement automatique des langues. Les traductions qui sont extraites de corpus parallèles sont généralement plus précises que celles extraites par d'autres ressources de traduction. Cependant, leur rareté reste le maillon faible de plusieurs applications d'intérêt (Azarbonyad et al., 2013 ; Patry et Langlais, 2005).

(Yang et al., 2005) ont suggéré une méthode qui combine les approches à base des corpus avec les approches à base de catégorisation afin d'améliorer la performance de la RIT. Plus précisément, ils ont utilisé la catégorie «*topic*» des sujets médicaux (MeSH : Medical Subject Headings). En outre, les auteurs ont évalué plusieurs méthodes d'apprentissage basées sur des corpus de haute-performance ainsi que des approches basées sur la traduction automatique, en exploitant le dictionnaire

8. http://www.speech.sri.com/projects/GALE/

bilingue *SYSTRAN*[9]. Le processus de recherche bilingue est effectué à l'aide de la collection *Springer* contenant 25 requêtes et un corpus parallèle de 9640 documents médicaux (des titres avec des résumés d'articles des revues médicales) en anglais et en allemand. Les résultats expérimentaux ont mis en évidence que les approches basées sur les corpus ont surpassé l'approche fondée sur la catégorisation, suite à la perte d'informations détaillées dans l'indexation au niveau des catégories. De plus, l'approche basée sur la traduction automatique a été inférieure à celle à base de catégorisation.

(Zhang et al., 2005) ont proposé une approche de TR basée sur un corpus Web. Cette approche est effectuée en deux étapes, à savoir : (1) l'extraction des traductions candidates et (2) la sélection de traduction. Dans l'extraction des traductions candidates, les auteurs ont utilisé le moteur de recherche *Google*, pour retrouver les données du corpus dans la langue chinoise, en soumettant une requête dans la langue anglaise. Chaque document retourné par le moteur de recherche comprend un titre et un résumé biaisé par la requête. Les traductions candidates doivent figurer à la fois dans le titre et dans le résumé biaisé par la requête. Dans la sélection de traduction, les auteurs ont déterminé la traduction possible parmi les traductions candidates, en utilisant une fonction de classement. Les résultats expérimentaux ont montré que l'approche de TR proposée est efficace pour trouver les traductions correctes, avec un taux d'inclusion de traduction de 75.57%.

Les performances des approches de TR à base de corpus sont affectées par trois qualités (Talvensaari, 2008). Premièrement*, la Similarité des sujets avec les requêtes traduites* : par exemple, un corpus parallèle, sur les actualités sportives, n'est pas susceptible de fournir des connaissances fiables pour la traduction d'une requête sur la physique quantique. Deuxièmement*, la qualité d'alignement* : un corpus parallèle est beaucoup mieux qu'un corpus comparable bruyant. Troisièmement, *la taille du corpus* : c'est un facteur très important, plus on a des documents alignés, plus on a des connaissances fiables pour la traduction. Ainsi, un grand corpus parallèle avec une bonne couverture du vocabulaire de domaine serait idéal pour la RIT. Les résultats expérimentaux de Talvensaari, ont montré que le facteur le plus pertinent parmi ces trois qualités est celui qui dépend de similarité des sujets des documents avec le sujet des requêtes. De plus, Talvensaari a suggéré l'exploitation des corpus comparables bruyants en tant que ressources complémentaires, quand les corpus parallèles ne sont pas disponibles dans un domaine donné. En dépit de ça, l'acquisition de tels corpus est très coûteuse (Kadri, 2008).

9. http://www.systran.fr/lp/dictionnaire-bilingue/

2.2.4 Les approches combinées

La combinaison de plusieurs techniques de traduction en utilisant diverses ressources, améliore considérablement l'efficacité de TR. Nous détaillons dans ce qui suit, les cas où les tâches de RIT ont été améliorées grâce à une combinaison entre les différentes ressources de traduction (Kadri, 2008) :

- *La sélection de la traduction appropriée en utilisant une probabilité raisonnable* : pour un terme donné d'une requête source, les traductions possibles existantes dans diverses ressources de traduction ont un différent degré de confiance. Pour certains termes, leurs traductions correspondantes peuvent être inappropriées, tandis que pour d'autres, elles peuvent être les bonnes. Cependant, le problème ici dépend du poids de la traduction. En effet, nous pouvons avoir la même traduction pour un terme donné à partir de différentes ressources de traduction, mais ayant des probabilités (ou des poids/pondérations) différentes. Par conséquent, un mixage de différentes ressources de traduction offre la possibilité de sélectionner la traduction appropriée en utilisant une probabilité raisonnable.
- *L'augmentation de couverture des ressources* : lorsque nous utilisons chaque ressource de traduction séparément, nous pouvons souffrir d'une faible couverture, car nous pouvons trouver des termes de requête ayant des traductions manquantes. Par exemple, les entités nommées comme les noms propres, les mots techniques, les noms des organisations, etc., ne sont pas généralement couvertes par des nombreux lexiques bilingues. Ainsi, le mixage de différentes ressources de traduction peut surmonter cette lacune et résoudre le problème de nombreuses traductions manquantes.
- *Expansion des requêtes* : la sélection de la traduction appropriée, conduit naturellement à un effet d'expansion de requête qui est une tâche souhaitable dans la RI.

(Xu et Weischedel, 2005) ont confirmé que la combinaison des lexiques manuels avec des textes parallèles améliore l'efficacité de TR dans la RIT Arabe-Anglais. De plus (Kadri et Nie, 2004) ont bénéficié du mixage des lexiques des termes bilingues avec des dictionnaires bilingues Anglais-Français et Anglais-Arabe, afin de bien identifier la traduction des entités nommées. En effet, (Kadri, 2008) a profité d'une combinaison linéaire de différentes ressources de traduction, en donnant un poids de confiance pour chacune d'elles. Les expériences ont montré que la combinaison augmente les performances de recherche des documents pertinents.

Certains travaux ont recommandé l'hybridation de différentes ressources de traduction telles que, les listes des termes générées à partir des corpus comparables ou parallèles, ou des lexiques bilingues, etc., pour pouvoir résoudre le problème

de couverture des dictionnaires. Par exemple, (Yahya et al., 2013) ont suggéré et évalué une méthode basée sur une ontologie bilingue de domaine, en utilisant des concepts coraniques afin de désambiguïser les traductions correspondantes à un terme donné dans une requête, et par conséquent, améliorer la tâche de TR basée sur les dictionnaires. Dans leurs expérimentations, les auteurs ont combiné nombreuses ressources et techniques de traduction telles que : (i) l'ontologie du Coran générée à partir d'un corpus parallèle bilingue anglais et malais, (ii) la technique de traduction à base de dictionnaire, et (iii) la technique translinguistique de TR en utilisant les concepts du Coran comme ressource linguistique. Les résultats ont mis en évidence que cette combinaison améliore considérablement l'efficacité de RIT, en particulier pour les collections des documents anglais, au contraire de la collection des documents malais.

(Azarbonyad et al., 2013) ont combiné différentes ressources afin de profiter de toutes les ressources de traduction, de réduire leurs lacunes et d'évaluer leurs influences sur la performance de la RIT Anglais-Persan. Par exemple, ils ont combiné des corpus comparables avec des corpus parallèles pour bénéficier des traductions précises extraites du corpus parallèle, aussi bien des relations entre les termes des requêtes extraites à partir du corpus comparable. Les expérimentations ont montré que les approches combinées (parallèle+dictionnaire, comparable+dictionnaire et parallèle+comparable) surpassent les approches de traduction basées sur une seule ressource (dictionnaire, comparable et parallèle). Les meilleurs résultats des approches combinées sont obtenus par la combinaison des corpus parallèles avec les corpus comparables. Les résultats ont montré que cette approche améliore l'approche à base de corpus comparable de 106%, et l'approche à base de corpus parallèle de 7%.

(Ben Khiroun et al., 2018) ont proposé une nouvelle approche possibiliste de TR qui combine des ressources statistiques et lexicales. D'une part, ils ont construit un dictionnaire bilingue Français-Anglais à partir des textes alignés de la collection *Europarl*. D'autre part, ils ont construit une structure de graphe de co-occurrence et ils ont utilisé le réseau lexical *BabelNet* pour traiter la désambiguïsation des traductions candidates des termes ambigus. Les auteurs ont étudié l'impact de l'expansion des requêtes en adoptant la technique de pseudo-rétroaction de pertinence (PRP). Les résultats ont mis en évidence l'impact positif de la technique PRP sur le processus de TR. En outre, l'approche possibiliste à base de graphe de co-occurrence a surpassé toutes les expérimentations à base des circuits.

2.2.5 Synthèse des approches de TR en RIT

Nous résumons dans la figure 2.2 les différents types d'approches étudiées pour la TR en RIT.

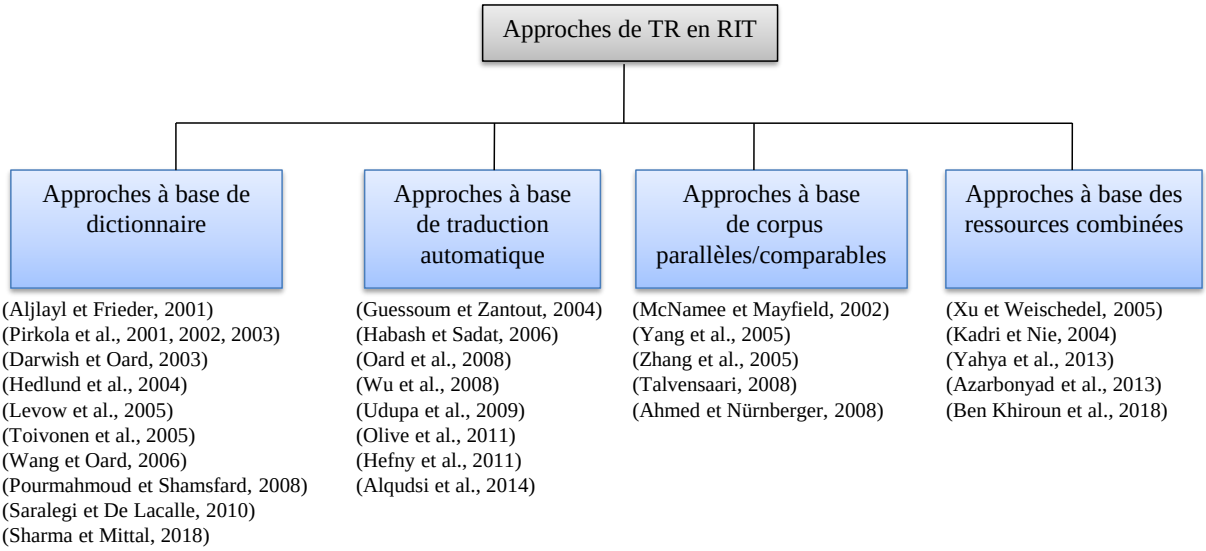

Figure 2.2 – Résumé des approches de TR en RIT

Les ontologies, *WordNet*, les graphes et les statistiques de co-occurrence sont utilisés pour résoudre les problèmes d'homonymie et de polysémie. La couverture et la qualité du dictionnaire, la traduction des syntagmes, l'homonymie, la polysémie et le manque de ressources sont les majeurs défis rencontrés par tous les types d'approches de TR en RIT (tableau 2.1).

Problèmes et défis		**Description**
Ambiguïté	**Homonymie**	un mot qui a deux ou plusieurs sens différents.
	Polysémie	un mot qui a plusieurs significations connexes.
	Mot flexionnel	un mot peut avoir différentes formes grammaticales.
	Traduction des syntagmes	la traduction des expressions multi-termes ou des syntagmes gardent le contexte de la requête mieux que la traduction de chaque terme indépendamment de l'autre.
Mauvaise couverture	**Les mots hors-vocabulaire**	un mot qui n'existe pas dans le dictionnaire tels que, les termes nouvellement inventés, les mots techniques, les noms propres, les acronymes, les abréviations, etc.
	Les entités nommées	consistent essentiellement à identifier les noms des personnes, les organisations et les localisations géographiques, les dates et les quantités.
Manque de ressources		non-disponibilité des ressources dans plusieurs paires de langues pour les expérimentations.

Table 2.1 – Les problèmes et les défis des approches de TR en RIT

2.3 La désambiguïsation des traductions de requêtes

La désambiguïsation est le but de la plupart des techniques des traductions de requêtes, elle vise à déterminer la traduction la plus appropriée pour un mot d'une langue source vers une langue cible. Au cours des dernières années, la désambiguïsation est devenue une nécessité dans les applications automatiques de traitement des langues. En fait, elle ne devra pas rester comme une tâche isolée, mais elle devrait être incorporée dans la traduction automatique ou la RI Translinguistique et Multilingue.

Par ailleurs, la RIT est une application qui se développe avec un rythme rapide, grâce à la croissance de la RI sur le Web, le commerce mondial et les besoins des communautés de renseignement. La désambiguïsation des sens de mots (WSD) pourrait bien avoir un grand potentiel dans la RIT que dans la RI Monolingue, en raison de l'interaction de l'ambiguïté des sens avec l'ambiguïté des traductions (Agirre et Edmonds, 2007).

La désambiguïsation en RIT est un cas particulier du problème de la WSD qui a reçu tant d'attention dans la communauté du traitement des langues naturelles (Ide et Véronis, 1998). La tâche de désambiguïsation translinguistique (CLWSD : Cross-Lingual WSD) a été organisée pour la première fois dans le cadre de SemEval-2010 (Lefever et Hoste, 2010). Afin de mieux comprendre la complexité et la praticabilité de la désambiguïsation translinguistique, les chercheurs ont proposé une deuxième édition SemEval-2013 (Lefever et Hoste, 2013). Les participants dans les ateliers ont confirmé que l'efficacité de la RIT a été améliorée grâce à la désambiguïsation translinguistique, en résolvant significativement le problème d'ambiguïté dans les traductions.

Nombreuses techniques ont été abordées pour résoudre la désambiguïsation translinguistique (tableau 2.2) telle que : les systèmes basés sur les graphes étant l'une des approches les plus réussies, comme il est indiqué dans les systèmes qui ont participé aux compétitions SemEval 2010 et 2013 (Duque et al., 2015), les techniques basées sur la co-occurrence et les techniques basées sur le calcul de similarité entre les termes.

Techniques	**Références**	**Langues**	**Description**	**Ressources**
à base des statistiques de co-occurrence	(Ballesteros et Croft, 1998)	Espagnol-Anglais	- Mesurer les statistiques de co-occurrence entre les termes qui existent dans un corpus. - Les résultats ont montré que la combinaison des méthodes de désambiguïsation améliore les performances de RIT de plus de 90% par rapport aux performances de RI Monolingue.	Corpus parallèle : *Nations Unies*
à base de concept de similarité des termes statistiques	(Adriani, 2000)	Anglais-Indonésien/ Indonésien-Anglais	- Mesurer l'efficacité de cette technique en termes de précision moyenne, ainsi que d'étudier son efficacité pour atténuer le problème d'ambiguïté dans la paire des langues Anglais-Indonésien. - Pour la RIT Indonésien-Anglais : traduire manuellement 24 requêtes de TREC pour des sujets translinguistiques en Indonésien. - Pour la RIT Anglais-Indonésien : créer manuellement 26 requêtes en Anglais qui couvrent un certain nombres d'évènements survenus en Indonésie.	- Dictionnaire bilingue Indonésien-Anglais - Corpus anglais : 242 918 documents des collections TREC Associated Press (AP) - Corpus indonésien : 20590 documents des articles de *Kompas* (un quotidien Indonésien).
à base des statistiques de co-occurrence	(Gao et al., 2001, 2006)	Anglais-Chinois	- Les auteurs ont présenté trois modèles probabilistes statistiques de TR pour la résolution des ambiguïtés de TR : (i) *modèle de co-occurrence* des termes pour calculer leurs poids, (ii) *modèle de traduction des syntagmes nominaux* et (iii) *modèle de traduction de dépendance*. - Les résultats ont montré que ces méthodes ont amélioré considérablement les performances de la RIT. - La combinaison de cette approche avec un système de traduction automatique conduit à une efficacité d'environ 105% par rapport à la RI Monolingue.	- Dictionnaire bilingue Anglais-Chinois généré automatiquement. - Corpus chinois de TREC contenant des articles publiés dans *Hong Kong Commercial Daily*, *Hong Kong Daily News* et *Takungpao*.
à base de co-occurrence	(Yuan et Yu, 2007)	Anglais-Chinois	- Le calcul de co-occurrence des paires de termes pour mesurer la qualité d'une traduction. - Les relations entre les termes cibles sont représentées sous forme de graphes pour comparer les différentes combinaisons de termes cibles. - Le résultat est présenté sous la forme d'une distribution des probabilités.	- Dictionnaire bilingue : *SYSTRAN*. - Corpus monolingue chinois : 2318 articles du ≪*People Daily*≫
à base des données annotées	(Lefever et Hoste, 2013)	Anglais-Français/Italien/ Allemand /Néerlandais	- C'est une tâche de désambiguïsation non supervisée pour les noms anglais. - Toutes les traductions d'un mot polysémique sont regroupées manuellement en groupes (clusters). - Pour les données de test, les examinateurs choisissent un ou plusieurs clusters de traduction pour chaque phrase de test, et fournissent leurs trois meilleures traductions de la liste prédéfinie des traductions d'*Europarl*.	- Corpus parallèle : *Europarl*
à base des graphes de co-occurrence	(Duque et al., 2015)	Anglais-Espagnol/Français/ Italien/ Allemand/ Néerlandais	- C'est une technique non supervisée pour construire automatiquement un dictionnaire bilingue et des graphes de co-occurrence. - Ils ont testé la robustesse de cette technique, en appliquant nombreuses langues différentes. - Ce système, appelé ≪ CO-Graph ≫, sélectionne, en tant que des nœuds de graphe, les termes du corpus qui ≪ co-occurrent ≫ de manière significative dans les mêmes documents.	- Corpus parallèle : *Europarl* - Dictionnaire bilingue extrait aussi de *Europarl*

Table 2.2 – Les techniques de désambiguïsation transliguistique

2.4 Les Systèmes de Recherche d'Information Translinguistique : SRIT

Dans cette section, nous présentons et nous discutons brièvement les principaux avantages et inconvénients des SRIT existants.

2.4.1 Mulinex

Pour le système *Mulinex*, (Capstick et al., 2000) ont fourni aux utilisateurs quelques opportunités de formuler/reformuler et désambiguïser leurs requêtes. En fait, l'utilisateur ne peut utiliser que sa langue maternelle pour lire et filtrer les documents retournés. *Mulinex* a profité de la TR à base de dictionnaire pour atteindre sa fonctionnalité multilingue. Il offre une traduction automatique des documents et leurs résumés. Le système supporte le Français, l'Allemand et l'Anglais, de sorte que l'utilisateur n'est pas obligé d'avoir une connaissance approfondie de la langue cible.

Le système *Mulinex* utilise six bases de données de dictionnaires bilingues avec 100000 à 200000 entrées pour chacune pour les six paires de langues supportées par le système (Allemand+Anglais, Allemand+Français, Français+Anglais et les couples inversés). Parmi aussi les fonctionnalités fournies par *Mulinex*, nous pouvons citer :

- La traduction de requête de l'utilisateur.
- L'expansion interactive de requête.
- Une recherche simultanée dans les collections des documents français, allemand et anglais.
- Une traduction automatique *LOGOS*[10] des résumés et des documents récupérés.
- Une filtration des résultats retournés selon la langue et la catégorie.
- Le système utilise «*l'assistant de requête*» (query-assistant) pour fournir une désambiguïsation interactive de la TR. Cette tâche est effectuée par la technique de «*traduction inverse*» de requête vers la langue d'origine. Cependant, cette approche a des limites telles que : (i) lorsqu'il n'y a pas de synonymie dans le dictionnaire, la technique n'est plus utile, ou (ii) lorsqu'il y a une homonymie significative dans la langue cible, elle peut entraîner une confusion dans la technique de «*traduction inverse*» (Farag et Nurnberger, 2012).

10. http://logos-os.dfki.de/

2.4.2 Keizai

Dans le projet *Keizai* (Ogden et Davis, 2000), le but ultime est de récupérer à partir des requêtes en anglais, des documents japonais et coréens sur le Web, à travers un système de recherche des textes translinguistiques. Les termes de requête en anglais sont traduits en japonais ou en coréen. En effet, si l'utilisateur ne comprend pas la langue japonaise ou coréenne, il peut choisir la traduction correcte parmi plusieurs traductions possibles, grâce aux définitions des traductions qui sont disponibles en anglais. Par conséquent, l'efficacité de la recherche est améliorée. En outre, le système retourne des résumés en anglais pour l'ensemble des premiers meilleurs documents retournés.

Keizai souffre de certains inconvénients tels que dans la tâche de désambiguïsation de traduction. L'utilisateur doit vérifier les définitions de toutes les traductions quand il choisit la traduction adéquate. Cette tâche nécessite beaucoup de temps au cas où les termes des requêtes ont un grand nombre de traductions possibles. Par ailleurs, *Keizai* a profité d'un dictionnaire bilingue dans ses tâches de traduction et de désambiguïsation. Néanmoins, il est rare de trouver ce type de dictionnaire bilingue dans lequel les définitions dans la langue source (anglaise) sont proposées pour chaque traduction dans les langues cibles (japonaise et/ou Coréenne).

2.4.3 WORDS

Dans l'outil *WORDS*, (Lopez-Ostenero et al., 2003) ont fourni à l'utilisateur la possibilité d'interagir avec le système par la formulation et la reformulation de sa requête. Cet outil comprend une assistance utilisateur de raffinement de TR afin d'accroître l'efficacité de la recherche. En fait, *WORDS* supporte l'utilisateur dans le choix des traductions adéquates pour chaque terme de requête. De plus, l'utilisateur profite d'un ensemble de phrases pertinentes, données par l'outil, lorsqu'il est en train de formuler ou d'affiner sa requête.

Le système *WORDS* est capable de fournir toutes les traductions anglaises possibles pour chaque terme de requête espagnole. Il fourni également la technique de «*traduction inverse*» de l'anglais vers l'espagnol, dans le but de renforcer la confiance des utilisateurs envers la traduction proposée. Par conséquent, l'utilisateur peut modifier certaines traductions déjà choisies avant de lancer le processus de recherche des documents. En outre, *WORDS* offre un résumé pour chaque document retourné (anglais) dans la langue de l'utilisateur (espagnol) dans lequel la traduction du «*corps*» des documents est basée sur la traduction de tous les syntagmes nominaux. Par contre, les «*titres*» des documents sont traduits auto-

matiquement à l'aide du système de traduction automatique *SYSTRAN*[11]. À l'aide des résumés, l'utilisateur peut décider quelle pertinence peut donner pour chaque document retourné. De plus, il peut vérifier la traduction en espagnol de chaque document anglais retourné. Puis, il peut affiner sa requête afin d'améliorer l'efficacité de sa recherche. Ensuite, le système pointe vers l'une des traductions possibles pour le terme de requête espagnole. L'utilisateur peut sélectionner ou désélectionner n'importe quel terme en anglais, mais il ne doit choisir que la traduction adéquate en anglais pour chaque terme de requête espagnole.

Pour désambiguïser ces traductions, l'utilisateur doit vérifier toutes les traductions possibles avec leurs définitions ; tâche évaluée complexe et coûteuse en termes du temps. D'autre part, la traduction automatique fournie par le système n'est pas disponible dans certains cas, ce qui peut avoir un impact direct sur le processus de raffinement des requêtes et par conséquent sur l'efficacité de la recherche.

2.4.4 UCLIR

Le système *UCLIR* (Unicode Cross-Language Information Retrieval) est proposé par Abdelali et son équipe (Abdelali et al., 2004), dans lequel la langue arabe est incluse. Ce système fourni trois différents modes de TR, tout dépend de la langue de la requête et du degré d'implication de l'utilisateur dans les processus de traduction et de recherche.

1. Dans le premier mode, il est possible d'exécuter une requête multilingue qui contient des termes en différentes langues.
2. Le second mode est un mode non-interactif qui profite d'un ensemble de dictionnaires bilingues et des listes des mots indexés pour traduire une requête en anglais vers différentes langues cibles. Les requêtes résultantes sont utilisées pour la recherche des documents à partir des listes des mots indexés. Dans ce mode, l'utilisateur n'a aucun contrôle sur le processus de formulation des requêtes multilingues.
3. Dans le troisième mode, la requête source est en anglais. Il s'agit d'un mode interactif dans lequel l'utilisateur est impliqué dans la décision pour sélectionner les termes des requêtes dans différentes langues.

Afin d'assurer une bonne qualité et un bon fonctionnement pour le système *UCLIR*, l'équipe a exploité un grand ensemble de ressources et techniques linguistiques, ceci inclut des analyseurs morphologiques, des dictionnaires bilingues annotés, une reconnaissance de haute-précision des noms propres, des procédures de translittération de haute-qualité pour les noms propres et une traduction automatique.

11. http://www.systransoft.com/

2.4.5 UTACLIR

UTACLIR (Hedlund et al., 2004) est un système de TR à base de dictionnaire. Les tests sont étendus d'une RIT pour les trois paires de langues : suédoise, finlandaise et allemande vers l'anglais, à six paires de langues, de l'anglais au français, allemand, espagnol, italien, néerlandais et finlandais, et de bilingue à multilingue. En outre, ce système comprend nombreuses tâches telles que :

- La traduction transitive ;
- La contribution des composants individuels ;
- L'efficacité de gestion des composants ;
- La correspondance des noms propres ;
- La structuration des requêtes.

Les résultats expérimentaux sont achevés via le standard de test CLEF 2000/2001/2002, et ils ont montré les effets individuels positifs des différents composants. Cependant, les auteurs ont confirmé que leurs résultats dépendent principalement de nombreux facteurs tels que : le thème de la collection, les dictionnaires bilingues et multilingues, les langues sources et cibles, le nombre des composants et des noms propres dans le sujet, etc.

2.4.6 MultiLexExplorer

Le système *MultiLexExplorer* (De Luca et al., 2006) a bénéficié de la distribution des traductions des termes des requêtes sur le Web pour aider les utilisateurs multilingues à améliorer leur recherche sur le Web. Leur objectif étant la résolution des problèmes liés à la désambiguïsation des traductions. Le système utilise certaines combinaisons des traductions des termes de requête en visualisant simultanément les relations d'*EuroWordNet* avec les documents retournés et en affichant des statistiques à partir des moteurs de recherche sur Web. Par ailleurs, ce système fournit d'autres fonctionnalités telles que :

- La visualisation d'une hiérarchie générale pour chaque contexte linguistique d'un mot donné ;
- La recherche dans différentes langues, par exemple par la traduction des sens de mots à l'aide des index inter-linguistiques d'*EuroWordNet* ;
- La désambiguïsation des sens de mots des différentes combinaisons de mots ;
- L'amélioration de l'interaction entre l'utilisateur et le système, en donnant à l'utilisateur la possibilité de modifier les termes récupérés ;
- La mise en œuvre d'un processus d'expansion de requête pour limiter le nombre de documents retournés ;

– La catégorisation automatique des documents retournés à partir du Web, à l'aide de différentes méthodes de catégorisation.

2.4.7 Patent CLIR

Le système *Patent CLIR* est proposé par (Bian et Teng, 2008). Ce système assure la recherche des patrons monolingues et translinguistiques correspondants aux langues anglaise et japonaise. Pour bien accomplir la tâche de RIT, ce système a utilisé trois unités principales : l'unité d'indexation, l'unité de traduction et l'unité de classification. Ce système profite de différentes traductions disponibles sur le Web pour traduire les termes des requêtes. D'abord, il accepte comme entrée un ensemble des termes de requête. Ensuite, l'utilisateur sélectionne la traduction adéquate pour chaque terme de requête, en utilisant une liste proposée. Il bénéficie aussi de la possibilité d'examiner et d'ajuster ces traductions.

Malgré que *Patent CLIR* a la possibilité d'examiner et de modifier les traductions Web fournies par le système, parfois l'utilisateur ne peut pas choisir les traductions adéquates de différents traducteurs en raison de sa connaissance limitée de la langue cible.

2.4.8 MIRACLE

Le système *MIRACLE* (Maryland Interactive Retrieval Advanced Cross-Language Engine) (Oard et al., 2008) a suggéré deux techniques de TR :

– La première est une TR assistée par l'utilisateur, dans ce cas, ce dernier peut choisir ou supprimer une traduction parmi toutes les traductions proposées. Le système avertit l'utilisateur lorsqu'il élimine une traduction correcte, il peut aussi reformuler et/ou affiner sa requête originale s'il n'est pas satisfait des résultats retenus.
– La seconde est une TR automatique dans laquelle chaque terme de la requête source est automatiquement traduit afin de retourner les documents correspondants.

Par ailleurs, le système offre plusieurs fonctionnalités telles que : (i) fournir le sens de chaque traduction en fonction des traductions passées, (ii) proposer un ensemble des synonymes possibles pour chaque terme et (iii) utiliser des textes liés à la traduction pour générer une collection d'exemples d'utilisation de traduction.

La langue des requêtes sources dans *MIRACLE* est l'anglais. Grâce à sa bonne conception, il peut être étendu pour accepter d'autres langues sources et offrir l'opportunité de profiter des ressources linguistiques existantes, indépendamment de

la langue du document. Toutefois, ce système n'utilise qu'une simple liste des termes bilingues, mais il est capable de travailler avec d'autres ressources supplémentaires.

Malgré que, *MIRACLE* surmonte certains inconvénients des outils d'interaction de RIT précédemment étudiés, ce système a plusieurs limites telles que :

- Toutes les traductions possibles doivent être vérifiées par l'utilisateur avant de choisir la bonne traduction, cette tâche est complexe et nécessite beaucoup du temps ;
- L'absence de toute information contextuelle explicative des traductions dans la langue de l'utilisateur ;
- La traduction automatique n'a pas permis à l'utilisateur d'améliorer le processus de TR avant de lancer le processus de recherche de documents.

2.4.9 SRIT de Kadri

(Kadri, 2008) a proposé un SRIT Anglais-Arabe, où la requête source est en anglais et la collection de documents en arabe. Ce système s'appuie sur deux contributions :

Dans la première, l'utilisateur profite d'une nouvelle méthode de lemmatisation basée sur des règles linguistiques pour trouver facilement le lemme correct d'un mot Arabe donné à travers des statistiques de corpus. Cette méthode est testée dans la RI monolingue avec des requêtes et documents en arabe et elle a donné des bons résultats par rapport à une méthode légère de lemmatisation qui ne collecte pas dans le même indexe les termes sémantiquement similaires.

Dans la deuxième contribution, Kadri a proposé une approche de TR pour combiner différentes ressources de traduction, telles que : les deux dictionnaires bilingues *Ajeeb* et *Almisbar*[12] qui sont extraits du Web et deux modèles de traduction statistiques (Kadri et Nie, 2004). Le premier modèle est formé à partir des pages Web parallèles, alors que le deuxième à partir des corpus parallèles des «*Nations Unies*». Cette approche est basée sur un facteur de confiance pour ajuster la pondération des traductions (Kadri et Nie, 2006, 2008). Les expérimentations sont achevées sur la collection TREC-2001, en utilisant 383 872 articles des journaux. Les résultats ont montré que le mixage de plusieurs ressources de traduction a donné une bonne efficacité pour la RIT.

2.4.10 [Mult]iSearcher

Dans le SRIT *[Mult]iSearcher*, (Farag et Nurnberger, 2013) ont intégré un algorithme de traduction automatique à l'approche proposée pour supporter l'interac-

12. http://www.almisbar.com/

tivité. Cette approche répond à plusieurs lacunes dans la TR, telle que la tâche de désambiguïsation dans le processus de RIT.

Ce système a considéré l'utilisateur comme une partie intégrante dans le processus de RIT : il peut interagir avec des informations contextuelles qui expliquent la traduction dans la langue désirée ce qui augmentera sa confiance. Le système *[Mult]iSearcher* intègre des interfaces simples qui fournissent à l'utilisateur un certain contrôle dès la soumission de sa requête jusqu'à l'obtention des documents pertinents. Cela améliorera la traduction et la performance globale du processus de RIT.

2.4.11 Sogou English

La compagnie *Sogou*[13] a construit un SRIT nommé *Sogou English*[14] pour franchir la barrière de langue et relier les chinois aux informations écrites en langues étrangères. Tout d'abord, le système traduit la requête de l'utilisateur du chinois vers l'anglais. Puis, il effectue une recherche sur le Web pour retrouver des documents anglais. Enfin, pour que les utilisateurs puissent lire et parcourir les informations du monde anglais, sans réellement connaître cette langue, *Sogou English* traduit les résultats de recherche vers le chinois.

(Xu et al., 2017) ont confirmé que *Sogou English* est le premier SRIT qui a intégré la technologie de *Traduction Automatique Neuronale* (TAN) dans son moteur de recherche. Cette technique a fait des progrès significatifs ces dernières années, en produisant des traductions de haute-qualité, voire meilleure que la traduction automatique statistique. Les auteurs ont confronté de nombreux défis techniques lors de l'intégration du système de traduction dans le moteur de recherche, à savoir : la polysémie des termes des requêtes et l'ambiguïté des traductions lors de la traduction des résultats de recherche en chinois.

2.4.12 SPEEDSER

Le système *SPEEDSER* (Système Possibiliste d'Expansion Et de Désambiguïsation SEmantique de Requêtes) est proposé par (Ben Khiroun, 2018). Ce système combine des méthodes de désambiguïsation et d'expansion de requêtes via des procédures automatisées et semi-automatisées en RI Monolingue ou Multilingue.

Le système *SPEEDSER* offre des interfaces Homme-Machine pour faciliter à l'utilisateur la reformulation en exploitant le module de navigation dans le graphe de

13. https://www.sogou.com/
14. http://english.sogou.com/

dictionnaire de type réseaux de petits mondes hiérarchiques (RPMH) pour un cadre de RI Monolingue. Le système consiste également à modéliser la pertinence en se basant sur la théorie des possibilités.

2.4.13 Synthèse et discussion

Nous présentons dans cette section une étude comparative (tableau 2.1) des SRIT étudiés précédemment, en citant plusieurs fonctionnalités et critères tels que : les langues supportées, l'adaptation du SRIT à des nouvelles langues, les ressources exploitées, le système fournit ou non un résumé pour chaque document récupéré dans la langue de l'utilisateur, le système fournit ou non une interface de visualisation, le système possède ou non un processus d'expansion de requête, le système supporte ou non la confiance envers la traduction, et existe-t-il une amélioration de traduction fournie par le SRIT ? Enfin, nous abordons les limitations de chaque système.

Nous nous focalisons, dans la suite, sur les principales caractéristiques qui doivent être prises en compte dans un processus de RIT (Elayeb et al., 2015 ; Farag et Nurnberger, 2012), telles que : (i) l'adaptation du SRIT à des nouvelles langues, (ii) la traduction automatique, (iii) la confiance envers la traduction et (iv) l'amélioration de la traduction suggérée par le SRIT.

Dans la conception d'un SRIT, nous devons prendre en compte que le système peut accepter d'autres langues. En effet, la fonctionnalité d'*adaptation à une nouvelle langue* n'existe que dans quelques SRIT que nous avons étudié, tels que, *UCLIR*, *UTACLIR, MIRACLE* et *SPEEDSER*. Cependant, si cette fonctionnalité existe déjà, elle n'était pas suffisamment expliquée et elle n'était pas décrite comment et dans quelle mesure un tel système peut être extensible (Farag et Nurnberger, 2013).

Les SRIT existants, comme *Keizai*, *UCLIR*, *MultiLex-Explorer* et *SRIT de Kadri*, n'ont pas supporté l'utilisateur par une *traduction automatique*, ni par une désambiguïsation de traduction automatique. En effet, des nombreux systèmes demandent à l'utilisateur d'effectuer le processus de désambiguïsation de traduction, en vérifiant manuellement toutes les traductions possibles afin de choisir celles qui lui conviennent. Ce processus nécessite beaucoup de temps et il est considéré comme une tâche pénible pour l'utilisateur. Dans d'autres systèmes, l'utilisateur doit vérifier toutes les traductions possibles ainsi que leurs définitions avant de choisir la traduction appropriée. Ce scénario est long et nécessite une connaissance approfondie de la langue cible.

La confiance envers la traduction fait partie des caractéristiques pertinentes qui doivent être impliquées dans tous les SRIT. Dans certains cas, l'utilisateur est in-

capable de lire ou de comprendre la traduction retournée par le système, de sorte qu'il ne peut pas utiliser ses traductions de façon correcte pour la désambiguïsation ou pour un processus d'expansion. Malheureusement, parmi les douze SRIT étudiés, cinq seulement (*Mulinex*, *WORDS*, *MultiLexExplorer*, *SRIT de Kadri* et *[Mult]iSearcher*) ont donné à l'utilisateur une certaine confiance envers les traductions retournées. Ces cinq systèmes ont bénéficié de la technique de «*traduction inverse*» vers la langue source. En effet, l'utilisateur sera mieux confiant de sa traduction lorsqu'il existe une similarité entre les termes de sa requête source et les traductions inverses. En outre, la technique de «*traduction inverse*» est sensible à la couverture limitée du lexique utilisé et à une certaine homonymie dans la langue cible lorsqu'aucun synonyme ne peut être trouvé pour un terme de requête source (Farag et Nurnberger, 2013).

Par ailleurs, plusieurs SRIT ont utilisé les définitions des dictionnaires pour donner à l'utilisateur une certaine confiance envers les traductions retournées. Ces dictionnaires bilingues fournissent, pour chaque traduction vers la langue cible, sa définition correspondante dans la langue source. Cependant, ce type de dictionnaire est souvent limité en termes de couverture et de disponibilité pour toutes les langues. De plus, ces définitions sont affichées pour chaque traduction indépendamment des autres termes de requête. Par conséquent, l'information contextuelle sera ignorée et ne peut pas être utile pour une tâche de désambiguïsation.

Par ailleurs, *l'amélioration de la traduction* nécessite que le SRIT doit considérer l'utilisateur comme une partie intégrante dans le processus de traduction, en lui fournissant des informations appropriées pour accomplir sa tâche. Parmi les douze SRIT étudiés, cinq systèmes seulement ont proposé une variété dans l'amélioration de traduction tels que *Patent CLIR*, *MultiLexExplorer*, *MIRACLE*, *SRIT de Kadri* et *[Mult]iSearcher*. Néanmoins, cette amélioration manquait différentes caractéristiques. Par exemple, certains systèmes fournissent d'autres informations pertinentes avec des résultats de recherche utiles pour les deux processus de désambiguïsation et d'expansion. En se basant sur les documents récupérés et examinés, l'utilisateur peut relancer un nouveau processus de TR. Par ailleurs, il risque de perdre beaucoup de temps et de se sentir frustré, lorsqu'il sent qu'il est incapable d'améliorer les résultats de traduction. D'autres systèmes ont profité des relations d'*EuroWordNet*, afin d'améliorer leurs traductions. Mais, cette ressource linguistique souffre de son espace de couverture linguistique limité et de la non-disponibilité pour toutes les langues (Elayeb et Bounhas, 2016).

Abréviations utilisées dans le tableau : *TA* Traduction automatique, *DB* Dictionnaire bilingue, *CP* Corpus parallèle, *DM* Dictionnaire multilingue, *ER* Expansion de requête, *CT* Confiance de traduction, *TI* Traduction inverse, *RD* Résumé

des documents, *VVD* Visualisation des vignettes de documents, *AT* Amélioration de traduction, *RR* Reformulation de requêtes, *Ar* Arabe, *En* Anglais, *Fr* Français, *De* Allemand, *It* Italien, *Es* Espagnol, *Ja* Japonnais, *Ko* Coréen, *Nl* Néerlandais, *Fi* Finnois, *Ch* Chinois.

Conclusion

Ce chapitre nous a servi à présenter un état de l'art sur le domaine de RIT. Tout d'abord, nous avons défini les différentes catégories d'approches de TR, telles que les approches à base de dictionnaire, les approches à base de traduction automatique, les approches à base des corpus et les approches à base des ressources combinées. Puis, nous avons présenté une synthèse sur ces approches, en notant les problèmes et les défis rencontrés par chaque type d'approches. Ensuite, nous avons présenté les différentes techniques de désambiguïsation, abordées par les chercheurs pour résoudre la désambiguïsation translinguistique, dans le but est de déterminer la traduction la plus appropriée. Enfin, nous avons cité et étudié les SRIT existants dans la littérature afin de synthétiser les principales caractéristiques qui doivent être prises en compte dans un processus de RIT.

À l'issue de cette étude, les chercheurs ont confirmé que les SRIT ne sont pas efficaces lorsque le besoin de l'utilisateur n'est pas bien exprimé. C'est pourquoi la plupart de ces systèmes consistent à traiter le problème d'ambiguïté qui est basée pratiquement sur des informations incertaines et imprécises. De ce fait, plusieurs théories ont été proposées pour la modélisation de ces types d'imperfection.

Dans le chapitre suivant, nous apportons une attention particulière aux théories des probabilités et des possibilités qui représentent un cadre prometteur pour la modélisation des informations entachées d'incertitude.

SRIT	**Langues supportées**	**Technique de traduction/ Ressources exploitées**	**Fonctionnalités**	**Limitations**
Mulinex	- Fr, De, et En -Adaptable à des nouvelles langues	- DB : 6 bases de données avec 100.000 à 200.000 entrées - TA : *LOGOS*	ER, CT, TI, raduction de requête assistée par l'utilisateur, RD	Synonymie et Homonymie
Keizai	- En-Ja - En-Ko	- DB : les actualités du Web archivées en Japonais à partir du ≪*Asahi Shimbun*≫ et en Coréen à partir du ≪*Dong-A Ilbo*≫. - CP à partir du Web	Traduction interactive des requêtes avec définition en anglais, RD en anglais, VVD	-Synonymie et Homonymie -Traduction de requête assistée par utilisateur nécessite beaucoup du temps
WORDS	Es-En	- TA : *SYSTRAN*	ER, CT , TI de l'anglais vers l'espagnol, RD	La tâche de désambiguïsation est couteuse et complexe.
UCLIR	- La langue arabe est incluse - Adaptable à des nouvelles langues	- DB à partir du Web - TA : *UNICODE* (Près de 1000 langues sont déjà incluses dans la norme	Requête multilingue, Requête anglaise interactive et non interactive, VVD	-L'approche de requête non interactive inclut la traduction non pertinente -L'approche de requête interactive est la traduction de requête assistée par utilisateur
UTACLIR	- De l'anglais au Fr, De, Es, It, Nl et Fi (bilingue à multilingue) - Adaptable à des nouvelles langues	- DB de plusieurs couples de langues : Motcom par Kielikone (https ://www.motstore.fi/) - *Oxford Duden* Allemand-Anglais (260 000 entrées) - DM de traduction Motcom GlobalDix (18 langues, nombre total des mots 665 000)	Traduction transitive, Gestion des composants, Correspondance des noms propres	Son efficacité dépend de plusieurs facteurs
MultiLex-Explorer	Support multilingue (Nl, Fr, En, Es, It)	- *EuroWordNet* - Moteur de recherche sur le Web : Google	ER,Visualisation des cercles, CT , AT, Catégorisation automatique	*EuroWordNet* n'est pas disponible pour toutes les langues
Patent CLIR	Ja-En	TA : *Google*, *Yahoo* et *Excite* (www.excite.co.jp/world/english/)	AT	Traduction de requête assistée par utilisateur
MIRACLE	- L'anglais et d'autres langues - Adaptable à des nouvelles langues	- TA - DB : Liste simple des termes bilingues	ER,WSD, AT, RR, TI, RD	-Les ressources ne sont pas disponibles -Synonymie et Homonymie
SRIT de Kadri	- En-Ar - RI monolingue en arabe	- DB : *Ajeeb* et *Almisbar* - CP : extrais automatiquement à partir du Web+ corpus des *Nations Unies*	ER, CT, AT	Les ressources ne sont pas disponibles
[Mult]i Searcher	Support multilingue	- DB - CP : *Europarl* - Information mutuelle	CT, AT, TI, RD	Corpus parallèles non disponibles pour toutes les langues
Sogou English	Ch-En	- Moteur de recherche sur le Web : *Sogou* - TA neuronale	WSD, Traductions des résultats de recherche vers le Chinois.	- Polysémie - Ambiguïté des traductions anglaises vers le chinois
SPEEDSER	- Fr-En -Adaptable à des nouvelles langues	- Liste bilingue de termes - CP : *Europarl*	ER, outils de visualisation (word sketch)	- La non disponibilité des ressources - Problème de couverture

Table 2.3 – Etude comparative entre les SRIT étudiés

Chapitre

3 Concepts de base des théories des probabilités et des possibilités

Sommaire

Introduction

Les approches de TR sont les plus populaires et les plus communément utilisées dans la RIT. Néanmoins, ces approches souffrent principalement des problèmes

de couverture des dictionnaires et d'ambiguïté des traductions, vu qu'elles sont basées sur des données pauvres, incertaines et imprécises. Plusieurs théories ont été proposées pour la modélisation de ces types d'imperfection comme la théorie des probabilités, la théorie des ensembles flous, la théorie des possibilités, les fonctions de croyance, etc.

En se basant sur notre connaissance et sur les approches de TR dans la littérature, l'unique approche disponible basée sur la théorie des possibilités dans le domaine de RIT est celle de (Ben Khiroun et al., 2018). Ce qui nous a motivé à pousser nos recherches dans cette direction et de proposer des nouvelles approches possibilistes de désambiguïsation des traductions dans la RIT.

Nous considérons l'ambiguïté des traductions d'un terme d'une requête source comme un cas d'imprécision. C'est pour cela que nous nous sommes inspirés de la théorie des possibilités qui s'applique naturellement à ce genre d'imperfection. Par contre, la théorie des probabilités n'est pas appropriée pour traiter ce type des données. En outre, et étant donné que la théorie des possibilités est le meilleur cadre approprié pour le traitement de l'imprécision, nous avons profité des distributions de possibilités pour résoudre le problème d'ambiguïté des traductions dans la tâche de RIT.

Dans ce chapitre, nous introduisons le cadre théorique sur lequel repose nos contributions : nous rappelons les notions de base des théories des probabilités, des possibilités et de transformation de la probabilité vers la possibilité. Nous détaillons aussi les méthodes de calcul pour chacune. Nous décrivons dans la section 3.1 la théorie des probabilités, et dans la section 3.2 la théorie des possibilités. Ensuite, nous présentons dans la section 3.3 une étude comparative entre ces deux théories. Enfin, nous détaillons dans la section 3.4 les transformations probabilité-possibilité.

3.1 La théorie des probabilités

La théorie des probabilités est une théorie mathématique, connue comme étant le cadre le plus ancien dans la modélisation des informations incertaines. Il s'agit principalement d'un outil quantitatif pour le traitement de l'incertitude qui touche souvent plusieurs domaines du monde réel tels que : la gestion, l'économie, les sciences, l'industrie, etc. Selon plusieurs chercheurs, la probabilité signifie un processus quantitatif (basé sur la fréquence) de traitement d'incertitude, alors que selon d'autres c'est plutôt une théorie de jeux ou une théorie aléatoire qui est indispensable pour modéliser n'importe quel problème réel (Bounhas, 2013).

La théorie des probabilités peut avoir deux interprétations principales (Jenhani,

2010), à savoir :

- *La théorie des probabilités objective* (fréquence classique), la probabilité d'un évènement est le rapport entre le nombre des cas favorables et le nombre de tous les cas possibles.
- *La théorie des probabilités subjective* (Bayésienne), la probabilité d'un évènement est le degré de confiance que l'on accorde à l'occurrence d'une réalisation d'un évènement A.

Nous nous focalisons dans cette section sur la distribution des probabilités (cf. section 3.1.1). Puis, nous présentons brièvement la probabilité conditionnelle (cf. section 3.1.2) et la modélisation statistique de la traduction (cf. section 3.1.3). Ensuite, nous présentons un résumé sur la méthode probabiliste naïve bayésienne pour la traduction des syntagmes nominaux (cf. section 3.1.4). Enfin, nous synthétisons cette section (cf. section 3.1.5).

3.1.1 Distribution des probabilités

Soit Ω l'univers du discours, une distribution de probabilité, associe à chaque singleton ω_i un degré dans l'intervalle $[0, 1]$:

$$P : \Omega \rightarrow [0, 1]$$
$$\forall \omega_i \in \Omega, \omega_i \rightarrow p(\omega_{i)}$$

Une mesure de probabilité $P(A)$ est une occurrence, associe à un évènement $A \subseteq \Omega$ un degré dans $[0, 1]$. $P(A)$ est représentée selon le suivant :

$$P(A) = \sum_{\omega_i \in A} p(\omega_i) \tag{3.1}$$

Une mesure de probabilité $P(A)$ doit vérifier les trois propriétés suivantes (Alsahwa, 2014) :

- *La normalisation* : $P(\phi) = 0$ et $P(\Omega) = \sum_{\omega_i \in A} p(\omega_i) = 1$
- *L'additivité* : $\forall A,\ B \subseteq \Omega,\ P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A) + P(\bar{A}) = 1$, avec $\bar{A}$ est l'évènement complémentaire de A.

3.1.2 La probabilité conditionnelle

Le raisonnement de *Bayes* est une continuation du raisonnement probabiliste classique qui permet de modéliser une situation incertaine. La règle de *Bayes* peut être

définie comme étant un modèle par lequel l'expert peut mettre à jour sa croyance sur un évènement A, en utilisant des probabilités conditionnelles (Bounhas, 2013).

La probabilité conditionnelle $P(A|B)$ estime la probabilité qu'un évènement A se produise si B s'est produit. Cette probabilité est formulée par *Bayes* comme suit :

$$\forall A,\ B \subseteq \Omega,\ P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{3.2}$$

3.1.2.1 Théorème de Bayes

Le théorème de *Bayes* est un théorème de la théorie des probabilités. Il offre une représentation mathématique entre la probabilité conditionnelle $P(A|B)$ et la probabilité conditionnelle inverse $P(B|A)$.

Considérons la définition de la probabilité conditionnelle :

$P(A \cap B) = P(A|B) * P(B)$

Suivant la règle du produit, nous avons :

$P(A \cap B) = P(B|A) * P(A) = P(A|B) * P(B)$

ce qui conduit au théorème de *Bayes* suivant :

$$\forall A,\ B \subseteq \Omega,\ P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \tag{3.3}$$

3.1.3 La modélisation statistique de traduction

Les méthodes statistiques permettent de traiter un processus linguistique comme la traduction. Au début des années 90, Brown et son équipe (Brown et al., 1990) ont développé à *IBM* des distributions et des estimations probables pour la modélisation statistique de traduction automatique. Cette traduction dépend principalement de la disponibilité des corpus parallèles. La traduction statistique se définit par la recherche d'une phrase cible qui a la plus grande probabilité d'être la traduction d'une phrase source. Lorsque nous appliquons le théorème de *Bayes* sur la paire des phrases (S, T), où la phrase S est dans la langue source et la phrase T est dans la langue cible, nous obtenons la probabilité $P(T|S)$ donnée par la formule (3.4) (Brown et al., 1993) :

$$P(T|S) = \frac{P(T) * P(S|T)}{P(S)} \tag{3.4}$$

Le dénominateur de cette équation ne dépend pas de T, donc il suffit de choisir le T qui maximise le produit $P(T) * P(S|T)$. Le but de la traduction automatique est de trouver la meilleure traduction T^*, nous obtenons l'équation suivante :

$$T^* = argmax_T P(T|S) = argmax_T P(T) * P(S|T) \tag{3.5}$$

Selon (Affi, 2014) le problème de traduction peut être divisé en trois sous problèmes, à savoir : Le calcul du modèle de langue $P(T)$ (section 3.1.3.1), le calcul du modèle de traduction $P(S|T)$ (section 3.1.3.2) et le mécanisme de maximisation de l'équation $argmax_T P(T|S)$ (section 3.1.3.3). Dans les sections suivantes, nous détaillons chacun de ces sous problèmes.

3.1.3.1 Modèle de langue

Le modèle de langue constitue un modèle de base pour les systèmes de traduction statistique. Aussi, il est utilisé pour plusieurs applications, telles que la reconnaissance automatique de la parole et la correction orthographique.

Cette tâche est réalisée au cours d'une phase d'apprentissage, elle permet de chercher la traduction la plus probable, en se basant sur des connaissances extraites à partir d'un corpus monolingue dans la langue cible. Ce modèle suit les contraintes syntaxiques et sémantiques exigées par la langue cible. Si nous considérons que T comme une suite de termes ($T = \{t_1, t_2, \ldots, t_n\}$), la probabilité de $P(T)$ est calculée comme suit (Boughanem et al., 2004) :

$$P(T) = P(t_1, t_2, \ldots, t_n) = \prod_{i=1}^{n} P(t_i|t_1, .., t_{i-1}) \tag{3.6}$$

Cette formule peut être écrite autrement, par l'hypothèse que le terme t_i de T ne dépends que des $(n-1)$ termes précédents :

$$P(T) = P(t_1)P(t_2|t_1)P(t_3|t_1t_2) * \ldots * P(t_1|t_1t_2, \ldots, t_{n-2}t_{n-1}) \tag{3.7}$$

À partir de la formule (3.7) le modèle de langue a une nouvelle appellation : *modèle de n-gramme*.

3.1.3.2 Modèle de traduction

Le modèle de traduction permet de déterminer qu'une phrase source donnée se traduit en une phrase cible équivalente (Blain, 2013). L'apprentissage d'un tel modèle

consiste à exploiter un corpus parallèle, composé de couples de textes alignés au niveau de phrase.

Notion d'alignement

L'apprentissage de modèle de traduction repose sur la notion d'alignement. Cette dernière consiste à associer chaque mot ou ensemble de mots de la phrase source le mot ou l'ensemble de mots de la phrase cible correspondant à sa traduction. La probabilité de traduction $P(S|T)$ est estimée par la somme des alignements entre les éléments de S et les éléments de T :

$$P(S|T) = \sum_{a \in A} P(S, a|T) \tag{3.8}$$

où a est un alignement possible entre le $i^{ième}$ mot de S et le $j^{ième}$ mot de T, tel que :

$$a : \ S_i \rightarrow T_j$$

Pour une phrase source N et une phrase source M, nous avons $(N+1)^M$ alignements possibles entre les mots. Si nous appliquons à notre exemple d'alignement de la figure 3.1, la fonction d'alignement a serait alors :

$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 5, 5 \rightarrow 4\}$$

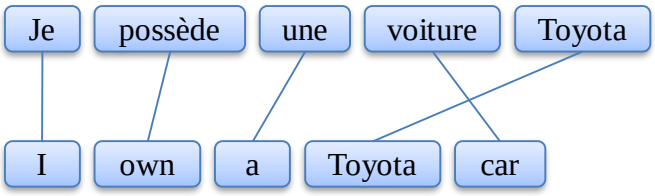

Figure 3.1 – Exemple d'alignement pour le couple de langue Français-Anglais

Il est possible d'avoir un mot qui n'a pas son correspondant dans l'autre phrase. Ce mot sera aligné par le mot spécial «*Null*». Par ailleurs, l'aspect asymétrique de l'alignement n'autorise pas que plusieurs mots de la langue source puissent s'aligner avec un même mot de la langue cible. Les modèles d'*IBM* permettent d'y remédier en réalisant un alignement bi-directionnel : cible vers source et source vers cible (Blain, 2013).

Le système d'alignement le plus utilisé est *Giza++* [1], il est disponible gratuitement et met en œuvre les modèles d'*IBM*.

Les modèles de traduction « IBM »

(Brown et al., 1993) ont développé une série de cinq modèles de traduction de complexité croissante avec les algorithmes nécessaires pour estimer leurs paramètres. Chaque modèle donne une démarche pour calculer la probabilité conditionnelle afin de modéliser la traduction statistiquement. Nous résumons dans la suite la fonctionnalité de chaque modèle.

– ***Le modèle IBM1 : Probabilité de traduction lexicale***

Dans ce modèle, les alignements sont équiprobables, cela veut dire que l'ordre des mots n'est pas considéré et tous les mots *f,* des phrases sources, sont alignés à tous les mots e des phrases cibles avec la même probabilité. Le modèle *IBM1* repose sur une seule probabilité de traduction lexicale [2] $P(f|e)$.

– ***Le modèle IBM2 : Ré-ordonnancement***

Dans ce modèle, l'ordre des mots est pris en compte. Ce modèle intègre un modèle de ré-ordonnancement qui présente la distance entre un mot de la phrase source s et le mot de la phrase cible t.

– ***Le modèle IBM3 : Fertilité***

IBM3 introduit la notion de fertilité qui permet à un mot source de générer non seulement un seul mot, mais plusieurs mots cibles.

– ***Le modèle IBM4 : Distorsion***

Ce modèle est pareil à *IBM3*, avec beaucoup plus de complexité au niveau de ré-ordonnancement. *IBM4* introduit alors une probabilité de distorsion relative qui dépend des structures des mots liés et aussi des positions des autres mots cibles connectés avec le même mot source.

– ***Le modèle IBM5 : Déficience***

IBM5 reste toujours le modèle le plus utilisé, il traite les déficiences des modèles *IBM3* et *IBM4*.

Le tableau 3.1 résume les avantages et les inconvénients des cinq modèles étudiés.

1. http://www.statmt.org/moses/giza/GIZA++.html
2. la probabilité de traduction lexicale : chaque mot est traduisible par n traductions équivalentes.

Modèle	**Avantages**	**Inconvénients**
IBM1	- Repose sur la probabilité de traduction lexicale (basée sur les mots). - Tout les modèles d'*IBM* reposent sur *IBM1*.	- Ne prend pas en considération l'ordre des mots dans la phrase. - ≪*garbage collector*≫ : permet d'aligner les mots rares de la phrase cible par nombreux mots de la phrase source. - Ne capte pas la dépendance entre les mots qui sont alignés, ce qui induit à des effets de distorsion.
IBM2	- Comble les défaillances d'*IBM1.* - L'équiprobabilité des alignements n'est plus considérée : un modèle de ré-ordonnancement. - Dispose d'une loi d'alignement ou de distorsion.	- C'est un modèle d'entraînement pour initialiser les paramètres du modèle *IBM3*.
IBM3	- Intègre la fertilité comme un paramètre.	- Il existe un problème avec le paramétrage des probabilités de distorsion (Brown et al., 1993).
IBM4	- Introduit la probabilité de distorsion relative. - Les résultats obtenus sont très bons.	- Utilise une modélisation un peu plus complexe pour le ré-ordonnancement des mots. - Mathématiquement, les résultats ne sont pas justes. - Le modèle de distorsion ne considère pas les positions cibles déjà sélectionnées pour un alignement et la masse des probabilités d'alignement s'en trouve tronquée.
IBM5	- Traite les déficiences d'*IBM3* et *IBM4*. - Corrige la déficience du modèle *IBM4*, en conservant un historique des positions toujours vacantes durant un processus d'alignement.	- Utilise une modélisation un peu plus complexe pour le ré-ordonnancement des mots. - La complexité du modèle rend son utilisation moins attractive.

Table 3.1 – Les avantages et les inconvénients des modèles d'IBM

3.1.3.3 Mécanisme de maximisation

Le mécanisme de maximisation (figure 3.2) s'appelle aussi un décodeur, son but est de trouver la traduction la plus probable pour une phrase source S en combinant les deux modèles de langue et de traduction. Ce mécanisme représente l'équation $argmax_T P(T|S)$ telle que, $P(T)$ correspond au modèle de langue et $P(S|T)$ correspond au modèle de traduction.

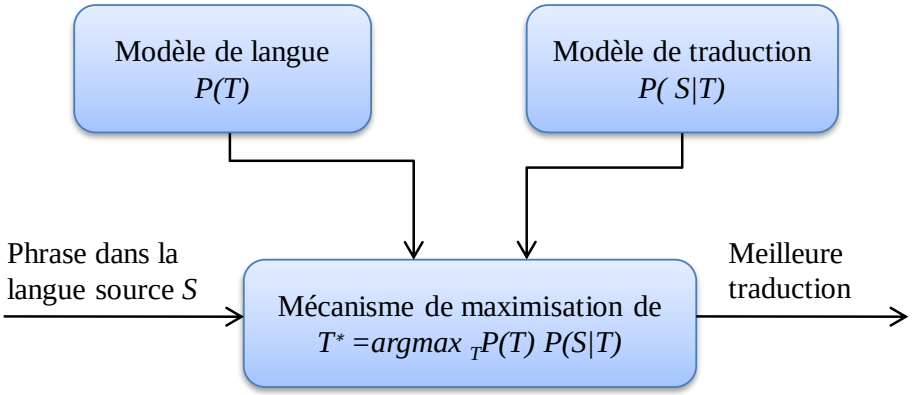

Figure 3.2 – Les étapes de modélisation d'une traduction automatique statistique

3.1.4 La traduction probabiliste des syntagmes nominaux

Les approches Naïves Bayésiennes pour la Désambiguïsation de Traduction (NBDT) sont inspirées de la règle de *Bayes*. Dans ces techniques, les variables d'entrée sont supposées indépendantes. En dépit de leurs facilités, les approches NBDT peuvent être plus efficaces que certaines approches de TR reconnues (Gao et al., 2006, 2001). En outre, nous considérons une approche NBDT comme un réseau bayésien, dans lequel les termes de la requête source sont censés être conditionnellement indépendants compte tenu de la traduction de chacun (Ben Amor et al., 2002).

Considérons un syntagme nominal NP (Noun Phrase) français modélisé comme un vecteur $FNP = \{f_1, \ldots, f_n\}$, avec le patron de son NP, FPT. À l'aide du dictionnaire bilingue Français-Anglais, nous cherchons toutes les traductions anglaises disponibles pour chaque terme français en FNP. Nous avons également profité de tous les patrons de traduction disponibles EPT pour FPT. Le meilleur syntagme traduit en anglais, $ENP^* = \{e_1, \ldots, e_m\}$, est celui qui maximise la formule (3.9) :

$$\begin{aligned} ENP^* = \underset{ENP}{argmax}(P(ENP|FNP)) = \underset{ENP}{argmax}(P(FNP|ENP) * P(ENP)) \\ = \underset{e_j}{argmax}(P(e_j) * \prod_{i=1}^{n} P(f_i|e_j)) \end{aligned} \quad (3.9)$$

En utilisant la règle de *Bayes* :

$$P(e_j|f_1, f_2, \ldots, f_n) = \frac{P(e_j) * P(f_1, f_2, \ldots, f_n|e_j)}{P(f_1, f_2, \ldots, f_n)} \quad (3.10)$$

Où : $P(FNP|ENP)$ est la probabilité de traduction, et $P(ENP)$ est une probabilité a priori des mots du NP anglais traduit.

Nous pouvons ignorer le facteur de normalisation $P(f_1, f_2, \ldots, f_n)$ puisqu'il ne dépend pas de la traduction.

Nous estimons le facteur le plus important de la formule (3.10), à savoir : $P(f_1, f_2, ..., f_n|e_j)$, en utilisant les données de traitement. Nous pouvons décomposer la vraisemblance en un produit des termes, comme indiqué dans la formule (3.11). En fait, l'approche *naïve bayésienne* suppose que les probabilités conditionnelles des termes de la requête source sont statistiquement indépendantes.

$$P(e_j) * P(f_1, f_2, \ldots, f_n|e_j) = \prod_{i=1}^{n} P(f_i|e_j) \tag{3.11}$$

Étant donné un NP (*FNP* ou *ENP*), comme un ensemble de termes (F ou E) regroupés par patron de NP (FPT ou EPT). Nous supposons que la traduction des termes et des patrons des NPs sont indépendants comme indiqué dans la formule (3.12) :

$$\begin{aligned} P(FNP|ENP) &= P(F, FPT|E, EPT) \\ &= P(F|E, EPT) * P(FPT|E, EPT) = P(F|E) * P(FPT|EPT) \end{aligned} \tag{3.12}$$

Remplacer la formule (3.12) dans la formule (3.9) comme suit :

$$ENP^* = \underset{ENP}{argmax}(P(F|E) * P(FPT|EPT) * P(ENP)) \tag{3.13}$$

Où, $P(F|E)$ est la probabilité de traduction des mots anglais E en *ENP* aux mots français F dans *FNP*. Compte tenu du patron anglais EPT, $P(FPT|EPT)$ est la probabilité du patron de traduction FPT. Ces probabilités sont estimées comme suit :

$$P(FPT|EPT) = \frac{Occ(FPT, EPT)}{Occ(EPT)} \tag{3.14}$$

Où, $Occ(EPT)$ est le nombre d'occurrence d'EPT dans la partie anglaise du corpus bilingue aligné.

$Occ(FPT, EPT)$ est le nombre de fois que FPT correspond à EPT dans les phrases alignées.

$P(ENP)$ est déterminé par le modèle de langue anglais de Trigramme tel que donné dans la formule (3.15) :

$$P(ENP) = P(e_1, \ldots, e_n) = \prod_{i=1}^{n} P(e_i | e_{i-2}, e_{i-1}) \tag{3.15}$$

3.1.5 Synthèse

La théorie des probabilités est un bon outil, basé sur des théories mathématiques solides et riches. Cette théorie permet de traiter l'imperfection, à savoir les informations entachées par l'incertitude (Jenhani, 2010). Néanmoins, cette théorie a montré des sérieux problèmes dans la modélisation de l'incertitude subjective, de l'imprécision et des situations d'incertitude totale (Dubois et Prade, 2006). Par ailleurs, la théorie des possibilités appartient à la famille des théories modernes de l'incertain, elle s'applique naturellement à ce genre d'imperfection.

Nous détaillons dans la section suivante la théorie des possibilités. Nous examinons les éléments fondamentaux de cette théorie tels que : la distribution des possibilités (cf. section 3.2.1), les mesures de possibilité et de nécessité (cf. section 3.2.2), et les réseaux possibilistes (cf. section 3.2.3).

3.2 La théorie des possibilités

(Zadeh, 1978) a introduit la théorie des possibilités pour traiter l'incertitude où les connaissances sont exprimées d'une manière ambiguë. La théorie des possibilités est relativement nouvelle par rapport aux autres théories du traitement de l'incertitude. En fait, c'est l'une des théories, les plus utilisées, qui fournit des outils mathématiques pour le traitement des informations incomplètes et des données imprécises et incertaines (Bounhas et al., 2013 ; Moussa et Alfred, 2014). Cette théorie a été développée et étendue de manière qualitative ou quantitative par d'autres auteurs tels que (Dubois et Prade, 1985, 1998) et (De Cooman et Aeyels, 1999, 2000).

Contrairement à la théorie des probabilités, la théorie des possibilités distingue entre *l'imprécision* et *l'incertitude* de la façon suivante :

- *L'incertitude* est expliquée comme étant l'information supplémentaire qui soutient le fait de ne pas fournir ou de connaître une déclaration (un état de la réalité) pour identifier la valeur de vérité d'une proposition (probabilité). Une information est considérée incertaine, si on ne sait pas si celle-ci est vraie ou fausse. Considérons les exemples suivants : «*La probabilité pour que l'opération prenne plus d'une heure est 0,7*», «*Il est très possible*

qu'il neige demain», «*Il n'est pas absolument certain que Jean vienne à la réunion*» (Dubois et Prade, 2000).

- *L'imprécision* correspond à l'idée de l'information incomplète ou lorsqu'un certain état de la réalité est ambigu, c'est-à-dire lorsqu'il est décrit par une variable propositionnelle de valeurs multiples (ensemble flou (Zadeh, 1978)) (Ayed et al., 2012). Une information imprécise prend la forme d'une telle disjonction de valeurs mutuellement exclusives. Par exemple, dire que «*Pierre a entre 20 et 25 ans*», soit $v = \hat{a}ge(Pierre) \in \{20, 21, 22, 23, 24, 25\}$, c'est supposer que $v = 20$ ou $v = 21$ ou $v = 22$ ou $v = 23$ ou $v = 24$ ou $v = 25$ (Dubois et Prade, 2000). Dans le cas de la TR, les traductions sont des cas d'imprécision et même pour désambiguïser une traduction donnée nous utilisons le contexte de la requête source qui est aussi ambigu.

Nous avons indiqué précédemment que la théorie des probabilités peut être interprétée de diverses manières (objective et subjective). Dans la théorie des possibilités, le terme possibilité peut aussi soutenir différentes interprétations (Dubois et Prade, 1998), à savoir :

- *Interprétation physique* : «*possible*» signifie facile à réaliser, réalisable, prenons l'exemple de la phrase suivante « *il est possible pour Hans de manger six œufs pour le petit-déjeuner*».
- *Interprétation épistémique* : «*possible*» signifie plausible, le cas dans la phrase suivante «*il est possible qu'il pleuve ce soir*».
- *Interprétation logique* : c'est la consistance. La vision logique de la théorie des possibilités est facilement introduite en présence des informations incomplètes. Par exemple, «*il est possible que Mike soit célibataire*», sachant qu'il est jeune et qu'il n'a pas d'enfants.

3.2.1 La distribution des possibilités

L'un des concepts fondamentaux de la théorie des possibilités est celui de la notion de distribution de possibilité, notée π. Considérons l'univers de discours $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$, π correspond à une fonction qui associe à chaque élément $\omega_i \in \Omega$ une valeur d'une échelle possibiliste L. L'échelle possibiliste peut être illustrée par l'intervalle $[0, 1]$. $\pi(\omega)$ est appelée le degré de possibilité qui encode une connaissance du monde réel. Selon (Borgelt et Gebhardt, 1999) la meilleure façon d'expliquer la différence, entre une information incertaine et imprécise, est de considérer la notion de degré de possibilité.

Par convention, $\pi(\omega_i) = 0$, signifie qu'il est impossible que ω_i soit le monde réel, $\pi(\omega_i) = 1$ signifie qu'il est plausible (complètement possible) que ω_i soit le monde

réel. Une distribution des possibilités π est considérée normalisée s'il existe au moins un état ω_i qui est totalement possible (Ben Amor et al., 2002).

L'échelle possibiliste a deux interprétations, à savoir : (i) *L'échelle numérique* : lorsque les valeurs traitées ont un sens réel, c'est le paramètre *quantitatif* qui peut être appliqué à l'aide de l'opérateur « *produit* ». Et, (ii) *l'échelle ordinale* : lorsque les valeurs traitées ne reflètent qu'un ordre entre les différents états du monde, c'est le paramètre *qualitatif* qui peut être appliqué à l'aide de l'opérateur « *minimum* ».

Dans la théorie des possibilités, il existe différents cas particuliers, qui sont cruciaux pour la situation de connaissance, et qui peuvent être représentés comme suit (Alsahwa, 2014 ; Bounhas, 2013 ; Dubois et Prade, 1998) :

- *La connaissance complète* : il existe un seul singleton qui dispose d'une valeur maximale de possibilité, alors que tous les autres singletons ont une possibilité d'occurrence nulle ; $\exists\, \omega_i \in \Omega$, $\pi(\omega_i) = 1$ and $\forall \omega_j \neq \omega_i$, $\pi(\omega_j) = 0$
- *L'ignorance partielle* ; $\forall \omega_i \in A$, $\pi(\omega_i) = 1$, $\forall \omega_i \notin A$, $\pi(\omega_i) = 0$, lorsque l'évènement A n'est pas un singleton.
- *L'ignorance totale* : nous n'avons aucune connaissance qui nous permette de favoriser (ou défavoriser) les différents singletons. Ces derniers sont considérés possibles et ont tous la même valeur maximale de possibilité ; $\forall \omega_i \in \Omega$, $\pi(\omega_i) = 1$

3.2.2 Les mesures de possibilité et de nécessité

Contrairement à la théorie des probabilités qui utilise une seule mesure qui est la mesure de Probabilité (P), la théorie des possibilités utilise deux mesures : la mesure de la Possibilité (Π) et la mesure de Nécessité (N). La Possibilité et la Nécessité sont deux mesures duales dans lesquelles une distribution de possibilité π sur Ω permet aux évènements d'être qualifiés en termes de *plausibilité* et de *certitude*. Les mesures de possibilité sont dites *maxitives* et les mesures de nécessité sont dites *minitives*, contrairement aux mesures de probabilité qui sont *additives* (Dubois et Prade, 2006).

Considérons une distribution de possibilité π sur l'univers du discours Ω. Ces deux mesures duales caractérisent l'incertitude liée à la réalisation de chaque sous-ensemble (un évènement) $A \subseteq \Omega$ (Alsahwa, 2014). Ces deux mesures sont définies comme suite (Bounhas et al., 2013) :

$$\Pi(A) = max_{\omega \in A} \pi(\omega) \tag{3.16}$$

$$N(A) = min_{\omega \notin A}(1 - \pi(\omega)) \tag{3.17}$$

$\Pi(A)$ évalue à quel niveau A correspond à notre connaissance exprimée par π. Alors que $N(A)$ évalue à quel niveau A est certainement impliqué par notre connaissance exprimée par π. $\Pi(A)$ évalue dans quelle mesure A est logiquement cohérente avec π. Alors que $N(A)$ évalue dans quelle mesure A est implicitement impliqué par π. La dualité possibilité-nécessité dit qu'une proposition est certaine si son contraire est impossible, et cela s'exprime par (Dubois et Prade, 2016) :

$$N(A) = 1 - \Pi(\bar{A}) \tag{3.18}$$

où : $\bar{A}$ est le complément de A.

Généralement, $\Pi(\Omega) = N(\Omega) = 1$ et $\Pi(\varnothing) = N(\varnothing) = 0$. Les mesures de possibilité satisfont la propriété de base « *maxitive* ».

$$\Pi(A \cup B) = max(\Pi(A), \Pi(B)) \tag{3.19}$$

Les mesures de nécessité répondent à un « *axiome de minitivité* » dual à celui des mesures de possibilité, à savoir :

$$N(A \cap B) = min(N(A), N(B)) \tag{3.20}$$

Exprimant qu'être certain de $(A \cap B)$, est le même d'être certain de A et de B (Dubois et Prade, 2016).

Les distributions des possibilités quantitatives peuvent représenter une famille de mesures de probabilité (Dubois et Prade, 1992). En effet, une mesure de possibilité peut être considérée comme *une limite supérieure* d'une mesure de probabilité, et aussi associée à la famille de mesures de probabilité, définie par :

$$\mathcal{F}(\pi) = \{P | \forall A, \Pi(A) \geq P(A)\} \tag{3.21}$$

Grâce à la dualité entre Π et N, et à l'auto-dualité de $P(P(A) = 1 - P(\bar{A}))$, il est évident que :

$$\forall P \in \mathcal{F}(\Pi), \forall A, \Pi(A) \geq P(A) \geq N(A) \tag{3.22}$$

Donc, nous avons (De Cooman et Aeyels, 1999 ; Dubois et Prade, 1992) :

$$sup_{P\in\mathcal{F}(\pi)}P(A) = P(\Pi) \; et \; inf_{P\in\mathcal{F}(\pi)}P(A) = N(A)$$

D'après (Baudrit, 2005), la famille $\mathcal{F}(\Pi)$ est déterminée par les intervalles de probabilité qu'elle génère et peut être codée par π.

3.2.3 Les Réseaux Possibilistes (RP)

Les réseaux possibilistes (Ben Amor et al., 2002 ; Benferhat et Smaoui, 2007 ; Borgelt et Gebhardt, 1999) sont des modèles graphiques utilisés pour traiter la représentation et le traitement d'informations incertaines. Par ailleurs, le développement de ces réseaux a été déclenché par le fait que les réseaux probabilistes sont bien adaptés pour représenter et traiter des informations incertaines, mais ne peuvent pas facilement être étendus pour traiter des informations imprécises (Borgelt et Gebhardt, 1999).

3.2.3.1 Notations et définitions des graphes

Les caractéristiques principales d'un réseau possibiliste sont les composants numériques et graphiques (Bounhas, 2013), à savoir :

- *Le composant graphique* est un Graphe Acyclique Dirigé (GAD) qui permet de représenter une dépendance conditionnelle entre les variables dépendantes ou indépendantes. Chaque nœud dans le graphe indique une variable du domaine étudié, et chaque lien indique une dépendance entre deux variables. La structure du graphe encode les ensembles de relations d'indépendance entre les nœuds.
- *Les liens du graphe* sont quantifiés via un *composant numérique*. Ces liens représentent une matrice de possibilité conditionnelle de chaque nœud, compte tenu du contexte de ses parents.

Les distributions des possibilités doivent respecter les conditions de la normalisation de chaque variable V du graphe (Benferhat et al., 1999 ; Benferhat et Smaoui, 2007) :

- Pour les nœuds V_i qui sont des racines ($Dom(V_i) = \phi$), la possibilité doit satisfaire la condition suivante :

$$max_{v\in Dom(V_i)}\Pi(v) = 1, \; \forall v_i \in Dom(V_i) \tag{3.23}$$

- Pour chaque autre nœud V_i, la distribution conditionnelle de V dans le contexte de ses parents (désigné par U_V) doit satisfaire :

$$max_{v\in Dom(V_i)}\Pi(v|u_V) = 1; u_V \in Dom(U_V) \tag{3.24}$$

Où : $Dom(U_V)$ est le domaine de l'ensemble de parents de V.

Sachant qu'il y a un conditionnement basé sur le *produit* (respectivement, *minimum*) pour le paramétrage *quantitatif* (respectivement, *qualitatif*), les réseaux possibilistes peuvent également être divisés en réseaux *quantitatifs* et *qualitatifs*.

3.2.3.2 Les réseaux possibilistes quantitatifs

Les réseaux possibilistes quantitatifs sont confortables pour le réglage numérique, où les mesures de possibilité sont des réels dans l'intervalle $[0, 1]$ (Dubois et Prade, 2006). Les réseaux possibilistes quantitatifs associent une distribution des possibilités π_p basée sur le «*produit*». Cette distribution est calculée sur l'ensemble des variables $V = \{V_1, ..., V_n\}$, en utilisant la règle de chaînage suivante (Bounhas, 2013) :

$$\pi_p(V_1, ..., V_n) = \prod_{i=1}^{n} \Pi(V_i|U_{V_i}) \tag{3.25}$$

3.2.3.3 Les réseaux possibilistes qualitatifs

Les réseaux possibilistes qualitatifs sont utilisés lorsque seul l'ordre induit par les mesures de possibilité est pris en compte. Les réseaux possibilistes qualitatifs associent une distribution des possibilités basée sur l'opérateur «*minimum*». Elle est calculée sur l'ensemble des variables $V = \{V_1, ..., V_n\}$, en utilisant l'équation suivante (Bounhas, 2013) :

$$\pi_p(V_1, ..., V_n) = min_{i=1,..,n}\pi(V_i|U_{V_i}) \tag{3.26}$$

3.3 Synthèse et étude comparative entre les théories des probabilités et des possibilités

Une différence principale entre une distribution des possibilités π et une distribution des probabilités p est que cette dernière est une mesure normalisée qui exige que la somme des probabilités des éléments dans l'univers du discours Ω soit égale à 1, alors que dans la théorie des possibilités, il n'existe aucune contrainte de ce type (Bounhas, 2013).

La théorie des probabilités prétend que la probabilité d'un évènement est déterminée par la probabilité de son complément, contrairement à la théorie des possibilités qui

n'exige pas des mesures additives. Ainsi, l'inconvénient crucial de l'utilisation des probabilités dans le traitement d'incertitude réside dans la nécessité d'énumérer un ensemble exhaustif d'alternatives mutuellement exclusives (Dubois et Prade, 2006). Dans le monde réel, il est difficile pour un expert de fournir des évènements exhaustifs et mutuellement exclusifs, parce que ses connaissances augmentent au fil du temps et que l'incertitude au sujet de la situation diminue (Bounhas, 2013).

Plusieurs chercheurs (par exemple (Zadeh, 1978)) ont montré la différence entre ces deux théories comme suit : considérons X le nombre d'œufs que *Hans* va manger demain matin. Soit $\pi(x)$ le degré de facilité avec lequel *Hans* peut manger x œufs, et $p(x)$ la probabilité que *Hans* mange x œufs demain. $X \in \Omega = \{1, 2, 3, 4, 5\}$. Selon les valeurs données dans le tableau 3.2 ci-dessous :

x	1	2	3	4	5
$\pi(x)$	1	1	$0,7$	$0,4$	0
$p(x)$	$0,3$	$0,5$	$0,1$	$0,1$	0

Table 3.2 – Valeurs de possibilité et de probabilité associées à X

Il est assez difficile de tirer des conclusions sur la relation entre les théories des probabilités et des possibilités. En fait, nous pouvons constater que les éléments hautement possibles ne sont pas nécessairement hautement probables (par exemple $\pi(x = 1) = 1$ et $p(x = 1) = 0,3$). La seule conclusion qui peut être tirée, est que les éléments impossibles sont nécessairement improbables (par exemple $\pi(x = 5) = 0$ et $p(x = 5) = 0$). Inversement, nous ne pouvons pas décider si les éléments peu probables sont peu possibles ou non.

Dans certains cas, la distribution des possibilités est plus expressive. En fait, les mesures de possibilité peuvent refléter *l'ignorance* : en particulier, la distribution $\pi(\omega) = 1; \forall\omega \in \Omega$ est une expression *d'ignorance totale* qui reflète l'absence de toute information pertinente, ce qui signifie que l'expert «*sait qu'il ne sait pas*». Néanmoins, dans la théorie des probabilités, l'ignorance totale est modélisée par une répartition uniforme qui se traduit par l'attribution des poids égaux pour un n évènements, tel que $p(\omega) = 1/n, \forall\omega \in \Omega$, alors qu'aucune raison ne peut clarifier cette obligation subjective. La situation d'ignorance totale n'est pas fidèlement rendue par une seule mesure de probabilité (Dubois et Prade, 2009).

Les règles de calcul dans la théorie des probabilités sont principalement basées sur les *produits* et les *sommes*. L'utilisation de ces opérateurs favorise la propagation des erreurs, essentiellement lorsque le processus de calcul est trop long. Comme conséquence, le modèle probabiliste distingue beaucoup entre les alternatives, alors que cette distinction peut être un manque de sens dans certaines situations. Nous

pouvons éviter ce problème, en utilisant les opérateurs *min* et *max* qui font le sens principal de la théorie des possibilités ordinales.

Transformer une mesure de probabilité en une mesure de possibilité est parfois nécessaire. Ce type de transformation est utile en présence de faibles sources d'informations qui rendent les données des probabilités irréalistes, ou dans le cas où le calcul des possibilités est plus simple que celui des probabilités (Dubois et al., 2004 ; Oussalah, 2000).

Plusieurs ponts entre ces deux cadres ont été établis. Surtout que, plusieurs recherches ont étudié le problème de la transformation des distributions de probabilités en possibilité et vice-versa. (Ben Slimen et al., 2013 ; Dubois et al., 1993, 2004). Nous nous concentrons, dans la section suivante, sur la transformation probabilité-possibilité. Nous présentons les règles de transformation les plus connues, dans la littérature, et nous citons les opportunités offertes par cette transformation légitime.

3.4 Transformation probabilité-possibilité

Le premier intérêt de la transformation probabilité-possibilité, c'est d'étudier la cohérence entre ces deux théorèmes, plus précisément la cohérence des distributions dérivées. La transformation probabilité-possibilité trouve son origine dans le «*principe de consistance*» de (Zadeh, 1978) avec la formule C_z suivante :

$$C_z = \sum_{i=1}^{n} \pi_i p_i \tag{3.27}$$

avec : $p = (p_1, p_2, \ldots, p_n)$ et : $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$

Ce principe de consistance affirme qu'un évènement probable est possible et les degrés de possibilité ne peuvent pas être inférieurs à des degrés de probabilité (Dubois et al., 2004). Nous citons dans ce qui suit, les règles de transformation, proposées dans la littérature.

3.4.1 Transformation de Klir (TK)

La transformation de (Klir et Geer, 1993) est basée sur trois hypothèses, qui sont :

1. *Une hypothèse d'échelle* qui force chaque valeur π_i à être une fonction de p_i/p_1 ($\forall i = 1, \ldots, n\, p_i > 0, p_i \geq p_{i+1}$). Cette fonction peut être une échelle de ratio, une échelle d'intervalle, des transformations d'échelle logarithmique, etc.

2. *Une hypothèse d'invariance d'incertitude,* selon laquelle la mesure de l'information de p doit être numériquement égale à la mesure de l'information de π contenue dans π de p, qui peut être exprimée par l'égalité de leurs incertitudes.
3. Les transformations doivent satisfaire à *la condition de consistance* $\pi(u) \geq p(u)$, $\forall u$, indiquant ce qui est probable doit être possible.

Supposons que les éléments de Ω sont ordonnés de telle sorte que : $\forall i = \{1, ..n\}$, $p_i > 0$ et $p_i \geq p_{i+1}$, $\pi_i > 0$ et $\pi_i \geq \pi_{i+1}$, avec $p_{n+1} = 0$ et $\pi_{n+1} = 0$.

Klir a examiné le principe de la préservation de l'incertitude sous deux échelles (Ben Slimen et al., 2013 ; Klir et Geer, 1993) :

– *Échelle de ratio* : les transformations $p \to \pi$ et $\pi \to p$, sont appelées les transformations normalisées et elles sont définies par :

$$\pi_i = \frac{p_i}{p_1},\ p_i = \frac{\pi_i}{\sum_{i=1}^{n} \pi_i} \tag{3.28}$$

– *Échelle logarithmique* : les transformations $p \to \pi$ et $\pi \to p$ sont définies par :

$$\pi_i = \left(\frac{p_i}{p_1}\right)^{\alpha},\ p_i = \frac{\pi_i^{1/\alpha}}{\sum_{i=1}^{n} \pi_i^{1/\alpha}} \tag{3.29}$$

α est un paramètre appartenant à l'intervalle ouvert $]0, 1[$.

3.4.2 Transformation optimale (TO)

(Dubois et al., 2004) ont proposé une transformation asymétrique, dite optimale, parce qu'elle donne la distribution des possibilités la plus spécifique, c'est-à-dire qu'elle perd moins d'informations et elle calcule les limites supérieures des évènements probabilistes.

Les auteurs ont affirmé que la transformation probabilité-possibilité doivent satisfaire les trois principes suivants :

– *Le principe de consistance* : $\Pi(A) \geq P(A)\ \forall A \subseteq \Omega$, avec P et Π sont respectivement une distribution des probabilités et une distribution des possibilités dans $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, dans ce cas on dit Π domine P.
– *La préservation de la préférence* : $\forall(\omega_1, \omega_2) \in \Omega^2$; si $p(\omega_1) > p(\omega_2)$ équivalent à $\pi(\omega_1) > \pi(\omega_2)$
– *Le principe du maximum de spécificité* : il faut chercher la distribution la plus spécifique qui satisfait les deux principes précédents.

Supposons le cas que les éléments de Ω sont ordonnés de sorte que $p(\omega_1) \geq p(\omega_2) \geq \ldots \geq p(\omega_n)$. Dans ce cas, ω_i peut avoir un degré de nécessité $p(\omega_i) - p(\omega_{i+1})$ supplémentaire par rapport à ω_{i+1}. Ainsi, le degré de nécessité $N(A_i)$ de l'évènement $A_i = \{1, 2, \ldots, n\} \subseteq \Omega$ se définie par la somme de $p(\omega_j) - p(\omega_{i+1})$; où ω_{i+1} a la plus grande probabilité parmi les éléments dans $\Omega - A_i$, çà veut dire :

$$N(A_i) = \sum_{j-1}^{i} \left(p(\omega_j) - p(\omega_{i+1})\right), \; o\grave{u} \; p(\omega_{n+1}) = 0 \tag{3.30}$$

Dubois et Prade ont prouvé que $N(A_i)$, définit dans (3.23), est une mesure de nécessité dans le sens mathématique. Ensuite, ils ont dérivé l'équation suivante pour la transformation de la probabilité vers la possibilité (Dubois et Prade, 1985) :

$$\pi(\omega_i) = \Pi\left(\{\omega_i\}\right) = 1 - \left(\Omega - \{\omega_i\}\right) \tag{3.31}$$

$$\pi(\omega_i) = i * p(\omega_i) + \sum_{j=i+1}^{n} p(\omega_j), \forall i = 1, \ldots, n \tag{3.32}$$

Où : $\sum_{j=n+1}^{n} p(\omega_j) = 0$; $p(\omega_{n+1}) = 0$ par convention.

(Dubois et al., 1993) ont suggéré une transformation appelée symétrique qui nécessite moins de calcul. La formule de transformation est applicable dans les deux sens $p \rightarrow \pi$et $\pi \rightarrow p$ et elle est définie comme suivant :

$$\pi_i = \sum_{j=i}^{n} \min(p_i, p_j) \tag{3.33}$$

La transformation «*inverse*» est appliquée à la distribution des possibilités et elle permet de retrouver la distribution des probabilités initiales :

$$p(\omega_i) = \sum_{j=1}^{n} \frac{\pi(\omega_j) - \pi(\omega_{j+1})}{j} \tag{3.34}$$

Où : $\pi(\omega_{n+1}) = 0$

En 1993 l'équipe de Dubois (Dubois et al., 1993) ont suggéré une idée de transformation symétrique (TS) de $p \rightarrow \pi$. Cette Transformation nécessite moins de calcul, mais elle est loin d'être optimale, car elle ne donne pas la meilleure spécificité (Ben Slimen et al., 2013). La TS est définie par :

$$\pi_i = \sum_{j=i}^{n} min(p_i, p_j) \tag{3.35}$$

3.4.3 Transformation variable (TV)

(Mouchaweh et al., 2006) ont proposé une transformation paramétrique $p \rightarrow \pi$, nommée la transformation variable (TV), et définie comme suit :

$$\pi_i = \left(\frac{p_i}{p_1}\right)^{k.(1-p_i)} \tag{3.36}$$

k est une constante qui doit garantir la condition de consistance suivante :

$$\forall \omega \in \Omega,\ \pi(\omega) \geq p(\omega) \tag{3.37}$$

Cette condition, qui est un cas particulier du principe de consistance de Dubois et Prade défini par $(\Pi(A) \geq P(A), \forall A \subseteq \Omega)$, implique une distribution des possibilités intéressantes pour le diagnostic de reconnaissance des formes. En effet (3.30) est un cas discret, c'est-à-dire la distribution contenant un ensemble de singletons bien fini de la condition de consistance de Dubois et Prade.

Pour garantir le principe de consistance défini par (3.30), la valeur de k doit appartenir à l'intervalle suivant :

$$0 \leq K \leq \frac{\log(p_n)}{(1-p_n) * \log\left(\frac{p_n}{p_1}\right)} \tag{3.38}$$

Quand on a la valeur de k est égale à sa valeur maximale : $k_{max} = \frac{\log(p_n)}{(1-p_n)*\log\left(\frac{p_n}{p_1}\right)}$, la possibilité π_n calculée selon la TV est égale à p_n. Lorsque la valeur de k est supérieure à k_{max} alors la valeur de π_n devient inférieure à celle de p_n, ce qui signifie que la TV ne satisfera plus la condition de consistance de Dubois et Prade (Dubois et al., 2004) $\forall i = 1, \ldots, n$. En effet, la TV transforme une distribution des probabilités de manière non-linéaire ; cela signifie que la TV ajoute plus d'informations aux probabilités élevées qu'aux plus petites. Ceci est dû au fait que les probabilités élevées ont une puissance $k.(1-p_i)$, plus petite que celle des petites probabilités.

La différence entre la TK et la TV est que la transformation de Klir a une puissance constante α qui appartient à l'intervalle ouvert $]0, 1[$ pour préserver l'incertitude, alors que la puissance $k.(1-p_i)$ dans la TV est variable pour la rendre plus spécifique.

3.4.4 Transformation de Masson et Denoeux (MD)

La transformation proposée par (Masson et Denoeux, 2006) permet de retrouver une distribution des possibilités dominant toutes les mesures de probabilité définies par des intervalles de probabilité. Lorsqu'on travaille avec un ensemble d'intervalles de probabilité $L = [l_i, u_i], \forall i = 1, \ldots, n$, l'ordre induit sur les masses de probabilité n'est plus complet et l'ordre partiel est réduit à :

$$p(\omega_i) \leq \pi(\omega_j) \leftrightarrow u_i \leq l_j \tag{3.39}$$

Les deux probabilités $p(\omega_i)$, $p(\omega_j)$ sont incomparables si les intervalles $[l_i, u_i]$ et $[l_j, u_j]$ se croisent d'une certaine manière. Notons, M est un ordre partiel sur Ω, et C est l'ensemble de ses extensions linéaires (une extension linéaire (ordre total) $C_l \in C$ est compatible avec l'ordre partiel M). La procédure de transformation de *MD*, se déroule comme suit : pour chaque élément $\omega_i \in \Omega$, on cherche la distribution des probabilités compatible qui donne le degré de possibilité le plus élevé, quand la transformation TO est utilisée pour obtenir la distribution des possibilités (Benferhat et al., 2017) :

$$\pi^{C_1}(\omega_i) = max_{p_1,\ldots,p_n}(\sum_{p_j \leq p_i} p_i) \tag{3.40}$$

L'avantage de la transformation de *MD* est non-paramétrique, elle ne nécessite pas la connaissance des probabilités (Moussa, 2014). Par contre, elle a deux inconvénients principaux à savoir : (i) son coût de calcul (il considère dans le pire des cas $n!$ extensions linéaires ; où n est la taille des distributions à transformer), et (ii) cette transformation ne garantit pas une distribution optimale en terme de spécificité (Haddad et al., 2016).

3.4.5 Synthèse et domaines d'application

Le tableau 3.3 (Ben Slimen et al., 2013) résume les caractéristiques des transformations étudiées : *TK, TO, TS, TV* et *MD*. Pour chaque transformation, il est mentionné s'il s'agit d'un cas discret (D) et/ou continu (C) et si elle satisfait le principe de consistance (PC), la préservation de préférences (PP) et la spécificité maximale (SM).

Transformations	$p \to \pi$	$\pi \to p$	**Propriétés**				
			D	C	PC	PP	SM
TK	✓	✓	✓	✓		✓	
TO	✓	✓	✓	✓	✓	✓	✓
TS	✓	✓	✓	✓	✓	✓	
TV	✓		✓			✓	
MD	✓		✓				

Table 3.3 – Résumé des transformations (Ben Slimen et al., 2013)

Étant une brève synthèse, la théorie des probabilités est le modèle le plus classique pour la modélisation des informations imparfaites. Cette théorie, ne permet pas la modélisation du raisonnement humain, où les informations et les connaissances fournies sont de nature ambiguë (Alsahwa, 2014 ; Mouchaweh, 2002). Par contre, la théorie des possibilités offre un outil flexible pour représenter une information imparfaite, incertaine et imprécise, exprimée par les humains. Cette dernière est destinée à la gestion de l'incertitude épistémique ; c'est-à-dire l'incertitude dans un contexte où les connaissances sont exprimées d'une façon ambiguë (Alsahwa, 2014).

Les chercheurs ont exploité la théorie des possibilités dans plusieurs axes du domaine de RI, tels que : la construction des systèmes multi-agent pour la recherche intelligente des documents Web (Elayeb, 2009), l'expansion sémantique des requêtes (Ben Khiroun et al., 2011), la désambiguïsation morphologique des textes Arabes (Bounhas et al., 2015), la proposition d'un modèle de recherche d'information basé sur les réseaux possibilistes (Garrouch et Omri, 2015), l'indexation des documents biomédicaux à partir d'un réseau possibiliste (Chebil et al., 2016), etc.

Ainsi, les transformations probabilité-possibilité ont été exploitées dans plusieurs domaines d'application, autres que la RI, tels que : (Mouchaweh et al., 2006) ont travaillé sur le diagnostic de reconnaissance des formes à l'aide d'une méthode de classification FPM (Fuzzy Pattern Matching) pour des données provenant des secteurs industriels et médicaux. (Moussa, 2014) ont adopté la *TO* dans le domaine de la Finance quantitative. (Benferhat et al., 2017) ont utilisé la transformation probabilité-possibilité pour donner une approximation d'inférence MAP (Maximum A Posteriori) dans les réseaux crédaux [3], afin de transformer un réseau crédal en un réseau possibiliste, etc.

Nous portons une attention particulière à la théorie des possibilités qui représente un cadre prometteur pour traiter le contexte d'ambiguïté des traductions dans le domaine de RIT. Nous nous focalisons aussi sur la règle de transformation probabilité-possibilité de Dubois et Prade, parce que d'une part, elle satisfait les principes

3. Les réseaux crédaux sont des modèles graphiques probabilistes, comme les réseaux Bayésiens. Les réseaux crédaux sont basés sur la théorie des probabilités imprécises (Benferhat et al., 2017).

de consistance et de préservation de préférence, d'autre part, elle est la transformation la plus utilisée et elle a donné des bons résultats dans les différents domaines.

Conclusion

Nous avons présenté dans ce chapitre un survol sur les différentes notions de base des théories des probabilités, des possibilités et de transformation d'un cadre probabiliste vers un cadre possibiliste. Nous avons présenté aussi une étude comparative entre ces théories afin de profiter de leurs avantages dans nos contributions.

La seconde partie de ce manuscrit aura pour finalité de présenter des nouvelles approches de désambiguïsation de TR en RIT. Ces approches sont basées sur les théories des possibilités et de transformation probabilité-possibilité afin de traiter l'imprécision et l'incertitude de manière innovante. Cette seconde partie comprend les trois chapitres suivants : dans le chapitre quatre, nous présentons trois approches possibilistes pour la désambiguïsation des traductions de requêtes, nous exposons les modèles conceptuels de ces approches, ainsi qu'un exemple illustratif pour chacune. Le chapitre cinq présente la validation et les évaluations des performances des approches proposées. Dans le chapitre six, nous présentons et validons une nouvelle approche à base de proximité bilingue.

Deuxième partie

CONTRIBUTIONS

Chapitre

4 Approches possibilistes proposées pour la désambiguïsation des traductions de requêtes

Sommaire

Intoduction

Lors du développement d'un SRIT, il est crucial de garantir une bonne traduction de la requête source, afin de retrouver des documents pertinents qui répondent aux besoins des utilisateurs. Donc, il est nécessaire de traiter l'ambiguïté des traductions qui est considérée comme un cas d'imprécision. En effet, la théorie des possibilités

est prête naturellement à ce genre d'application (Dubois et Prade, 2015). Cette théorie, présentée dans le chapitre 3, est considérée à la base de notre travail pour différentes raisons : elle a une capacité de réaliser la représentation des connaissances imparfaites et de proposer des outils de raisonnement qui permettent de traiter l'incertitude et l'imprécision dans le cas où les connaissances sont de nature ambiguë (Alsahwa, 2014).

Par ailleurs, les dictionnaires bilingues sont de plus en plus disponibles pour de nombreuses langues, ce qui a amélioré davantage l'utilité des approches de TR à base des dictionnaires. Malgré cela, ces derniers souffrent principalement de problème de couverture et d'ambiguïté. Plusieurs approches (Kwok, 2000 ; Nie, 1999 ; Xu et al., 2001) ont proposé de collecter des grandes ressources lexicales pour augmenter la couverture des dictionnaires, que ce soit manuellement ou automatiquement. D'autres approches (Davis et Ogden, 1997 ; MarkW. Davia, 1997) ont utilisé un dictionnaire des phrases pour traduire les requêtes sources en unités. En outre, l'efficacité acquise, en utilisant une traduction manuelle des phrases, est largement meilleure qu'une traduction mot-à-mot à base de dictionnaire (Hull et Grefenstette, 1996). De plus, la traduction des concepts multi-termes comme des phrases permet d'injecter des termes supplémentaires ayant des traductions non-ambiguës, en fournissant le contexte correct pour la désambiguïsation des traductions (Ballesteros et Croft, 1998). Cette méthode peut supporter l'efficacité de la RIT, en résolvant le problème d'ambiguïté de traduction mieux qu'une traduction mot-à-mot.

Jusqu'à présent, il est difficile d'obtenir un dictionnaire exhaustif des phrases. En réalité, plusieurs nouvelles phrases se forment chaque jour, donc il est possible d'avoir de nombreuses phrases non-existantes dans le dictionnaire utilisé. Par conséquent, le problème de couverture du dictionnaire augmente ce qui posera de nouveau le problème d'identification et de traduction des phrases inconnues.

Dans ce chapitre, nous proposons d'exploiter la théorie des possibilités pour traiter l'ambiguïté de TR afin de franchir les difficultés retrouvées dans les approches à base de dictionnaire. Notre contribution inclut trois parties : (i) une approche possibiliste à base d'une transformation probabilité-possibilité (section 4.1), qui se focalise sur la traduction des syntagmes nominaux ; (ii) une approche possibiliste discriminative basée sur un réseau possibiliste pour améliorer la qualité de traduction des mots restants (section 4.2) ; et (iii) une approche possibiliste hybride qui combine les deux premières approches en profitant de ses avantages (section 4.3).

4.1 Proposition d'une approche de TR à base de transformation probabilité-possibilité

Nous proposons dans cette section notre approche possibiliste à base d'une transformation probabilité-possibilité. Cette approche est un moyen pour introduire plus de tolérance dans le processus de TR. L'idée principale est d'identifier, en premier lieu, les syntagmes nominaux (Noun Phrases : NPs) existants dans la requête source et les traduire en unités. En traduisant un terme, les autres termes (ou leurs traductions) représentent un «*contexte*» qui aide à déterminer la traduction correcte du terme donné (Gao et al., 2001). Dans notre cas, nous supposons que la traduction d'un terme dépend uniquement des autres termes du même syntagme nominal. Ainsi, nous avons intégré un modèle de langue avec les patrons de traduction. En deuxième lieu, les termes de la requête source, qui ne font pas partie des syntagmes nominaux sélectionnés, seront traduits mot-à-mot, en utilisant une approche possibiliste de traduction des mots simples (Ben Romdhane et al., 2013). Dans cette étape, nous partons de l'idée que les traductions appropriées des termes de la requête source ont tendance à coexister dans les documents de la langue cible, contrairement à celles qui ne conviennent pas.

D'abord, nous présentons dans la figure 4.1, une vue d'ensemble de notre approche. Puis, nous détaillons les scénarios effectués. Dans la section 4.1.1, nous présentons le scénario de traduction des syntagmes nominaux. Ensuite, nous détaillons le modèle de notre approche possibiliste *naïve bayésienne* (cf. 4.1.2) et les étapes de traduction des mots simples (cf. 4.1.3). Enfin, nous présentons un exemple illustratif (cf. 4.1.4) récapitulant les étapes de traduction, à l'anglais, d'une requête source française.

4.1.1 Scénario de traduction des syntagmes nominaux (Noun Phrases : NPs)

Nous avons généré les NPs de base et complexes, en utilisant l'analyseur du langage naturel *Stanford Parser*[1]pour diminuer le calcul et la complexité du processus de TR. Contrairement au modèle probabiliste, proposé par (Gao et al., 2006), qui a appliqué l'algorithme de *Viterbi*. Ce dernier est basé sur un long calcul de probabilité conditionnelle et d'estimation, ce qui a augmenté sa complexité.

Après avoir identifié les NPs de la requête source, la technique de traduction mot-à-mot n'est plus adaptable pour les traduire. Nous avons profité d'un dictionnaire

1. http://nlp.stanford.edu/software/lex-parser.shtml, c'est un programme qui élabore la structure grammaticale des phrases, par exemple, quels groupes de mots vont ensemble (en tant que «phrases») et quels mots sont le sujet ou l'objet d'un verbe.

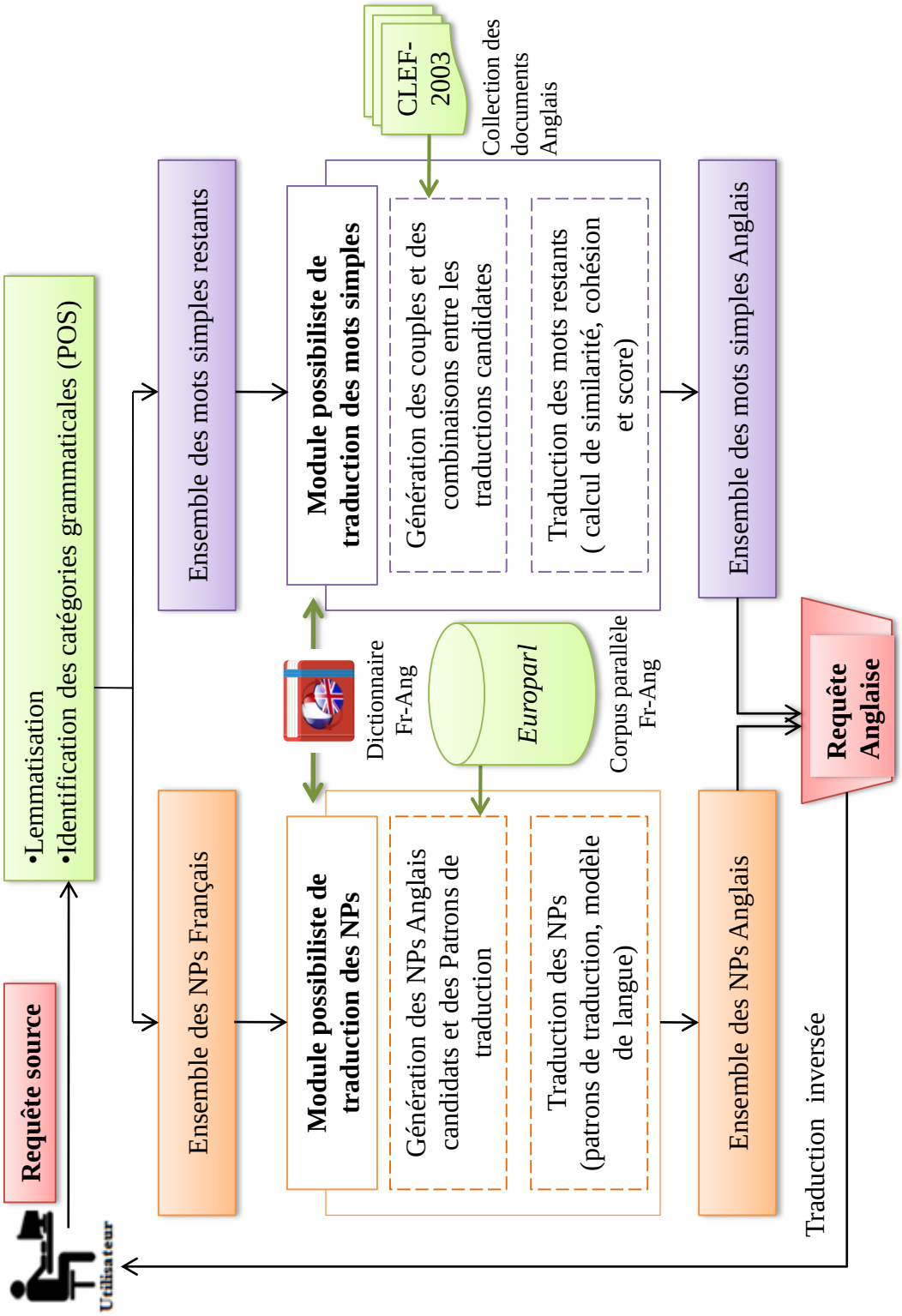

Figure 4.1 – Modèle conceptuel de l'approche possibiliste à base de transformation probabilité-possibilité

bilingue Fr-Ang qui comprend certaines NPs et leurs traductions. Cependant, ce dictionnaire souffre d'une couverture limitée, car plusieurs NPs identifiés n'y figurent pas. Ainsi, nous avons profité des patrons de traduction des NPs français et anglais pour accomplir la tâche. Par exemple, un NP français [NN-1, NN-2] est généralement traduit en anglais comme suit : [NN-2, NN-1] (par exemple, ≪Voiture de Tom≫ peut être traduite par ≪Tom car≫). Par conséquent, pour chaque phrase française correspondante à un tel patron de traduction, nous identifions sa traduction possible.

En plus, nous avons utilisé *Europarl*, un corpus parallèle bilingue aligné par mots, afin d'extraire tous les patrons de traduction possibles. Premièrement, nous avons commencé par identifier les NPs des phrases françaises et étiqueter leurs catégories grammaticales (POS : Part-of-Speech). Deuxièmement, nous avons généré l'ensemble des patrons de traduction anglais (EPT : English Translation Pattern) correspondants à chaque patron des NPs français (FPT : French NP Pattern). Troisièmement, nous avons fourni la catégorie grammaticale et la position de chaque mot dans une phrase française donnée. Quatrièmement, nous avons segmenté la phrase anglaise alignée en une séquence des mots liés à leurs positions. En effet, un alignement de mots (a, b) désigne qu'un mot français en position a est associé à un mot anglais en position b. Cependant, nous désignons par l'alignement de mot (a, f) s'il n'y a pas de mots en anglais dans lesquels est lié un mot français donné en position a.

Le tableau 4.1 donne un exemple (Elayeb et al., 2018) des NPs français ainsi que leurs éventuels patrons de traduction anglais. Malheureusement, quelques NPs traduits, particulièrement les NPs complexes, ne correspondent à aucun index des documents anglais. Dans ce cas, nous les re-segmentons en des NPs plus basiques afin de résoudre partiellement le problème d'ambiguïté. Nous détaillons dans la section suivante l'estimation des possibilités des patrons de traduction.

Phrase française	L'/DT/1 aéroport/NN/2 a/VV/3 changé/VV/4 ses/PRO/5 circuits/NN/6 de/PRE/7 vol/NN/8 afin/ADV/9 de/PRE/10 réduire/VV/11 le/ART/12 bruit/NN/13
Phrase anglaise	The/DT/1 airport/NN/2 changed/VV/3 its/PRO/4 flight/NN/5 patterns/NN/6 to/PRE/7 decrease/VV/8 the/DT/10 noise/NN/11
Paire de mots alignés	(1, 1) (2, 2) (3, f) (4, 3) (5, 4) (6, 6) (7, f) (8, 5) (9, 7) (10, f) (11, 8) (12, 10) (13, 11)
Patrons de traductions	[DT NN-1 VV-1 VV-2 PRO NN-2 PRE NN-3] [DT NN-1 VV PRO NN-3 NN-2] [ADV PRE VV ART NN] [PRE VV DT NN]

Table 4.1 – Exemple des patrons des NPs français et ses patrons de traduction anglais

Les NPs anglais et des patrons de traduction sont générés grâce à un processus de traitement de données en utilisant un ensemble des phrases alignées Français-

Anglais extraites du corpus *Europarl*. Ces étapes sont illustrées par la figure 4.2 (Elayeb et al., 2018). D'abord, nous avons exploité le système statistique de traduction automatique *Moses*[2] pour nettoyer le corpus. Puis, nous avons aligné le corpus au niveau des mots en utilisant *GIZA ++*[3]. L'extraction des NPs français a été faite via l'outil *Stanford Parser*, afin d'analyser 50 556 phrases françaises. Ensuite, nous avons extrait les couples de mots Français-Anglais à partir du fichier de sortie de *GIZA ++* en utilisant *Berkeley Aligner*[4] avec quelques modifications nécessaires. Enfin, l'outil *TreeTagger*[5] est utilisé pour l'étape d'étiquetage des catégories grammaticales (POS). Cet outil a généré 285111 patrons de traduction.

Il est difficile de capter les dépendances «*non locales*» de la langue avec les modèles « *locaux*» tels que les modèles *n-gramme*. Même si les patrons de traduction génèrent l'ensemble correct des mots, le modèle de langue ne peut pas collecter correctement la traduction. Dans notre approche, nous avons intégré le modèle de langue avec les patrons de traduction. En effet, le modèle de langue capte la dépendance locale, par contre les patrons de traduction fournissent des informations sur la dépendance globale d'une phrase. Bien que cette méthode ne soit pas assez puissante pour la traduction au niveau des phrases, elle fonctionne bien pour la traduction des syntagmes nominaux (Gao et al., 2006).

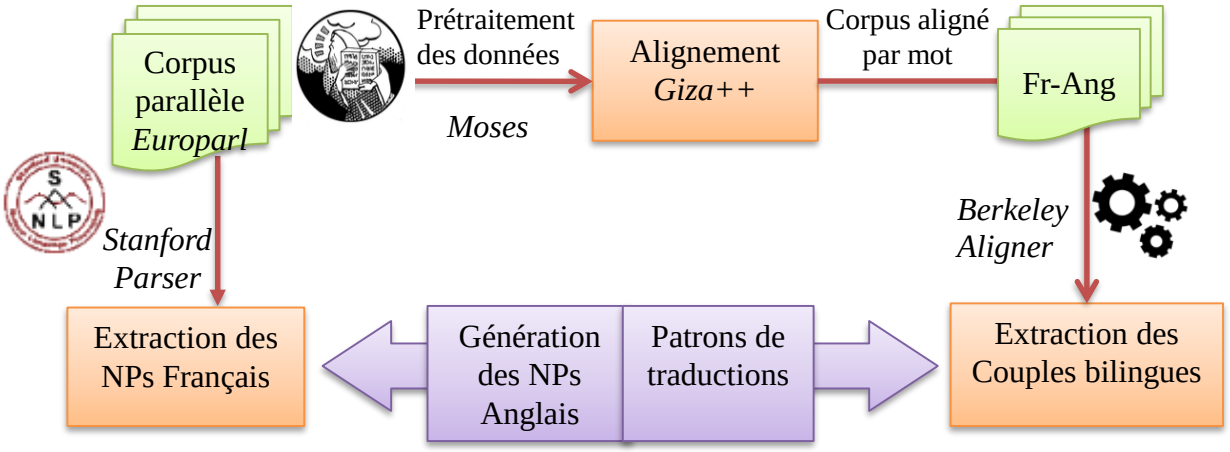

Figure 4.2 – Principales étapes pour la génération des NPs Anglais et des patrons de traduction

2. *http://www.statmt.org/moses/*
3. http://www.statmt.org/moses/giza/GIZA++.html
4. http://berkeleyaligner.sourceforge.net/, Berkeley Aligner est un logiciel d'alignement de mots qui met en œuvre des innovations récentes dans l'alignement non supervisé des mots.
5. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

4.1.2 Le modèle possibiliste

Étant donné un vecteur de NP français, $FNP = \{f_1, \ldots, f_n\}$ de n variables observées $F_1, \ldots, F_n$, et FPT est le patron de NP. Pour chaque terme français f_i dans FNP, nous récupérons toutes les traductions anglaises e_j disponibles à partir d'un dictionnaire bilingue Fr-Ang. Nous obtenons également tous les patrons de traductions disponibles EPT pour FPT.

Le problème de TR consiste à estimer une distribution des possibilités pour ENP (NP anglais). Nous appliquons l'équation (4.1) pour choisir le NP anglais qui a la possibilité la plus élevée pour FNP :

$$\pi(e_j|f_1, f_2, ..., f_n) = \frac{\pi(e_j) * \pi(f_1, f_2, ..., f_n|e_j)}{\pi(f_1, f_2, ..., f_n)} \tag{4.1}$$

Selon l'équation (4.1), la composante quantitative possibiliste de TR comprend une distribution des possibilités associée aux variables d'entrée et une distribution des possibilités associée aux traductions. Le terme $\pi(f_1, f_2, \ldots, f_n)$ est un facteur de normalisation et c'est le même pour toutes les traductions. Nous avons aussi $\pi(e_j) = 1$ et $\pi(f_1, f_2, \ldots, f_n) = 1$, au cas où nous supposons qu'il n'y ait aucune connaissance à priori sur le vecteur d'entrée à traduire ainsi que ses traductions correspondantes.

Par une analogie à la TR *naïve bayésienne* ((revenir en détails, dans le chapitre 3, vers la section 3.1.4 page 60), la TR possibiliste *naïve bayésienne* se base sur une hypothèse d'indépendance sur les variables f_i dans le contexte de ses traductions (Ben Amor et al., 2002). La plausibilité, de chaque traduction e_j pour un terme donné d'une requête source $(f_1, f_2, \ldots, f_n)$ est calculée en appliquant l'équation (4.2) :

$$\pi(e_j|f_1, f_2, ..., f_n) = \frac{\pi(e_j) * \prod_{i=1}^{n} \pi(f_i|e_j)}{\pi(f_1, f_2, ..., f_n)} \tag{4.2}$$

Les possibilités conditionnelles $\pi(f_i|e_j)$, dans l'équation (4.2), indiquent dans quelle mesure f_i est une valeur possible pour la variable F_i avec l'existence de la traduction e_j. Puisque la théorie des possibilités a deux types de conditionnement, cela conduit à deux définitions d'indépendance possibilistes causales : l'opérateur $*$ (ou son extension Π) peut être utilisé comme un opérateur *produit* ou *min*. Les valeurs partiellement possibles sont rendues conjointement moins possibles grâce à l'utilisation de l'opérateur *produit*, tandis que l'opérateur *min* correspond à l'indépendance logique complète. En outre, le facteur $\pi(e_j)$ peut être ignoré si on suppose qu'il n'y a pas une connaissance préalable sur les traductions.

En utilisant un contexte à base du produit, nous attribuons à un terme donné d'une requête source française, la phrase anglaise traduite la plus plausible, ENP^*. Le meilleur syntagme nominal anglais traduit, $ENP^* = \{e_1, ..., e_m\}$, est celui qui maximise la formule (4.3) :

$$ENP^* = \underset{ENP}{argmax}\,(\pi(ENP|FNP)) = \underset{ENP}{argmax}\,(\pi(FNP|ENP) * \pi(ENP)) \\ = \underset{e_j}{argmax}\left(\pi(e_j) * \prod_{i=1}^{n} \pi(f_i|e_j)\right) \quad (4.3)$$

Où : $\pi(FNP|ENP)$ est la possibilité de traduction et $\pi(ENP)$ est une possibilité à priori des mots du NP anglais traduit.

Notons que l'équation (4.2) a une signification théorique. Dans le cas où les distributions des possibilités n'ont que les valeurs 1 et 0, cette équation signifie qu'un terme de la requête source peut avoir une traduction dans e_j tant que les termes restants de la requête sont compatibles avec cette traduction. Par conséquent, la TR possibiliste peut être considérée comme un intermédiaire entre une TR purement paramétrée et une TR *naïve bayésienne*. En effet, pour chaque terme de la requête source, la TR possibiliste utilise les valeurs des données comme une distribution des possibilités afin de sélectionner les traductions appropriées, ce qui conduit principalement aux différentes traductions.

Nous considérons un NP (FNP ou ENP) comme étant un ensemble de mots (F ou E) assemblés par des patrons de NP (FPT ou EPT). Supposons que les traductions des termes et les patrons de NP sont indépendants, nous avons :

$$\pi(FNP|ENP) = \pi(F, FPT|E, EPT) \\ = \pi(F|E, EPT) * \pi(FPT|E, EPT) = \pi(F|E) * \pi(FPT|EPT) \quad (4.4)$$

En remplaçant l'équation (4.4) dans l'équation (4.3), nous avons :

$$ENP^* = \underset{ENP}{argmax}\,(\pi(F|E) * \pi(FPT|EPT) * \pi(ENP)) \quad (4.5)$$

Où : $\pi(F|E)$ est la possibilité de traduction des mots anglais E dans ENP aux mots français F dans FNP ; et $\pi(FPT|EPT)$ est la possibilité des patrons de traduction FPT (c'est-à-dire l'ordre des mots de traduction), étant donné le patron anglais EPT.

$\pi(F|E)$ et $\pi(FPT|EPT)$ sont calculées en appliquant la transformation probabilité-possibilité des probabilités $P(F|E)$ et $P(FPT|EPT)$ qui sont détaillées la section 3.1.4 du chapitre 3. Alors que, $\pi(ENP)$ est déterminée par le modèle de langue Trigramme anglais comme suit :

$$\pi(ENP) = \pi(e_1, \ldots, e_n) = \prod_{i=1}^{n} \pi(e_i|e_{i-2}, e_{i-1}) \tag{4.6}$$

Pour la traduction des NPs, il existe trois possibilités à estimer : $\pi(FPT|EPT)$, $\pi(e_i|e_{i-2}, e_{i-1})$ et $\pi(F|E)$. Pour l'estimation de $\pi(FPT|EPT)$, nous avons utilisé le corpus de texte parallèle *Europarl*. Les patrons de traduction sont d'abord générés automatiquement à partir de ce corpus bilingue, puis filtrés. Ensuite, nous avons estimé $\pi(e_i|e_{i-2}, e_{i-1})$ en utilisant la transformation probabilité-possibilité sur $P(e_i|e_{i-2}, e_{i-1})$. En effet, le calcul de ces probabilités conditionnelles est détaillé dans la section 3.1.4 du chapitre 3. Enfin, lorsque nous estimons $\pi(F|E)$, nous supposons dans nos tests des distributions des possibilités ayant des valeurs égales sur la traduction d'un mot. Autrement dit, si un mot anglais e a n traductions dans notre dictionnaire bilingue, chacun d'entre eux aura une distribution de possibilité égale $\pi(f|e) = 1/n$.

4.1.3 La traduction possibiliste des mots simples

Suite à la traduction des syntagmes nominaux, il reste des termes dans la requête source non inclus dans aucun syntagme, donc nous devons les traduire mot-par-mot. Nous avons proposé dans (Ben Romdhane et al., 2013) une approche possibiliste pour la traduction des mots simples. Cette approche a comme principe que les mots simples ne peuvent pas être traduits les uns indépendamment des autres. Ainsi, la meilleure traduction est celle qui coexiste fréquemment dans les documents de la langue cible, avec les traductions des autres mots de la requête source.

En effet, nous avons exploité la transformation probabilité-possibilité proposée par (Dubois et Prade, 1985) pour l'appliquer à notre approche possibiliste de traduction des mots simples. D'une part, le calcul du score de similarité entre deux termes et la cohésion d'un terme r avec un ensemble R des autres termes sont deux étapes essentielles avant de sélectionner la meilleure traduction. La similarité probabiliste entre deux termes est calculée comme il est détaillé dans (Gao et al., 2001). D'autre part, nous avons étendu ce calcul vers un cadre possibiliste, en utilisant la transformation probabilité-possibilité pour calculer une similarité possibiliste notée $SIMPOSS$. Formellement, la similarité possibiliste entre deux termes r et s est donnée par l'équation (4.7). Alors que, la cohésion d'un terme r avec un ensemble R des autres

termes représente la similarité possibiliste maximale ($\max_{s \in R}(SIMPOSS(r,s))$) de ce terme avec chaque terme de l'ensemble R, tel qu'il est donné par l'équation (4.10).

$$SIMPOSS(r,s) = \pi(r,s) * log_2(\frac{\pi(r,s)}{\pi(r) * \pi(s)}) - k * log_2(Distance(r,s)) \quad (4.7)$$

$$p(r,s) = \frac{c(r,s)}{c(s)} + \frac{c(r,s)}{c(r)} \quad (4.8)$$

$$p(r) = \frac{c(r)}{\sum\limits_{r} c(r)}; p(s) = \frac{c(s)}{\sum\limits_{s} c(s)} \quad (4.9)$$

$$Cohesion(r,R) = \underset{s \in R}{Max}(SIMPOSS(r,s)) \quad (4.10)$$

En utilisant la transformation probabilité-possibilité, les distributions des possibilités $\pi(r,s)$, $\pi(r)$ et $\pi(s)$ sont respectivement dérivées à partir de $p(r,s)$, $p(r)$ et $p(s)$ qui sont données par les équations (4.8) et (4.9) avec :

- $C(r,s)$ est la fréquence que le terme r et le terme s coexistent dans les mêmes phrases anglaises de la collection de test.
- $C(r)$ est le nombre d'occurrence du terme r dans la collection de test.
- $Distance(r,s)$ est la distance moyenne, en termes de nombre de mots, entre les termes r et s dans une phrase.
- k est un coefficient constant qui est choisi empiriquement ($k = 0.8$ dans nos tests).

Le processus de traduction subit une phase de prétraitement pour la requête source. Tout d'abord, nous commençons par une élimination des «*mots vides*», en utilisant une liste prédéfinie des termes considérés non significatifs pour les documents français. Cette étape réduit le bruit dans les résultats de recherche, et par conséquent, elle augmente l'efficacité de la recherche monolingue et translinguistique. Ensuite, nous effectuons une étape de lemmatisation pour trouver la forme canonique de chaque mot. Enfin, nous sélectionnons pour chaque terme de la requête source sa meilleure traduction, parmi plusieurs traductions candidates qui coexistent fréquemment avec les traductions des autres termes dans les documents anglais.

Par ailleurs, il est très coûteux d'identifier un ensemble optimal des traductions (Gao et al., 2001). Pour cette raison, nous avons profité d'un algorithme approximatif Glouton. Nous résumons brièvement le principe de cet algorithme comme suit :

Algorithme 4.1 Algorithme Glouton pour la sélection de la meilleure traduction

Entrées : requête Française $RS = \{f_1, \ldots, f_n\}$, sélectionner l'ensemble des traductions candidates, associées aux termes Français, à partir du dictionnaire bilingue Fr-Ang

Sorties : requête traduite en Anglais

1 début
2 pour chaque *ensemble de* T_i*(i=1 à n)* **faire**
3 **pour** chaque *traduction* t_{ij} *dans* T_i **faire**
4 **pour** chaque *ensemble* T_k *(k=1 à n &* $k \neq i$*)* **faire**
5 *calculer la cohésion*(t_{ij}, T_k)
6 *calculer le score de* t_{ij} *qui est la somme de cohésion*(t_{ij}, T_k)*(* $k=1$ *à* n *&* $k \neq 1$*)*
7 *sélectionner le terme* t_{ij} *qui a le score maximal*
8 *ajouter la traduction sélectionnée dans l'ensemble* $T = \{e_1, ..., e_n\}$
9 **fin**
10 **fin**
11 fin
12 fin

4.1.4 Exemple illustratif

Considérons l'exemple de la requête source française (RS) suivante (Elayeb et al., 2018) :

RS : « *Trouvez [des rapports] sur [la fragmentation]* ».

Nous avons choisi une partie de la requête (Query_ID : 176-AH) du standard CLEF-2003 dans le but de simplifier le calcul.

Nous commençons par la lemmatisation en utilisant l'outil *TreeTagger* pour fournir une nouvelle forme pour RS *: (Trouver [du rapport] sur [la fragmentation])*. Ensuite, nous identifions les NPs (entre crochets) et les mots simples restants (soulignés).

Afin de rendre le calcul dans cet exemple encore plus simple, nous fournissons les détails pour la traduction du second NP uniquement (*la fragmentation*). Tout d'abord, nous sélectionnons toutes les traductions candidates pour chaque mot dans ce NP. Ainsi, le mot «*la*» a les traductions possibles suivantes (*a, an, it, the, those, them*) et le mot «*fragmentation*» possède l'ensemble des traductions possibles suivantes (*piecemeal, fragmentation*).

En outre, *TreeTagger* est utilisé pour identifier la catégorie grammaticale (POS) de chaque terme : Les termes (*a, an, it, the, them*) ont respectivement les POS suivants (*DT, DT, PP, DT, DT, PP*) et pour l'ensemble (*piecemeal, fragmentation*), les POS correspondants sont (*RB, NN*). Le nombre total des combinaisons est la multiplication de la taille de tous les ensembles de POS des traductions ($n = 12 = 6 * 2$).

Pour le modèle probabiliste, nous calculons les probabilités suivantes :

- $P(F|E) = 1/n = 1/12 = 0.0833$. Nous avons la même distribution des possibilités $\pi(F|E) = 0.0833$.
- $P(FPT|EPT) = Occ(FPT, EPT)/Occ(EPT)$. Le FPT de « *la fragmentation* » est [DET : ART NOM]. Nous sélectionnons 10 patrons parmi 12 combinaisons possibles ayant $P(FPT|EPT)$ supérieur à zéro. Par exemple, nous avons : (FPT (DET : ART NOM) - EPT (DT RB)), (FPT (DET : ART NOM) - EPT (DT NN)) et ainsi de suite.
- $P(ENP) = P(e_1, \ldots, e_n) = \prod_{i=1}^{n} P(e_i|e_{i-2}, e_{i-1})$ est calculée en utilisant le modèle de langue *Trigramme* pour la partie anglaise (la langue cible) du corpus *Europarl*. Par conséquent, la meilleure traduction correspond au plus haut score de : $ENP^* = \underset{ENP}{argmax}\,(P(F|E) * P(FPT|EPT) * P(ENP))$

Cependant, les distributions des possibilités correspondantes sont obtenues en utilisant la formule de transformation probabilité-possibilité.

Le tableau 4.2 fournit les résultats de calcul de la traduction probabiliste et possibiliste du syntagme nominal. Nous concluons que la meilleure traduction choisit du syntagme nominal «*la fragmentation*» en utilisant le modèle probabiliste est «*a piecemeal*» correspond au score probabiliste le plus élevé ($ENP^* = 2.29E-07$), alors que sa traduction possibiliste est «*a fragmentation*» correspond au plus haut score possibiliste ($ENP^* = 2.73E-02$).

Combinaisons	***P(ENP)***	***π((ENP)***	***P(FPT|EPT)***	***π(FPT|EPT)***	***Calcul Probabiliste***	***Calcul Possibiliste***
a,piecemeal,	6.101E-05	1.095E-04	0.044	2066.937	**2.29E-07**	1.89E-02
a,fragmentation,	1.284E-05	4.163E-05	0.206	7856.303	2.21E-07	**2.73E-02**
an,piecemeal,	0	0	0	0	0	0
an,fragmentation,	0	0	0	0	0	0
it,fragmentation,	0	0	0	0	0	0
the,piecemeal,	3.102E-06	1.240E-05	0.142	5709.351	3.69E-08	5.90E-03
the,fragmentation,	3.257E-05	8.109E-05	0.044	2066.937	1.22E-07	1.40E-02
those,piecemeal,	0	0	0	0	0	0
those,fragmentation	0	0	0	0	0	0
them,fragmentation,	0	0	0	0	0	0

Table 4.2 – Résultats de calcul de la traduction probabiliste et possibiliste

Après la traduction des syntagmes nominaux, nous passons à la traduction des termes restants de la requête source. Dans notre exemple, nous avons les deux termes restants « *trouver* » et « *sur* » et leurs traductions candidates sont respectivement (*come, find, get, pinpoint, think, strike*) et (*about, along, on, upon,*

to, at, onto). A partir de l'ensemble {*trouver, sur*}, nous générons l'ensemble de combinaisons des traductions qui sont 42 combinaisons ($n = 7 * 6 = 42$).

Nous présentons dans le tableau 4.4, les résultats de calcul de la fréquence $C(x, y)$ des deux traductions candidates x et y de coexister dans les mêmes phases dans la collection, la distance moyenne entre x et y, la distribution des probabilités $P(x, y)$, la similarité probabiliste, la distribution des possibilités $\pi(x, y)$ et la similarité possibiliste. En outre, le tableau 4.3 présente les résultats de calcul de nombre d'occurrences $C(x)$ de chaque traduction candidate dans les documents anglais de la collection de test, $P(x)$, $\pi(x)$ et les scores probabilistes et possibilistes (Ben Romdhane et al., 2013). Nous concluons que la meilleure traduction sélectionnée pour le terme «*trouver*» dans le modèle probabiliste est «*get*» et pour le modèle possibiliste «*find*». Nous avons «*to*» est la meilleure traduction pour le terme «*sur*» pour les deux modèles : probabiliste et possibiliste.

Les traductions du terme x	***C(x)***	***P(x)***	***π(x)***	***Prob_Score (x)***	***Poss_Score (x)***
come	31187	0.001	0.144	0	5002.054
find	21635	0.001	0.122	0	**9539.895**
get	32749	0.001	0.148	**0.632**	6327.86
pinpoint	164	0	0.003	0	498.705
think	20196	0.001	0.118	0	2650.245
strike	4665	0	0.051	0	3342.081
about	44993	0.002	0.17	36.415	273874.718
along	5209	0	0.055	83.329	285766.024
on	191368	0.009	0.301	33.272	335981.385
upon	2344	0	0.03	82.353	283083.465
to	656959	0.03	0.446	**148.548**	**457309.126**
at	155582	0.007	0.281	21.961	299223.866
onto	259	0.002	0.004	142.283	346464.998

Table 4.3 – Résultats des calculs probabiliste et possibiliste des termes restants

Couple (x, y)	***C(x, y)***	***Distance moyenne***	***P(x,y)***	***π(x, y)***	***SIM_Prob (x, y)***	***SIM_Poss (x, y)***
come,about,	1403	4.833	0.038	127.774	-1.296	1575.685
come,along,	439	2.075	0.049	156.289	0.003	2229.018
come,on,	5322	6.507	0.099	252.773	-0.871	3159.093
come,upon,	92	4.489	0.021	79.324	-1.372	1121.575
come,to,	17631	6.876	0.296	432.306	1.566	5493.175
come,at,	4113	6.397	0.079	219.823	-1.115	2725.047
come,onto,	18	3.833	0.035	119.566	-0.814	2099.956
find,about,	1000	5.804	0.034	117.338	-1.548	1460.301
find,along,	126	6.69	0.015	59.95	-1.953	784.639
find,on,	3964	6.898	0.102	256.862	-0.845	3277.638
find,upon,	43	10.907	0.01	43.101	-2.589	579.989
find,to,	11554	7.039	0.276	423.736	1.396	5473.468
find,at,	3308	6.165	0.087	233.486	-0.911	2970.836
find,onto,	5	9.6	0.01	41.628	-2.418	675.648
get,about,	1656	4.527	0.044	142.48	-1.138	1775.175
get,along,	180	4.861	0.02	76.006	-1.508	1001.136
get,on,	5583	5.525	0.1	253.672	-0.681	3163.758
get,upon,	33	5.697	0.008	33.118	-1.891	424.19
get,to,	18323	5.829	0.294	431.407	1.702	5466.942
get,at,	3779	6.055	0.07	202.418	-1.19	2478.653
get,onto,	14	7.357	0.027	97.649	-1.742	1682.361
pinpoint,about,	2	11	0.006	27.422	-2.653	431.671
pinpoint,along,	0	0	0	0	0	0
pinpoint,on,	20	5.95	0.061	184.066	-0.845	3267.928
pinpoint,upon,	0	0	0	0	0	0
pinpoint,to,	93	6.925	0.284	427.286	3.527	7864.779
pinpoint,at,	21	7.048	0.064	190.675	-0.956	3414.193
pinpoint,onto,	0	0	0	0	0	0
think,about,	1954	4.222	0.07	202.909	-0.596	2696.873
think,along,	86	6.953	0.01	43.899	-2.076	556.237
think,on,	2439	8.261	0.067	196.213	-1.565	2435.886
think,upon,	43	8.907	0.01	43.361	-2.353	586.126
think,to,	9267	9.067	0.236	399.94	0.555	5150.932
think,at,	2041	7.351	0.057	175.186	-1.553	2163.858
think,onto,	5	7.8	0.01	41.658	-2.177	678.37
strike,about,	152	5.132	0.018	69.58	-1.611	900.957
strike,along,	19	4.474	0.004	17.906	-1.666	224.747
strike,on,	1047	5.226	0.115	275.224	-0.074	3888.554
strike,upon,	7	6.286	0.002	10.648	-2.084	133.86
strike,to,	2095	8.492	0.226	391.921	0.958	5514.576
strike,at,	773	6.003	0.085	230.547	-0.718	3221.339
strike,onto,	4	2.5	0.008	35.488	-0.881	613.97

Table 4.4 – Résultats des calculs probabilistes et possibiliste des syntagmes nominaux

4.2 Proposition d'une approche possibiliste discriminative de désambiguïsation de TR

L'approche à base de transformation probabilité-possibilité (Elayeb et al., 2018) traite le problème d'incertitude et d'imprécision des traductions candidates de chaque terme de la requête source. Cette approche n'évalue pas la discrimination des valeurs des traductions, car elle utilise une seule mesure de possibilité (formule 4.3). Cependant, la théorie des possibilités modélise l'effet discriminatif à l'aide de la mesure de nécessité. D'après l'état de l'art, et en se basant sur les techniques de traduction existantes en RIT, il n'existe pas dans la littérature des approches pour la TR qui modélise la discrimination. Ce fait nous a motivé à pousser la recherche dans cette direction en proposant une nouvelle approche possibiliste discriminative de désambiguïsation de TR. Nous optons, dans cette approche d'exploiter les deux mesures de possibilité et de nécessité afin d'étudier l'impact de ces deux dernières sur la désambiguïsation des traductions.

Nous illustrons, à travers la figure 4.3, le modèle conceptuel de notre approche possibiliste discriminative.

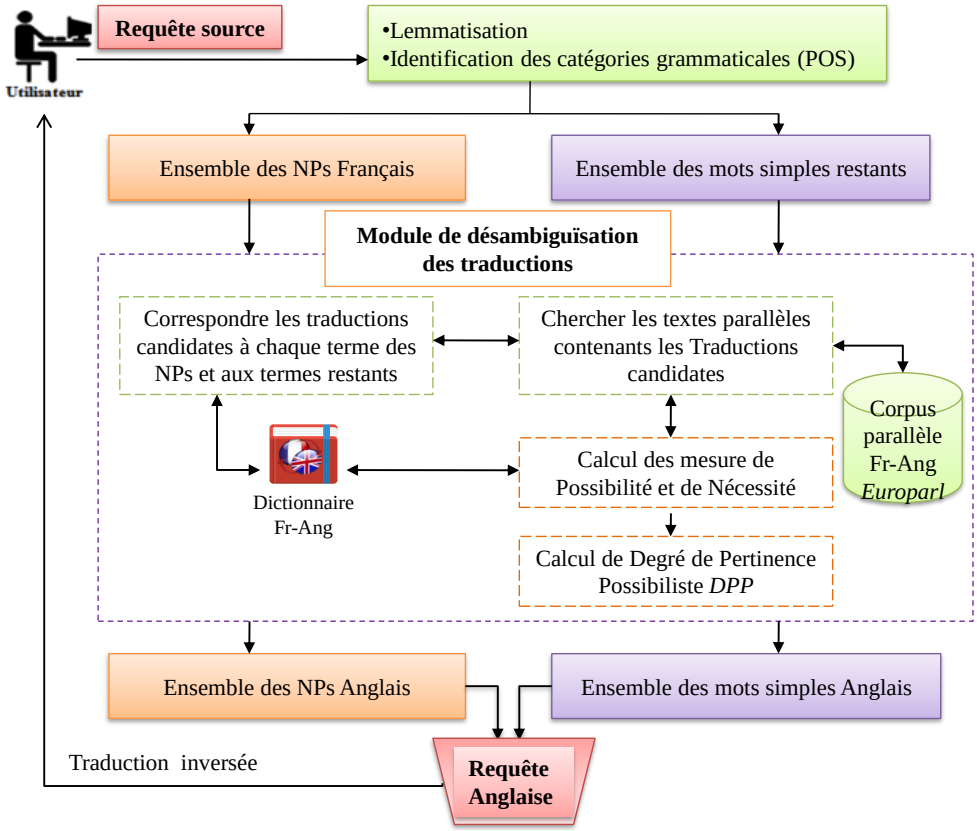

Figure 4.3 – Modèle conceptuel de l'approche possibiliste discriminative

Avant de procéder au module de désambiguïsation des traductions, la requête source

subit une étape de lemmatisation et d'identification des catégories grammaticales de chaque terme. D'autre part, nous avons gardé le même principe de traduction que l'approche à base de transformation. En effet, nous avons traduit les syntagmes nominaux de la requête en unités ainsi que les mots restants en appliquant pour les deux les mêmes formules. Nous présentons, dans la section 4.2.1, les formules de calcul de degré de pertinence possibiliste (DPP), et dans la section 4.2.2, un exemple illustratif avec un calcul détaillé.

4.2.1 Le degré de pertinence possibiliste

Considérons la requête source $RS = (t_1, t_2, \ldots, t_P)$ contenant P termes. La figure 4.4 présente un réseau possibiliste qui relie l'ensemble des traductions candidates (T_j) aux termes de RS. Le résultat du processus de désambiguïsation est une requête cible $RC = (T_1, T_2, \ldots, T_N)$. Nous évaluons la pertinence d'une traduction candidate en utilisant deux mesures de pertinence : la pertinence possible et la pertinence nécessaire. Les traductions non-pertinentes sont rejetées par la pertinence possible. Tandis que, la pertinence nécessaire renforce les traductions restantes qui n'ont pas été rejetées par la possibilité. Nous notons $DPP(T_j|RS)$ le Degré de Pertinence Possibiliste (DPP) d'une traduction T_j sachant RS.

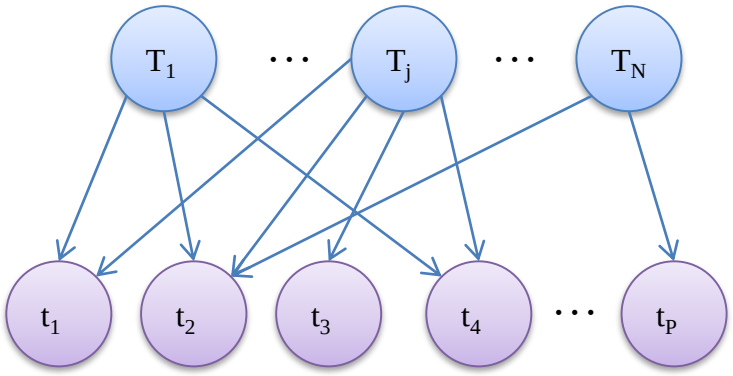

Figure 4.4 – Réseau possibiliste de l'approche discriminative de désambiguïsation

La pertinence possible de chaque traduction T_j est calculée comme suit (Ben Romdhane et al., 2017) :

La possibilité $\Pi(T_j|RS)$ est proportionnelle à :

$$\Pi'(T_j|RS) = \Pi(t_1|T_j) * \ldots * \Pi(t_p|T_j) = nft_{1j} * \ldots * nft_{Pj} \tag{4.11}$$

Où :

- $nft_{ij} = tf_{ij}/max(tf_{kj})$: est la fréquence normalisée du terme source t_i dans le texte parallèle (les phrases alignées) de la traduction candidate T_j.

– tf_{ij} : est le nombre d'occurrence du terme source t_i dans le texte parallèle de la traduction candidate T_j divisé par le nombre de termes dans le texte parallèle de la traduction candidate T_j.

La nécessité, désignée par $N(T_j|RS)$, est calculée pour restituer la traduction pertinente T_j donnée pour la requête source RS. Cette mesure est calculée comme suit :

$$N(T_j|RS) = 1 - \Pi(\neg T_j|RS) \tag{4.12}$$

Où :

$$\Pi(\neg T_j|RS) = (\Pi(RS|\neg T_j) * \Pi(\neg T_j))/\Pi(RS) \tag{4.13}$$

De la même manière $\Pi(\neg T_j|RS)$ est proportionnelle à :

$$\Pi(\neg T_j|RS) = \Pi(t_1|\neg T_j) * \ldots * \Pi(t_P|\neg T_j) \tag{4.14}$$

Ce numérateur peut être exprimé comme suit :

$$\Pi'(\neg T_j|RS) = (1 - \phi T_{1j}) * \ldots * (1 - \phi T_{Pj}) \tag{4.15}$$

Où :

$$\phi T_{ij} = Log_{10}(nCT/nT_j) * (nft_{ij}) \tag{4.16}$$

Avec :

– nCT est le nombre de traductions candidates dans le dictionnaire bilingue Fr-Ang.

– nT_j est le nombre de textes parallèles (phrases parallèles françaises) de la traduction T_j contenant le terme source t_i.

Nous calculons le degré de pertinence possibiliste (DPP) de chaque traduction T_j des termes de RS via l'équation suivante (4.17) :

$$DPP(T_j|RS) = \Pi(T_j|RS) + N(T_j|RS) \tag{4.17}$$

Les traductions appropriées sont celles qui ont le score de DPP le plus élevé.

4.2.2 Exemple illustratif

Considérons la requête source $RS = (W, t_2, t_4, t_5, t_7)$ (Ben Romdhane et al., 2017). Pour simplifier le calcul dans cet exemple, nous supposons que la requête n'a qu'un

seul terme polysémique W. Nous supposons que W a deux traductions candidates T_1 et T_2. Nous supposons également que le texte parallèle de T_1 est indexé par les termes $\{t_1, t_2, t_3, t_4\}$ et le texte parallèle de T_2 est indexé par $\{t_1, t_4, t_5, t_6, t_7\}$.

La mesure de Π est calculée comme suit :

$\Pi(T_1|RS) = nf_{(W,T_1)} * nf_{(t_2,T_1)} * nf_{(t_4,T_1)} * nf_{(t_5,T_1)} * nf_{(t_7,T_1)} = 0*(1/4)*(1/4)*0*0 = 0$

Avec $nf_{(W,T_1)}$ est la fréquence normalisée de W dans le texte parallèle de la traduction T_1.

$\Pi(T_2|RS) = nf_{(W,T_2)} * nf_{(t_2,T_2)} * nf_{(t_4,T_2)} * nf_{(t_5,T_2)} * nf_{(t_7,T_2)} = 0 * 0 * (1/5) * (1/5) * (1/5) = 0$

Nous avons souvent $\Pi(T_j|RS) = 0$; sauf si tous les mots de la requête source existent dans l'index du texte parallèle de la traduction T_j.

D'autre part, nous n'avons pas des valeurs nulles pour la mesure $N(T_j|RS)$:

$N(T_1|RS) = 1 - ((1 - \phi(T_1, W)) * (1 - \phi(T_1, t_2)) * (1 - \phi(T_1, t_4)) * (1 - \phi(T_1, t_5)) * (1 - \phi(T_1, t_7)))$;

Avec :

- $\phi(T_1, W) = 0$ parce que $nf_{(T_1,W)} = 0$;
- $\phi(T_1, t_2) = log_{10}(2/1) * (1/4) = 0.075$;
- $\phi(T_1, t_4) = log_{10}(2/2) * (1/4) = 0$;
- $\phi(T_1, t_5) = 0$;
- $\phi(T_1, t_7) = 0$

Donc :

$N(T_1|RS) = 1-((1-0)*(1-0.075)*(1-0)*(1-0)*(1-0)) = 1-(1*0.925*1*1*1) = 0.075$

Par conséquent, $DPR(T_1|RS) = 0.075$.

$N(T_2|RS) = 1 - ((1 - \phi(T_2, W)) * (1 - \phi(T_2, t_2)) * (1 - \phi(T_2, t_4)) * (1 - \phi(T_2, t_5)) * (1 - \phi(T_2, t_7)))$

Avec :

- $\phi(T_2, W) = 0$ parce que $nf(T_2, W) = 0$;
- $\phi(T_2, t_2) = 0$;
- $\phi(T_2, t_4) = log_{10}(2/2) * (1/5) = 0$;
- $\phi(T_2, t_5) = log_{10}(2/1) * (1/5) = 0.06$;
- $\phi(T_2, t_7) = log_{10}(2/1) * (1/5) = 0.06$

Donc :

$N(T_2|RS) = 1 - ((1-0) * (1-0) * (1-0) * (1-0.06) * (1-0.06)) = 1 - (1 * 1 * 1 * 0.94 * 0.94) = 1 - 0.8836 = 0.1164$

Ainsi, $DPP(T_2|RS) = 0.1164 > DPP(T_1|RS) = 0.075$

Nous remarquons que la requête source RS est plus pertinente pour T_2 que T_1, car elle renferme trois termes (t_4, t_5, t_7) de l'index du texte parallèle de T_2 et seulement deux termes (t_2, t_4) de l'index du texte parallèle de T_1.

4.3 Approche possibiliste hybride

Nous proposons dans cette section une approche possibiliste hybride pour la désambiguïsation de TR (Ben Romdhane et al., 2019), en utilisant à la fois un dictionnaire bilingue et un corpus de texte parallèle. L'idée principale de notre l'approche est de combiner l'approche possibiliste à base d'une transformation probabilité-possibilité (Elayeb et al., 2018) (section 4.1) avec l'approche possibiliste discriminative (Ben Romdhane et al., 2017) (section 4.2) pour bénéficier des avantages des deux.

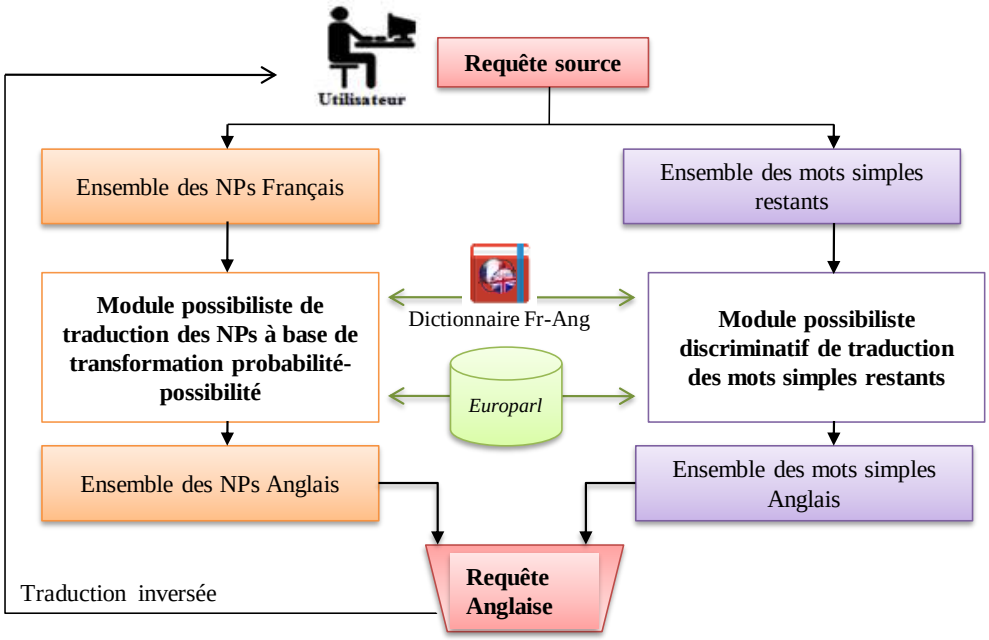

Figure 4.5 – Modèle conceptuel de l'approche possibiliste hybride

Étant donné les termes d'une requête source, la première étape consiste à sélectionner les syntagmes nominaux et les traduire en unités. Dans cette étape, nous avons profité du module possibiliste de traduction des NPs de l'approche à base de transformation probabilité-possibilité. En effet, nous avons introduit une tolérance supplémentaire dans la traduction des NPs en utilisant les patrons de traduction et un modèle de langue. Dans la deuxième étape, les termes restants de la requête source qui ne sont pas inclus dans les NPs sélectionnés, sont traduits en utilisant l'approche possibiliste discriminative. Cette dernière modélise la pertinence de

traduction d'un terme donné de la requête source à travers deux mesures : la pertinence possible qui permet de rejeter les traductions candidates non-pertinentes, alors que la pertinence nécessaire permet de renforcer les traductions non-éliminées par la possibilité. Les principales étapes de cette approche sont schématisées dans la figure 4.5.

4.3.1 Exemple illustratif

Considérons l'exemple de la requête source (RS), extraite du standard de test CLEF-2003 (Query_ID : 162-AH) :

« Trouvez [des documents] traitant [des problèmes] posés par [la Grèce] concernant [l'abolissement] [des restrictions douanières] entre [l'Union Européenne] et [la Turquie] ».

Tout d'abord, nous avons lemmatisé RS comme suit : « *Trouver [du document] traiter [du problème] poser par [la Grèce] concerner [l'abolissement] [de restriction douanière] entre [l'Union Européenne] et [la Turquie]* ».

Puis, nous identifions les NPs [entre crochets] et les mots simples restants (soulignés). Nous détaillons dans la suite uniquement le calcul lié au syntagme nominal «*du problème*». Nous sélectionnons toutes les traductions possibles pour chaque mot de ce NP, en utilisant le dictionnaire bilingue Fr-Ang. Par exemple, le mot «*du*» possède 4 traductions candidates (*from, of, the, of the*). Le mot «*problème*» possède 4 traductions candidates (*difficulty, issue, problem, trouble*).

Ensuite, nous identifions la catégorie grammaticale (POS) de chaque mot. Les termes (*from, of, the, of the*) ont respectivement les POS suivants (*IN, IN, DT, IN DT*), tandis que les POS correspondants pour l'ensemble (*difficulty, issue, problem, trouble*) sont (*NN, NN, NN, NN*). Par conséquent, le nombre total des combinaisons est la multiplication de la taille de tous les ensembles de POS ($n = 16 = 4 * 4$).

À l'aide du modèle probabiliste, nous calculons la première probabilité $P(F|E) = 1/n$ correspondant au premier NP (*du problème*) est égal à $1/16 = 0.0625$. Nous avons la même distribution des possibilités $\pi(F|E) = 0.0625$.

Après, nous calculons la deuxième probabilité : $P(FPT|EPT) = Occ(FPT, EPT)/Occ(EPT)$. Le FPT de « *du problème* » est [PRP : DET NOM]. Nous avons sélectionné tous les patrons de traduction correspondants aux 16 combinaisons possibles ayant $P(FPT|EPT)$ supérieur à zéro. Par exemple, nous avons les combinaisons telles que (FPT(PRP :det NOM) – EPT(IN NN)), (FPT(PRP :det NOM) – EPT(DT NN)),)), (FPT(PRP :det NOM) – EPT(IN DT NN)), etc.

La troisième probabilité $P(ENP) = P(e_1, ..., e_n) = \prod_{i=1}^{n} P(e_i|e_{i-2}, e_{i-1})$ est calculée en utilisant le modèle de langue Trigramme sur la partie anglaise du corpus *Europarl*. Par conséquent, la meilleure traduction correspondante au plus haut score de : $ENP^* = argmax_{ENP}(P(F|E) \times P(FPT|EPT) \times P(ENP))$

Néanmoins, les distributions des possibilités correspondantes sont obtenues en utilisant la formule de transformation probabilité-possibilité. Le tableau 4.5 illustre les résultats des calculs probabilistes et possibilistes des NPs.

Selon les résultats du tableau (4.5), nous notons que la meilleure traduction sélectionnée du NP « *du problème* » en utilisant le modèle probabiliste est « *of the problem* » correspond au score probabiliste le plus élevé $ENP^* = 3.16E-06$, alors que « *of issue* » est la traduction sélectionnée pour le modèle possibiliste qui correspond au score possibiliste le plus élevé $ENP^* = 8.56E-01$.

La dernière étape de l'approche hybride est la traduction des mots simples restants. Nous fournissons dans le tableau (4.6) le calcul possibiliste discriminatif qui correspond aux 5 premiers mots restants. Nous nous concentrons sur le *degré de pertinence possibiliste* (DPP) des meilleures traductions (gras et souligné).

Combinaisons	***Couple (FPT, EPT)***	***P(FPT|EPT)***	***π(FPT|EPT)***	***P(ENP)***	***π(ENP)***	***Score Probabiliste***	***Score Possibiliste***
(1). from, difficulty	PRP :det NOM , IN NN	0.055	2495.358	1.97E-05	2.87E-04	6.83E-08	4.48E-02
(2). from, problem	PRP :det NOM , IN NN	0.055	2495.358	3.93E-05	5.43E-04	1.37E-07	8.47E-02
(3). from, trouble	PRP :det NOM , IN NN	0.055	2495.358	0	0	0	0
(4). from, issue	PRP :det NOM , IN NN	0.055	2495.358	9.84E-05	0.001	3.42E-07	1.82E-01
(5). of, difficulty	PRP :det NOM , IN NN	0.055	2495.358	1.29E-04	0.001	4.47E-07	2.25E-01
(6). of, problem	PRP :det NOM , IN NN	0.055	2495.358	5.65E-04	0.004	1.96E-06	7.64E-01
(7). of, trouble	PRP :det NOM , IN NN	0.055	2495.358	4.87E-05	6.55E-04	1.69E-07	1.02E-01
(8). **of, issue**	PRP :det NOM , IN NN	0.055	2495.358	6.64E-04	0.005	2.31E-06	**8.56E-01**
(9). the, difficulty	PRP :det NOM , DT NN	1.32E-04	8.674	5.31E-04	0.004	4.39E-09	2.53E-03
(10). the, problem	PRP :det NOM , DT NN	1.32E-04	8.674	0.005	0.022	4.58E-08	1.21E-02
(11). the, trouble	PRP :det NOM , DT NN	1.32E-04	8.674	6.44E-05	8.28E-04	5.32E-10	4.49E-04
(12). the, issue	PRP :det NOM , DT NN	1.32E-04	8.674	0.005	0.021	4.22E-08	1.16E-02
(13). of the, difficulty	PRP :det NOM , IN DT NN	0.008	462.528	7.71E-04	0.006	4.12E-07	1.74E-01
(14). **of the, problem**	PRP :det NOM , IN DT NN	0.008	462.528	0.005	0.022	**3.16E-06**	6.54E-01
(15). of the, trouble	PRP :det NOM , IN DT NN	0.008	462.528	1.19E-05	1.78E-04	6.34E-09	5.14E-03
(16). of the, issue	PRP :det NOM , IN DT NN	0.008	462.528	0.003	0.015	1.67E-06	4.46E-01

Table 4.5 – Résultats de calcul de traduction des NPs par les modèles probabiliste et possibiliste

Mots simples	***Traductions candidates possibles***	$\pi(T_j\|RS)$	$N(T_j\|RS)$	$DPP(T_j\|RS)$
trouver	come, find, get, **pinpoint**, strike, think.	0	4.71E-05	4.71E-05
traiter	address, consider, discuss, examine, handle, include, maintain, process, resolve, tackle, trade, treat, treaty, **unfold**.	2.13E-11	8.88E-06	8.88E-06
poser	ask, cause, decide, establish, **make**, pose, propose, put, raise, stipulate, table.	1.01E-31	-3.03E-07	-3.03E-07
par	by, on, per, through, **to**, via.	7.48E-45	-3.98E-08	-3.98E-08
concerner	**about**, concern, for, intend, involve, regard, relate, surround.	1.75E-30	1	1

Table 4.6 – Calcul discriminatif des 5 premiers mots simples restants

Conclusion

Ce travail constitue la première tentative de l'application de la théorie des possibilités pour la désambiguïsation translinguistique. En effet, nous avons proposé, dans ce chapitre, trois approches possibilistes de désambiguïsation de TR dans la RIT.

La première est une approche à base d'une transformation probabilité-possibilité, dans laquelle nous avons suivi trois étapes principales. Tout d'abord, nous avons construit un dictionnaire bilingue Français-Anglais, à l'aide du corpus parallèle *Europarl*, et nous l'avons enrichi en utilisant des correcteurs intelligents en ligne. Puis, nous avons identifié les syntagmes nominaux français des requêtes sources. Ensuite, nous l'avons traduit en utilisant des patrons de traduction et un modèle de langue. Enfin, nous avons traduit les termes de requête source qui ne font pas partie des syntagmes nominaux en utilisant une approche possibiliste de traduction des mots simples.

La seconde est une approche possibiliste discriminative basée sur un réseau possibiliste. La pertinence de traduction d'un terme d'une requête source est modélisée par deux mesures, à savoir : la possibilité et la nécessité afin d'étudier ses impacts sur la désambiguïsation des traductions.

La troisième est une approche hybride qui représente une combinaison des deux premières approches. Nous avons profité dans cette dernière des points forts de l'approche à base de transformation probabilité-possibilité, en utilisant son module possibiliste de traduction des syntagmes nominaux et de l'approche possibiliste discriminative, en utilisant son module de traduction des termes restants de la requête source.

Nous présenterons, dans le chapitre suivant, la validation expérimentale de ces trois approches et la proposition d'un SRIT afin de récupérer des documents pertinents

qui répondent aux besoins des utilisateurs.

Chapitre

5 Validation des approches possibilistes pour la désambiguïsation des traductions de requêtes

Sommaire

Introduction

Ce chapitre présente un récapitulatif pour la validation de nos approches possibilistes de désambiguïsation de TR dans la RIT, à savoir : l'approche possibiliste à base de transformation probabilité-possibilité, l'approche possibiliste discriminative et l'approche hybride. En fait, nous détaillons les expérimentations et les résultats

fournis par nos approches proposées afin de mesurer la performance de chacune par rapport aux approches probabiliste *naïve bayésienne* et monolingue.

Nous commençons dans la section 5.1 par introduire le cadre du travail, en dérivant les différentes ressources exploitées. Dans la section 5.2, nous mettons en œuvre un SRIT, dans lequel nous avons intégré nos approches. Dans la section 5.3, nous exposons les résultats globaux et les scénarios d'évaluation en termes de précision aux top documents, de précision moyenne et de précision exacte. Enfin, dans la section 5.4, nous synthétisons et nous discutons les expériences effectuées.

5.1 Cadre du travail

Nos expériences ont été réalisées via notre SRIT possibiliste SPORT (Système POssibiliste de tRaduction de requêTes) qui est implémenté en langage Java et à l'aide de la plate-forme *Terrier* pour la phase d'indexation. SPORT fournit des nombreux modèles d'appariement existants entre les requêtes et les documents tels que OKAPI BM25 et le modèle d'appariement possibiliste initialement proposé par (Elayeb, 2009 ; Elayeb et al., 2009) et étendu par (Ben Khiroun et al., 2011 ; Elayeb et al., 2011). Nous avons comparé les résultats de notre SRIT en utilisant le modèle d'appariement OKAPI BM25.

Nous présentons brièvement dans la suite la collection de test utilisée CLEF-2003 [1] (cf. section 5.1.1), le corpus parallèle *Europarl* (cf. section 5.1.2) et le dictionnaire bilingue Français-Anglais (cf. section 5.1.3).

5.1.1 La collection de test CLEF-2003

Dans les expérimentations, nous avons utilisé un sous ensemble de la collection de test CLEF-2003. Cette partie comprend des articles publiés en 1995 dans le journal «*Glasgow Herald*». Elle comprend 56472 documents et 54 requêtes, formant 154 Mo.

La collection de test constitue les éléments nécessaires pour l'évaluation d'un SRI, tels que : un ensemble de requêtes, un ensemble de documents et une liste de jugements de pertinence des documents pour chaque requête. Les requêtes sont représentées sous format XML. Suivant le modèle de la campagne TREC, chaque requête de test a un identifiant <*identifier*> qui contient le numéro de la requête. En fait, les requêtes débutent au numéro 141 et s'achèvent avec le numéro 200. De plus, chaque requête possède trois champs, à savoir : un titre <*titre*>, une phrase

1. https://www.clef-campaign.org/

qui exprime le besoin d'un utilisateur <*description*>, et une partie <*narration*> qui comprend plus de détail et de contexte pour la requête. La figure 5.1 présente un exemple de requête en français.

```
<topic lang="fr">
<identifier>199-AH</identifier>
<title>L'épidémie d'Ebola au Zaïre</title>
<description>Trouvez des documents sur les mesures préventives prises après le déclenchement de
l'épidémie d'Ebola au Zaïre.</description>
<narrative>En mai 1995, dans la ville zairoise de Kikwit, est survenue l'une des épidémies les plus
redoutées: l'épidémie d'Ebola. Les documents pertinents portent sur les mesures prises pour empêcher la
propagation de cette maladie.</narrative>
</topic>
```

Figure 5.1 – Exemple de requête en Français du standard CLEF-2003

5.1.2 Le corpus parallèle Europarl

Le corpus *Europarl* est un corpus de textes parallèles parlementaires, généré gratuitement sur le Web par les travaux du Parlement Européen dans 11 langues européennes (Français, Italien, Portugais, Espagnole, Anglais, Allemand, Néerlandais, Danois, Suédois, Finnois et grec). Ce corpus a établi une large utilisation dans la communauté du traitement des langues naturelles (Koehn, 2005).

Pour préparer les données, nous avons utilisé le script Perl *clean-corpus-n.perl*. Ce dernier nettoie le corpus en supprimant les lignes vides, les espaces redondants ainsi que les phrases trop courtes et trop longues par rapport aux phrases correspondantes. De plus, nous avons ôté la casse en mettant les données en minuscule. Au-delà, pour la préparation du corpus au format *Giza++*, nous avons filtré les phrases en éliminant les phrases trop longues (plus de 30 mots dans notre configuration) à cause de *Giza++* qui met beaucoup du temps à aligner les phrases longues. Le corpus contient, désormais, 1.150.437 phrases alignées.

5.1.3 Le dictionnaire bilingue Français-Anglais

D'abord, nous avons généré notre dictionnaire bilingue à partir du corpus parallèle *Europarl* en utilisant tous les mots français avec leurs traductions existantes dans ce corpus pour élargir sa couverture. Puis, nous avons vérifié ce dictionnaire à l'aide du correcteur intelligent d'orthographe et de grammaire *Reverso*[2], afin de tenir compte des mots techniques, des noms propres, etc. Enfin, nous avons enrichi et vérifié ce dictionnaire à l'aide du traducteur en ligne *Google Translation*[3] dans l'objectif de s'assurer de son contenu ainsi qu'élargir sa couverture. Nos 54 requêtes de

2. http://www.reverso.net/spell-checker/english-spelling-grammar/
3. *https://translate.google.fr/?hl=fr*

test contiennent 717 mots français ayant 2324 traductions anglaises possibles dans le dictionnaire bilingue. Les étapes de préparation de notre dictionnaire bilingue Français-Anglais, sont présentées dans la figure 4.2 dans le chapitre 4.

5.2 SPORT : Système POssibiliste de tRaduction de requêTes

Nous mettons en place un système de RIT, intitulé SPORT, qui tire profil des nos approches de désambiguïsation de TR proposées.

Nous présentons dans la figure 5.2 l'interface principale du SPORT (Elayeb et al., 2018). L'interface utilisateur est composée de deux parties : Premièrement, la partie supérieure est pour la traduction, elle contient : (i) une zone de texte pour écrire la requête source initiale de l'utilisateur après avoir choisi la langue source «*Source language*» ; et (ii) une zone du texte réservée à la requête cible traduite après avoir choisi la langue cible «*Target language*». La traduction sera démarrée via le bouton «*Translate*». Tandis que, le bouton «*Back Translation*» est dédié pour la traduction inverse. D'autre part, la reformulation de requête est assurée par le bouton «*Expansion*», alors que la visualisation des liaisons entre les termes est assurée par le bouton «*Visualization*».

Deuxièmement, la partie inférieure représente les résultats de recherche. Cette partie contient : (i) une liste des modèles d'appariement de RI, tels que le modèle possibiliste (Elayeb et al., 2011) ou le modèle probabiliste OKAPI-BM25 qui est intégré à partir de la plate-forme Terrier ; (ii) une zone réservée pour afficher les résultats de recherche avec un score de pertinence pour chaque document. Les documents sont affichés après la validation de l'utilisateur en utilisant le bouton «*Search*».

En fait, la plate-forme d'expérimentation Terrier-3.5 (Ounis et al., 2006) est un cadre robuste pour la recherche et l'expérimentation en RI. Cette plate-forme assure l'appariement entre requête/documents et l'indexation que nous avons utilisé pour la réalisation de notre système.

L'architecture modulaire du Terrier permet une flexibilité dans le processus d'indexation à plusieurs étapes : Dans le traitement d'un corpus de documents, dans le traitement et l'analyse de chaque document, dans le traitement des termes à partir des documents et dans l'écriture des structures de données d'index.

Chaque terme extrait d'un document possède trois propriétés fondamentales, à savoir : (i) la forme textuelle réelle du terme (ii), la position dans laquelle le terme apparaît dans le document et (iii), les champs dans lesquels le terme apparaît (les

champs sont généralement utilisés pour définir les balises HTML dans lesquelles un terme apparaît).

Terrier ajoute la configuration « *TermPipeline* » à l'indexation, ce qui permet de transformer les termes des documents de différentes manières. Dans notre cas, le traitement des termes est effectué au préalable avec d'autres outils linguistiques, mais la suppression de mots vides et la réduction sous forme canonique peuvent aussi être effectuer par Terrier.

Nous avons utilisé divers modèles d'appariement. Chaque combinaison, d'un modèle d'appariement et des indexes générés, donne des échantillons qui correspondent aux documents ayant les meilleurs scores par rapport à la requête insérée. Le score de pertinence représente le jugement de pertinence de l'utilisateur par rapport au document.

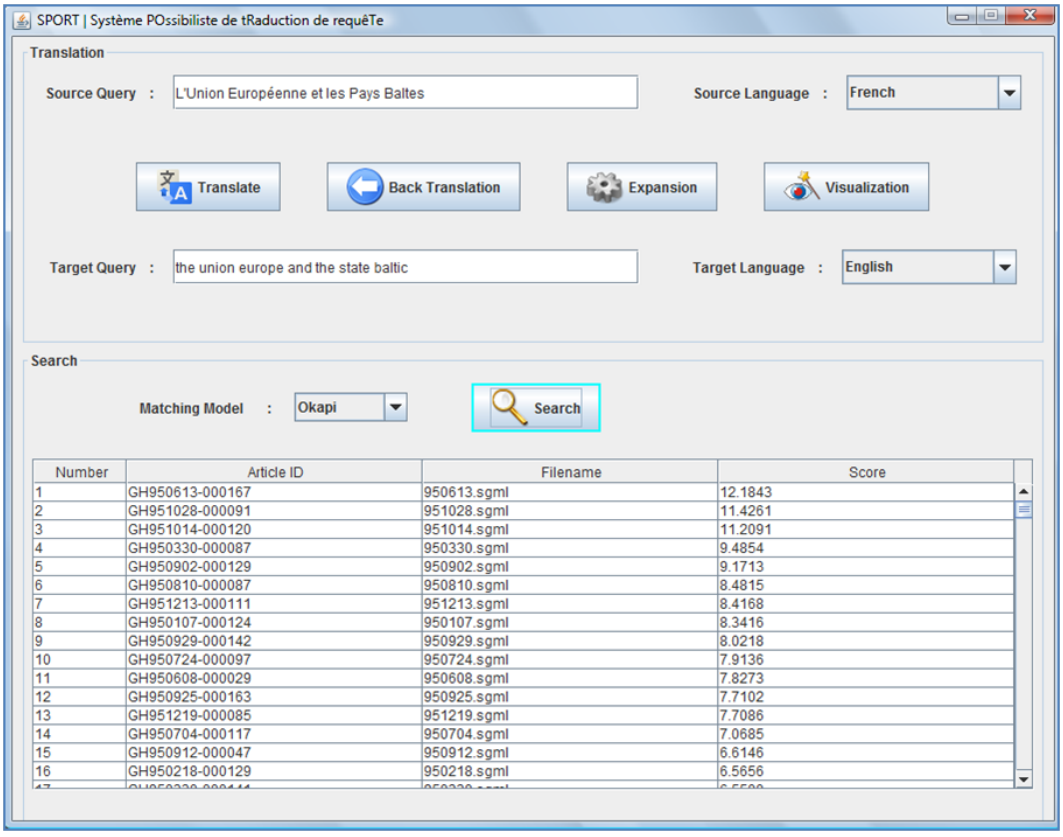

Figure 5.2 – Interface principale du système SPORT

En outre, le système SPORT fournit plusieurs fonctionnalités telles que :

– Un résumé, dans la langue de l'utilisateur, pour chaque document récupéré à l'aide de l'API *custom-translate-api*[4] et *Classifier4J-0.6*[5] (figure 5.3). En fait, l'objectif de ces résumés est de renforcer la confiance de l'utilisateur

4. https://code.google.com/p/java-google-translate-text-to-speech/
5. http://classifier4j.sourceforge.net/

envers les documents retournés, et ils peuvent être utilisés dans le processus d'expansion de la requête source.

– Une technique de *traduction inverse* pour mieux convaincre l'utilisateur envers la traduction de sa requête.
– Une expansion de requête donnant la possibilité à l'utilisateur de reformuler sa requête source ou/et cible avant/après le processus de traduction. En fait, nous avons intégré dans SPORT un outil existant pour l'expansion sémantique de requête, nommé SPORSER [6] (Elayeb et al., 2011 ; Khiroun et al., 2011). Cet outil est exploité pour les deux processus d'expansion et de visualisation.
– Une interface de visualisation pour supporter l'utilisateur via un composant de navigation (graphe de dictionnaire monolingue/bilingue). Cette technique de visualisation interactive a pour objectif d'aider l'utilisateur à identifier les relations sémantiques entre les mots et ses traductions. En conséquence, il pourra choisir les traductions les plus proches sémantiquement, utile pour le processus d'expansion de requête source et/ou cible. Nous avons profité de l'outil disponible de visualisation *Prefuse* [7] (Heer et al., 2005) pour enrichir la visualisation et la navigation dans des données interactives.

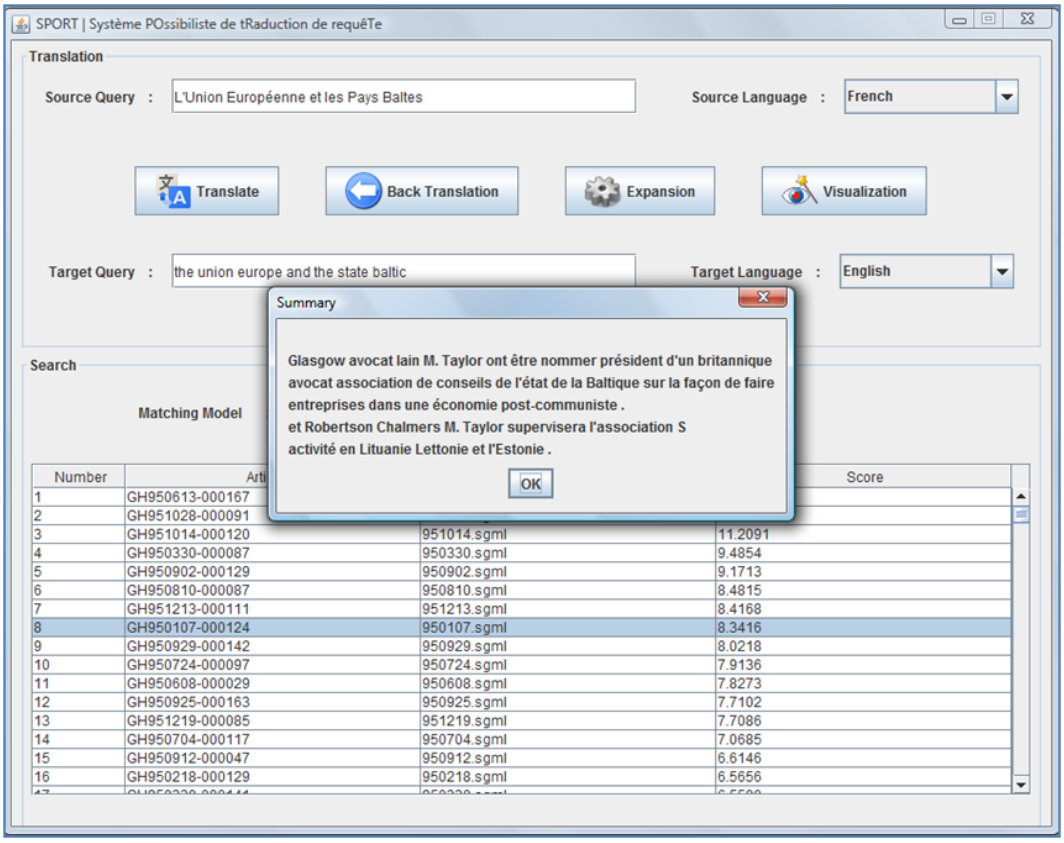

Figure 5.3 – Interface SPORT affichant le résumé d'un document résultat

6. SPORSER : Système POssibiliste de Reformulation SEmantique de Requêtes.
7. http://prefuse.org/

5.3 Résultats expérimentaux

Notre évaluation est effectuée conformément au protocole TREC. Plus précisément, chaque requête est soumise au système SPORT avec des paramètres fixes (*titre*, *description*, *narration*, etc.), et le modèle d'appariement retourne l'ensemble des documents pertinents correspondants à chaque requête. Des valeurs de précision sont calculées aux top documents $P@5, P@10, \ldots, P@1000$, *MAP*, *R-Précision*. Le ratio des documents pertinents parmi les 5 premiers documents retournés est la précision au point 5, notée $P@5$. Certaines évaluations sont également mesurées sur 11 points de rappel $0.1, 0.2, \ldots, 1.0$.

Le but de ces expériences est d'évaluer nos diverses contributions, en se basant principalement sur les quatre points suivants : les différents scénarios des courbes de rappel-précision (section 5.3.1), les points de précision aux top documents et les métriques *MAP* et *R-Précision* (section 5.3.2), le pourcentage d'amélioration (section 5.3.3), et enfin le test de significativité statistique de *Wilcoxon* (section 5.3.4).

En fait, nos résultats dépendent de la capacité de la requête de test à fournir le maximum d'informations contextuelles adaptées pour traduire les syntagmes nominaux (NPs) en première étape et les mots simples restants dans la deuxième étape. Les requêtes longues impliquent : «*titre+description+narration*» ou «*titre+description*» ou «*titre+narration*» ou «*description+narration*», alors que les requêtes courtes sont celles qui utilisent «*titre*» ou «*description*» ou «*narration*». Ces différentes requêtes sont utilisées comme des données d'entrée dans le SRIT (Ben Romdhane et al., 2017 ; Elayeb et al., 2018).

5.3.1 Scénarios des courbes de rappel-précision

Dans le but de mener une étude détaillée, nous avons évalué nos expérimentations en se référant aux résultats des courbes de *Rappel-Précision*.

Cette section est divisée en trois parties. La première section est basée sur l'évaluation de l'approche possibiliste à base de transformation probabilité-possibilité par rapport à l'approche probabiliste *naïve bayésienne* (cf. section 5.3.1.1). La deuxième section est basée sur l'évaluation de l'approche discriminative par rapport à l'approche probabiliste (cf. section 5.3.1.2). La troisième section est basée sur l'évaluation de l'approche hybride par rapport aux approches possibiliste, discriminative et probabiliste (cf. section 5.3.1.3).

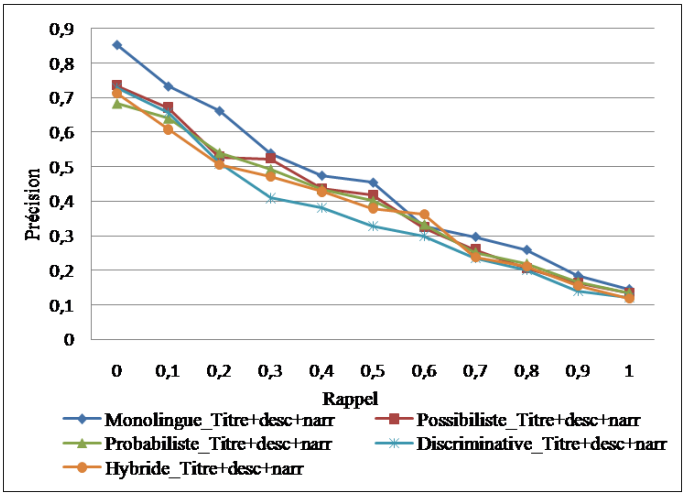

Figure 5.4 – Courbes de Rappel-Précision utilisant « *Titre+desc+narr* »

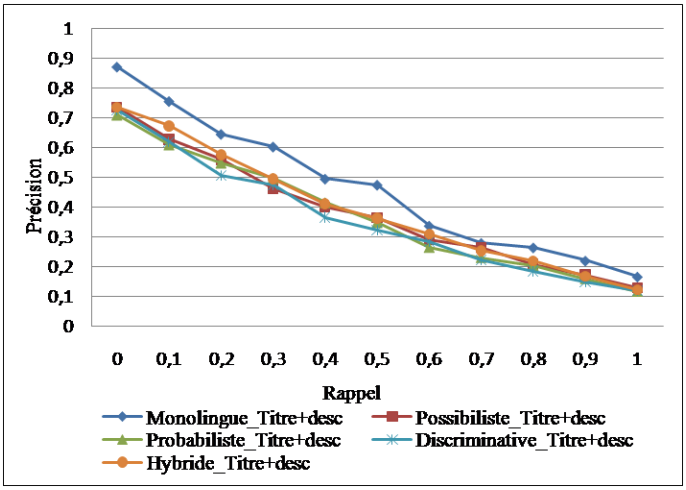

Figure 5.5 – Courbes de Rappel-Précision utilisant « *Titre+desc* »

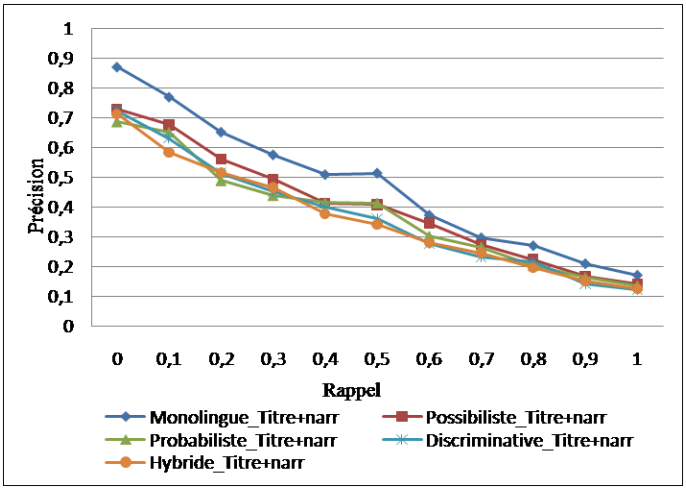

Figure 5.6 – Courbes de Rappel-Précision utilisant « *Titre+narr* »

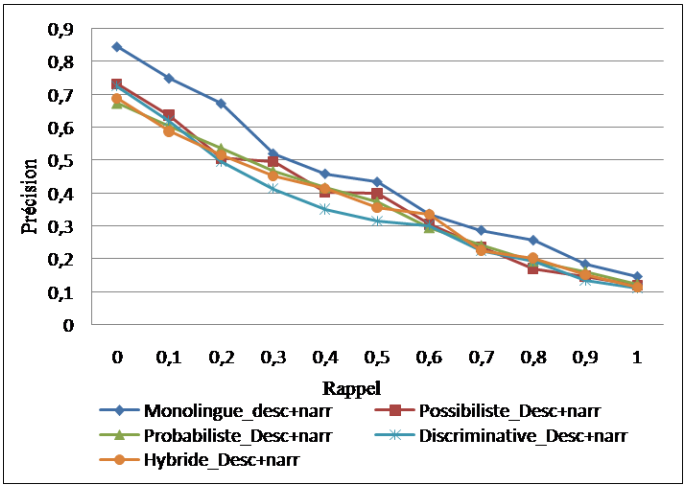

Figure 5.7 – Courbes de Rappel-Précision utilisant « *Desc+narr* »

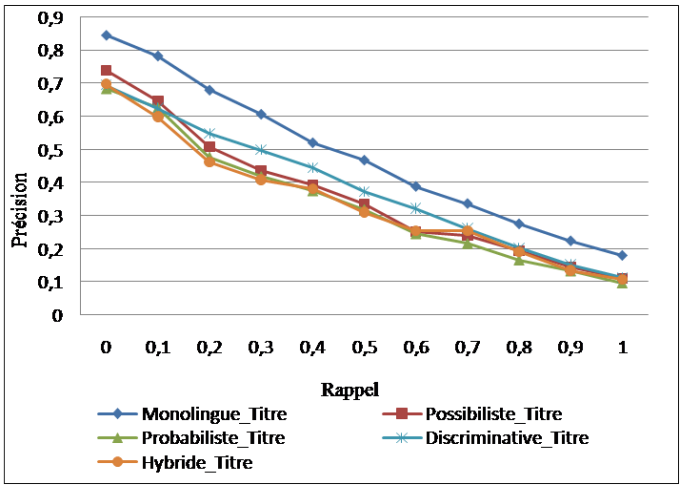

Figure 5.8 – Courbes de Rappel-Précision utilisant « *Titre* »

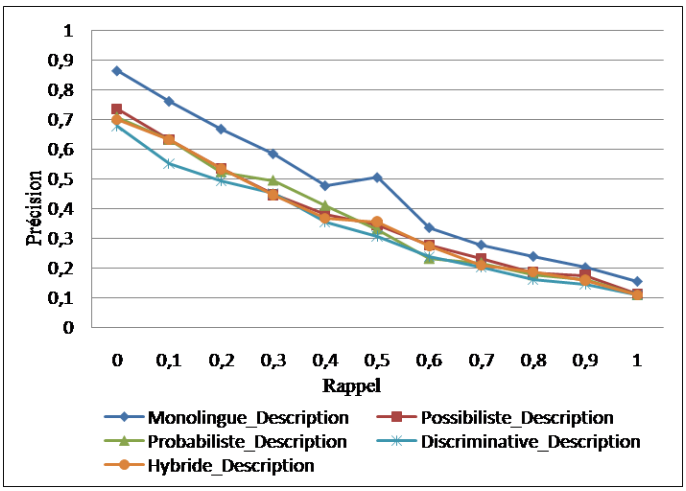

Figure 5.9 – Courbes Rappel-Précision utilisant « *Description* »

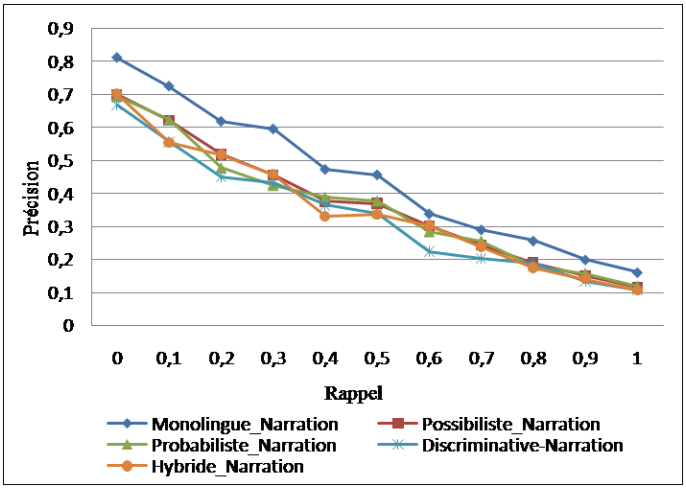

Figure 5.10 – Courbes Rappel-Précision utilisant « *Narration* »

5.3.1.1 Comparaison de l'approche possibiliste à base de transformation à l'approche probabiliste

Nous nous focalisons sur les différents scénarios des courbes de *rappel-précision* qui comparent les exécutions de l'approche possibiliste, probabiliste et monolingue (Elayeb et al., 2018).

Si nous utilisons uniquement les requêtes «*Titre*», le contexte est limité à un nombre réduit de termes dans lesquels l'identification des NPs n'est pas fréquente. Par conséquent, la traduction mot-à-mot est fréquente dans ce cas. Nous remarquons que notre approche possibiliste est légèrement supérieure à l'approche probabiliste surtout dans certains points de rappel de bas- et haut-niveau. Cependant, l'exécution monolingue a surpassé les deux approches dans tous les points de rappel (figure 5.8).

Le contexte de la requête est amélioré lorsque nous utilisons «*Description*», puisqu'elle contient quelques syntagmes nominaux. Dans ce cas, l'approche possibiliste est supérieure à l'approche probabiliste dans certains points de rappel de haut-niveau grâce à l'identification et la traduction des NPs fournies par le contexte étendu de la requête. Par conséquent, l'approche possibiliste est de plus en plus proche de monolingue à partir du point de rappel 0.5 (figure 5.9).

Lorsque nous nous concentrons sur la partie «*Narration*» des requêtes, le contexte est riche, amélioré et plus détaillé. Ainsi, l'identification des NPs est plus fréquente et l'approche possibiliste est supérieure à l'approche probabiliste. En outre, les écarts entre les exécutions de monolingue et des approches possibiliste et probabiliste sont de plus en plus réduits à partir du point de rappel 0.5 (figure 5.10).

D'une manière appropriée, lorsque nous utilisons à la fois «*Titre+narr*», l'approche possibiliste est significativement supérieure à la probabiliste, surtout sur les points

de rappel de bas- et haut-niveau, et elle est proche de monolingue dans certains points de rappel de haut-niveau (figure 5.6).

Néanmoins, «*Titre+desc*» semblent moins confortables pour l'approche possibiliste lorsque sa performance dépasse légèrement l'approche probabiliste dans certains points de rappel. Dans le point de rappel 0.7, l'approche possibiliste semble très proche de monolingue (figure 5.5).

En outre, l'utilisation de «*Desc+narr*» fournit un maximum de contexte dans lequel l'approche possibiliste est supérieure à l'approche probabiliste et elle est plus proche de monolingue dans plusieurs points de rappel (figure 5.7).

Lorsque nous exécutons les requêtes complètes «*Titre+desc+narr*», la performance de l'approche possibiliste est légèrement sous l'approche probabiliste et elle a la même performance que la monolingue dans certains points de rappel de haut-niveau (figure 5.4).

5.3.1.2 Comparaison de l'approche discriminative à l'approche probabiliste

Nous comparons dans cette section les courbes de *rappel-précision* de l'approche possibiliste discriminative avec celles de l'approche probabiliste et monolingue (Ben Romdhane et al., 2017).

Pour les requêtes longues utilisant «*Titre+desc+narr*», les courbes de rappel-précision de la figure 5.4 ont montré que l'approche discriminative est légèrement inférieure à l'approche probabiliste, particulièrement pour les points de rappel de bas-niveau, alors qu'elle est proche de l'approche probabiliste dans certains points de rappel de haut-niveau. En outre, les écarts entre ces deux approches, par rapport au monolingue, sont de plus en plus réduits à partir du point de rappel 0.6. Ces interprétations restent valables pour les trois autres scénarios de requêtes longues utilisant «*Titre+desc*» (figure 5.5), «*Titre+narr*» (figure 5.6) et «*Desc+narr*» (figure 5.7).

Si nous utilisons la partie «*Description*» des requêtes (figure 5.9), les deux approches discriminative et probabiliste se croisent dans certains points de rappel de haut-niveau à partir de 0.5. Ce croisement a eu lieu de nouveau aux points de rappel 0.8 et 1 en utilisant plus de contexte dans la partie «*Narration*» (figure 5.10). À partir du point de rappel 0.6, les écarts entre la monolingue et ces deux approches sont de plus en plus réduits.

Enfin, si nous nous concentrons uniquement sur la partie «*Titre*» des requêtes, le contexte semble être plus limité à quelques termes dans lesquels l'identification des

syntagmes nominaux est rare. L'approche discriminative surpasse significativement l'approche probabiliste, sauf dans certains points de rappel de bas-niveau (0 et 0.1). La monolingue reste supérieure en dépassant la discriminative et la probabiliste dans tous les points de rappel (figure 5.8).

Nous confirmons de nouveau que les résultats fournis par les courbes de *rappel-précision* dépendent de la capacité des requêtes à fournir un maximum d'information contextuelle, utiles pour l'identification et la traduction des NPs. Pour ces raisons, l'approche probabiliste semble plus confortable en utilisant les requêtes longues. Par contre, l'approche discriminative a montré son efficacité dans les requêtes courtes ; où le contexte est limité à un petit ensemble des termes dans lesquels l'identification des NPs est limitée (Ben Romdhane et al., 2017).

5.3.1.3 Comparaison de l'approche hybride aux approches possibiliste, discriminative et probabiliste

Nous illustrons dans cette section les scénarios des courbes de rappel-précision des requêtes longues et courtes, en comparant l'approche possibiliste hybride avec les approches probabiliste, discriminative, possibiliste et monolingue (Ben Romdhane et al., 2019). Notre objectif ici est d'étudier la sensibilité de ces approches au contexte fourni par la requête source.

Les requêtes longues utilisant «*Titre+desc+narr*» contiennent une information contextuelle complète adéquate pour l'identification des syntagmes nominaux. Si nous étudions les courbes de rappel-précision de la figure 5.4, nous remarquons que l'approche hybride est légèrement au-dessous des approches probabiliste, possibiliste et discriminative, particulièrement dans les points de rappel de bas-niveau (0 à 0.2). En outre, l'approche hybride surpasse l'approche discriminative à partir du point de rappel 0.2 et elle est au-dessus de toutes les approches au point de rappel 0.6. L'approche hybride est très proche de l'approche possibiliste dans certains points de rappel de haut-niveau (0.7 à 1.0). Elle surpasse également la monolingue dans le point de rappel 0.6. L'écart entre l'approche hybride et la monolingue est devenu de plus en plus réduit à partir du point de rappel 0.7.

D'autre part, et en utilisant «*Titre+desc*», l'approche hybride surpasse les trois autres approches, en particulier dans les points de rappel de bas-niveau (0 à 0.3). Ensuite, elle dépasse légèrement ces approches et devient très proche de monolingue à partir du point de rappel 0.3. Les écarts entre ces approches et la monolingue sont progressivement réduits à partir du point de rappel 0.6 (figure 5.5).

Lorsque nous nous focalisons sur les résultats en utilisant «*Titre+narr*», l'approche hybride est légèrement inférieure aux trois autres approches, en particulier dans

les points de rappel de bas-niveau (0 et 0.1). Elle surpasse les deux approches probabiliste et discriminative, mais reste sous l'approche possibiliste, dans les points de rappel de bas-niveau (0.2 et 0.3). Les écarts entre elles ont été réduits dans certains points de rappel de haut-niveau (0.7 à 1.0) (figure 5.6).

Enfin, les requêtes longues en utilisant «*Desc+narr*» fournissent beaucoup d'informations contextuelles et ont montré que l'approche hybride se situe légèrement sous les trois approches dans certains points de rappel de bas-niveau (0 à 0.2). Elle surpasse l'approche discriminative, mais elle reste sous les deux approches possibiliste et probabiliste dans les points de rappel entre 0.2 et 0.5. De plus, l'approche hybride surpasse toutes les trois approches dans certains points de rappel de haut-niveau (0.6 et 0.8). Elle a également réalisé la même performance que la monolingue au point de rappel 0.6. Les écarts entre toutes ces approches et la monolingue diminuent progressivement à partir du point de rappel 0.7 (figure 5.7).

Les requêtes courtes utilisant «*Titre*» semblent plus adaptées à l'approche discriminative parce qu'elles ont une information contextuelle réduite. En conséquence, l'approche discriminative a principalement surpassé les trois autres approches dans des nombreuses parties du rappel (0.2 à 0.7). Cependant, à partir du point 0.6, l'approche hybride a légèrement surpassé les approches probabiliste et possibiliste. Elle a également obtenu une performance très proche de la discriminative dans certains points de rappel de haut-niveau (0.7 à 1.0) (figure 5.8).

Pour les requêtes sources utilisant «*Description*», qui fournissent quelques informations contextuelles, l'approche hybride est dotée d'une légère surperformance par rapport aux trois autres approches dans certains points de rappel de bas-niveau (0.1 et 0.2) et de haut-niveau (0.5 et 0.6). En partant du point de rappel 0.6, l'écart entre la monolingue et les autres approches a considérablement diminué notamment entre les points 0.9 et 1 (figure 5.9).

Enfin, si nous nous concentrons sur les parties «*Narration*» des requêtes sources, le contexte est plus large et donc plus approprié pour l'identification des syntagmes nominaux. Dans ce cas, l'approche hybride est légèrement inférieure à l'approche probabiliste et l'approche possibiliste, en particulier dans certains points de rappel de bas-niveau (0 à 0.2) et de haut-niveau (0.7 à 1.0), mais elle est supérieure à l'approche discriminative dans la plupart des points de rappel. De plus, l'approche hybride semble meilleure que les trois autres approches dans certains points de rappel tels que 0.2, 0.3 et 0.6. L'écart entre ces approches est réduit à partir du point de rappel 0.8 (figure 5.10). D'autre part, la monolingue surpasse toutes ses concurrentes dans la plupart des points de rappel pour la même raison indiquée précédemment : nous n'avons pas impliqué dans nos expériences un processus d'expansion de requêtes avant et/ou après le processus de traduction.

Les résultats fournis par les courbes de rappel-précision dépendent de la capacité de la requête source à fournir un maximum d'information contextuelle. Pour ces raisons, l'approche hybride semble plus confortable en utilisant des requêtes longues, où le contexte est important. Cependant, l'approche discriminative a montré son efficacité dans le cas des requêtes courtes (utilisant «*Titre*»), où le contexte est limité à un petit ensemble des termes dans lesquels l'identification des syntagmes nominaux n'est pas fréquente.

5.3.2 Les points de précision aux top documents, les métriques MAP et R-Précision

Nous présentons dans cette section la performance globale des approches étudiées en utilisant les valeurs de précision aux top documents, ainsi que les métriques *MAP* et *R-Précision* afin de confirmer davantage les observations précédentes.

Cette section est divisée en trois parties. La première est basée sur la comparaison de l'approche possibiliste à base de transformation à l'approche probabiliste (cf. section 5.3.2.1) (Elayeb et al., 2018). La deuxième est basée sur la comparaison de l'approche discriminative à l'approche probabiliste (cf. section 5.3.2.2) (Ben Romdhane et al., 2017). La troisième est basée sur la comparaison de l'approche hybride avec les approches possibiliste, discriminative et probabiliste (cf. section 5.3.2.3) (Ben Romdhane et al., 2019).

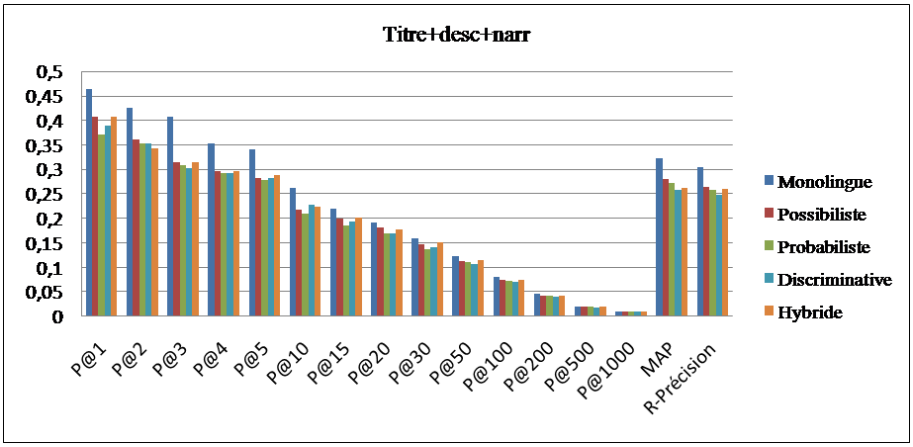

Figure 5.11 – Les valeurs de précision aux top documents, *MAP* et *R-Précision* utilisant « *Titre+desc+narr* »

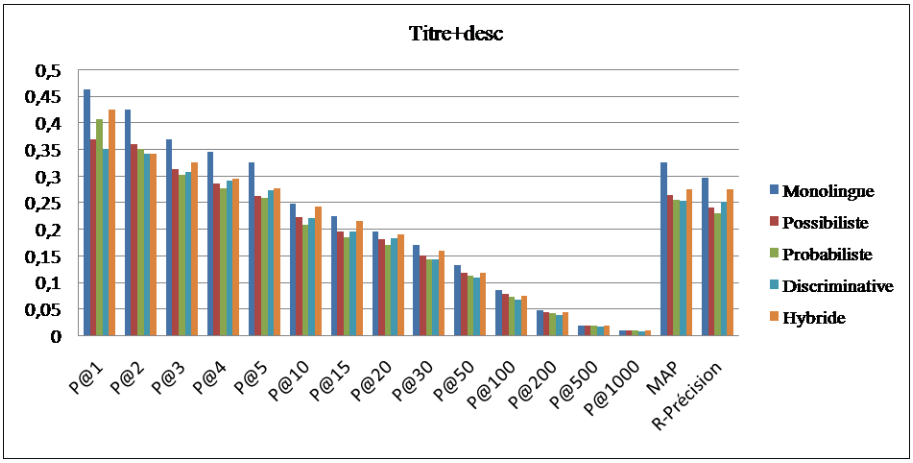

Figure 5.12 – Les valeurs de précision aux top documents, *MAP* et *R-Précision* utilisant « *Titre+desc* »

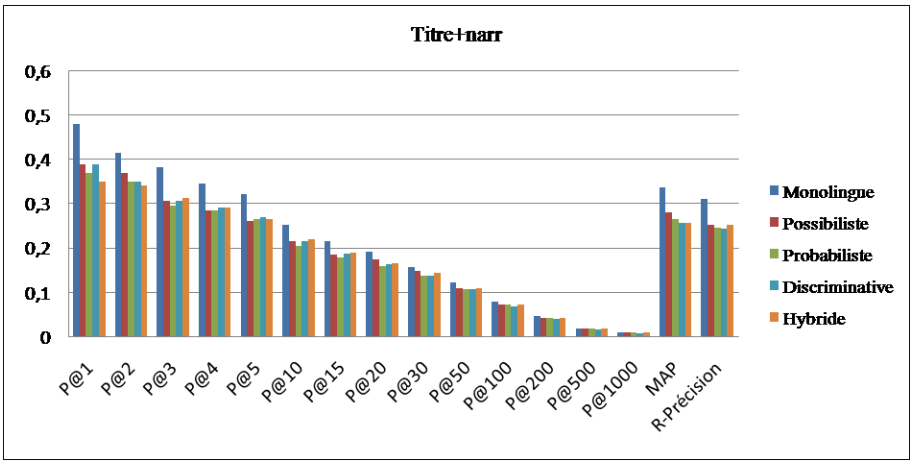

Figure 5.13 – Les valeurs de précision aux top documents, *MAP* et *R-Précision* utilisant « *Titre+narr* »

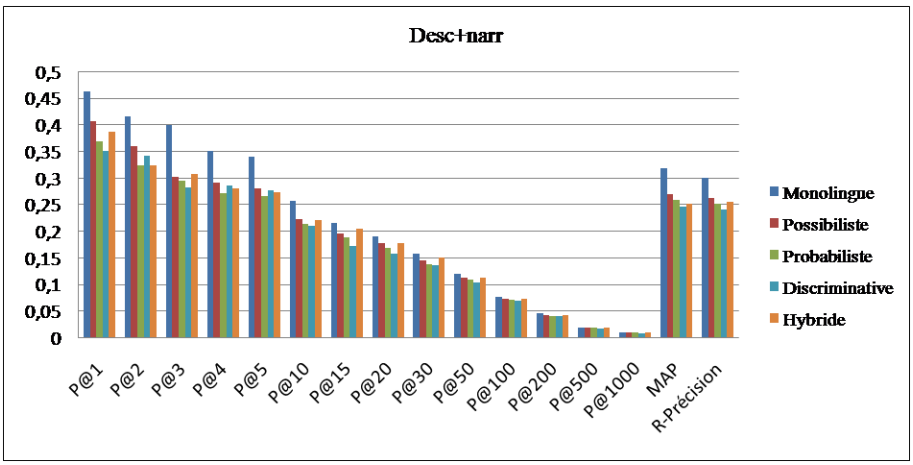

Figure 5.14 – Les valeurs de précision aux top documents, *MAP* et *R-Précision* utilisant « *Desc+narr* »

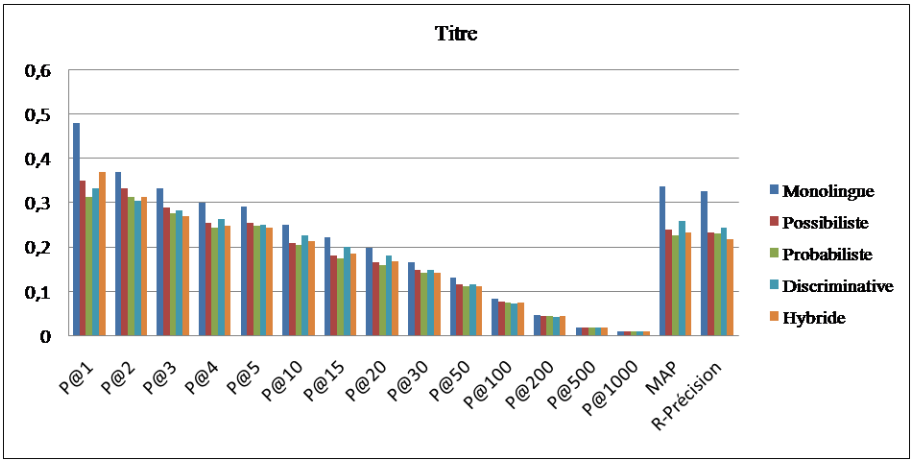

Figure 5.15 – Les valeurs de précision aux top documents, *MAP* et *R-Précision* utilisant « *Titre* »

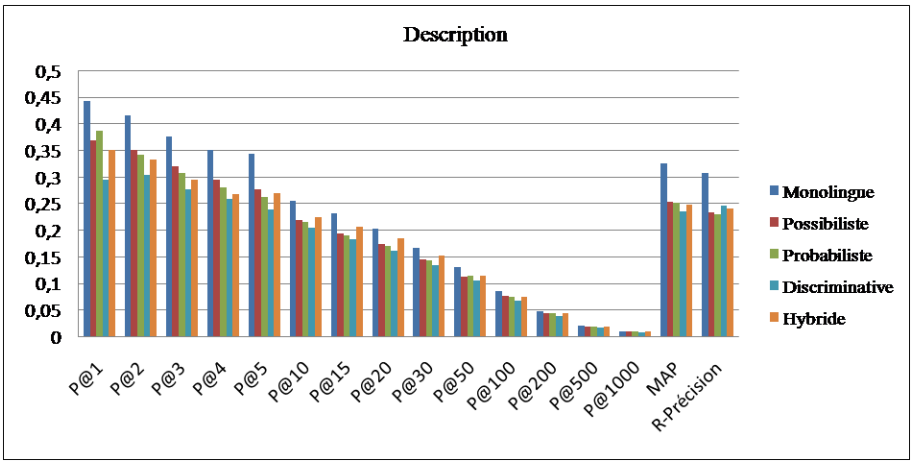

Figure 5.16 – Les valeurs de précision aux top documents, *MAP* et *R-Précision* utilisant « *Description* »

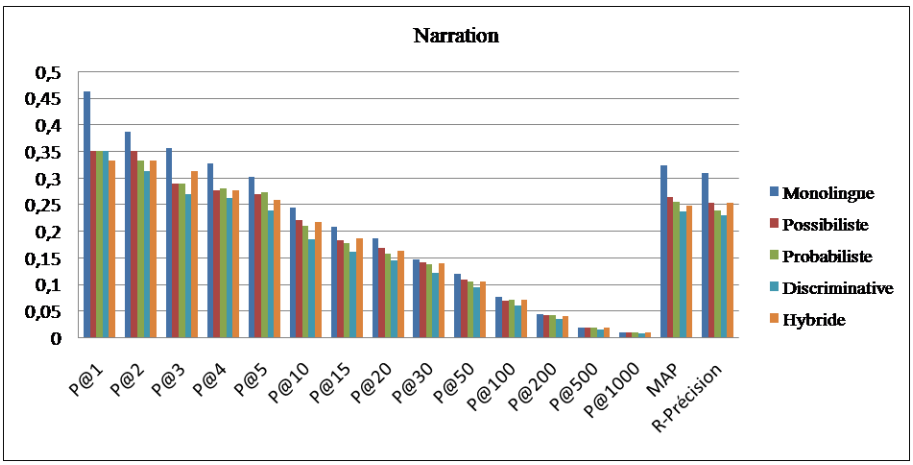

Figure 5.17 – Les valeurs de précision aux top documents, *MAP* et *R-Précision* utilisant « *Narration* »

5.3.2.1 Comparaison de l'approche possibiliste à base de transformation à l'approche probabiliste

De manière générale, la précision diminue lorsque le nombre de documents retournés augmente. L'approche possibiliste est supérieure à l'approche probabiliste en termes de valeurs de précision aux top documents, sauf dans certains cas tels que $P@5$ en utilisant «*Titre+narr*» (figure 5.13) et $P@50$ en utilisant «*Description*» (figure 5.16) dans lesquels l'approche probabiliste dépasse légèrement l'approche possibiliste.

Tandis que, les métriques *MAP* et *R-Précision* de l'approche possibiliste sont toujours plus élevés que l'approche probabiliste. Dans le cas d'un grand nombre de documents retournés tel que $P@1000$, l'écart entre l'approche possibiliste et l'approche probabiliste semble plus étendu pour les requêtes longues et courtes, ce qui montre que l'approche possibiliste est plus efficace que l'approche probabiliste dans le cas des grandes collections. Malheureusement, la performance de monolingue est souvent plus supérieure. Une raison essentielle, est que nous n'avons pas impliqué un processus d'expansion de requêtes dans nos expériences avant et/ou après le processus de traduction.

Globalement, notre approche possibiliste est supérieure à l'approche probabiliste pour les requêtes courtes (*Titre*, *Description*, *Narration*) et les requêtes longues (*Titre+desc+narr*, *Titre+desc*, *Titre+narr*, *Desc+narr*). Elle est trop proche de monolingue dans le cas des requêtes longues lorsque le contexte est plus adapté pour identifier et traduire le maximum de syntagmes nominaux avant de passer au processus de traduction des mots simples. Nous notons également que la performance générale de notre approche possibiliste est souvent meilleure par rapport à l'approche probabiliste avec un écart clair pour les premières valeurs de rappel correspondantes aux premiers documents sélectionnés.

5.3.2.2 Comparaison de l'approche discriminative à l'approche probabiliste

Pour confirmer davantage les conclusions indiquées précédemment, nous fournissons une deuxième étude comparative entre les approches discriminative, probabiliste, et monolingue en utilisant les valeurs de précision aux top documents ainsi que les métriques *MAP* et *R-Précision* (figure 5.11).

Lorsque nous utilisons les requêtes longues, l'approche discriminative a surpassé l'approche probabiliste en terme de précision aux top documents retournés ($P@5$, $P@10$, $P@15$ et $P@20$) sauf dans certains cas tels que $P@10$, $P@15$ et $P@20$ en uti-

lisant «*Desc+narr*» (figure 5.14) et P@20 en utilisant «*Titre+desc+narr*» (figure 5.11) dans lesquels l'approche probabiliste dépasse légèrement la discriminative.

Cependant, si nous exécutions les requêtes courtes, en utilisant «*Description*» ou «*Narration*», le contexte est plus adapté pour l'approche probabiliste qui surpasse la discriminative en terme de précision sur la majorité des valeurs de précision aux top documents (figures 5.16 et 5.17). Les requêtes courtes en utilisant uniquement «*Titre*» sont plus efficaces pour l'approche discriminative pour la majorité des valeurs de précision aux top documents, à l'exception dans $P@100$ et $P@1000$, où le nombre de documents retournés est important, ce qui augmente le bruit dans les résultats récupérés (figure 5.15).

Pour les requêtes longues et courtes, les écarts entre l'approche discriminative et l'approche probabiliste sont plus étendus lorsque nous utilisons un grand nombre de documents retournés. Par conséquent, l'approche probabiliste est plus appropriée en cas de grande collection en utilisant des requêtes longues. Tandis que, l'approche discriminative est plus efficace en cas des requêtes courtes en utilisant «*Titre*» (figure 5.15) ou «*Description*» (figure 5.16). Globalement, la performance de l'approche discriminative est souvent meilleure que celle de probabiliste avec un écart clair pour les premières valeurs de rappel correspondantes aux premiers documents sélectionnés.

Par ailleurs, la monolingue est la limite supérieure de performance de la RIT dans toutes nos expériences, car nous n'impliquons pas dans nos tests une étape d'expansion de requête avant et/ou après le processus de traduction.

Pour les requêtes longues, l'approche probabiliste est supérieure à l'approche discriminative, à l'exception dans les requêtes «*Titre+desc*» (figure 5.12) dans lesquelles l'approche discriminative semble meilleure que l'approche probabiliste. Cependant, les métriques *MAP* et *R-Précision* confirment que les requêtes courtes en utilisant «*Titre*», restent mieux adaptées à l'approche discriminative que l'approche probabiliste.

5.3.2.3 Comparaison de l'approche hybride aux approches possibiliste, discriminative et probabiliste

L'objectif dans cette section est de confirmer nos conclusions précédemment déduites pour l'approche hybride par rapport aux approches possibiliste, discriminative et probabiliste. Nous utilisons, cette fois, dans notre étude comparative les valeurs de précision aux top documents, *MAP* et *R-Précision*.

Pour les requêtes longues (figures 5.11, 5.12, 5.13 et 5.14), l'approche hybride surpasse les deux approches probabiliste et discriminative, dans les premières valeurs

de rappel qui correspondent aux premiers documents sélectionnés, sauf dans certains cas tels que :

- Pour $P@1000$ en utilisant «*Titre+narr*», l'approche probabiliste est légèrement meilleure que l'hybride.
- L'approche discriminative surpasse l'approche hybride dans $P@10$ en utilisant «*Titre+desc+narr*» et dans $P@5$ en utilisant «*Titre+narr*» ou «*Desc+narr*».
- L'approche possibiliste surpasse l'approche hybride dans $P@20$ et $P@1000$ en utilisant «*Titre+desc+narr*», dans $P@100$ et $P@1000$ en utilisant «*Titre+desc*», dans $P@20$, $P@30$, $P@50$ et $P@1000$ en utilisant «*Titre+narr*», et dans $P@5$, $P@10$ et $P@1000$ en utilisant «*Desc+narr*».

Si nous nous limitons aux requêtes courtes (figures 5.15, 5.16, et 5.17), nous remarquons que l'approche hybride semble meilleure que l'approche probabiliste ainsi que l'approche discriminative en utilisant les premières valeurs du rappel qui correspondent aux premiers documents sélectionnés, sauf dans certains cas tels que :

- L'approche probabiliste surpasse l'approche hybride dans $P@5$ et $P@100$ en utilisant «*Titre*», $P@100$ en utilisant «*Description*» et dans $P@5$, $P@100$ et $P@1000$ en utilisant «*Narration*».
- L'approche discriminative semble meilleure que l'approche hybride dans $P@5$, $P@10$, $P@15$, $P@20$, $P@30$ et $P@50$ en utilisant «*Titre*».
- L'approche hybride n'a pas pu surpasser l'approche possibiliste en termes de précision aux top documents sélectionnés en utilisant «*Titre*». En fait, l'approche hybride a obtenu des bons résultats uniquement dans $P@10$, $P@15$ et $P@20$, en plus de $P@15$ et $P@100$ en utilisant «*Narration*». Cependant, l'approche possibiliste a obtenu de meilleurs résultats que l'approche hybride, en utilisant «*Description*», uniquement dans $P@5$ et $P@100$.

Globalement, les valeurs de précision aux top documents retournés confirment que l'approche hybride est principalement supérieure aux approches possibiliste, probabiliste et discriminative en utilisant des requêtes courtes et longues avec un écart clair pour les premières valeurs de rappel correspondantes aux premiers documents sélectionnés. Toutefois, l'approche discriminative surpasse l'approche hybride en cas des requêtes courtes en utilisant «*Titre*», parce que cette dernière est plus appropriée lorsque le nombre des termes simples est fréquent. D'autre part, l'approche possibiliste a obtenu des meilleurs résultats dans le cas des requêtes courtes utilisant «*Narration*». En fait, la transformation probabilité-possibilité a donné plus de tolérance dans la traduction des termes simples restants de la requête source, au contraire des scores de traduction calculés par l'approche discriminative via les pertinences possible et nécessaire.

Par ailleurs, nous nous focalisons dans la suite sur les deux métriques *MAP* et *R-Précision* pour comparer ces approches. Pour les requêtes longues, l'approche

hybride surpasse l'approche probabiliste en terme de *R-Précision* et de *MAP* uniquement dans les requêtes utilisant «*Titre+desc*». De plus, l'approche hybride a obtenu des meilleurs résultats que l'approche discriminative en termes de *MAP* et de *R-Précision* dans toutes les combinaisons possibles des requêtes longues, sauf dans le cas de *MAP* pour les requêtes utilisant «*Titre+narr*» où l'approche discriminative dépasse légèrement l'approche hybride. Cette dernière semble meilleure que l'approche possibiliste en terme de *MAP* et de *R-Précision* pour les requêtes utilisant «*Titre+desc*».

En général, ces deux métriques confirment que les requêtes courtes en utilisant «*Titre*» sont encore mieux adaptées à l'approche discriminative par rapport à toutes les autres approches. Tandis que, les parties «*Narration*» des requêtes sources sont plus appropriées pour l'approche possibiliste. Pour ces raisons, l'approche hybride a bénéficié à la fois des points forts de l'approche discriminative ainsi que les atouts de l'approche possibiliste. Cela est confirmé par l'efficacité de l'approche hybride en cas des requêtes longues en utilisant «*Titre+desc*», mais elle a encore des lacunes en terme de *MAP* et de *R-Précision* par rapport à ses approches concurrentes.

5.3.3 Le pourcentage d'amélioration

Pour comparer davantage nos approches possibilistes, nous calculons leurs pourcentages d'amélioration, en utilisant les valeurs de précision aux top documents et les métriques *MAP* et *R-Précision*.

Nous divisons cette section en trois parties : tout d'abord, nous étudions l'amélioration de l'approche possibiliste à base de transformation par rapport à l'approche probabiliste (cf. section 5.3.3.1) (Elayeb et al., 2018). Ensuite, l'amélioration de l'approche discriminative par rapport à l'approche probabiliste (cf. section 5.3.3.2) (Ben Romdhane et al., 2017). Enfin, l'amélioration de l'approche hybride par rapport aux approches possibiliste, discriminative et probabiliste (cf. section 5.3.3.3) (Ben Romdhane et al., 2019).

5.3.3.1 L'amélioration de l'approche possibiliste à base de transformation par rapport à l'approche probabiliste

Nous constatons une claire amélioration des performances de l'approche possibiliste pour les top documents retournés au début de la liste (tableau 5.1). Compte tenu de ces résultats, les valeurs aux points de précision $P@5, \ldots, P@30$ enregistrées par la possibiliste sont les plus élevées, sauf dans des cas tels que $P@5$ en utilisant «*Titre+narr*» et $P@5$ en utilisant «*Narration*».

Nous avons obtenu une amélioration de plus de 7.5% pour $P@30$ pour les requêtes longues en utilisant «*Titre+desc+narr*». La moyenne des pourcentages d'amélioration aux points de précision aux différents top documents est d'environ 4.3% pour les requêtes longues en utilisant «*Titre+desc+narr*». En outre, pour les requêtes longues, la moyenne des pourcentages d'amélioration du MAP est d'environ 3.87% et la *R-Précision* est d'environ 3.33%.

Pour les requêtes courtes, nous avons noté une amélioration d'environ 7% pour $P@20$ en utilisant «*Narration*», plus de 5.6% pour $P@5$ en utilisant «*Description*» et plus de 4.6% pour $P@20$ en utilisant «*Titre*». La moyenne des pourcentages d'améliorations du MAP est d'environ 3.11% et de *R-Précision* est d'environ 2.83%. Cependant, dans certains cas, le pourcentage d'amélioration est nul ($P@1000$ en utilisant «*Description*»). Ces résultats confirment l'efficacité de l'approche possibiliste par rapport à l'approche probabiliste pour les requêtes longues et courtes en exploitant différents scénarios et métriques d'évaluation (Elayeb et al., 2018).

Requêtes longues	**Valeurs de précision**	**Poss vs. Proba**	**Requêtes courtes**	**Valeurs de précision**	**Poss vs. Proba**
Titre + Desc + narr	P@5	1.33	**Titre**	P@5	3.02
	P@10	4.4		P@10	2.68
	P@15	6.65		P@15	3.54
	P@20	7.67		P@20	4.62
	P@30	7.55		P@30	3.91
	P@50	3		P@50	4.01
	P@100	1.79		P@100	2.88
	P@1000	1.98		P@1000	3
	MAP	2.67		MAP	5.4
	R-Précision	1.7		R-Précision	0.82
Titre + desc	P@5	1.43	**Description (desc)**	P@5	5.63
	P@10	7.07		P@10	1.71
	P@15	5.95		P@15	1.93
	P@20	6.48		P@20	2.15
	P@30	4.69		P@30	1.74
	P@50	4.59		P@50	-0.61
	P@100	6.7		P@100	3.31
	P@1000	0.97		P@1000	0
	MAP	3.51		MAP	0.47
	R-Précision	4.82		R-Précision	1.73
Titre + narr	P@5	-1.39	**Narrative (narr)**	P@5	-1.35
	P@10	5.4		P@10	5.26
	P@15	2.77		P@15	2.79
	P@20	8.63		P@20	6.97
	P@30	7.13		P@30	3.54
	P@50	2.01		P@50	2.43
	P@100	0		P@100	-2.22
	P@1000	0.98		P@1000	0.98
	MAP	5.79		MAP	3.48
	R-Précision	2.27		R-Précision	5.95
Desc + narr	P@5	5.55			
	P@10	4.33			
	P@15	3.92			
	P@20	5.49			
	P@30	6.22			
	P@50	2.73			
	P@100	1.24			
	P@1000	1.98			
	MAP	3.53			
	R-Précision	4.55			

Table 5.1 – Le pourcentage d'amélioration de l'approche possibiliste par rapport à l'approche probabiliste

5.3.3.2 L'amélioration de l'approche discriminative par rapport à l'approche probabiliste

Nous présentons dans le tableau 5.2 les pourcentages d'amélioration de l'approche discriminative par rapport à l'approche probabiliste pour les requêtes longues et courtes et en utilisant les valeurs de précision aux top documents, la *MAP* et la *R-Précision*.

L'approche discriminative a réalisé une amélioration significative en termes de précision aux documents retournés au début de la liste dans le cas de requête à base de «*Titre*». Par exemple, nous avons enregistré un pourcentage d'amélioration supérieur à 15% pour *P*@15 et *P*@20 en utilisant «*Titre*» et plus de 7% pour *P*@5, *P*@10 et *P*@20 en utilisant «*Titre+desc*». Pour les requêtes longues, la moyenne d'amélioration est d'environ 2.5% si nous considérons les 5, 10, 15 et 20 premiers documents retournés. Mais, pour les requêtes courtes, en utilisant uniquement «*Titre*», la moyenne d'amélioration est d'environ 12.23% pour les mêmes documents retournés.

Si nous nous concentrons sur la métrique *MAP*, la moyenne d'amélioration pour les requêtes longues est d'environ−3.26%, alors qu'elle est d'environ 1.03% pour les requêtes courtes. Néanmoins, la moyenne d'amélioration de la métrique *R-Précision* pour les requêtes courtes est d'environ 3.16%, tandis qu'elle est d'environ 0.15% pour les requêtes longues.

Ces résultats confirment de nouveau nos constats conclus précédemment sur l'efficacité de l'approche discriminative dans la traduction de requêtes courtes. Réellement, ce genre de requête courte est fréquent, car l'utilisateur propose dans la majorité des cas un nombre limité des termes dans sa requête source, en raison de ses lacunes linguistiques ou de ses connaissances limitées sur son domaine de recherche (Ben Romdhane et al., 2017).

5.3.3.3 L'amélioration de l'approche hybride par rapport aux approches possibiliste, discriminative et probabiliste

Nous présentons dans le tableau 5.3 le pourcentage d'amélioration de l'approche hybride par rapport aux approches possibiliste, discriminative et probabiliste. À l'aide des requêtes longues, l'approche hybride effectue une amélioration significative en termes de précision aux top documents. Par exemple, si nous comparons l'approche hybride avec l'approche probabiliste, nous avons obtenu un pourcentage d'amélioration de l'approche hybride supérieur à 16% pour *P*@10 et *P*@15 en utilisant «*Titre+desc*» et plus de 10% pour *P*@30 en utilisant «*Titre+desc+narr*».

Requêtes longues	Valeurs de précision	Discri vs. Proba	Requêtes courtes	Valeurs de précision	Discri vs. Proba
Titre + Desc + narr	P@5	1.33	**Titre**	P@5	6.01
	P@10	7.93		P@10	11.67
	P@15	2.68		P@15	16.2
	P@20	-1.06		P@20	15.04
	P@30	2.69		P@30	5.59
	P@50	-1.36		P@50	3.65
	P@100	-3.02		P@100	-3.66
	P@1000	-5.94		P@1000	0
	MAP	-5.49		MAP	15.94
	R-Précision	-4.13		R-Précision	5.97
Titre + desc	P@5	7.13	**Description (desc)**	P@5	-7.07
	P@10	7.93		P@10	-3.41
	P@15	6.65		P@15	-3.24
	P@20	8.11		P@20	-4.82
	P@30	0		P@30	-5.15
	P@50	-3.27		P@50	-7.06
	P@100	-7.91		P@100	-10.05
	P@1000	-3.88		P@1000	-6.73
	MAP	0.39		MAP	-5.77
	R-Précision	10.02		R-Précision	7.35
Titre + narr	P@5	1.39	**Narrative (narr)**	P@5	-12.19
	P@10	4.47		P@10	-13.17
	P@15	4.16		P@15	-9.66
	P@20	2.86		P@20	-8.73
	P@30	0		P@30	-10.7
	P@50	0.27		P@50	-10.37
	P@100	-4.22		P@100	-13.85
	P@1000	-5.88		P@1000	-9.8
	MAP	-2.74		MAP	-7.08
	R-Précision	-1.17		R-Précision	-3.83
Desc + narr	P@5	4.16			
	P@10	-2.56			
	P@15	-8.52			
	P@20	-6.55			
	P@30	-1.37			
	P@50	-4.73			
	P@100	-4.28			
	P@1000	-5.94			
	MAP	-5.22			
	R-Précision	-4.12			

Table 5.2 – Le pourcentage d'amélioration de l'approche discriminative par rapport à l'approche probabiliste

En outre, la moyenne d'amélioration est d'environ 9% si nous considérons les valeurs

de précision aux top documents retournés en utilisant «*Titre+desc*», et la moyenne d'amélioration de *R-Précision* est d'environ 6%.

Cependant, si nous comparons l'approche hybride avec l'approche discriminative en utilisant «*Titre+desc*», nous avons obtenu un pourcentage d'amélioration supérieur à 12% pour P@30, une amélioration moyenne d'environ 7.75% pour les valeurs de précision aux top documents. L'amélioration moyenne de *MAP* et de *R-Précision* sont respectivement 3% et 5.75%.

Par ailleurs, si nous comparons l'approche hybride avec l'approche possibiliste en utilisant «*Titre+desc*», nous avons noté un pourcentage d'amélioration dépassant le 10% pour P@15, une amélioration moyenne d'environ 4% pour les valeurs de précision aux top documents et une amélioration moyenne de *R-Précision* d'environ 2.6%.

Si nous nous concentrons sur la comparaison de l'approche hybride avec l'approche probabiliste, nous avons enregistré des pourcentages d'amélioration encourageants pour les requêtes courtes. Par exemple, dans P@15 nous avons des améliorations de plus de 6% (*Titre*), plus de 8.3% (*Description*) et plus de 4.8% (*Narration*). Pour les valeurs de précision aux top documents, la moyenne de pourcentage d'amélioration est d'environ 2% en utilisant «*Titre*», d'environ 3.7% en utilisant «*Description*» et d'environ 0.75% en utilisant «*Narration*». L'amélioration moyenne de *R-Précision* est d'environ 1.8%.

Si nous comparons l'approche hybride avec l'approche discriminative, nous remarquons que le meilleur pourcentage d'amélioration est en P@100 en utilisant «*Titre*» avec 1.9%, alors que plus de 14.3% dans P@20 en utilisant «*Description*» et d'environ 18% dans P@10 en utilisant «*Narration*». La moyenne de pourcentage d'amélioration des valeurs de précision aux top documents est d'environ 11% en utilisant «*Description*» et de 13.4% en utilisant «*Narration*».

Globalement, nous remarquons que l'approche hybride est meilleure que l'approche possibiliste en utilisant la partie «*Description*». Elle dépasse 6% de pourcentage d'amélioration dans P@15. Tandis que, la moyenne de pourcentage d'amélioration est d'environ 1.7% pour les valeurs de précision aux top documents retournés.

Ces résultats confirment davantage nos conclusions à propos l'efficacité de l'approche hybride dans les cas des requêtes longues et courtes.

Requêtes longues	**Valeurs de précision**	**Hybride vs. Poss**	**Hybride vs. Discri**	**Hybrid vs. Proba**	**Requêtes courtes**	**Valeurs de précision**	**Hybride vs. Poss**	**Hybride vs. Discri**	**Hybrid vs. Proba**
Titre+desc +narr	P@5	2.63	2.63	4	**Titre**	P@5	-4.38	-2.98	-1.49
	P@10	2.56	-1.62	7.07		P@10	1.75	-5.71	4.47
	P@15	1.86	5.14	8.64		P@15	2.7	-7.95	6.33
	P@20	-2.52	5.52	4.96		P@20	0.54	-7.62	5.18
	P@30	2.57	6.97	10.31		P@30	-3.36	-4.13	0.42
	P@50	0.97	6.12	4		P@50	-3.17	-3.17	0.71
	P@100	1.21	5.78	3.02		P@100	-3.81	1.88	-1.05
	P@1000	-1.94	6.32	0		P@1000	-1.94	1	1
	MAP	-6.85	1.36	-4.36		MAP	-2.71	-10.36	2.55
	R-Précision	-1.21	4.62	0.46		R-Précision	-5.87	-10.59	-5.1
Titre+desc	P@5	5.63	1.35	7.13	**Description**	P@5	-2.66	12.34	2.81
	P@10	9.06	9.99	16.77		P@10	2.5	9.87	4.25
	P@15	10.03	10.03	16.58		P@15	6.3	12.72	8.36
	P@20	5.1	4.58	11.91		P@20	5.29	14.32	7.55
	P@30	6.06	12.03	11.03		P@30	5.06	13.68	6.88
	P@50	0.34	8.78	4.94		P@50	0.61	8	0
	P@100	-4.4	11.09	2.01		P@100	-3.46	11.21	-0.26
	P@1000	-0.96	4.04	0		P@1000	0	7.22	0
	MAP	4.26	8.39	7.93		MAP	-1.85	5.58	-1.38
	R-Précision	14.11	8.84	19.61		R-Précision	2.97	-2.26	4.76
Titre+narr	P@5	1.41	-1.37	0	**Narrative**	P@5	-4.11	7.73	-5.4
	P@10	1.71	1.71	7.2		P@10	-1.67	17.98	3.51
	P@15	2.65	1.28	5.49		P@15	2.01	16.08	4.86
	P@20	-4.74	1.15	3.48		P@20	-3.29	13.34	3.45
	P@30	-2.96	4.41	3.96		P@30	-1.26	15.15	2.24
	P@50	-0.72	1.65	1.28		P@50	-2.01	12.82	0.37
	P@100	0.82	5.56	0.82		P@100	1.27	15.51	-0.97
	P@1000	-1.94	5.21	-0.98		P@1000	-2.91	8.7	-1.96
	MAP	-8.45	-0.19	-3.16		MAP	-6.01	4.8	-2.74
	R-Précision	0.55	3.97	2.84		R-Précision	-0.16	9.9	5.78
Desc+narr	P@5	-2.63	-1.33	2.77					
	P@10	-0.85	5.26	3.45					
	P@15	4.38	17.69	8.47					
	P@20	0.5	12.74	6.02					
	P@30	3.4	10.88	9.83					
	P@50	1.24	9.89	4					
	P@100	1.5	6.9	2.76					
	P@1000	-1.94	6.32	0					
	MAP	-6.04	2.8	-2.72					
	R-Précision	-2.99	5.61	1.43					

Table 5.3 – Le pourcentage d'amélioration de l'approche hybride par rapport aux approches possibiliste, discriminative et probabiliste

5.3.4 Le test de significativité statistique de *Wilcoxon*

Le test des rangs signés de *Wilcoxon*, pour deux échantillons appariés (*Wilcoxon Matched-Pairs Signed-Ranks*) (Hull, 1993) est une alternative non-paramétrique au *t-test* apparié qui permet de décider si l'amélioration de l'approche 1 par rapport à l'approche 2 est significative ou non, tout en se basant sur plusieurs mesures d'évaluation. En effet, le *t-test* calcule la $p-valeur$ qui est basée sur les données de performance de 1 et 2. Généralement, l'amélioration est statistiquement significative si la $p-valeur$ est assez petite ($p-valeur < 0.05$).

Nous étudions, dans la section 5.3.4.1, la significativité statistique de l'approche possibiliste à base de transformation par rapport à l'approche probabiliste (Elayeb et al., 2018). Dans la section 5.3.4.2, la significativité statistique de l'approche discriminative par rapport à l'approche probabiliste (Ben Romdhane et al., 2017). Enfin, dans section 5.3.4.3, la significativité statistique de l'approche hybride par rapport aux approches possibiliste, discriminative et probabiliste (Ben Romdhane et al., 2019).

5.3.4.1 Possibiliste à base de transformation vs. Probabiliste

Le tableau 5.4 montre que l'amélioration des valeurs de précision aux top documents de l'approche possibiliste est plus significative dans les points de précision $P@1000$ ($p-valeur = 0.0001$) et au point de précision $P@10$ ($p-valeur = 0.013$). En général, l'amélioration de l'approche possibiliste est statistiquement significative puisque la $p-valeur < 0.05$ pour les différents points de précision, MAP et *R-Précision*, sauf dans $P@100$ dans laquelle $p-valeur = 0.115 > 0.05$. Ceci est dû au faible pourcentage d'amélioration de $P@100$ dans certains cas, par exemple 0% pour $P@100$ en utilisant «*Titre+narr*», et -2.22% pour $P@100$ en utilisant «*Narration*» (figure 5.17). L'amélioration obtenue est statistiquement significative pour les requêtes courtes, par exemple $p-valeur = 0.005$ en utilisant «*Titre*», et pour les requêtes longues, par exemple $p-valeur = 0.002$ en utilisant «*Titre+desc*».

	P@5	**P@10**	**P@15**	**P@20**	**P@30**	**P@50**	**P@100**	**P@1000**	**MAP**	**R-Précision**
Poss vs. Proba	0.037	0.013	0.017	0.017	0.017	0.027	0.115	0.0001	0.017	0.017
	Requêtes longues					**Requêtes courtes**				
	Titre+desc+narr		**Titre + desc**	**Titre + narr**	**Desc + narr**	**Titre**	**desc**		**narr**	
	0.005		0.002	0.020	0.004	0.005	0.009		0.028	

Table 5.4 – Le test de significativité statistique de *Wilcoxon* de Possibiliste vs. Probabiliste

Ces résultats confirment de nouveau la performance de notre approche possibiliste par rapport à l'approche probabiliste en utilisant différentes mesures d'évaluation pour les requêtes courtes et longues.

5.3.4.2 Discriminative vs. Probabiliste

Dans le but de prouver que les résultats de l'approche discriminative sont statistiquement significatifs par rapport à sa concurrente probabiliste, nous calculons les $p-valeurs$ associées au test de *Wilcoxon* (tableau 5.5). Nous constatons que les $p-valeurs$ de l'approche discriminative ne sont pas significatives ($p-valeur > 0.05$) dans la plupart des cas, à l'exception de $P@100$ et $P@1000$, où l'approche probabiliste dépasse légèrement l'approche discriminative.

L'approche probabiliste semble être meilleure que la discriminative pour les requêtes courtes en utilisant «*Narration*» ($p - valeur = 0.005 < 0.05$), ainsi que pour les requêtes longues utilisant «*Desc+narr*» ($p-valeur = 0.036 < 0.05$) ; où le contexte est plus riche.

Par ailleurs, au niveau des requêtes courtes, l'amélioration de l'approche discriminative est statistiquement significative, en particulier pour «*Titre*» ($p - valeur = 0.010 < 0.05$). Cette conclusion est valable aussi pour «*Description*» ($p-valeur \cong 0.05$) ; grâce à une amélioration de 7.35% de la *R-Précision*.

	P@5	**P@10**	**P@15**	**P@20**	**P@30**	**P@50**	**P@100**	**P@1000**	**MAP**	**R-Précision**
Discr	0.865	0.479	0.865	0.865	0.685	0.128	0.013	0.0013	0.310	0.498
vs.	**Requêtes longues**					**Requêtes courtes**				
Proba	**Titre+desc+narr**		**Titre + desc**	**Titre + narr**	**Desc + narr**	**Titre**	**desc**		**narr**	
	0.959		0.085	0.441	0.036	0.010	0.055		0.005	

Table 5.5 – Le test de significativité statistique de *Wilcoxon* de Discriminative vs. Probabiliste

5.3.4.3 Significativité statistique de l'approche hybride

Dans le but de comparer davantage l'approche hybride avec les approches proposées, nous étudions, dans cette section, les mesures de $p - valeurs$ associées au test de significativité statistique de *Wilcoxon*.

Comme le montre les tableaux 5.6, 5.7 et 5.8, l'amélioration de l'approche hybride par rapport à l'approche possibiliste est statistiquement significative pour les valeurs de précision aux top documents, les deux scores de *MAP* et de *R-Précision*. Par exemple, dans $P@10$ ($p - valeur = 0.037 < 0.05$), dans $P@15$

($p - valeur = 0.009 < 0.05$). Ceci restera valable pour les requêtes longues en utilisant «*Titre+desc*» ($p - valeur = 0.016 < 0.05$).

D'une manière générale, l'amélioration de l'approche hybride par rapport à l'approche discriminative est statistiquement significative. En fait, dans $P@30$ ($p - valeur = 0.042 < 0.05$), dans $P@50$ ($p - valeur = 0.042 < 0.05$), dans $P@100$ ($p - valeur = 0.013 < 0.05$) et dans $P@1000$ ($p - valeur = 0.013 < 0.05$).

Globalement, les résultats de l'approche hybride sont statistiquement significatifs dans les cas des requêtes courtes et pour toutes les combinaisons des requêtes longues. Néanmoins, l'approche discriminative est statistiquement significative par rapport à l'approche hybride dans les requêtes courtes utilisant «*Titre*» ($p - valeur = 0.012 < 0.05$).

Enfin, l'amélioration de l'approche hybride, par rapport à l'approche probabiliste, est statistiquement significative dans $P@10$ ($p - valeur = 0.013 < 0.05$), dans $P@15$, $P@20$, et $P@30$ ($p - valeur = 0.017 < 0.05$) et dans $P@50$ ($p - valeur = 0.018 < 0.05$). Cette amélioration est aussi significative en utilisant toutes les combinaisons des requêtes longues, sauf les cas des requêtes basées sur «*Titre+narr*» ou «*Titre*» ou «*Narration*». Cette conclusion est toujours valable pour «*Titre+desc+narr*» ($p - valeur \cong 0.05$).

	P@5	**P@10**	**P@15**	**P@20**	**P@30**	**P@50**	**P@100**	**P@1000**	**MAP**	**R-Précision**
Hybride	0.479	0.037	0.009	0.723	0.330	0.735	0.479	1.00	0.062	0.865
vs.	**Requêtes longues**					**Requêtes courtes**				
Poss	**Titre+desc+narr**		**Titre + desc**	**Titre + narr**	**Desc + narr**	**Titre**	**desc**		**narr**	
	0.575		0.016	0.798	0.721	0.120	0.213		0.062	

Table 5.6 – Le test de significativité statistique de Wilcoxon de l'approche hybride vs. possibiliste

	P@5	**P@10**	**P@15**	**P@20**	**P@30**	**P@50**	**P@100**	**P@1000**	**MAP**	**R-Précision**
Hybride	0.297	0.132	0.062	0.090	0.042	0.042	0.013	0.013	0.310	0.310
vs.	**Requêtes longues**					**Requêtes courtes**				
Dicscri	**Titre+desc+narr**		**Titre + desc**	**Titre + narr**	**Desc + narr**	**Titre**	**desc**		**narr**	
	0.012		0.005	0.036	0.009	0.012	0.009		0.005	

Table 5.7 – Le test de significativité statistique de Wilcoxon de l'approche hybride vs. discriminative

	P@5	**P@10**	**P@15**	**P@20**	**P@30**	**P@50**	**P@100**	**P@1000**	**MAP**	**R-Précision**
Hybride	0.345	0.013	0.017	0.017	0.017	0.018	0.310	0.285	0.398	0.128
vs.	**Requêtes longues**					**Requêtes courtes**				
Proba	**Titre+desc+narr**		**Titre + desc**	**Titre + narr**	**Desc + narr**	**Titre**	**desc**		**narr**	
	0.050		0.007	0.085	0.025	0.351	0.035		0.444	

Table 5.8 – Le test de significativité statistique de Wilcoxon de l'approche hybride vs. probabiliste

5.4 Synthèse et discussion

Nous présentons, dans cette section, une synthèse des expériences effectuées sur nos approches de désambiguïsation de TR dans la RIT. Nous avons travaillé sur le couple de langue Français-Anglais en utilisant la collection de test du standard CLEF-2003 et le corpus parallèle Europarl.

Dans la première étude expérimentale nous avons comparé l'approche possibiliste à base de transformation probabilité-possibilité avec l'approche probabiliste *naïve bayésienne* (Elayeb et al., 2018). Les résultats globaux ont montré que l'approche possibiliste est légèrement supérieure à l'approche probabiliste. Les scénarios d'évaluation des requêtes «*Narration*» et «*Desc+narr*» montrent une avance en faveur de l'approche possibiliste par rapport à l'approche probabiliste. Ces requêtes fournissent des informations contextuelles qui favorisent l'identification et la traduction des syntagmes nominaux. Néanmoins, lorsque nous avons utilisé le modèle de langue Trigramme dans le corpus *Europarl*, nous constatons une rareté pour les syntagmes nominaux qui se composent de plus de quatre termes. Nos résultats sont influencés aussi par la structure des requêtes du standard de test CLEF-2003, qui sont parfois sous la forme d'une question telle que «*Qui a tué Yitzhak Rabin et pourquoi ?*» ou elles décrivent un besoin d'information tel que «*Trouvez des documents qui...*». Ces structures existent dans toutes les requêtes ce qui réduit l'efficacité de la RIT (Maisonnasse, 2008).

Dans la deuxième étude expérimentale, nous avons mis en évidence la performance de l'approche possibiliste discriminative par rapport à l'approche probabiliste (Ben Romdhane et al., 2017). En fait, l'approche discriminative est statistiquement meilleure que l'approche probabiliste en termes de précision aux top documents, de précision moyenne et de précision exacte en utilisant les requêtes courtes, surtout la partie «*Titre*». En dépit de ses avantages, l'approche discriminative souffre de quelques faiblesses, surtout au cas où les requêtes sont dans un domaine spécifique ; où les sujets du corpus *Europarl* sont différents des sujets de la collection de test

CLEF-2003. En conséquence, l'identification des traductions appropriées nécessite une traduction spécifique au domaine (Ture et Boschee, 2014) et un modèle de langue. Néanmoins, combiner un modèle de langue avec une traduction spécifique au domaine ainsi que leur intégration ne sont pas des tâches faciles.

Afin de pallier aux certaines des lacunes ci-dessus identifiés, nous avons proposé une approche possibiliste hybride (Ben Romdhane et al., 2019), combinant l'approche possibiliste à base de transformation probabilité-possibilité avec l'approche possibiliste discriminative. En fait, nous avons proposé une traduction des syntagmes nominaux via la technique de transformation probabilité-possibilité, alors que les mots simples restants des requêtes sont traduits via la technique de discrimination. Nous avons comparé cette approche aux approches possibiliste, discriminative et probabiliste. Les résultats globaux confirment que l'approche hybride est statistiquement meilleure que les autres approches. Néanmoins, l'approche discriminative reste la meilleure pour les requêtes courtes utilisant $\ll Titre \gg$.

Ces expérimentations constituent la première tentative de l'application de la théorie des possibilités pour la désambiguïsation de TR. Les approches ont été intégrées dans notre SRIT SPORT pour la sélection des meilleurs documents en anglais en réponse aux besoins des utilisateurs sous forme des requêtes françaises. Notre but est d'accomplir une amélioration dans la qualité de traduction en RIT.

Comme nous avons noté précédemment, SPORT diffère des SRIT étudiés dans l'état de l'art. Il est supporté par un processus de traduction automatique basé sur le calcul des distributions des possibilités.

En tenant compte du caractère générique de nos approches possibilistes et de leur indépendance de la langue, elles peuvent être extensibles à d'autres langues comme l'Arabe. Cependant, la désambiguïsation des traductions en arabe semble être le principal défi pour cette extension. Cela nécessite des dictionnaires exhaustifs bilingues de large couverture, un corpus parallèle comprenant des textes alignés en arabe, ainsi qu'un standard de test arabe.

Notons qu'une étude comparative objective entre nos approches proposées avec d'autres approches de désambiguïsation de TR en RIT, nécessite l'utilisation des mêmes ressources linguistiques (ensemble de données de *Glasgow*, *Europarl*, etc.) ainsi que les mêmes langues (Français-Anglais) que notre SRIT SPORT. Malheureusement, et à notre connaissance, il n'y a pas un SRIT en Français-Anglais (dans les campagnes CLEF ou ailleur) qui a été testé en utilisant les mêmes langues en RIT et les mêmes ressources linguistiques.

Cependant, (Nie, 1999) a suggéré une approche de désambiguïsation de TR en RIT testée sur le couple de langues Français-Anglais. Cette approche est la plus proche

du SPORT dans l'utilisation des ressources linguistiques. En effet, Nie a testé en 1998 un modèle de traduction probabiliste estimé à partir d'un corpus parallèle de traitement (*HANSARD* canadien) en utilisant le SRI *SMART*. La meilleure précision moyenne obtenue est de 0.3229. Bien qu'en 1999, l'auteur a obtenu la meilleure précision moyenne de 0.3748 pour les requêtes longues en utilisant des modèles probabilistes de traduction estimés à partir d'un corpus de traitement parallèle extrait du Web et des données de TREC-7.

Il est également pertinent d'évaluer l'impact de la désambiguïsation de TR avant et/ou après l'expansion de requête sur l'efficacité de RIT en utilisant des techniques récentes telles que (Ben Khiroun et al., 2014 ; Elayeb et al., 2015).

Conclusion

Dans le présent chapitre, nous avons exposé les résultats expérimentaux de nos approches possibilistes qui s'inscrivent dans le domaine de désambiguïsation de TR dans la RIT. Notre objectif est d'améliorer la performance globale de notre SRIT en renforçant la confiance envers la traduction de requêtes tout en garantissant une récupération des documents qui répondent aux besoins des utilisateurs.

Nous avons prouvé que l'approche possibiliste à base de transformation probabilité-possibilité a donné des bons résultats en utilisant les requêtes longues. Ces requêtes sont riches de contexte où nous pouvons extraire et traduire un nombre maximal de syntagmes nominaux. En outre, nos résultats montrent une avance en faveur de l'approche possibiliste discriminative par rapport à l'approche probabiliste dans le cas des requêtes courtes utilisant «*Titre*» ou «*Description*». Ces dernières sont caractérisées par une structure simple qui se compose de quelques mots dans lesquels l'identification des syntagmes nominaux semble rare. De plus, nous avons proposé une approche hybride afin de profiter des avantages conclus à partir des expérimentations réalisées des deux premières approches. L'approche hybride semble plus confortable par rapport à l'approche possibiliste à base de transformation en utilisant des requêtes longues. Cependant, l'approche discriminative est la meilleure dans les requêtes utilisant «*Titre*».

Nous avons utilisé plusieurs scénarios et métriques d'évaluation, à savoir : les courbes de *rappel-précision*, les valeurs de précision aux top documents, les métriques de *MAP* et *R-Précision*, le pourcentage d'amélioration et le test de significativité statistique de *Wilcoxon*. Les expérimentations menées montrent que les résultats sont encourageants et prouvent l'apport de la théorie des possibilités dans la désambiguïsation de TR.

Loin de la théorie des possibilités, nous proposons et testons, dans le prochain chapitre, une nouvelle approche de désambiguïsation de TR à base de proximité bilingue. Dans cette approche, nous estimons construire une nouvelle ressource linguistique supportant la couverture des ressources classiques utilisées en désambiguïsation de TR en RIT.

Chapitre

6 Proposition et validation d'une approche à base de la proximité bilingue pour la désambiguïsation des traductions de requêtes

Sommaire

Introduction

L'ambiguïté dans la traduction de requêtes est considérée comme un cas d'imprécision puisque la majorité des termes dans un dictionnaire bilingue possèdent plus d'une traduction candidate possible. Diverses tâches, telles que la traduction automatique, la recherche d'informations et l'extraction d'informations, doivent faire face aux nombreux sens différents que possède chaque élément lexical. Ce problème a été abordé de différentes manières, à savoir : en examinant les contextes d'utilisation (avec l'apprentissage supervisé ou le regroupement des sens sans supervision), ou en utilisant des ressources lexicales telles que les dictionnaires ou les thésaurus. Le premier type d'approche repose sur des données difficiles à collecter (supervisées) ou très sensibles au type de corpus (non supervisé). Le deuxième type d'approche tente d'exploiter les connaissances lexicales représentées dans les dictionnaires ou les thésaurus avec des résultats variés. Dans tous les cas, la distance, entre les mots ou les sens des mots, est utilisée pour déterminer le bon sens dans un contexte donné (Banerjee et Pedersen, 2003).

Gaume et al. (2004) ont présenté une méthode de désambiguïsation, dans laquelle les sens des mots sont déterminés à l'aide d'un dictionnaire monolingue. Ils ont utilisé une mesure de proximité sémantique entre les termes du dictionnaire. De plus, ils ont tenu compte de la topologie complète du dictionnaire, en transformant ses entrées en un graphe. Ainsi, nous étendons, cette étude, en proposant une approche à base de proximité bilingue pour la désambiguïsation de traduction de requêtes dans la RIT. En effet, nous définissons une nouvelle ressource linguistique, nommée, le Dictionnaire Sémantique Bilingue de Contextes (*DSBC*), pour soutenir et assurer l'apprentissage automatique dans une plate-forme sémantique de TR. La construction de ce *DSBC* repose sur la combinaison d'un corpus parallèle avec un dictionnaire bilingue. Un modèle d'appariement proxémique, basé sur ce *DSBC*, a été utilisé pour sélectionner la traduction adéquate correspondante à chaque terme d'une requête source polysémique. Au-delà, le *DSBC* et le graphe sémantique bilingue du dictionnaire ont été utilisés pour calculer la similarité entre les termes de requête et leurs traductions candidates à l'aide d'une distance sémantique probabiliste existante.

Nous introduisons, dans la section 6.1, le *DSBC*. Dans la section 6.2, nous présentons notre approche à base de proximité bilingue. Dans la section 6.3, nous récapitulons les résultats globaux des expérimentations de cette approche tout en comparant avec l'approche probabiliste (revenir en détails, dans le chapitre 3, vers la section 3.1.4 page 60) et l'approche possibiliste à base de transformation probabilité-possibilité (Elayeb et al., 2018).

6.1 Le Dictionnaire Sémantique Bilingue de Contextes : DSBC

La robustesse des connaissances disponibles sous forme de lexiques ou de corpus (Vidhu Bhala et Abirami, 2012) est l'un des facteurs les plus pertinents qui affectent les tendances de recherche dans le domaine de désambiguïsation des sens de mots (WSD). En conséquence, le problème de désambiguïsation de TR dans la RIT nécessite une modélisation efficace pour ces connaissances.

Dans ce chapitre, nous proposons un *DSBC* comme une solution générique basée sur les graphes, afin de modéliser les connaissances requises pour la TR. Cependant, la construction et la représentation automatique d'un graphe $G = (T, E)$, associé à un ensemble des termes d'une requête source et ses traductions possibles, nécessite de définir l'ensemble des nœuds T et l'ensemble des arêtes E. Au cours de l'étape de traitement de notre approche, les arêtes et les nœuds sont modélisés à partir des ressources suivantes : un dictionnaire bilingue traditionnel et un corpus bilingue parallèle (Ben Romdhane et al., 2018).

6.1.1 L'ensemble bilingue des nœuds

Les nœuds du graphe sont les termes de la requête source (RS) et ses traductions possibles existantes dans le dictionnaire bilingue. Le terme qui possède plus qu'une traduction est considéré comme un terme polysémique. Les entrées du *DSBC* sont le contexte du mot polysémique et les entrées du dictionnaire bilingue traditionnel. Nous les modélisons comme suit (Ben Romdhane et al., 2018) :

- *Polysémie (RS)* : l'ensemble des termes polysémiques de RS : $p_1, p_2, \ldots, p_k$.
- *Significatif (RS)* : l'ensemble des termes significatifs de RS qui ne sont pas polysémiques : $m_1, m_2, \ldots, m_i$.
- *Contexte (RS)* = $\{Significatif(RS) \cup Polysémie(RS)\}$
- *Contexte (p, RS)* = $\{Significatif(RS) \cup Polysémie(RS)\} \backslash \{p\}$; Avec p est un terme polysémique $\in RS$.
- *Traduction (t_i)* : l'ensemble des traductions des termes significatifs t_i de RS dans le dictionnaire bilingue traditionnel d_{t_i} si le terme t_i n'est pas polysémique ; et $p_i^1, p_i^2, \ldots, p_i^a$ si le terme t_i est polysémique. Cet ensemble comprend uniquement les traductions qui sont réellement utilisées dans l'ensemble d'apprentissage (c'est-à-dire dans le *DSBC*).

6.1.2 L'ensemble des arêtes

Nous pouvons distinguer plusieurs types de réseaux lexicaux, selon la nature des relations sémantiques qui définissent les arêtes du graphe bilingue. Nous avons utilisé dans la construction de notre graphe du *DSBC* les relations suivantes (Ben Romdhane et al., 2018) :

- *Les relations de traduction* : nous créons une arête entre un terme polysémique t_i de RS et t_j, si t_j est une traduction de t_i. Nous avons modélisé ces relations comme suit :

$$\forall t_i \in Polysémie\,(RS), \forall t_j \in Traduction\,(t_i) \Rightarrow < t_j, t_i > \in E$$

- *Les relations syntagmatiques* : ils sont également connus comme les relations de co-occurrence. Nous créons dans le graphe une arête entre deux termes s'ils co-occurrent dans le même contexte. En effet, nous nous inspirons des approches qui apprennent des dépendances entre les termes qui apparaissent dans un contexte donné et qui représentent les traductions des termes (Yuret et Yatbaz, 2010). Ces relations sont modélisées comme suit :

$$\forall t_i, t_j \in Contexte\,(RS)\ si\ i \neq j \Rightarrow < t_j, t_i > \in E$$

- *Les relations paradigmatiques* : nous considérons ici la relation de synonymie entre deux nœuds. Nous construisons une arête entre deux termes s'ils conservent une relation de synonymie entre eux, tel que le cas de *WordNet* (Ploux et Victorri, 1998). En d'autres termes, s'ils partagent des mots communs dans leurs traductions. Nous avons modélisé ces relations comme suit :

$$\forall t_i, t_j\ si\ \{Traduction\,(t_i) \cap Traduction\,(t_j)\} \neq \varnothing \Rightarrow < t_j, t_i > \in E$$

- *Les relations de proximité sémantique* : cette relation prend en considération l'aspect syntagmatique et paradigmatique. Nous nous sommes inspirés des travaux de (Veronis et Ide, 1990) qui ont testé la construction automatique de réseaux de neurones à partir des définitions des termes des dictionnaires bilingues dans le cas de désambiguïsation des sens de mots. Dans notre cas, nous construisons des liens indépendants du contexte entre toutes les traductions possibles dans le graphe (Yuret et Yatbaz, 2010). En effet, nous construisons une arête entre deux traductions t_i et t_j si t_j apparaît dans la définition de t_i. Nous avons modélisé ces relations comme suit :

$$\forall t_j \in Traduction\,(t_i)\ si\ i \neq j \Rightarrow < t_j, t_i > \in E$$

La pondération est attribuée pour chaque arête dans le graphe sémantique bilingue. Nous détaillons dans les sections suivantes les formules de calcul (cf. section 6.2.1) et nous présentons un exemple illustratif de calcul (cf. section 6.2.2).

6.2 Proposition d'une approche à base de proximité bilingue pour la désambiguïsation de TR

La technique de proximité a été initialement utilisée par (Gaume et al., 2004), quand ils ont abordé le problème de désambiguïsation sémantique des sens de mots. Les auteurs ont présenté et testé une méthode de désambiguïsation par proximité structurelle dans laquelle les sens des mots sont déterminés à l'aide d'un dictionnaire. En fait, ils ont utilisé une mesure de proximité sémantique entre mots, appelée *proxémie*, à partir de laquelle ils ont défini une distance proxémique entre les nœuds des graphes de dictionnaire. Elayeb et al. (2015) ont étendu et exploité cette technique pour l'appliquer dans la désambiguïsation sémantique monolingue des requêtes. Ils ont proposé une approche probabiliste pour la désambiguïsation automatique des sens de mots français. Dans notre cas, nous présentons dans cette section une nouvelle approche de proximité bilingue, inspirée des précédentes afin de l'appliquer à la désambiguïsation de TR dans le domaine de RIT (Ben Romdhane et al., 2018).

6.2.1 Calcul sémantique

Le processus de modélisation d'un problème de désambiguïsation de TR à travers un graphe de *DSBC* nécessite une étape d'apprentissage à partir de données. La sélection de la traduction appropriée, correspondante à un terme donné d'une requête source, nécessite les connaissances existantes dans sa définition figurant dans le dictionnaire bilingue. D'abord, nous générons la matrice de transition ou d'adjacence (6.2.1.1) à partir du graphe du *DSBC*. Puis, nous transformons cette matrice d'adjacence en matrice de Markov (6.2.1.2) ayant les nœuds du graphe comme des états et les transitions possibles comme des arêtes. Enfin, nous exécutons l'approche à base de proximité bilingue (6.2.1.3) afin de sélectionner la meilleure traduction.

6.2.1.1 Construction de matrice d'adjacence d'un graphe

La matrice d'adjacence est générée à partir du graphe $G =< T, E >$. Nous notons $[G]$ la matrice carrée $n * n$ telle que :

$\forall r, s \in T$, $[G]_{r,s} = 1$ si $(r, s) \in E$ et $[G]_{r,s} = 0$ si $(r, s) \notin E$

Nous appelons aussi $[G]$ la matrice de transition [1] de G. Comme G n'est pas orientée,

1. La matrice de transitions a les propriétés suivantes : tous les coefficients sont compris entre 0 et 1 ; pour chaque ligne, la somme des coefficients vaut à 1.

elle est donc symétrique :

$\forall r, s \in T$, si $(r, s) \in E$ alors $(s, r) \in E$

6.2.1.2 Construction de matrice de Markov

On considère $G =< T, E >$ un graphe réflexif. La matrice de Markov est générée à partir de matrice d'adjacence de G. Notons $\left[\hat{G}\right]$ la matrice Markovienne du graphe G définie comme suit :

$$\forall r, s \in T, \ \left[\hat{G}\right]_{r,s} = \frac{[G]_{r,s}}{\sum_{x \in T} [G]_{r,x}} \tag{6.1}$$

Avec $\forall r, s \in T$, $[G]_{r,s} \geq 0$

Et comme G est réflexif, $\forall r \in T$, $[G]_{r,r} = 1$ alors $\forall r \in T$, $\sum_{x \in T} [G]_{r,x} > 0$

6.2.1.3 Approche basée sur la proximité bilingue

Dans cette section, nous présentons en premier lieu l'idée fondamentale de l'approche de proximité et en deuxième lieu nous proposons une nouvelle version de cette approche appliquée à la désambiguïsation de TR dans la RIT.

La technique de proximité est une méthode stochastique pour étudier la structure des graphes de petits mondes hiérarchiques, en transformant les graphes en chaînes de Markov dont les états sont les nœuds du graphe et les transitions sont ses arêtes. L'idée globale est d'étudier la relation sémantique entre les termes et calculer la *distance proxémique* entre les sommets des graphes qui sont les entrées du dictionnaire français.

Par ailleurs, les chaînes de Markov ont obtenu des résultats significatifs dans la désambiguïsation des sens de mots, mais elles ne sont pas suffisamment utilisées dans le problème de désambiguïsation de TR dans la RIT. Dans cette technique, le graphe du dictionnaire est transformé en une chaîne de Markov, où les nœuds du graphe (mots et leurs traductions possibles) sont des états et les relations entre les nœuds sont des arêtes. Étant donné une particule, elle peut partir d'un nœud s_0 à un instant $t = 0$, et se déplace à un nœud s_1 l'un des voisins de s_0 sélectionnée au hasard ; la particule se déplace de nouveau à un nœud s_2, l'un des voisins de s_1 sélectionnée au hasard, etc. Si à le $t^{ième}$ déplacement, la particule est sur le nœud s_t, elle se déplace alors à un nœud s_{t+1} qui est sélectionné au hasard parmi les voisins qui ont la même probabilité. Une trajectoire sélectionnée $s_1, s_2, \ldots, s_t, \ldots$

est une «*balade aléatoire*» sur le graphe et les dynamiques de ces trajectoires des particules donnent les propriétés structurelles de ces graphes (Gaume, 2004) .

Gaume et al. (2004) ont défini $PROX\,(G, i, s, r) = [\hat{G}^i]_{r,s}$ comme la probabilité qu'une particule passe aléatoirement du nœud r à l'instant $t = 0$, au nœud s à l'instant $t = i$. Notons que $[\hat{G}^i]$ est la matrice $[\hat{G}]$ multiplié i fois par elle-même.

6.2.1.4 Calcul dynamique des traductions

Nous proposons ici une technique dynamique qui calcule les scores des traductions correspondantes aux termes polysémiques d'une requête source. Nous utilisons les graphes sémantiques bilingues du *DSBC* ainsi que la distance sémantique entre les nœuds du graphe. La meilleure traduction est celle qui a le meilleur score parmi toutes les traductions possibles existantes dans le graphe. Nous nous basons ici sur le principe de la méthode *PROX*, qui calcule un score de similarité sémantique (*proximité*) entre les nœuds du graphe bilingue de manière originale et innovante comme suit (Ben Romdhane et al., 2018) :

Compte tenu du terme polysémique de requête source t_i ; ayant plus d'une traduction possible dans le lexique bilingue Français-Anglais. Notons par :

- t_i est un nœud du graphe G.
- *Tranduction* $(t_i) = p_i^1, p_i^2, ..., p_i^a$
- $\lim_{z\to\infty}[\hat{G}^z] = G^\infty$, G^∞ est un vecteur de $\mathbb{R}^a$
- $f_\infty(r, s) = \lim_{z\to\infty} PROX\,(G, z, s, r)$

f_∞ indique la proximité sémantique entre les termes polysémiques français et leurs traductions anglaises dans le *DSBC*. Nous avons donc la propriété suivante :

Comme G est un graphe réflexif et fortement connexe, alors :

$$\forall a \in T \lim_{z\,\to\,\infty} PROX\,(G, z, s, r) = \lim_{z\to\infty} PROX\,(G, i, a, r) \tag{6.2}$$

Cela signifie que la probabilité pour un temps z assez long d'atteindre un nœud r ne dépend pas du nœud de départ (s ou a), mais seulement du nœud r.

Nous disons aussi que α est fortement liée à α^i si et seulement si :

$$\forall j \in \mathbb{N}\backslash\{i\},\ f_\infty(\alpha, \alpha^i) > f_\infty(\alpha, \alpha^j)$$

Dans ce cas, la traduction appropriée de α est α^i. Cela permet de définir que la proximité entre deux nœuds n'est pas par la simple existence d'une arête entre ces deux derniers, mais par la structure globale du graphe (Gaume, 2004).

Le terme β porte des informations qui vérifie les propriétés suivantes :

- $\beta \in Contexte\,(\alpha, RS)$.
- $\forall \delta \in Contexte\,(\alpha, RS) \setminus \{\beta\},\ f_\infty(\alpha, \beta) = \max_\delta(f_\infty(\alpha, \delta))$; où β est la définition de α. Dans ce cas, β est la définition sémantique de α dans le *DSBC*.

6.2.2 Exemple illustratif

Considérons l'exemple de la première requête source polysémique française *RS1* : « *L'Union Européenne et les Pays Baltes* ».

RS1 est représentée comme suit :

- ***Contexte(RS1)*** = {union, européenne, pays, Balte}.
- ***Significatif (RS1)*** = {européenne, Balte}.
- ***Polysémie(RS1)*** = {union, pays}.
- ***Traduction(union)*** =

  Union : « the condition of being united, the act of uniting, or a conjunction formed by such an act ... »

  Unity : « the state or quality of being one ; oneness ... »
- ***Traduction(pays)*** =

  Country : « an **area** of land distinguished by its political autonomy ; state... »

  Land : « The solid part of the surface of the earth as distinct from seas, lakes, etc... »
- ***Traduction(Européenne)*** = **European** : « a supporter of the European Union or of political union of the **countries** of Europe or a part of it... »
- ***Traduction(Balte)*** = **Baltic** : « a branch of the Indo-European family of languages consisting of Lithuanian, Latvian, and old Prussian... ».
- ***Espace sémantique bilingue*** : {union, union, unity, européenne, european, pays, country, land, balte, baltic}.

La figure 6.1 schématise le graphe sémantique bilingue qui correspond à l'espace sémantique bilingue de la requête source *RS1* .

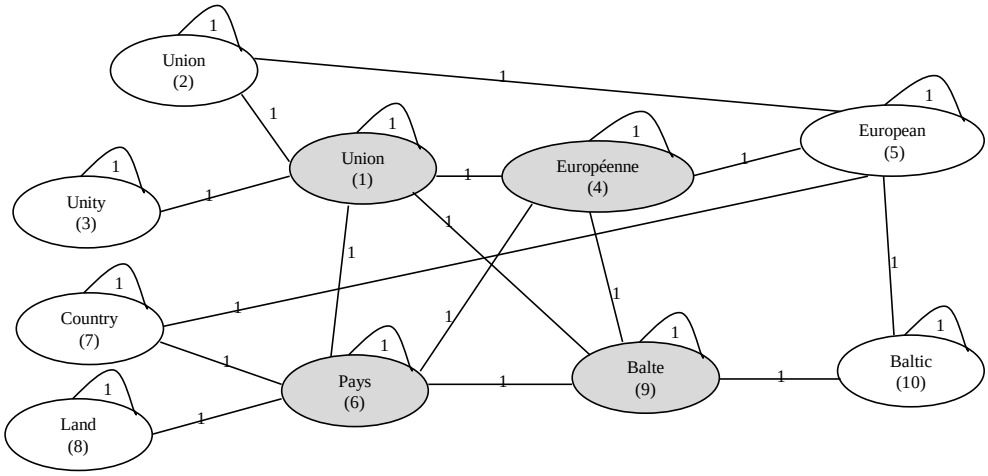

Figure 6.1 – Graphe sémantique bilingue de la requête source *RS1*

Le calcul de similarité sémantique (proximité), pour l'identification de la traduction appropriée, est basé sur la matrice Markovienne suivante de *RS1*. Sachant que le graphe est construit à partir d'un dictionnaire bilingue, avec $\left[\hat{G}\right]_{r,s} = 0$ s'il n'y a pas d'arête entre les nœuds r et s, et $1/D$ sinon, où D est le nombre de voisins de r.

$$\left[\hat{G}\right] = \begin{pmatrix} (0) & (1) & (2) & (3) & (4) & (5) & (6) & (7) & (8) & (9) & (10) \\ (1) & 1/6 & 1/6 & 1/6 & 1/6 & 0 & 1/6 & 0 & 0 & 1/6 & 0 \\ (2) & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ (3) & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ (4) & 1/5 & 0 & 0 & 1/5 & 1/5 & 1/5 & 0 & 0 & 1/5 & 0 \\ (5) & 0 & 1/5 & 0 & 1/5 & 1/5 & 0 & 1/5 & 0 & 0 & 1/5 \\ (6) & 1/6 & 0 & 0 & 1/6 & 0 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \\ (7) & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 1/3 & 0 & 0 & 0 \\ (8) & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ (9) & 1/5 & 0 & 0 & 1/5 & 0 & 1/5 & 0 & 0 & 1/5 & 1/5 \\ (10) & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 1/3 & 1/3 \end{pmatrix}$$

$$\left[\hat{G}\right]^{\infty} = \begin{pmatrix} (0) & (1) & (2) & (3) & (4) & (5) & (6) & (7) & (8) & (9) & (10) \\ (1) & 0.1528 & \mathbf{0.0755} & 0.0528 & 0.1249 & 0.1236 & 0.1489 & 0.0736 & 0.0490 & 0.1249 & 0.0740 \\ (2) & 0.1510 & 0.0766 & 0.0509 & 0.1250 & 0.1267 & 0.1472 & 0.0747 & 0.0471 & 0.1249 & 0.0759 \\ (3) & 0.1585 & 0.0764 & 0.0588 & 0.1247 & 0.1204 & 0.1470 & 0.0707 & 0.0472 & 0.1245 & 0.0718 \\ (4) & 0.1499 & 0.0750 & 0.0499 & 0.1250 & 0.1251 & 0.1499 & 0.0750 & 0.0499 & 0.1251 & 0.0751 \\ (5) & 0.1483 & 0.0760 & 0.0481 & 0.1251 & 0.1280 & 0.1483 & 0.0760 & 0.0481 & 0.1252 & 0.0769 \\ (6) & 0.1489 & 0.0736 & 0.0490 & 0.1249 & 0.1236 & 0.1528 & \mathbf{0.0755} & 0.0528 & 0.1249 & 0.0740 \\ (7) & 0.1472 & 0.0747 & 0.0471 & 0.1250 & 0.1267 & 0.1510 & 0.0766 & 0.0509 & 0.1249 & 0.0759 \\ (8) & 0.1470 & 0.0707 & 0.0472 & 0.1247 & 0.1204 & 0.1585 & 0.0764 & 0.0588 & 0.1245 & 0.0718 \\ (9) & 0.1498 & 0.0749 & 0.0498 & 0.1251 & 0.1252 & 0.1498 & 0.0749 & 0.0498 & 0.1252 & 0.0753 \\ (10) & 0.1480 & 0.0759 & 0.0478 & 0.1252 & 0.1282 & 0.1480 & 0.0759 & 0.0478 & 0.1255 & 0.0774 \end{pmatrix}$$

L'observation de dynamique de la particule à partir d'un nœud r vers un nœud s indique la relation sémantique entre les deux nœuds. Par exemple, la courbe $f_z(union, union) = [\hat{G}^z]_{union,union}(resp. f_z(union, unity) = [\hat{G}^z]_{union,unity})$ représente la probabilité de la particule en partant du nœud «*union*», qui est le terme polysémique de *RS1,* pour atteindre le nœud «*union*» (resp. «*unity*») qui est sa traduction à l'instant z. D'après la propriété du calcul dynamique des traductions,

quand le temps z tend vers l'infini, donc la courbe tend vers sa limite, dans notre cas la courbe tend vers le nœud qui correspond à la traduction «*union*» (resp. «*unity*»).

Alors :

$f_\infty(union, union) = \lim_{z\to\infty}[\hat{G}^z] = 0.0755$

$f_\infty(union, unity) = \lim_{z\to\infty}[\hat{G}^z]_{union,unity} = 0.0528$

Notons que dans $\left[\hat{G}\right]^\infty$, «*union*» puis «*pays*» sont les termes les plus importants parmi tous les autres nœuds du graphe sémantique bilingue. Nous distinguons :

- «*union*» est plus similaire à «*union*» qu'à «*unity*» car $f_\infty(1,2) = 0.0755 > f_\infty(1,3) = 0.0528$.

- «*pays*» est plus similaire à «*country*» qu'à «*land*» parce que $f_\infty(6,7) = 0.0755 > f_\infty(6,8) = 0.0528$.

L'espace d'apprentissage bilingue sera : «*union*» = «*union*» et «*pays*» = «*country*». Les deux définitions de «*union*» et «*country*» sont ajoutées au *DSBC*. En outre, ce résultat sera pris en compte dans les prochaines étapes de traitement.

Considérons une seconde requête source française polysémique :

RS2 : « *Le pays du zone politique* ».

Le graphe sémantique bilingue correspondant à la deuxième requête *RS2* est présenté par la figure 6.2.

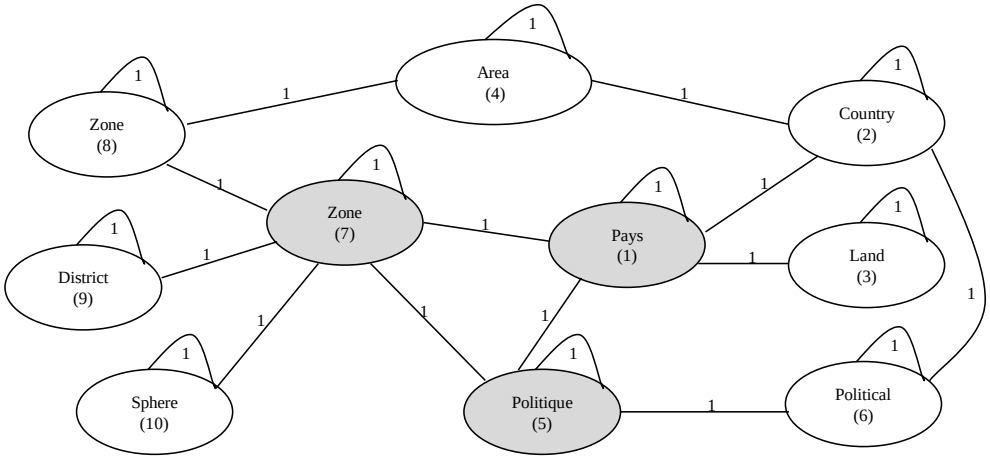

Figure 6.2 – Graphe sémantique bilingue la requête source *RS2*

La requête *RS2* est représentée comme suit :

- ***Contexte(pays, RS2)*** = {zone, politique}.
- ***Polysémie (RS2)*** = {zone}.
- ***Traduction(politique)*** = **political** : « of or relating to the state, government, the public administration, policy making... »

- ***Traduction(zone)*** =

  Zone : « a region, **area**, or section characterized by some distinctive feature or quality... »

  District : « any subdivision of any territory, region, etc.... »

  Sphere : « any object having approximately this shape ; globe... »

- «Pays» existe déjà dans le *DSBC*. Le terme «area» existe dans la définition de «zone» et la définition de «country» («zone» $\bigcap$ «country» = {area}). Donc, il y'a un enrichissement du contexte de « Pays (country) » : ***Contexte***(pays) = {zone, country, area}.
- ***Espace sémantique bilingue*** : {pays, country, land, area, politique, political, zone, zone, district, sphere}.

Le calcul de similarité sémantique (proximité) pour l'identification de la traduction appropriée est basé sur la matrice Markovienne suivante de *RS2* :

$$\left[\hat{G}\right] = \begin{pmatrix}
(0) & (1) & (2) & (3) & (4) & (5) & (6) & (7) & (8) & (9) & (10) \\
(1) & 1/5 & 1/5 & 1/5 & 0 & 1/5 & 0 & 1/5 & 0 & 0 & 0 \\
(2) & 1/4 & 1/4 & 0 & 1/4 & 0 & 1/4 & 0 & 0 & 0 & 0 \\
(3) & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
(4) & 0 & 1/3 & 0 & 1/3 & 0 & 0 & 0 & 1/3 & 0 & 0 \\
(5) & 1/4 & 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 0 & 0 & 0 \\
(6) & 0 & 1/3 & 0 & 0 & 1/3 & 1/3 & 0 & 0 & 0 & 0 \\
(7) & 1/6 & 0 & 0 & 0 & 1/6 & 0 & 1/6 & 1/6 & 1/6 & 1/6 \\
(8) & 0 & 0 & 0 & 1/3 & 0 & 0 & 1/3 & 1/3 & 0 & 0 \\
(9) & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 \\
(10) & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 1/2
\end{pmatrix}$$

$$\left[\hat{G}\right]^{\infty} = \begin{pmatrix}
(0) & (1) & (2) & (3) & (4) & (5) & (6) & (7) & (8) & (9) & (10) \\
(1) & 0.1528 & \mathbf{0.1192} & 0.0642 & 0.0852 & 0.1200 & 0.0905 & 0.1722 & 0.0838 & 0.0560 & 0.0560 \\
(2) & 0.1490 & 0.1267 & 0.0598 & 0.0943 & 0.1188 & 0.0945 & 0.1653 & 0.0888 & 0.0514 & 0.0514 \\
(3) & 0.1605 & 0.1197 & 0.0721 & 0.0809 & 0.1222 & 0.0914 & 0.1681 & 0.0785 & 0.0533 & 0.0533 \\
(4) & 0.1420 & 0.1257 & 0.0539 & 0.1005 & 0.1143 & 0.0908 & 0.1691 & 0.0961 & 0.0537 & 0.0537 \\
(5) & 0.1501 & 0.1188 & 0.0611 & 0.0857 & 0.1200 & 0.0907 & 0.1740 & 0.0848 & 0.0573 & 0.0573 \\
(6) & 0.1508 & 0.1260 & 0.0609 & 0.0908 & 0.1210 & 0.0958 & 0.1657 & 0.0856 & 0.0518 & 0.0518 \\
(7) & 0.1435 & 0.1102 & 0.0560 & 0.0845 & 0.1160 & 0.0828 & 0.1866 & \mathbf{0.0893} & 0.0655 & 0.0655 \\
(8) & 0.1397 & 0.1184 & 0.0524 & 0.0961 & 0.1131 & 0.0856 & 0.1785 & 0.0963 & 0.0600 & 0.0600 \\
(9) & 0.1401 & 0.1028 & 0.0533 & 0.0806 & 0.1146 & 0.0777 & 0.1966 & 0.0900 & 0.0726 & 0.0716 \\
(10) & 0.1401 & 0.1028 & 0.0533 & 0.0806 & 0.1146 & 0.0777 & 0.1966 & 0.0900 & 0.0716 & 0.0726
\end{pmatrix}$$

Selon la dernière matrice, nous remarquons que «country» puis «zone» recueillent plus de votes que tous les autres nœuds du graphe sémantique bilingue. Nous distinguons cela :

- «Pays» est plus similaire à «country» qu'à «land» ; car $f_{\infty}(1,2) = 0.1192 > f_{\infty}(1,3) = 0.0642$.

- «Zone» est plus similaire à «zone» qu'à «district» et qu'à «sphere» ; car $f_{\infty}(7,8) = 0.0893 > f_{\infty}(7,9) = f_{\infty}(7,10) = 0.0655$.

Nous remarquons également que «Pays», «Zone», «country», «zone» et «area» sont fortement liés.

La mise à jour du *DSBC* est la suivante :

- Ajouter à l'entrée «Pays» le concept «zone». Donc, «Pays» = «country», «Zone» = « zone » et « Union » = « union ».

- Ajouter une nouvelle entrée «Zone» = «zone» et «Pays» = «country».

Nous avons montré à travers ces deux exemples la pertinence de la nouvelle ressource linguistique (*DSBC*) dans le processus de désambiguïsation de TR. En effet, nous avons mis en évidence l'impact de l'étape d'apprentissage sur l'efficacité de l'approche de proximité bilingue.

Dans la section expérimentale suivante, nous montrerons l'efficacité de notre approche en la comparant aux approches probabiliste ((revenir en détails, dans le chapitre 3, vers la section 3.1.4 page 60) et possibiliste à base de transformation (Elayeb et al., 2018), qui ont utilisé seulement le dictionnaire bilingue traditionnel ainsi que le corpus parallèle.

6.3 Résultats expérimentaux de l'approche à base de proximité bilingue

Dans cette partie nous présentons les expérimentations et les évaluations de notre approche à base de proximité bilingue pour la désambiguïsation de TR. Notre approche a profité d'une combinaison d'un corpus parallèle et d'un dictionnaire bilingue Français-Anglais afin de construire un Dictionnaire Sémantique Bilingue de Contextes (*DSBC*). Ce dernier est utile pour sélectionner la traduction appropriée pour un terme polysémique de la requête source en utilisant un modèle d'appariement à base de proximité. La couverture du *DSBC* augmente de façon continue au cours de la phase d'apprentissage, qui prend principalement en considération les informations contextuelles requises pour la désambiguïsation des traductions. Pendant la phase d'apprentissage, le *DSBC* a inclus 1323 mots français et 7705 mots anglais, alors que le dictionnaire bilingue traditionnel a inclus 717 mots français et 2324 traductions anglaises.

En effet, nous suivons les mêmes scénarios et métriques d'évaluation effectués dans les expérimentations du chapitre 5. Nous nous concentrons sur les métriques de performance telles que : le rappel et la précision (section 6.3.1), les valeurs de précision aux top documents, ainsi que la précision moyenne et la précision exacte (section 6.3.2), le pourcentage d'amélioration de l'approche à base de proximité par rapport à ses concurrentes (section 6.3.3), et le test de significativité statistique de *Wilcoxon* (6.3.4) (Ben Romdhane et al., 2018).

6.3.1 Scénarios des courbes de Rappel-Précision

Dans cette section, nous exploitons diverses métriques et scénarios, exécutés à travers notre SRIT en utilisant nombreuses combinaisons des paramètres (*titre*, *description*, *narration*, *titre+description*, *description+narration*, etc.), et testés sur la collection du standard de test CLEF-2003. En effet, nous suivons le protocole TREC recommandé pour l'évaluation des SRIT. Nous comparons la performance globale de cette approche aux approches probabiliste, possibiliste et monolingue.

Nous exposons, dans cette section, les courbes de rappel-précision pour comparer l'approche de proximité avec les approches probabiliste, possibiliste et monolingue en utilisant différentes combinaisons des éléments des requêtes sources. Notre objectif est d'évaluer la sensibilité de notre approche à l'information contextuelle fournie par ces éléments.

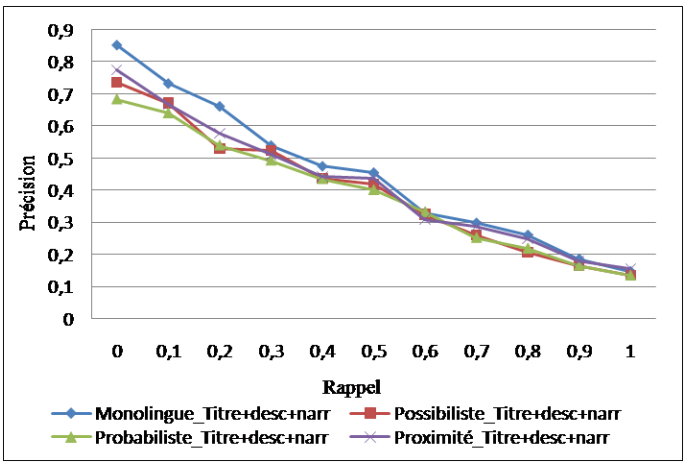

Figure 6.3 – Courbes de Rappel-Précision utilisant « *Titre+desc+narr* »

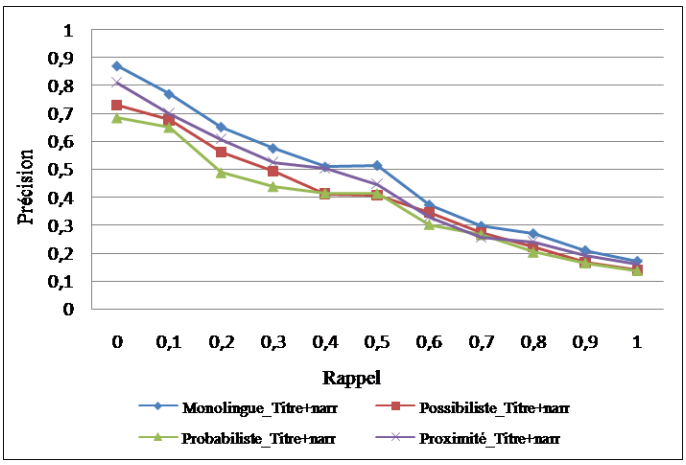

Figure 6.4 – Courbes de Rappel-Précision utilisant « *Titre+narr* »

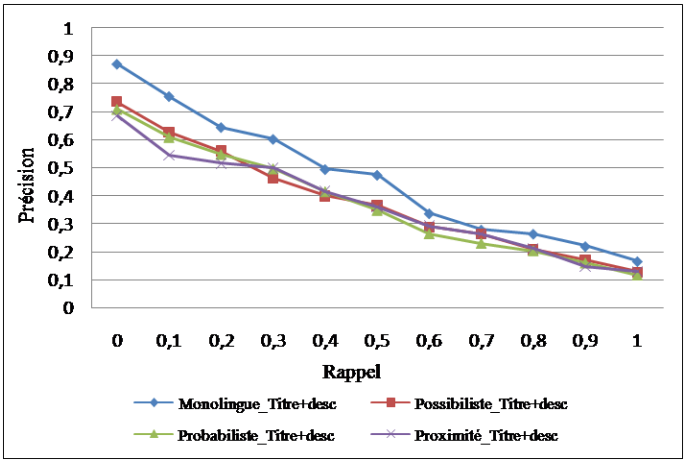

Figure 6.5 – Courbes de Rappel-Précision utilisant « *Titre+desc* »

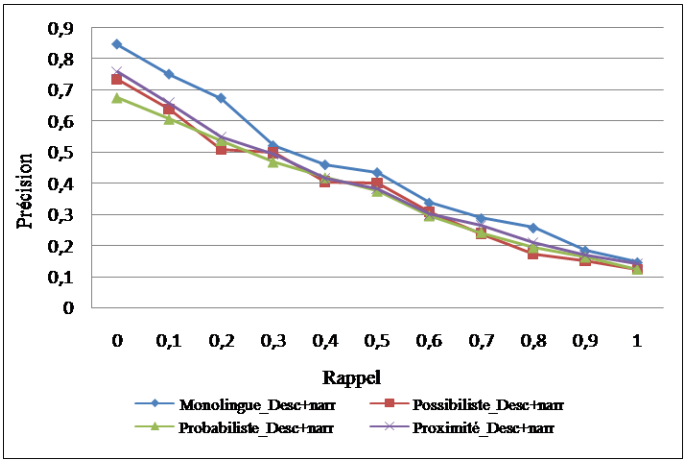

Figure 6.6 – Courbes de Rappel-Précision utilisant « *Desc+narr* »

En effet, si nous exécutons uniquement la partie «*Titre*» de la requête source (figure 6.7), le contexte est limité à un petit ensemble des termes. Par conséquent, l'approche probabiliste manque des syntagmes nominaux identifiés, elle sera alors limitée à un processus de traduction mots-à-mots. C'est pourquoi notre approche à base de proximité a surpassé les approches probabiliste et possibiliste à partir du point de rappel 0.2. Cependant, elle est légèrement sous la probabiliste et la possibiliste surtout dans certains points de rappel de bas-niveau (0 et 0.1). Entre les points de rappel 0.3 et 0.6, l'approche à base proximité a surpassé les deux approches avec un écart important entre elles. À partir du point de rappel 0.7, cette supériorité devient légère et l'écart entre la proximité et la possibiliste est limitée pour des points de rappel de haut-niveau. En outre, la performance de monolingue a surpassé toutes les approches dans tous les points de rappel. Mais, l'écart entre la monolingue et la proximité est de plus en plus réduit à partir du point de rappel 0.5. Cette supériorité de la proximité peut s'expliquer par l'information contextuelle

supplémentaire fournie par le *DSBC* qui surmonte le manque du contexte causé par les parties «*Titre*» des requêtes sources.

Lorsque nous exécutons la partie «*Description*» des requêtes sources (figure 6.8), le contexte est légèrement élargi et les approches probabiliste et possibiliste peuvent identifier et traduire davantage des syntagmes nominaux. Par conséquent, l'approche possibiliste à base de transformation a surpassé à la fois les deux approches : probabiliste et proximité, en particulier dans le point 0.6 et dans certains points de rappel de bas-niveau (de 0 à 0.2) et de haut-niveau (0.9 et 1), alors que la probabiliste semble meilleure dans les points de rappel 0.3 et 0.4. Néanmoins, l'approche à base de proximité a légèrement surpassé la probabiliste et la possibiliste dans certains points de rappel (0.5 et 0.7). En outre, l'écart entre la monolingue et les autres approches devient de plus en plus réduit à partir du point de rappel 0.7.

D'autre part, le contexte des requêtes sources est plus amélioré grâce à sa partie «*Narration*» (figure 6.9) qui est plus adaptée aux approches probabiliste et possibiliste pour l'identification et la traduction des syntagmes nominaux les plus fréquents. C'est pourquoi elles ont réalisé des performances presque équivalentes à l'approche à base de proximité dans le point de rappel 0.5 et les ont surpassés au point 0.7. Ces deux approches sont aussi légèrement sous l'approche de proximité dans certains points de rappel de haut-niveau (0.8 à 1). Cependant, la proximité a surpassé les deux, en particulier dans certains points de rappel (de 0 à 0.4). Les écarts entre la monolingue et toutes les autres approches sont plus compacts à partir du point de rappel 0.8. La supériorité de l'approche de proximité s'explique par l'information contextuelle supplémentaire issue de la partie «*Narration*» de la requête source et du *DSBC*.

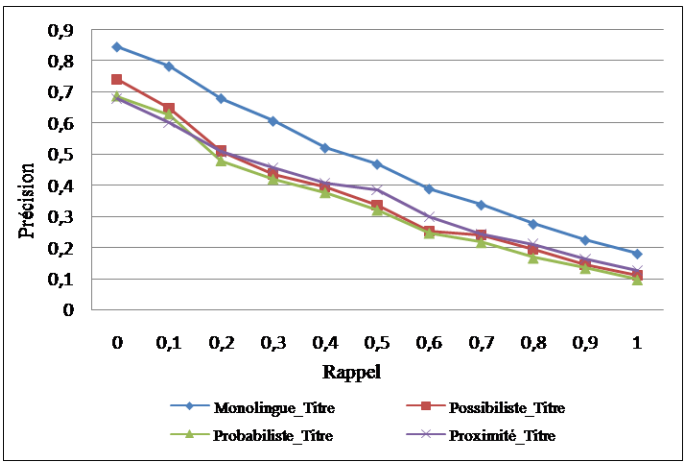

Figure 6.7 – Courbes de Rappel-Précision utilisant « *Titre* »

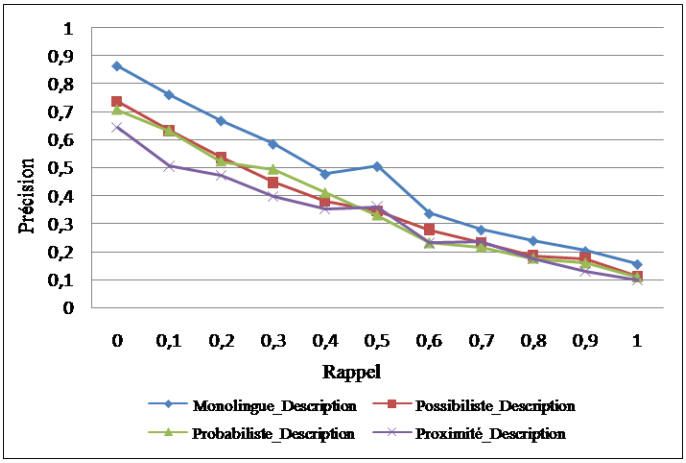

Figure 6.8 – Courbes de Rappel-Précision utilisant « *Description* »

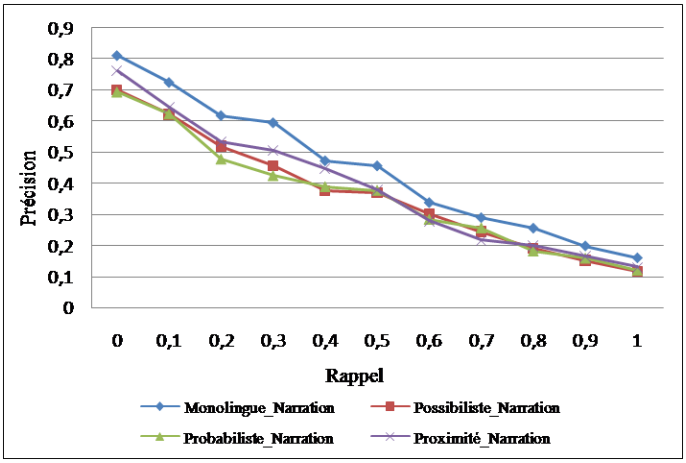

Figure 6.9 – Courbes de Rappel-Précision utilisant « *Narration* »

L'utilisation des requêtes sources complètes «*Titre+desc+narr*» (figure 6.3) est la plus populaire dans les évaluations des SRIT. Dans ce cas, chaque approche peut bénéficier de la disponibilité des informations contextuelles complètes. En fait, l'approche à base de proximité a surpassé la probabiliste et la possibiliste, en particulier dans certains points de rappel de bas-niveau (0 à 0.3). Ces deux approches sont toujours inférieures ou égales à la proximité à partir du point de rappel 0.4. Nous enregistrons également les mêmes performances de la proximité et la monolingue dans certains points de rappel de haut-niveau (0.7 à 1). Les écarts en termes de performances deviennent plus réduits entre toutes ces approches par rapport aux trois premiers cas.

Nous étudions également les combinaisons possibles des parties des requêtes sources. En effet, lorsque nous nous concentrons sur «*Titre+narr*», la proximité a considérablement surpassé la probabiliste et la possibiliste, en particulier dans les points de rappel de bas-niveau (0 à 0.5). De plus, elle est légèrement au-dessus des deux approches,

mais au-dessous de la monolingue dans certains points de rappel de haut-niveau (0.8 à 1). Les écarts entre ces approches sont progressivement réduits à partir du point de rappel 0.6 (figure 6.4).

D'autre part, «*Titre+desc*» semblent être moins adaptées à l'approche à base de proximité lorsque sa performance dépasse légèrement l'approche probabiliste et égale à l'approche possibiliste dans certains points de rappel (0.5 à 0.8). Néanmoins, la proximité est légèrement inférieure aux approches probabiliste et possibiliste dans certains points de rappel de bas-niveau (0 à 0.2) et de haut-niveau (0.8 à 1), mais elles sont égales entre 0.4 et 0.5 (figure 6.5).

En outre, «*Desc+narr*» fournissent plus d'informations contextuelles pour l'approche à base de proximité pour obtenir des meilleures performances par rapport à ses concurrentes, en particulier dans certains points de rappel de bas-niveau (0 à 0.4). Mais, cette supériorité est considérablement diminuée à partir du point de rappel 0.6 jusqu'au point 1 (figure 6.6).

Cette perturbation dans la performance de l'approche de proximité dans ces deux dernières combinaisons peut être causée par la partie «*Description*», qui comprend des expressions répétitives, telles que «*Trouvez des documents qui...*» ou «*Trouvez des informations sur...*».

6.3.2 Les points de précision aux top documents, les métriques MAP et R-Précision

Nous comparons, dans cette section, les performances globales des approches suivantes : monolingue, proximité, probabiliste et possibiliste en utilisant les valeurs de précision aux top documents ($P@5, P@10, \ldots, P@1000$) et les métriques *MAP* et *R-Précision*. Notre objectif est d'élargir et de renforcer l'étude comparative de l'approche de proximité par rapport à ses concurrentes (Ben Romdhane et al., 2018).

Lorsque nous nous concentrons sur les valeurs de précision aux top documents, nous remarquons que la précision a diminué lorsque le nombre des documents retournés a augmenté.

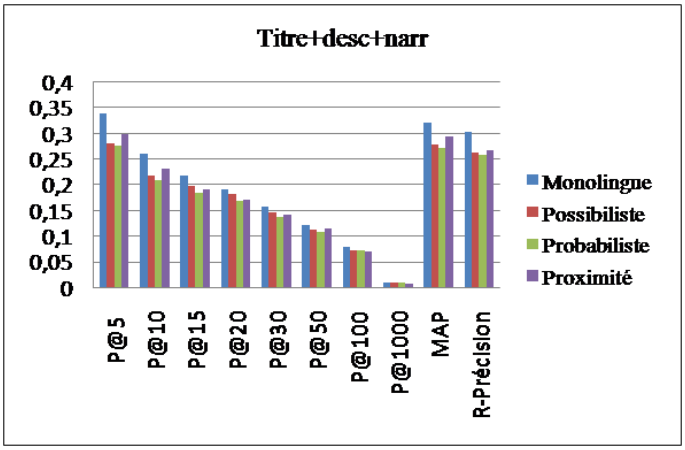

Figure 6.10 – Les valeurs de précision aux top documents, *MAP* et *R-Précision* utilisant «*Titre+desc+narr*»

Si nous exécutons le contexte complet de la requête source «*Titre+desc+narr*» (figure 6.10), l'approche à base de proximité a surpassé à la fois la probabiliste et la possibiliste en termes de valeurs de précision aux top documents retournés, sauf dans certains cas tels que :

- Dans $P@100$ et $P@1000$ dans lesquelles l'approche probabiliste est légèrement supérieure par rapport à la proximité.
- Dans $P@15$, $P@20$, $P@30$ et $P@1000$ dans lesquelles l'approche possibiliste est légèrement supérieure par rapport à la proximité.

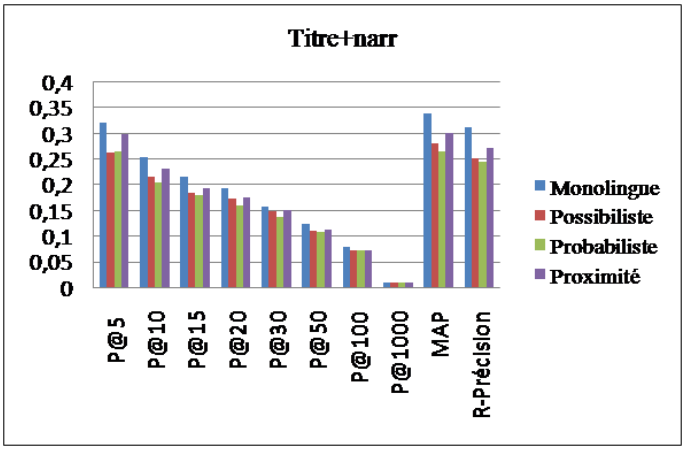

Figure 6.11 – Les valeurs de précision aux top documents, *MAP* et *R-Précision* utilisant «*Titre+narr*»

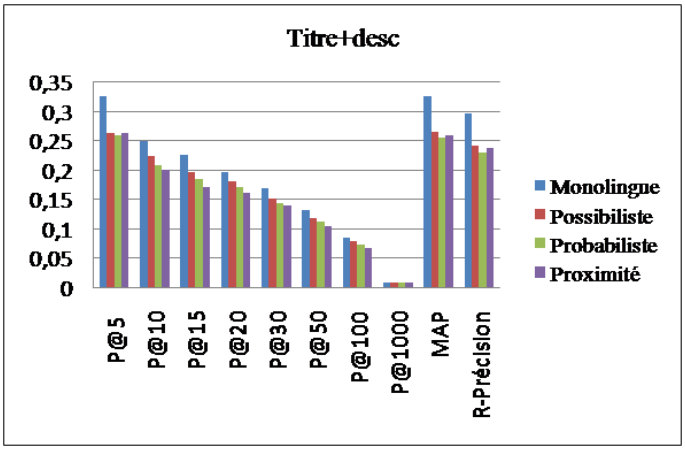

Figure 6.12 – Les valeurs de précision aux top documents, *MAP* et *R-Précision* utilisant «*Titre+desc*»

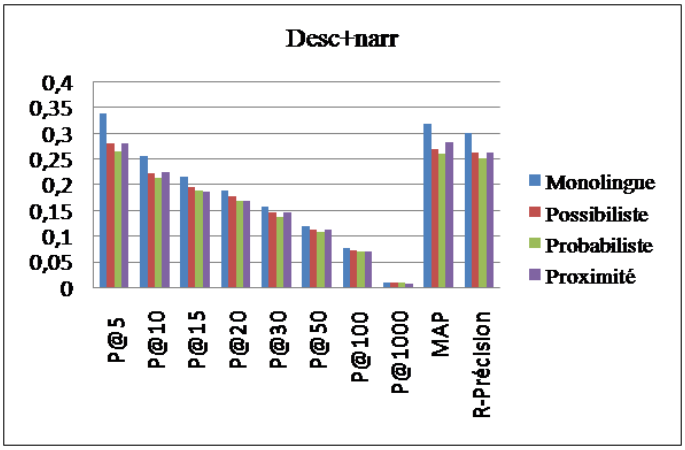

Figure 6.13 – Les valeurs de précision aux top documents, *MAP* et *R-Précision* utilisant «*Desc+narr*»

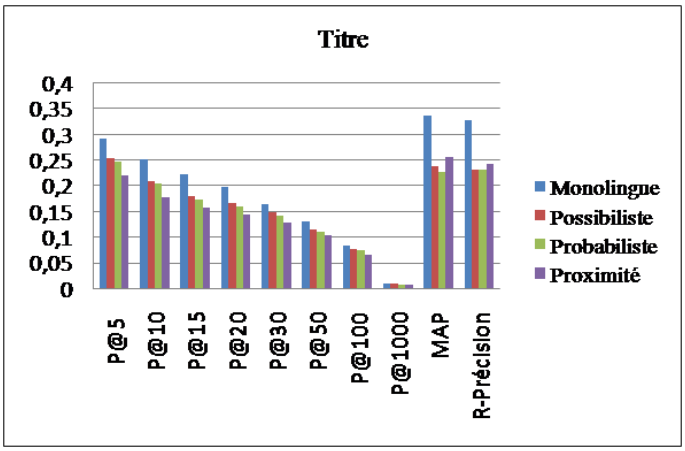

Figure 6.14 – Les valeurs de précision aux top documents, *MAP* et *R-Précision* utilisant «*Titre*»

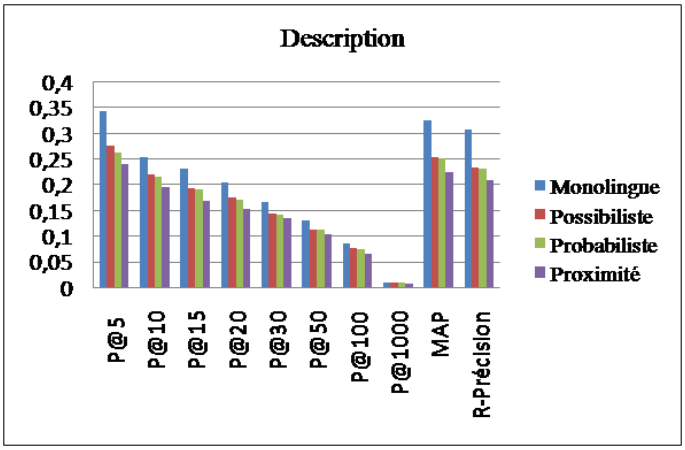

Figure 6.15 – Les valeurs de précision aux top documents, *MAP* et *R-Précision* utilisant «*Description*»

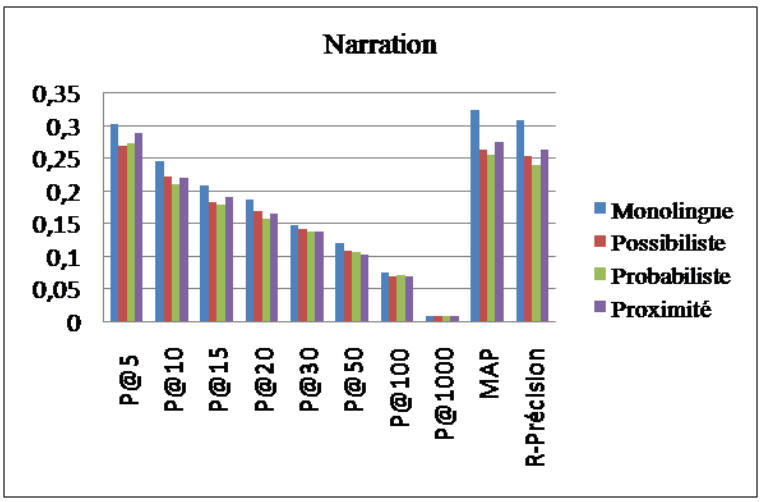

Figure 6.16 – Les valeurs de précision aux top documents, *MAP* et *R-Précision* utilisant «*Narration*»

Par ailleurs, lorsque le contexte des requêtes sources augmente progressivement en utilisant «*Narration*» (figure 6.16), «*Titre+narr*» (figure 6.11) ou «*Desc+narr*» (figure 6.13) :

- L'approche à base de proximité a donné une meilleure précision sur tous les valeurs de précision aux top documents que l'approche probabiliste, sauf dans certains cas tels que : $P@50$, $P@100$ et $P@1000$ en utilisant «*Narration*», $P@15$, $P@100$ et $P@1000$ en utilisant «*Desc+narr*» et $P@1000$ en utilisant «*Titre+narr*».
- La proximité est également la meilleure en utilisant «*Titre+narr*», tandis que la possibiliste est la meilleure en utilisant «*Titre+desc*». Néanmoins, nous remarquons quelques exceptions dans lesquelles la proximité est encore au-dessous de l'approche possibiliste telles que $P@20$, $P@30$, $P@50$, $P@100$ et $P@1000$ en utilisant «*Narration*» ou «*Desc+narr*».

Selon les métriques *MAP* et *R-Précision*, l'approche basée sur la proximité est la meilleure dans toutes les combinaisons des requêtes sources, sauf lorsque nous exécutons «*Description*» (figure 6.15) ou «*Titre+desc*» (figure 6.12) où la possibiliste est légèrement meilleure que la proximité, ce qui confirme nos conclusions faites précédemment.

En général, l'approche à base de proximité a obtenu des meilleurs résultats que l'approche probabiliste pour les requêtes longues en utilisant «*Titre+desc+narr*», «*Titre+narr*» ou «*Desc+narr*» et pour les requêtes courtes en utilisant «*Titre*» et «*Narration*».

Par rapport à l'approche possibiliste, l'approche à base de proximité a donné des bons résultats pour les requêtes longues en utilisant «*Titre+narr*» et les requêtes courtes en utilisant «*Narration*», mais, elle a des faiblesses avec la partie «*Description*» comme nous avons expliqué.

Globalement, lorsque le contexte de l'exécution est plus riche grâce au double support provenant du contexte maximal des requêtes sources et du *DSBC*, la performance de l'approche de proximité est plus proche de l'approche monolingue. De plus, l'approche à base de proximité est souvent meilleure que la probabiliste et la possibiliste avec un écart clair pour les premières valeurs de rappel correspondantes aux premiers documents sélectionnés, sauf dans «*Description*» ou «*Titre+desc*». Nous notons également que la monolingue est régulièrement la limite supérieure de la performance de RIT. Toutes nos approches restent au-dessous de cette limite parce que nous n'avons pas impliqué dans nos tests aucun processus d'expansion ou de reformulation de requêtes (Ben Romdhane et al., 2018).

6.3.3 Le pourcentage d'amélioration

Nous montrons dans le tableau 6.1 le pourcentage d'amélioration de l'approche de proximité par rapport aux approches probabiliste et possibiliste, en utilisant les valeurs de précision aux top documents, *MAP* et *R-Précision* pour les requêtes longues et courtes (Ben Romdhane et al., 2018). Nous remarquons qu'il y a une amélioration significative de l'approche à base de proximité, en particulier pour les valeurs de précision aux top documents tels que $P@5, \ldots, P@30$ en utilisant les requêtes longues (sauf le cas de «*Titre+desc*») ou les requêtes courtes (sauf le cas de «*Titre*» ou «*Description*»).

Requêtes longues	**Valeurs de précision**	**Prox vs. Proba**	**Prox vs. Poss**	**Requêtes courtes**	**Valeurs de précision**	**Prox vs. Proba**	**Prox vs. Poss**
Titre+ desc+ narr	P@5	7.99	6.57	**Titre**	P@5	-10.44	-13.07
	P@10	10.61	5.95		P@10	-13.52	-15.77
	P@15	2.68	-3.72		P@15	-9.13	-12.23
	P@20	1.65	-5.59		P@20	-8.68	-12.71
	P@30	4.43	-2.9		P@30	-9.08	-12.5
	P@50	5.09	2.03		P@50	-5.61	-9.25
	P@100	-0.27	-2.02		P@100	-11.9	-14.36
	P@1000	-1.98	-3.88		P@1000	-7	-9.71
	MAP	8.06	5.24		MAP	12.77	7
	R-Précision	3.17	1.44		R-Précision	5.36	4.5
Titre+desc	P@5	1.43	0	**Description**	P@5	-8.48	-13.35
	P@10	-3.54	-9.91		P@10	-9.41	-10.93
	P@15	-7.3	-12.51		P@15	-11.65	-13.33
	P@20	-4.85	-10.64		P@20	-10.74	-12.62
	P@30	-3.45	-7.77		P@30	-5.15	-6.77
	P@50	-6.53	-10.63		P@50	-8.01	-7.45
	P@100	-9.38	-15.08		P@100	-9.92	-12.8
	P@1000	-2.91	-3.85		P@1000	-4.81	-4.81
	MAP	1.29	-2.15		MAP	-11.26	-11.68
	R-Précision	3.43	-1.32		R-Précision	-9.6	-11.13
Titre+narr	P@5	12.49	14.07	**Narrative**	P@5	5.4	6.84
	P@10	12.6	6.83		P@10	4.41	-0.81
	P@15	7.55	4.64		P@15	6.93	4.02
	P@20	9.81	1.09		P@20	4.02	-2.76
	P@30	9.79	2.49		P@30	0	-3.42
	P@50	4.39	2.33		P@50	-2.71	-5.02
	P@100	0.82	0.82		P@100	-3.32	-1.13
	P@1000	-0.98	-1.94		P@1000	-2.94	-3.88
	MAP	13.64	7.42		MAP	7.7	4.08
	R-Précision	10.81	8.36		R-Précision	10.07	3.89
desc+narr	P@5	5.55	0				
	P@10	5.17	0.8				
	P@15	-0.64	-4.38				
	P@20	0.59	-4.64				
	P@30	6.22	0				
	P@50	2.73	0				
	P@100	-2.49	-3.68				
	P@1000	-3.96	-5.83				
	MAP	8.52	4.82				
	R-Précision	4.32	-0.23				

Table 6.1 – Le pourcentage d'amélioration de l'approche à base de proximité par rapport aux approches probabiliste et possibiliste

Tout d'abord, si nous comparons la proximité à la probabiliste, nous remarquons que la proximité a atteint une amélioration de plus de 10.6% pour $P@10$ pour les requêtes longues en utilisant «*Titre+desc+narr*». Le pourcentage d'amélioration moyen en terme de valeurs de précision aux top documents est d'environ 3.77% en utilisant «*Titre+desc+narr*», 7.05% en utilisant «*Titre+narr*» et 1.64% en utilisant «*Desc+narr*». Pour les requêtes longues, le pourcentage d'amélioration moyen du *MAP* est d'environ 7.87% et pour *R-Précision* est d'environ 5.43%.

L'approche à base de proximité a obtenu un pourcentage d'amélioration significatif dans les requêtes courtes, en particulier en utilisant la partie «*Narration*» qui fournisse le maximum d'information contextuelle. Par exemple, nous avons réalisé plus de 5% pour $P@5$ et d'environ 7% pour $P@15$, alors que le pourcentage moyen d'amélioration en terme de valeurs de précision aux top documents est d'environ 1.47%. Mais, le pourcentage moyen d'amélioration du MAP est d'environ 3.07% et de $R - Précision$ est d'environ 1.94% en utilisant les requêtes courtes.

Ensuite, si nous comparons à la possibiliste, la proximité a dépassé le 14% pour P@5 et environ 7% pour P@10 pour les requêtes longues en utilisant «*Titre+narr*». Tandis que, le pourcentage moyen d'amélioration en termes de valeurs de précision aux top documents est d'environ 3.8% pour les requêtes longues en utilisant «*Titre+narr*», alors que cette moyenne est d'environ 3.83% pour la *MAP* et de 2.06% pour la *R-Précision*. Cependant, lorsque nous nous concentrons sur les requêtes courtes en utilisant «*Narration*», la proximité a atteint une amélioration de plus de 6.8% pour $P@5$ et plus de 4% pour $P@15$. Ces résultats ont montré et confirmé l'efficacité de l'approche de proximité par rapport à la probabiliste et la possibiliste pour les requêtes longues en utilisant «*Titre+narr*» ainsi que les requêtes courtes en utilisant «*Narration*» en testant des différents métriques et scénarios d'évaluation.

6.3.4 Le test de significativité statistique de *Wilcoxon*

Il est également pertinent d'étudier la signification statistique des résultats de la proximité par rapport aux approches possibiliste et probabiliste. Pour ce faire, nous avons profité du test de significativité statistique de *Wilcoxon* (tableau 6.2). Notre objectif ici est de renforcer l'étude comparative entre ces trois approches en termes de précision aux top documents et des deux métriques *MAP* et *R-Précision* en utilisant des requêtes longues et courtes. Ainsi, nous comparons l'amélioration de la proximité par rapport à l'approche probabiliste, puis à la possibiliste (Ben Romdhane et al., 2018).

	P@5	P@10	P@15	P@20	P@30	P@50	P@100	P@1000	MAP	R-Précision
Prox	0.612	0.735	0.498	0.735	0.753	0.398	0.042	0.013	0.128	0.128
vs.	**Requêtes longues**					**Requêtes courtes**				
Proba	**Titre+desc+narr**		**Titre + desc**	**Titre + narr**	**desc + narr**	**Titre**	**desc**		**narr**	
	0.012		0.139	0.006	0.046	0.168	0.004		0.050	

	P@5	P@10	P@15	P@20	P@30	P@50	P@100	P@1000	MAP	R-Précision
Prox	0.878	0.330	0.132	0.027	0.046	0.115	0.027	0.003	0.310	0.498
vs.	**Requêtes longues**					**Requêtes courtes**				
Poss	**Titre+desc+narr**		**Titre + desc**	**Titre + narr**	**desc + narr**	**Titre**	**desc**		**narr**	
	0.507		0.007	0.006	0.332	0.036	0.005		0.507	

Table 6.2 – Le test de significativité statistique de l'approche de proximité par rapport aux approches probabiliste et possibiliste

Si nous nous focalisons sur les valeurs de précision aux top documents, nous notons que les améliorations réalisées par l'approche probabiliste sont statistiquement significatives, en particulier dans $P@100$ ($p - valeur = 0.042 < 0.05$) et dans $P@1000$ ($p - valeur = 0.013 < 0.05$). Tandis que l'approche possibiliste a permis des améliorations statistiquement significatives en particulier dans $P@20$ ($p - valeur = 0.027 < 0.05$), $P@30$ ($p - valeur = 0.046 < 0.05$), $P@100$ ($p - valeur = 0.027 < 0.05$) et $P@1000$ ($p - valeur = 0.003 < 0.05$). Ces résultats ont montré la faiblesse de l'approche à base de proximité dans le cas d'un grand nombre de documents retournés. De plus, les améliorations de cette dernière par rapport à la probabiliste et la possibiliste ne sont pas statistiquement significatives ni pour le reste des valeurs de précision aux top documents, ni pour la *MAP* et la *R-Précision*.

Cependant, la proximité a montré sa signification statistique par rapport à l'approche probabiliste, en particulier pour les requêtes longues en utilisant «*Titre+desc +narr*» ($p - valeur = 0.012 < 0.05$) ou «*Titre+narr*» ($p - valeur = 0.006 < 0.05$) ou «*Desc+narr*» ($p - valeur = 0.046 < 0.05$). D'autre part, et en utilisant la partie «*Description*» de la requête source, l'amélioration obtenue par l'approche probabiliste par rapport à l'approche à base de proximité est statistiquement significative avec une $p - valeur = 0.004 < 0.05$. Enfin, la proximité a montré sa signification statistique par rapport à l'approche possibiliste, en particulier pour les requêtes longues en utilisant «*Titre+narr*» ($p - valeur = 0.006 < 0.05$). Par contre, l'approche possibiliste a assuré une amélioration statistiquement significative par rapport à la proximité particulièrement pour les requêtes longues utilisant «*Titre+desc*» ($p - valeur = 0.007 < 0.05$) et pour les requêtes courtes en utilisant «*Titre*» ($p - valeur = 0.036 < 0.05$) ou «*Description*» ($p - valeur = 0.005 < 0.05$).

Conclusion

Dans ce chapitre, nous avons présenté et testé une nouvelle approche pour la désambiguïsation de TR, en utilisant une technique de proximité bilingue. Dans cette approche, nous avons combiné un corpus parallèle avec un dictionnaire bilingue afin de constituer une nouvelle ressource linguistique à forte couverture, nommée, un dictionnaire sémantique bilingue de contexte (*DSBC*). Ce dernier est utile pour sélectionner la meilleure traduction correspondante à un terme polysémique dans une requête source, en utilisant un modèle de correspondance de proximité. En effet, ce modèle est basé sur le calcul de distance proxémique entre les nœuds d'un graphe sémantique bilingue, en étudiant la relation sémantique entre les termes et les définitions des traductions. Ce graphe nous a permis de calculer les scores de similarité sémantique des traductions afin de sélectionner la plus appropriée.

Les expérimentations de notre approche ont été réalisées à l'aide de la collection de test CLEF-2003 Français-Anglais. Les résultats ont montré que l'amélioration de l'approche à base de proximité bilingue est statistiquement significative lorsque nous avons comparé ses performances aux deux approches probabiliste et possibiliste à base de transformation probabilité-possibilité (Elayeb et al., 2018), en particulier lorsque nous utilisons de grandes informations contextuelles issues des requêtes longues.

Conclusion générale et perspectives

Synthèse des contributions

La recherche d'information est un domaine multidisciplinaire, il n'a pas cessé de s'améliorer afin de garantir l'information pertinente qui répond au besoin d'un utilisateur via un système appelé : Système de Recherche d'Information (SRI). Nous avons introduit au début de ce manuscrit le cadre général où nous avons présenté les concepts fondamentaux de la RI à savoir : les mots clés qui s'articulent autour de la définition d'un SRI, les étapes principales de fonctionnement d'un SRI pour l'automatisation de la recherche classique, les modèles existants et la description des mesures et des compagnes d'évaluation.

Avec l'augmentation exponentielle des informations multilingues suite à l'avènement du Web, les documents en langue étrangère sont considérés comme un «*bruit*» indésirable (Abusalah et Oakes, 2005). Ainsi, la nécessité de gérer plusieurs langues introduit des nouveaux domaines dans la RI tels que la RI multilingue et translinguistique. Ces derniers, prennent en compte tous les documents indépendamment des langues utilisées. Dans la RI classique, la requête et les documents sont dans la même langue. Dans la RI multilingue, les documents retournés sont écrits en plusieurs langues. Alors que la RIT permet à l'utilisateur d'énoncer sa requête dans une langue et de récupérer des documents dans une autre langue. Dans la présente thèse, nous nous sommes intéressés à la RIT. Cette dernière représente une intersection entre la RI et la traduction, donc plusieurs questions qui se posent dans ce cas à savoir : quelle méthode à adopter pour la traduction ? Comment sélectionner la meilleure traduction ? Est-ce qu'il faut traduire la requête ou les documents ou les deux ?

En effet, la plupart des études ont montré que la traduction des requêtes (TR) est la méthode la plus utilisée dans la RIT grâce à sa disponibilité linguistique, et son faible coût de calcul par rapport à la traduction d'une grande collection

de documents. Néanmoins, cette méthode soufre de plusieurs problèmes tels que l'ambiguïté des traductions suite à la quantité réduite du contexte dans les requêtes courtes et le manque de couverture dans les dictionnaires bilingues utilisés.

La thématique principale de cette thèse traite la TR dans les Systèmes de Recherche d'Information Translinguistique (SRIT). Nous avons essayé dans ce manuscrit d'améliorer la TR pour accéder aux informations textuelles pertinentes en anglais à l'exigence d'un utilisateur, exprimé sous forme d'une requête en français, à partir d'une grande collection électronique de documents. Nous avons, donc, étudié les différentes approches de TR existantes dans la littérature, les SRIT existants, ainsi une discussion des principaux avantages et inconvénients de chacun d'eux. En outre, nous avons présenté une étude sur la désambiguïsation monolingue et translinguistique. Cette étude nous a permis de révéler les pistes de recherche qui permettraient d'améliorer les travaux proposés.

Dans ce contexte, nous avons remarqué qu'il n'existe pas d'approches de TR basées sur la théorie des possibilités, alors que cette dernière s'applique naturellement à l'imprécision due au problème d'ambiguïté des traductions. La théorie des possibilités est relativement nouvelle et elle est parmi les théories les plus utilisées. Elle fournit des outils mathématiques pour le traitement des informations incomplètes et des données imprécises et incertaines (Moussa, 2014). Toutefois, nous avons exploité aussi la transformation probabilité-possibilité vu que ces deux théories sont similaires et il est difficile d'inférer empiriquement une distribution possibiliste. Parmi plusieurs transformations abordées, nous nous sommes inspirés de la transformation de Dubois et Prade (Dubois et al., 2004). Nous avons consacré tout un chapitre pour étudier les caractéristiques des théories des probabilités, des possibilités et les transformations probabilité-possibilité. De plus, nous avons présenté une étude comparative entre ces théories et ses domaines d'application.

Dans cette thèse, nous avons proposé quatre approches de TR, dans lesquelles nous avons essayé d'améliorer la qualité de traduction dans le domaine de RIT. Nous avons essayé de franchir les difficultés retrouvées dans les approches existantes dans la littérature, surtout au niveau de la désambiguïsation des traductions, le manque de couverture des dictionnaires bilingues et le manque de contexte provoqué par les termes des requêtes sources. Ces approches aspirent garantir une bonne performance au niveau des SRIT, en retournant des documents pertinents en réponse à un besoin d'un utilisateur.

Les trois premières approches sont basées principalement sur les mesures des possibilités et la transformation probabilité-possibilité. La quatrième approche est une approche à base de proximité, initialement proposée par (Gaume et al., 2004) et nous l'avons appliquée dans la désambiguïsation des traductions dans la RIT. Nous

rappelons ces approches brièvement ci-dessous :

Approche de TR à base de transformation probabilité-possibilité (Elayeb et al., 2018) : nous avons proposé une approche possibiliste de TR à base d'un dictionnaire bilingue Français-Anglais. Tout d'abord, nous avons enrichi la couverture du dictionnaire en générant les mots et leurs traductions automatiquement à partir d'un corpus parallèle. Le dictionnaire obtenu a été vérifié et enrichi via des correcteurs intelligents en ligne. En outre, nous avons pensé à garder le maximum d'informations contextuelles existantes dans les requêtes sources pour garantir des documents pertinents. Puis, nous avons identifié à partir de la requête source les syntagmes nominaux afin de les traduire en unités en utilisant des patrons de traduction et un modèle de langue (trigramme). Les termes de requête source qui ne font pas partie des syntagmes nominaux sélectionnés sont traduits mot-à-mot en utilisant une approche possibiliste de traduction des mots simples.

Approche possibiliste discriminative de désambiguïsation de TR (Ben Romdhane et al., 2017) : nous avons adopté une autre technique pour la sélection de la meilleure traduction. Dans cette approche, la pertinence des traductions est modélisée par deux mesures, à savoir : la pertinence possible et la pertinence nécessaire. La pertinence possible permet de rejeter les traductions non-pertinentes, alors que la pertinence nécessaire permet de renforcer les traductions non-éliminées par la possibilité.

Approche possibiliste hybride de TR (Ben Romdhane et al., 2019) : cette approche présente une combinaison des deux premières approches. Nous avons profité des points forts de l'approche à base de transformation probabilité-possibilité dans la traduction des syntagmes nominaux et les points forts de l'approche discriminative dans la traduction des termes restants de la requête.

Approche de désambiguïsation de TR à base de proximité bilingue (Ben Romdhane et al., 2018) : dans cette approche nous avons combiné un dictionnaire bilingue Français-Anglais avec un corpus parallèle Français-Anglais (*Europarl*), pour construire un *DSBC*. Cette nouvelle ressource linguistique assure un apprentissage automatique des requêtes. Notre approche est basée sur des méthodes stochastiques, où les termes des requêtes et leurs traductions sont représentés sous forme des graphes sémantiques bilingues. Ces graphes sont transformés en des chaînes de Markov, où les états sont les nœuds des graphes et les transitions sont ses arêtes. Nous avons utilisé le principe de *PROX* qui calcule un score de similarité sémantique (proximité) entre les termes polysémiques des requêtes sources afin d'identifier la traduction appropriée de chaque terme de la requête source.

Système possibiliste de traduction de requêtes SPORT en RIT : nous avons conçu un SRIT SPORT (Système POssibiliste de tRaduction de requêTes) (Elayeb et al., 2018) pour évaluer nos contributions. Nous avons implémenté les interfaces de ce

système en langage java et nous avons intégré la plate-forme *Terrier* qui fournit des modèles d'indexation et d'appariement. Dans notre cas, nous avons évalué les résultats en utilisant le modèle OKAPI BM25 et la collection de test CLEF-2003.

Dans les différentes validations expérimentales, nous avons utilisé le corpus parallèle Français-Anglais *Europarl* et un dictionnaire bilingue Français-Anglais.

Afin d'évaluer la performance de nos approches proposées et de les positionner par rapport à ses concurrentes, nous avons appliqué différents scénarios et métriques, tels que : les courbes de rappel-précision, les valeurs de précision aux top documents, la précision moyenne, la précision exacte, le pourcentage d'amélioration ainsi qu'une étude de significativité statistique des améliorations assurées par nos approches.

Nous avons établi des études comparatives entre nos approches proposées avec les approches probabiliste naïve bayésienne et monolingue. Les évaluations ont montré que l'approche possibiliste à base de transformation probabilité-possibilité est légèrement meilleure à l'approche probabiliste. En fait, les requêtes longues montrent une avance en faveur de l'approche possibiliste, surtout dans les parties «*narration*» et «*description+narration*», grâce à ses informations contextuelles qui favorisent l'identification et la traduction des syntagmes nominaux. De plus, les expérimentations menées prouvent que l'amélioration de l'approche discriminative est statistiquement significative par rapport à l'approche probabiliste en utilisant les requêtes courtes, surtout la partie «*Titre*» ; où le contexte est limité à un petit ensemble des termes dans lesquels l'identification des syntagmes nominaux semble rare. Au-delà, nous avons comparé l'approche possibiliste hybride aux approches possibiliste à base de transformation, discriminative et probabiliste. Les résultats globaux confirment que l'amélioration de l'approche hybride est statistiquement significative par rapport aux autres approches. De plus, l'approche hybride semble plus confortable en utilisant les requêtes longues ; où le contexte est important. D'autre part, l'amélioration de l'approche à base de proximité bilingue est statistiquement significative par rapport aux approches probabiliste et possibiliste à base de transformation probabilité-possibilité. Grâce au double support provenant du contexte maximal des requêtes sources et du *DSBC*, la performance de l'approche à base de proximité est plus proche de l'approche monolingue, surtout, lorsque nous utilisons des grandes informations contextuelles issues des requêtes longues.

Perspectives de recherche

Les résultats des expérimentations et les évaluations menées, découlent plusieurs perspectives que nous synthétisons dans ce qui suit :

Nous estimons tester nos approches de TR sur le couple de langue Anglais-Arabe. Cependant, cette extension semble difficile parce que la désambiguïsation des traductions est encore un défi majeur (Elayeb et Bounhas, 2016) pour la langue arabe. Pour ce faire, nous avons besoin non seulement d'un standard de test pertinent pour l'arabe, mais aussi d'une structuration des dictionnaires exhaustifs bilingues en arabe avec une large couverture.

En fait, quelques travaux existants suggèrent des solutions encourageantes pour structurer les dictionnaires arabes (Khemakhem et al., 2013) et pour la désambiguïsation des sens de mots arabes (Zouaghi et al., 2012). De plus, (Soudani et al., 2014) ont proposé une approche de normalisation générique des dictionnaires arabes utiles pour la désambiguïsation des requêtes arabes. Par ailleurs, (Ayed et al., 2012 ; Bounhas et al., 2015b, 2015a) ont proposé une approche de classification possibiliste de désambiguïsation morphologique des textes arabes, qui peut être utile pour la traduction et la désambiguïsation des requêtes arabes. Bounhas et al. (2011, 2014) et Lahbib et al. (2013) ont proposé et évalué plusieurs approches pour l'extraction de terminologies en langue arabe, qui sont utiles pour construire une ontologie bilingue à partir d'un corpus Arabe-Anglais. Ces terminologies bilingues peuvent être utilisées dans une plate-forme Arabe-Anglais dans le domaine de RIT. En fait, cette plate-forme peut aider l'utilisateur à formuler, désambiguïser et traduire sa requête avant de retourner les documents pertinents. Nous pouvons profiter à court terme des autres collections des textes parallèles, à savoir : (i) de l'ONU collecté, nettoyé et aligné à partir du site de l'ONU par (Xu et al., 2001) et (ii) le corpus Arabe TREC-2001/2002.

Il est également intéressant d'évaluer l'impact de la désambiguïsation de TR avant/après l'expansion de la requête sur les performances de notre SRIT en exploitant les approches récentes telles que (Ben Khiroun et al., 2014 ; Elayeb et al., 2015a, 2015b). D'un autre côté, nous prévoyons prendre en considération un feedback direct de l'utilisateur, notamment sur les requêtes d'un domaine spécifique, pour ce faire, nous pouvons tirer parti des techniques d'évaluation directe de la tâche de traduction comme BLEU (Papineni et al., 2002).

À long terme, nous envisagerons des nouvelles approches de désambiguïsation des traductions. Ces approches se basent sur des ressources non-classiques telles que le DSBC. L'hybridation de plusieurs ressources s'avère enrichissante pour la désambiguïsation dans la RIT.

Bibliographie

(Abdelali et al. 2004) Abdelali, Ahmed, James R. Cowie, David Farwell, et William Ogden. 2004. « UCLIR : a Multilingual Information Retrieval Tool ». *Inteligencia artificial revista iberoamericana de inteligencia artificial* 8 (22) : 103–110.

(Abusalah et Oakes 2005) Abusalah, Mustafa, John Tait et Michael Oakes. 2005. « Literature Review of Cross Language Information Retrieval ». In *The Second World Enformatika Conference (WEC'05)*, 4 :175-177. Istanbul, Turkey.

(Adriani 2000) Adriani, Mirna. 2000. « Using statistical term similarity for sense disambiguation in cross-language information retrieval ». *Information retrieval 2* (1) : 71–82.

(Afli 2014) Afli, Haithem. 2014. « La Traduction automatique statistique dans un contexte multimodal ». Thèse, France : Université du Maine.

(Agirre et Edmonds 2007) Agirre, Eneko, et Philip Edmonds. 2007. *Word Sense Disambiguation Algorithms and Applications*. 1re éd. Springer Publishing Company.

(Agirre et al., 2014) Agirre, Eneko, Oier Lopez de Lacalle, et Aitor Soroa. 2014. « Random walks for knowledge-based word sense disambiguation ». *Computational Linguistics* 40 (1) : 57–84.

(Aljlayl et Frieder 2001) Aljlayl, Mohammed, et Ophir Frieder. 2001. « Effective Arabic-English Cross-Language Information Retrieval via Machine Readable Dictionaries and Machine Translation ». In *Proceedings of the tenth international conference on Information and knowledge management (CIKM'01)*, 295-302. Atlanta, Georgia, USA : ACM.

(Alsahwa 2014) Alsahwa, Bassem. 2014. « Représentation d'un environnement par un système multi-capteurs : fusion et interprétation de scène ». Thèse, Télécom Bretagne ; Université de Rennes 1.

(Audibert 2003a) Audibert, Laurent. 2003. « Étude des critères de désambiguïsation sémantique automatique : résultats sur les cooccurrences ». In *10ème conférence sur le Traitement Automatique des Langues Naturelles (TALN-2003)*, 415-524.

(Audibert 2003b) Audibert, Laurent. 2003. « Outils d'exploration de corpus et désambiguïsation lexicale automatique ». Thèse, France : Université de Provence-Aix-Marseille I.

(Ayed et al. 2012) Ayed, Raja, Ibrahim Bounhas, Bilel Elayeb, Fabrice Evrard, et Narjès Bellamine Ben Saoud. 2012. « Arabic Morphological Analysis and Disambiguation Using a Possibilistic Classifier ». In *Proceedings of the 8th international conference on Intelligent Computing Theories and Applications (ICIC'12)*, 274–279. Huangshan, China : Intelligent Computing Theories and Applications.

(Azarbonyad, Shakery, et Faili 2013) Azarbonyad, Hosein, Azadeh Shakery, et Heshaam Faili. 2013. « Exploiting multiple translation resources for english-persian cross language information retrieval ». In *Proceedings of the 4th International Conference of the Cross-Language Evaluation*, 8138 :93–99. Valencia, Spain : Springer.

(Baccini et al. 2010) Baccini, Alain, Sébastien Déjean, Désiré Kompaoré, et Josiane Mothe. 2010. « Analyse des critères d'évaluation des systèmes de recherche d'information. » *Technique et Science Informatiques* 29 (3) : 289–308.

(Baeza-Yates et Ribeiro-Neto 2011) Baeza-Yates, Ricardo, et Berthier Ribeiro-Neto. 2011. *Modern information retrieval*. Pearson Education Limited 2011.

(Ballesteros et Croft 1998) Ballesteros, Lisa, et W. Bruce Croft. 1998. « Resolving ambiguity for cross-language retrieval ». In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 64–71. ACM.

(Banerjee et Pedersen 2002) Banerjee, Satanjeev, et Ted Pedersen. 2002. « An adapted Lesk algorithm for word sense disambiguation using WordNet ». In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'02)*, 136–145. Springer.

(Banerjee et Pedersen 2003) Banerjee, Satanjeev, et Ted Pedersen. 2003. « Extended Gloss Overlaps as a Measure of Semantic Relatedness ». In *Proceedings of the Eighteenth International Conference on Artificial Intelligence (IJCAI'03)*. Acapulco, Mexico : Morgan Kaufmann Publishers.

(Barathi et Valli 2010) Barathi, M., et S. Valli. 2010. « Ontology based query expansion using word sense disambiguation ». *International Journal of Computer Science and Information Security* 7 (2) : 22-27.

(Basile, Caputo, et Semeraro 2014) Basile, Pierpaolo, Annalina Caputo, et Giovanni Semeraro. 2014. « An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model ». In *Proceedings of the 25th International Conference on Computational Linguistics : Technical Papers*, 1591–1600. Dublin, Ireland.

(Baudrit 2005) Baudrit, Cédric. 2005. « Représentation et propagation de connaissances imprécises et incertaines : Application à l'évaluation des risques liés aux sites et aux sols pollués ». Thèse, France : IRIT - Institut de recherche en informatique de Toulouse.

(Ben Amor et al. 2002) Ben Amor, Nahla, Khaled Mellouli, Salem Benferhat, Didier Dubois, et Henri Prade. 2002. « A theoretical framework for possibilistic independence in a weakly ordered setting ». *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10 (02) : 117–155.

(Ben Khiroun 2018) Ben Khiroun, Oussama. 2018. « Recherche d'Information Monolingue & Translinguistique : de la Désambiguïsation vers l'Expansion Sémantique de Requêtes ». La Manouba, Tunisie : Université de Manouba ; Ecole Nationale des Sciences de l'Informatique.

(Ben Khiroun et al., 2018) Ben Khiroun, Oussama, Bilel ElAyeb, et Narjès Bellamine Ben Saoud. 2018. « Towards a Query Translation Disambiguation Approach using Possibility Theory ». In *Proceedings of the 10th International Conference on Agents and Artificial Intelligence*, 606-13. Funchal, Madeira, Portugal.

(Ben Khiroun et al., 2014) Ben Khiroun, Oussama, Bilel Elayeb, Ibrahim Bounhas, Fabrice Evrard, et Narjès Bellamine Ben Saoud. 2014. « Improving Query Expansion by Automatic Query Disambiguation in Intelligent Information Retrieval ». In *Proceedings of the 6th International Conference*

on Agents and Artificial Intelligence (ICAART'2014), 1 :153–160. Angers, Loire Valley, France : ESEO.

(Ben Khiroun et al., 2011) Ben Khiroun, Oussama, Bilel Elayeb, Ibrahim Bounhas, Fabrice Evrard, et Narjès Bellamine Ben Saoud. 2011. « A possibilistic approach for semantic query expansion ». In *Proceedings of the 4th international conference on Internet Technologies and Applications (ITA'2011)*, 308–316. Wrexham Wales, UK.

(Ben Romdhane et al., 2019) Ben Romdhane, Wiem, Bilel Elayeb et Narjès Bellamine Ben Saoud. 2019. « A Comparative Study Between Possibilistic and Probabilistic Approaches for Query Translation Disambiguation ». In *Proceedings of the 11th International Conference on Agents and Artificial Intelligence (ICAART), NLPinAI session*, 932-943, Prague, Czech Republic.

(Ben Romdhane et al., 2017) Ben Romdhane, Wiem, Bilel Elayeb, et Narjès Bellamine Ben Saoud. 2017. « A Discriminative Possibilistic Approach for Query Translation Disambiguation ». In *Proceedings of the 22th International Conference on Application of Natural Language to Information Systems (NLDB)*, 366-79. Liège, Belgium : Springer International Publishing Switzerland, LNCS 10260.

(Ben Romdhane et al., 2018) Ben Romdhane, Wiem, Bilel ElAyeb, et Narjès Bellamine Ben Saoud. 2018. « Automatic Query Translation Disambiguation using Bilingual Proximity-based Approach ». In *OTM Conferences & Workshops : Proceedings of 7th International Workshop on Methods, Evaluation, Tools and Applications for the Creation and Consumption of Structured Data for the e-Society (Meta4eS'18)*, 218-229. Valletta, Malta : Springer, LNCS.

(Ben Romdhane et al., 2013) Ben Romdhane, Wiem, Bilel Elayeb, Ibrahim Bounhas, Fabrice Evrard, et Narjès Bellamine Ben Saoud. 2013. « A Possibilistic Query Translation Approach for Cross-Language Information Retrieval ». In *Proceedings of the 9th International Conference on Intelligent Computing Theories and Technology, ICIC'2013*, 73–82. Nanning, China : Springer-Verlag Berlin.

(Ben Slimen et al., 2013) Ben Slimen, Yosra, Raouia Ayachi, et Nahla Ben Amor. 2013. « Probability-Possibility Transformation : Application to Bayesian and Possibilistic Networks ». In *Fuzzy logic and applications*, 122-30. Lecture Notes in Computer Science. Genoa, Italy.

(Benferhat et al., 1999) Benferhat, Salem, Didier Dubois, Laurent Garcia, et Henri Prade. 1999. « Possibilistic logic bases and possibilistic graphs ». In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, 57–64. Morgan Kaufmann Publishers Inc.

(Benferhat et al., 2017) Benferhat, Salem, Amélie Levray, et Karim Tabia. 2017. « Approximation de l'inférence MAP via les transformations probabilistes-possibilistes ». In *11ème Journées d'Intelligence Artificielle Fondamentale*. Caen, France.

(Benferhat et Smaoui 2007) Benferhat, Salem, et Salma Smaoui. 2007. « Hybrid Possibilistic Networks ». *International Journal of Approximate Reasoning* 44 (3) : 224-43.

(Bian et Teng 2008) Bian, Shun-Yuan, et Shun-Yuan Teng. 2008. « Integrating query translation and text classification in a cross-language patent access system ». In *NTCIR-7 Workshop Meeting*. Tokyo, Japan.

(Billami 2015) Billami, Mokhtar. 2015. « Désambiguisation lexicale à base de connaissances par sélection distributionnelle et traits sémantiques ». In *22ème Conférence sur le Traitement Automatique des Langues Naturelles et 17ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, 13--24. Caen, France.

(Blain 2013) Blain, Frédéric. 2013. « Modèles de traduction évolutifs ». Thèse, France : Université du Maine.

(Blei et al., 2003) Blei, David M., Andrew Y. Ng, et Michael I. Jordan. 2003. « Latent dirichlet allocation ». *Journal of machine Learning research* 3 (3-1) : 993–1022.

(Borgelt et Gebhardt 1999) Borgelt, Christian, et Jorg Gebhardt. 1999. « A naive bayes style possibilistic classifier ». In *Proceedings of the 7th European Congress on Intelligent Techniques and Soft Computing(EUFIT'99)*. Aachen, Germany.

(Boubekeur 2008) Boubekeur, Fatiha. 2008. « Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets ». Thèse, France : Université Paul Sabatier-Toulouse III.

(Boughanem et al., 2004) Boughanem, Mohand, Wessel Kraaij, et Jian-Yun Nie. 2004. « Modèles de langue pour la recherche d'information ». In *Les systèmes de recherche d'informations*, 163-82. Hermès, Paris.

(Bouhini et al., 2014) Bouhini, Chahrazed, Mathias Géry, et Christine Largeron. 2014. « Integrating user's profile in the query model for Social Information Retrieval ». In *2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS)*, 1–2. Marrakech, Morocco : IEEE.

(Bounhas et al., 2015) Bounhas, Ibrahim, Raja Ayed, Bilel Elayeb, et Narjès Bellamine Ben Saoud. 2015. « A Hybrid Possibilistic Approach for Arabic Full Morphological Disambiguation ». *Data & Knowledge Engineering* 100 (PB) : 240-54.

(Bounhas et al., 2015) Bounhas, Ibrahim, Raja Ayed, Bilel Elayeb, Fabrice Evrard, et Narjès Bellamine Ben Saoud. 2015. « Experimenting a Discriminative Possibilistic Classifier with Reweighting Model for Arabic Morphological Disambiguation ». *Computer Speech & Language* 33 (1) : 67-87.

(Bounhas et al., 2011) Bounhas, Ibrahim, Bilel Elayeb, Fabrice Evrard, et Yahya Slimani. 2011. « ArabOnto : experimenting a new distributional approach for building Arabic ontological resources ». *International Journal of Metadata, Semantics and Ontologies* 6 (2) : 81–95.

(Bounhas 2013) Bounhas, Myriam. 2013. « Possibilistic Classifiers for Certain/Uncertain Numerical Data ». Thèse, Tunisie : Institut Supérieur de Gestion.

(Bounhas et al,. 2014) Bounhas, Myriam, Mohammad Ghasemi Hamed, Henri Prade, Mathieu Serrurier, et Khaled Mellouli. 2014. « Naive possibilistic classifiers for imprecise or uncertain numerical data ». *Fuzzy Sets and Systems*, Elsevier North-Holland, 239 (mars) : 137-56.

(Bounhas et al., 2013) Bounhas, Myriam, Khaled Mellouli, Henri Prade, et Mathieu Serrurier. 2013. « Possibilistic Classifiers for Numerical Data ». *Soft Computing* 17 (5) : 733-51.

(Bouramoul 2011) Bouramoul, Abdelkrim. 2011. « Recherche d'information contextuelle et sémantique sur le web ». Thèse, Algérie : Université MENTOURI de Constantine.

(Brini 2005) Brini, Asma. 2005. « Un Modèle de Recherche d'Information basé sur les Réseaux Possibilistes ». Thèse, France : Université Paul Sabatier de Toulouse.

(Brown et al., 1990) Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John Djohn D. fff Lafferty, Robert L Mercer, et Paul S. Roossin. 1990. « A Statistical Approach To Machine Translation ». *Computational Linguistics*, MIT Press Cambridge, MA, USA, 16 (2) : 79-85.

(Brown et al., 1993) Brown, Peter, F., S A Pietra, V. J Pietra, et R. L. Mercer. 1993. « The Mathematics Of Machine Translation : Parameters Estimation ». *Computational Linguistics - Special issue on using large corpora*, 1993.

(Brun et al., 2001) Brun, Caroline, Bernard Jacquemin, et Frédérique Segond. 2001. « Exploitation de dictionnaires électroniques pour la désambiguïsation sémantique lexicale ». *Traitement Automatique des Langues*, ATALA 42 (3) : 667–690.

(Buckley et Voorhees, 2000) Buckley, Chris, et Ellen M. Voorhees. 2000. « Evaluating evaluation measure stability ». In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 33-40. Athens, Greece : ACM.

(Callan et al., 1992) Callan, James P, W. Bruce Croft, et Stephen M Harding. 1992. « The INQUERY Retrieval System ». In *Proceedings of the Third International Conference on Database and Expert Systems Applications*. Valencia, Spain : Springer-Verlag.

(Capstick et al., 2000) Capstick, Joanne, Abdel Kader Diagne, Gregor Erbach, Hans Uszkoreit, Anne Leisenberg, et Manfred Leisenberg. 2000. « A system for supporting cross-lingual information retrieval ». *Information processing & management* 36 (2) : 275–289.

(Chaplot et Salakhutdinov, 2018) Chaplot, Devendra Singh, et Ruslan Salakhutdinov. 2018. « Knowledge-based Word Sense Disambiguation using Topic Models ». *Association for the Advancement of Artificial Intelligence*, janvier.

(Chebil et al., 2016) Chebil, Wiem, Lina Fatima Soualmia, Mohamed Nazih Omri, et Stéfan Jacques Darmoni. 2016. « Indexing biomedical documents with a possibilistic network ». *Journal of the Association for Information Science and Technology* 67 (4) : 928–941.

(Chen et al., 2015) Chen, Tao, Ruifeng Xu, Yulan He, et Xuan Wang. 2015. « Improving Distributed Representation of Word Sense via WordNet Gloss Composition and Context Clustering ». In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 2 :15–20. Beijing, China : Association for Computational Linguistics.

(Cohen, 1968) Cohen, Jacob. 1968. « Weighted kappa : Nominal scale agreement provision for scaled disagreement or partial credit. » *Psychological bulletin*, 1968.

(Darwish et Magdy, 2014) Darwish, Kareem, et Douglas W. Oard. 2003. « Probabilistic structured query methods ». In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 338–344. Toronto, Canada : ACM New York, NY, USA.

(Davis et Ogden 1997) Davis, Mark W., et William C. Ogden. 1997. « Free resources and advanced alignment for cross-language text retrieval ». In *TREC*, 1997 :385–395. Citeseer.

(De Cooman et Aeyels, 1999) De Cooman, Gert, et Dirk Aeyels. 1999. « Supremum preserving upper probabilities ». *Information Sciences*, Elsevier Science Inc. New York, NY, USA, 118 (1-4) : 173–212.

(De Cooman et Aeyels, 2000) De Cooman, Gert, et Dirk Aeyels. 2000. « A random set description of a possibility measure and its natural extension ». *IEEE Transactions on Systems Man and Cybernetics-Part A : Systems and Humans* 30 (2) : 124–130.

(De Luca et al., 2006) De Luca, Ernesto William, Ernesto William, Stefan Hauke De Luca, Andreas Nürnberger, et Stefan Schlechtweg. 2006. « Multilexexplorer : Combining multilingual web search

with multilingual lexical resources ». In *Proceedings of the Combined Workshop on Language-Enhanced Educational Technology and Development and Evaluation of Robust Spoken Dialogue Systems*, 17–21.

(Dubois et al., 2004) Dubois, Didier, Laurent Foulloy, Gilles Mauris, et Henri Prade. 2004. « Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities ». *Reliable computing* 10 (4) : 273–297.

(Dubois et Prade, 1985) Dubois, Didier, et Henri Prade. 1985. « Unfair coins and necessity measures towards a possibilistic interpretation of histograms ». *Fuzzy Sets and Systems* 10 (1-3) : 15-20.

(Dubois et Prade, 1992) Dubois, Didier, et Henri Prade. 1992. « When upper probabilities are possibility measures ». *Fuzzy sets and systems*, Elsevier North-Holland, Inc. Amsterdam, The Netherlands, The Netherlands, 49 (1) : 65–74.

(Dubois et Prade, 1998) Dubois, Didier, et Henri Prade. 1998. « Possibility Theory : Qualitative and Quantitative Aspects ». In *Quantified Representation of Uncertainty and Imprecision*, 169-226.

(Dubois et Prade, 2000) Dubois, Didier, et Henri Prade. 2000. « An Overview of Ordinal and Numerical Approaches to Causal Diagnostic Problem Solving ». In *Abductive Reasoning and Learning*, 4 :231-80. Dordrecht : Handbook of Defeasible Reasoning and Uncertainty Management Systems.

(Dubois et Prade, 2006) Dubois, Didier, et Henri Prade. 2006. « La théorie des possibilités ». Université Paul Sabatier.

(Dubois et Prade, 2009) Dubois, Didier, et Henri Prade. 2009. « Formal representations of uncertainty ». In *Decision-Making Process : Concepts and Methods*, 85–156.

(Dubois et Prade, 2015) Dubois, Didier, et Henri Prade. 2015. « Possibility Theory and Its Applications : Where Do We Stand ? » In *Springer Handbook of Computational Intelligence*, Springer Handbook of Computational Intelligence, 31-60. Berlin, Heidelberg.

(Dubois et Prade, 2016) Dubois, Didier, et Henri Prade. 2016. « Practical Methods for Constructing Possibility Distributions ». *International Journal of Intelligent Systems* 31 (3) : 215-39.

(Dubois et al., 2004) Dubois, Didier, Henri Prade, et Sandra Sandri. 1993. « On possibility/probability transformations ». *Fuzzy logic*, 103–112.

(Duque et al., 2015) Duque, Andres, Lourdes Araujo, et Juan Martinez-Romo. 2015. « CO-Graph : A New Graph-Based Technique for Cross-Lingual Word Sense Disambiguation ». *Natural Language Engineering* 21 (5) : 743-72.

(Elayeb, 2009) Elayeb, Bilel. 2009. « SARIPOD : Système multi-agent de recherche intelligente possibiliste de documents web ». Thèse, Institut National Polytechnique de Toulouse-INPT ; Université de Manouba ; Ecole Nationale des Sciences de l'Informatique.

(Elayeb et al., 2018) Elayeb, Bilel, Wiem Ben Romdhane, et Narjès Bellamine Ben Saoud. 2018. « Towards a new possibilistic query translation tool for cross-language information retrieval ». *Multimedia Tools and Applications*, Springer, 77 (2) : 2423–2465.

(Elayeb et Bounhas, 2016) Elayeb, Bilel, et Ibrahim Bounhas. 2016. « Arabic Cross-Language Information Retrieval - A Review ». *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, ACM, 15 (3) : 1-44.

(Elayeb et al., 2011) Elayeb, Bilel, Ibrahim Bounhas, Oussama Ben Khiroun, Fabrice Evrard, et Narjès Bellamine-BenSaoud. 2011. « Towards a Possibilistic Information Retrieval System Using Semantic Query Expansion ». *International Journal of Intelligent Information Technologies* 7 (4) : 1-25.

(Elayeb et al., 2015a) Elayeb, Bilel, Ibrahim Bounhas, Oussama Ben Khiroun, Fabrice Evrard, et Narjès Bellamine Ben Saoud. 2015. « A Comparative Study between Possibilistic and Probabilistic Approaches for Monolingual Word Sense Disambiguation ». *Knowledge and Information Systems*, Springer-Verlag, 44 (1) : 91-126.

(Elayeb et al., 2015b) Elayeb, Bilel, Ibrahim Bounhas, Oussama Ben Khiroun, et Narjès Bellamine Ben Saoud. 2015. « Combining Semantic Query Disambiguation and Expansion to Improve Intelligent Information Retrieval ». In *Revised Selected Papers of the 6th International Conference on Agents and Artificial Intelligence (ICAART'2014)*, 8946 :280-95. Angers, France : Springer-Verlag Berlin.

(Elayeb et al., 2009) Elayeb, Bilel, Fabrice Evrard, Montaceur Zaghdoud, et Mohamed Ben Ahmed. 2009. « Towards an Intelligent Possibilistic Web Information Retrieval Using Multiagent System ». *Interactive Technology and Smart Education* 6 (1) : 40-59.

(Ellman et al., 2000) Ellman, Jeremy, Ian Klincke, et John Tait. 2000. « Word sense disambiguation by information filtering and extraction ». *Computers and the Humanities* 34 (1-2) : 127–134.

(Farag et Nurnberger, 2008) Farag, Ahmed, et Andreas Nurnberger. 2008. « Arabic/english word translation disambiguation using parallel corpora and matching schemes ». In *Proceedings of the 12th EAMT conference*, 8 :28. Hamburg, Germany.

(Farag et Nurnberger, 2012) Farag, Ahmed, et Andreas Nurnberger. 2012. « Literature review of interactive cross-language information retrieval tools ». *The international Arab journal of information technology (IAJIT)* 9 (5) : 479-86.

(Farag et Nurnberger, 2013) Farag, Saad, et Andreas Nurnberger. 2013. « Translation Ambiguity Resolution Using Interactive Contextual Information ». *Computational Linguistics*, 219-40.

(Fuhr, 1992) Fuhr, Norbert. 1992. « Probabilistic models in information retrieval ». *The Computer Journal*, Oxford University Press Oxford, UK, 35 (3) : 243–255.

(Fung et Del Favero, 1995) Fung, Robert, et Brendan Del Favero. 1995. « Applying Bayesian networks to information retrieval ». *Communications of the ACM*, ACM New York, NY, USA, 38 (3) : 42–50.

(Gao et al., 2001) Gao, Jianfeng, Jian-Yun Nie, Endong Xun, Jian Zhang, Ming Zhou, et Changning Huang. 2001. « Improving query translation for cross-language information retrieval using statistical models ». In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 96–104. New Orleans, Louisiana, USA : ACM.

(Gao et al., 2006) Gao, Jianfeng, Jian-Yun Nie, et Ming Zhou. 2006. « Statistical query translation models for cross-language information ». *ACM Transactions on Asian Language Information Processing (TALIP)* 5 (4) : 323-59.

(Garrouch et Omri, 2015) Garrouch, Kamel, et Mohamed Nazih Omri. 2015. « Possibilistic network based information retrieval model ». In *Proceedings of the 15th International Conference on Intelligent Systems Design and Applications (ISDA)*, 25–30. Marrakech, Morocco : IEEE.

(Gaume, 2004) Gaume, Bruno. 2004. « Balades aléatoires dans les petits mondes lexicaux ». *I3 information interaction intelligence*, Cépaduès 4 (2) : 39–96.

(Gaume et al., 2004) Gaume, Bruno, Nabil Hathout, et Philippe Muller. 2004. « Word sense disambiguation using a dictionary for sense similarity measure ». In *Proceedings of the 20th international conference on Computational Linguistics*. Geneva, Switzerland : Association for Computational Linguistics.

(Habash et Sadat, 2006) Habash, Nizar, et Fatiha Sadat. 2006. « Arabic preprocessing schemes for statistical machine translation ». In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume : Short Papers*, 49–52. Association for Computational Linguistics.

(Haddad et al., 2016) Haddad, Maroua, Philippe Leray, Amélie Levray, et Karim Tabia. 2016. « Possibilistic networks parameter learning : Preliminary empirical comparison ». In *8èmes journées francophones de réseaux bayésiens (JFRB 2016)*. Clermont-Ferrand, France.

(Harman, 1992) Harman, Donna. 1992. « The DARPA tipster project ». ACM.

(Hedlund et al., 2004) Hedlund, Turid, Eija Airio, Heikki Keskustalo, Raija Lehtokangas, Ari Pirkola, et Kalervo Järvelin. 2004. « Dictionary-based cross-language information retrieval : Learning experiences from CLEF 2000–2002 ». *Information retrieval*, Kluwer Academic Publishers Hingham, MA, USA, 7 (1-2) : 99–119.

(Heer et al., 2005) Heer, Jeffrey, Stuart K. Card, et James A. Landay. 2005. « Prefuse : a toolkit for interactive information visualization ». In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 421–430. Portland, Oregon, USA : ACM.

(Hefny et al., 2011) Hefny, Ahmed, Kareem Darwish, et Ali Alkahky. 2011. « Is a Query Worth Translating : Ask the Users ! » In *Advances in Information Retrieval*, édité par Paul Clough, Colum Foley, Cathal Gurrin, Gareth J. F. Jones, Wessel Kraaij, Hyowon Lee, et Vanessa Mudoch, 6611 :238-50. Berlin, Heidelberg : Springer Berlin Heidelberg.

(Huang et al., 2013) Huang, Po-Sen, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, et Larry Heck. 2013. « Learning deep structured semantic models for web search using clickthrough data ». In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, 2333–2338. San Francisco, CA, USA : ACM.

(Hull, 1993) Hull, David. 1993. « evaluation of retrieval experiments ». *In Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, 329–338. Pittsburgh, Pennsylvania, USA : ACM.

(Hull et Grefenstette, 1996) Hull, David A., et Gregory Grefenstette. 1996. « Querying across Languages : A Dictionary-Based Approach to Multilingual Information Retrieval ». In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96*, 49-57. Zurich, Switzerland : ACM Press.

(Ide et Véronis, 1998) Ide, Nancy, et Jean Véronis. 1998. « word sense disambiguation : the state of the art ». *Computational linguistics* 24 : 1–40.

(Jenhani, 2010) Jenhani, Ilyes. 2010. « From Possibilistic Similarity Measures To Possibilistic Decision Trees ». Thèse, France : Université d'Artois.

(Kadri, 2008) Kadri, Youssef. 2008. « Recherche d'information translinguistique sur les documents en arabe ». Thèse, Canada : Université de Montréal Faculté des études supérieures.

(Kadri et Nie, 2004) Kadri, Youssef, et Jian-Yun Nie. 2004. « Traduction des requêtes pour la recherche d'information translinguistique anglais-arabe ». In *JEP-TALN, session sur le traitement automatique de la langue arabe écrite et parlée*, 291. Fès, Maroc.

(Kadri et Nie, 2006) Kadri, Youssef, et Jian-Yun Nie. 2006. « Effective stemming for Arabic information retrieval ». In *proceedings of the Challenge of Arabic for NLP/MT Conference*. Londres, Royaume-Uni.

(Kadri et Nie, 2008) Kadri, Youssef, et Jian-Yun Nie. 2008. « A Comparative Study for Query Translation using Linear Combination and Confidence Measure ». In *IJCNLP*, 181-88.

(Kilgarriff, 1998) Kilgarriff, Adam. 1998. « SENSEVAL : An Exercise in Evaluating Word Sense Disambiguation Programs ». In *Proceedings of the European Conference on Lexicography (EURALEX)*, 174-76. Liège, Belgium.

(Klir et Geer, 1993) Klir, George J., et Games F. Geer, éd. 1993. « Information-Preserving Probability- Possibility Transformations : Recent Developments ». *Fuzzy Logic*, 417–428.

(Knight et Graehl, 1998) Knight, Kevin, et Jonathan Graehl. 1998. « Machine transliteration ». *Computational Linguistics*, MIT Press Cambridge, MA, USA, 24 (4) : 599–612.

(Koehn, 2005) Koehn, Philipp. 2005. « Europarl : A parallel corpus for statistical machine translation ». In *Proceedings of the tenth Machine Translation Summit*, 5 :79–86. Phuket, Thailand : Citeseer.

(Kompaoré, 2008) Kompaoré, Nongdo Désiré. 2008. « Fusion de systèmes et analyse des caractéristiques linguistiques des requêtes : vers un processus de RI adaptatif ». Thèse, France : Université Paul Sabatier-Toulouse III.

(Kwok, 2000) Kwok, K. L. 2000. « Exploiting a Chinese-English bilingual wordlist for English-Chinese cross language information retrieval ». In *Proceedings of the fifth international workshop on on Information retrieval with Asian languages*, 173–179. Hong Kong, China : ACM.

(Lahbib et al., 2013) Lahbib, Wiem, Ibrahim Bounhas, Bilel Elayeb, Fabrice Evrard, et Yahya Slimani. 2013. « A Hybrid Approach for Arabic Semantic Relation Extraction. » In *Proceedings of the 26th International Florida Artificial Intelligence Research Society (FLAIRS) Conference*. Florida, USA.

(Lee, 1998) Lee, Joon Ho. 1998. « Combining the evidence of different relevance feedback methods for Information Retrieval ». *Information Processing and Management : an International Journal*, Pergamon Press, Inc. Tarrytown, NY, USA, 34 (6) : 681-91.

(Lefever et Hoste, 2010) Lefever, Els, et Véronique Hoste. 2010. « SemEval-2010 Task 3 : Cross-Lingual Word Sense Disambiguation ». In *Proceedings of the 5th International Workshop on Semantic Evaluation* (*SemEval'10*), 15-20. Los Angeles, California : Geie-Ercim.

(Lefever et Hoste, 2013) Lefever, Els, et Véronique Hoste. 2013. « Semeval-2013 task 10 : Cross-lingual word sense disambiguation ». In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval'13)*, 158–166. Atlanta, Georgia, USA.

(Lesk, 1986) Lesk, Michael. 1986. « 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* (*SIGDOC'86*), 24–26. Toronto, Ontario, Canada : ACM.

(Levow et al., 2005) Levow, Gina-Anne, Douglas W. Oard, et Philip Resnik. 2005. « Dictionary-Based Techniques for Cross-Language Information Retrieval ». *Information Processing & Management* 41 (3) : 523-47.

(Lopez-Ostenero et al., 2003) Lopez-Ostenero, Fernando, Julio Gonzalo, Anselmo Penas, et Felisa Verdejo. 2003. « Interactive Cross-Language Searching : phrases are better than terms for query formulation and refinement ». In *Workshop of the Cross-Language Evaluation Forum for European Languages*.

(Manning et al., 2008) Manning, Christopher D., Prabhakar Raghavan, et Hinrich Schütze. 2008. *Introduction to information retrieval*. New York : Cambridge University Press.

(MarkW. Davia, 1997) MarkW. Davia. 1997. « QUILT Implementing a Large-scale Cross-language Text Retrieval System ». In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Philadelphia, PA, USA.

(Masson et Denoeux, 2006) Masson, Marie-Hélène, et Thierry Denoeux. 2006. « Inferring a possibility distribution from empirical data ». *Fuzzy sets and systems* 157 (3) : 319–340.

(Masterman, 1961) Masterman, Margaret. 1961. « Semantic message detection for machine translation, using an interlingua ». In *Proceedings of the International Conference on Machine Translation of Languages and Applied Language Analysis*, 438–475. Teddington, UK.

(McNamee et Mayfield, 2002) McNamee, Paul, et James Mayfield. 2002. « Comparing cross-language query expansion techniques by degrading translation resources ». In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 159–166. Tampere, Finland : ACM.

(Merhbene et al., 2014) Merhbene, Larousi, Anis Zouaghi, et Mounir Zrigui. 2014. « Approche basée sur les arbres sémantiques pour la désambiguïsation lexicale de la langue arabe en utilisant une procédure de vote ». In *21ème Traitement Automatique des Langues Naturelles*, 281–290. Marseille,France.

(Mihalcea et Moldovan, 1998) Mihalcea, Rada, et Dan Moldovan. 1998. « Word sense disambiguation based on semantic density ». In *Usage of wordnet in natural language processing systems*, 16–22. Montreal, Quebec, Canada.

(Mouchaweh, 2002) Mouchaweh, Moamar Sayed. 2002. « Conception d'un système de diagnostic adaptatif et prédictif basé sur la méthode Fuzzy Pattern Matching pour la surveillance en ligne des systèmes évolutifs Application à la supervision et au diagnostic d'une ligne de peinture au trempé ». Thèse, France : Université de Reims Champagne-Ardenne.

(Mouchaweh et al., 2006) Mouchaweh, Moamar Sayed, Mohamed Saïd Bouguelid, Patrice Billaudel, et Bernard Riera. 2006. « Variable probability-possibility transformation ». In *Proceedings of the 25th European Annual Conference on Human Decision-Making and Manual Control*, 122-130.

(Moulahi, 2015) Moulahi, Bilel. 2015. « Définition et évaluation de modèles d'agrégation pour l'estimation de la pertinence multidimensionnelle en recherche d'information ». Thèse, France : Université de Toulouse, Université Toulouse III-Paul Sabatier.

(Moussa, 2014) Moussa, Alfred Mbairadjim. 2014. « Transformation Probabilité-Possibilité et raisonnement statistique : Application à la Finance ».

(Navigli et Lapata, 2010) Navigli, Roberto, et Mirella Lapata. 2010. « An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation ». *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32 (4) : 678-92.

(Navigli, 2009) Navigli, Roberto. 2009. « Word Sense Disambiguation : A Survey ». *ACM Computing Surveys* 41 (2) : 1-69.

(Navigli et Ponzetto, 2012) Navigli, Roberto, et Simone Paolo Ponzetto. 2012. « Joining forces pays off : Multilingual joint word sense disambiguation ». In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, 1399–1410. Jeju Island, Korea : Association for Computational Linguistics.

(Nguyen et al., 2017) Nguyen, Gia-Hung, Laure Soulier, Lynda Tamine, et Nathalie Bricon-Souf. 2017. « Modèle Neuronal de Recherche d'Information Augmenté par une Ressource Sémantique ». In *Conférence francophone en Recherche d'Information et Applications (CORIA 2017)*. Marseille, France.

(Nie, 1999) Nie, Jian-Yun. 1999. « CLIR using a Probabilistic Translation Model based on Web Documents ». In *TREC*.

(Nie, 2010) Nie, Jian-Yun. 2010. Cross-Language Information Retrieval. Morgan and Claypool Publishers.

(Oard et al., 2008) Oard, Douglas W., Daqing He, et Jianqiang Wang. 2008. « User-Assisted Query Translation for Interactive Cross-Language Information Retrieval ». *Information Processing & Management* 44 (1) : 181-211.

(Ogden et Davis, 2000) Ogden, William C., et Mark W. Davis. 2000. « Improving cross-language text retrieval with human interactions ». In *Proceedings of the 33rd Hawaii International Conference on System Sciences*, 3 :3044. IEEE.

(Olive et al., 2011) Olive, Joseph, Caitlin Christianson, et John McCary, éd. 2011. *Handbook of Natural Language Processing and Machine Translation*. New York, NY : Springer New York.

(Ounis et al., 2006) Ounis, Iadh, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald et Christina Lioma. 2006. « Terrier : A high performance and scalable information retrieval platform ». In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*.

(Oussalah, 2000) Oussalah, Mourad. 2000. « On The Probability/Possibility Transformations :A Comparative Analysis ». *International Journal of General Systems* 29 (5) : 671-718.

(Papineni et al., 2002) Papineni, Kishore, Salim Roukos, Todd Ward, et Wei-Jing Zhu. 2002. « BLEU : a method for automatic evaluation of machine translation ». In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Philadelphia, Pennsylvania : Association for Computational Linguistics.

(Patry et Langlais, 2005) Patry, Alexandre, et Philippe Langlais. 2005. « Paradocs : un système d'identification automatique de documents parallèles ». In *12e Conference sur le Traitement Automatique des Langues Naturelles (TALN)*, 223–232. Dourdan, France.

(Pirkola et al., 2001) Pirkola, Ari, Turid Hedlund, Keskustalo Heikki, et Kalevro Jarvelin. 2001. « Dictionary-Based Cross-Language Information Retrieval Problems, Methods, and Research Findings ». *Information retrieval*, Kluwer Academic Publishers Hingham, MA, USA, 4 (3-4) : 209-30.

(Pirkola et al., 2002) Pirkola, Ari, Heikki Keskustalo, Erkka Leppänen, Antti-Pekka Känsälä, et Kalervo Järvelin. 2002. « OPAC3.html ». 2002.

(Pirkola et al., 2003) Pirkola, Ari, Deniz Puolamäki, et Kalervo Järvelin. 2003. « Applying Query Structuring in Cross-Language Retrieval ». *Information Processing & Management* 39 (3) : 391-402.

(Ploux et Victorri, 1998) Ploux, Sabine, et Bernard Victorri. 1998. « Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes ». *Traitement automatique des langues* 39 (1) : 161–182.

(Ponzetto et Navigli, 2010) Ponzetto, Simone Paolo, et Roberto Navigli. 2010. « Knowledge-rich word sense disambiguation rivaling supervised systems ». In *Proceedings of the 48th annual meeting of the association for computational linguistics*, 1522–1531. Uppsala, Sweden : Association for Computational Linguistics.

(Pourmahmoud et Shamsfard, 2008) Pourmahmoud, Solmaz, et Mehrnoush Shamsfard. 2008. « Semantic cross-lingual information retrieval ». In *Proceedings of the 23rd International Symposium on Computer and Information Sciences*, 1–4. Istanbul, Turkey : IEEE.

(Ranjan Pal et Saha, 2015) Ranjan Pal, Alok, et Diganta Saha. 2015. « Word Sense Disambiguation : A Survey ». *International Journal of Control Theory and Computer Modeling* 5 (3) : 1-16.

(Reymond, 2001) Reymond, Delphine. 2001. « Dictionnaires distributionnels et étiquetage lexical de corpus ». In *Actes de la Conférence Traitement Automatique des Langues (RECITAL'2001)*, 479–488. Tours, France.

(Reymond, 2002) Reymond, Delphine. 2002. « Méthodologie pour la création d'un dictionnaire distributionnel dans une perspective d'étiquetage lexical semi-automatique ». In *Actes de RECITAL (TALN)*, 1 :405–414. Nancy, France.

(Robertson et al., 1995) Robertson, Stephen E., Steve Walker, Susan Jones, Micheline M. Hancock-Beaulieu, et Mike Gatford. 1995. « Okapi at TREC-3 ». In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, 109-26. National Institute of Standards and Technology (NIST).

(Robertson et al., 2004) Robertson, Stephen, Hugo Zaragoza, et Michael Taylor. 2004. « Simple BM25 extension to multiple weighted fields ». In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, 42–49. Washington, D.C., USA : ACM.

(Salton, 1971) Salton, Gerard. 1971. *The SMART retrieval system ; experiments in automatic document processing*. New Jersey.

(Salton et al., 1983) Salton, Gerard, Edward A. Fox, et Harry Wu. 1983. « Extended Boolean information retrieval ». *Communications of the ACM*, 1983.

(Saralegi et De Lacalle, 2010) Saralegi, Xabier, et Maddalen Lopez De Lacalle. 2010. « Dictionary and Monolingual Corpus-based Query Translation for Basque-english ». In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta.

(Sarmah et Sarma, 2016) Sarmah, Jumi, et Shikhar Kumar Sarma. 2016. « Survey on Word Sense Disambiguation : An Initiative towards an Indo-Aryan Language ». *IJEM* 6 (3) : 37-52.

(Schutze, 1992) Schutze, Hinrich. 1992. « Dimensions of meaning ». In *Proceedings of the 1992 ACM/IEEE conference on Supercomputing*, 787–796. Minneapolis, Minnesota, USA : IEEE.

(Segond, 2000) Segond, Frédérique. 2000. « Framework and results for French ». *Computers and the Humanities*, Springer, 34 (1-2) : 49–60.

(Severyn et Moschitti, 2015) Severyn, Aliaksei, et Alessandro Moschitti. 2015. « Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks ». In *Proceedings of the 38th*

International ACM SIGIR Conference on Research and Development in Information Retrieval, 373-82. Santiago, Chile : ACM Press.

(Sharma et Mittal, 2018) Sharma, Vijay Kumar, et Namita Mittal. 2018. « Cross-Lingual Information Retrieval : A Dictionary-Based Query Translation Approach ». In *Advances in Computer and Computational Sciences*, édité par Sanjiv K. Bhatia, Krishn K. Mishra, Shailesh Tiwari, et Vivek Kumar Singh, 554 :611-18. Singapore : Springer Singapore.

(Soudani et al., 2014a) Soudani, Nadia, Ibrahim Bounhas, Bilel ElAyeb, et Yahya Slimani. 2014. « A Generic Normalization Approach of Arabic Dictionaries for Arabic Word Sense Disambiguation ». In *Actes des 5èmes journées Francophones sur les Ontologies*. Hammamet, Tunisie.

(Soudani et al., 2014b) Soudani, Nadia, Ibrahim Bounhas, Bilel ElAyeb, et Yahya Slimani. 2014. « Toward an Arabic ontology for Arabic word sense disambiguation based on normalized dictionaries ». In *OTM Confederated International Conferences « On the Move to Meaningful Internet Systems »*, 655–658. Springer.

(Stevenson et Wilks, 2001) Stevenson, Mark, et Yorick Wilks. 2001. « The Interaction of Knowledge Sources in Word Sense Disambiguation ». *Computational Linguistics*, MIT Press Cambridge, MA, USA, 27 (3) : 321–349.

(Talvensaari, 2008) Talvensaari, Tuomas. 2008. « Effects of aligned corpus quality and size in corpus-based CLIR ». In *Proceedings of the 30th European conference on Advances in information retrieval*, 114–125. Glasgow, UK : Springer.

(Tchechmedjiev, 2012) Tchechmedjiev, Andon. 2012. « État de l'art : mesures de similarité sémantique locales et algorithmes globaux pour la désambiguïsation lexicale à base de connaissances ». In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012*, 3 :295-308. Grenoble, France : RECITAL,ATALA/AFCP.

(Toivonen et al., 2005) Toivonen, Jarmo, Ari Pirkola, Heikki Keskustalo, Kari Visala, et Kalervo Järvelin. 2005. « Translating Cross-Lingual Spelling Variants Using Transformation Rules ». *Information Processing & Management* 41 (4) : 859-72.

(Ture et Boschee, 2014) Ture, Ferhan, et Elizabeth Boschee. 2014. « Learning to translate : a query-specific combination approach for cross-lingual information retrieval ». In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 589–599. Doha, Qatar.

(Turtle et Croft, 2017) Turtle, Howard, et W. Bruce Croft. 2017. « Inference networks for document retrieval ». *ACM SIGIR Forum - SIGIR Test-of-Time Awardees 1978-2001*, ACM, 51 (2) : 1–24.

(Udupa et al., 2009) Udupa, Raghavendra, Anton Bakalov, et Abhijit Bhole. 2009. « They Are Out There, If You Know Where to Look : Mining Transliterations of OOV Query Terms for Cross-Language Information Retrieval ». In *Proceedings of the 31th European Conference on IR Research*, 437–448. Toulouse,France.

(Véronis, 2003) Véronis, Jean. 2003. « Cartographie lexicale pour la recherche d'information ». In *10e Conférence sur le TraitementAutomatique des Langues Naturelles*, 265–274. Batz-sur-Mer, France : ATALA.

(Véronis et Ide, 1990) Véronis, Jean, et Nancy M. Ide. 1990. « Word sense disambiguation with very large neural networks extracted from machine readable dictionaries ». In *Proceedings of the 13th conference on Computational linguistics*, 2 :389–394. Helsinki, Finland : Association for Computational Linguistics.

(Vidhu Bhala et Abirami, 2012) Vidhu Bhala, R. V., et S. Abirami. 2012. « Trends in Word Sense Disambiguation ». *Artificial Intelligence Review*, Kluwer Academic Publishers Norwell, MA, USA, 42 (2) : 159-71.

(Wang et Oard, 2006) Wang, Jianqiang, et Douglas W. Oard. 2006. « Combining bidirectional translation and synonymy for cross-language information retrieval ». In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 202–209. Seattle, Washington, USA : ACM.

(Wilks et al., 1990) Wilks, Yorick, Dan Fass, Guo Cheng-Ming, McDonald James E., Plate Tony, et Slator Brian M. 1990. « Providing machine tractable dictionary tools ». *Machine Translation*, Springer, 5 (2) : 99-154.

(Wu et al., 2008) Wu, Dan, Daqing He, Heng Ji, et Ralph Grishman. 2008. « A study of using an out-of-box commercial MT system for query translation in CLIR ». In *Proceedings of the 2nd ACM workshop on Improving non english web searching*, 71–76. Napa Valley, California, USA : ACM.

(Xu et al., 2017) Xu, Jingfang, Feifei Zhai, et Zhengshan Xue. 2017. « Cross-Lingual Information Retrieve in Sogou Search ». In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'17)*, 1361-1361. Shinjuku, Tokyo, Japan : ACM Press.

(Xu et al., 2001) Xu, Jinxi, Alexander M. Fraser, et Ralph M. Weischedel. 2001. « TREC 2001 Cross-lingual Retrieval at BBN ». In *Proceedings of the Tenth Text REtrieval Conference*. Gaithersburg, Maryland, USA.

(Xu et Weischedel, 2005) Xu, Jinxi, et Ralph Weischedel. 2005. « Empirical Studies on the Impact of Lexical Resources on CLIR Performance ». *Information Processing and Management : An International Journal - Special Issue : Cross-Language Information Retrieval* 41 (3) : 475-87.

(Yang et al., 2005) Yang, Yiming, Monica Rogati, et Bryan Kisiel. 2005. « Combining Categorization-based and Corpus-based Approaches for CLIR ». In *FLAIRS Conference*, 295–300.

(Yen, 2013) Yen, Christine. 2013. « Évaluation de la production de quatre systèmes traduction automatique ». Thèse, Canada : Dalhousie University Halifax, Nova Scotia.

(Yuan et Yu, 2007) Yuan, Song An, et Song Nian Yu. 2007. « A new method for cross-language information retrieval by summing weights of graphs ». In *Proceedings of Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, 2 :326–330. Haikou, China : IEEE.

(Yuret et Yatbaz, 2010) Yuret, Deniz, et Mehmet Ali Yatbaz. 2010« The noisy channel model for unsupervised word sense disambiguation ». *Computational Linguistics*, MIT Press Cambridge, MA, USA, 36 (1) : 111–127.

(Zadeh, 1978) Zadeh, Lotfi Asker. 1978. « Fuzzy sets as a basis for a theory of possibility ». *Fuzzy sets and systems*, North-Holland Publishing Company, 1 : 3-28.

(Zhang et al., 2005) Zhang, Junlin, Le Sun, et Jinming Min. 2005. « Using the web corpus to translate the queries in cross-lingual information retrieval ». In *International Conference on Natural Language Processing and Knowledge Engineering*, 493–498. Wuhan, China, China : IEEE.

(Zhong and Ng, 2010) Zhong, Zhi, et Hwee Tou Ng. 2010. « It makes sense : A wide-coverage word sense disambiguation system for free text ». In *Proceedings of the 48th Annual Meeting of*

the Association for Computational Linguistics (ACL), 78–83. Uppsala, Sweden : Association for Computational Linguistics.

(Zhou et al., 2012) Zhou, Dong, Mark Truran, Tim Brailsford, Vincent Wade, et Helen Ashman. 2012. « Translation Techniques in Cross-Language Information Retrieval ». *ACM Computing Surveys* 45 (1) : 1-44.

(Znaidi, 2016) Znaidi, Eya. 2016. « Contribution à l'analyse et l'évaluation des requêtes expertes : cas du domaine médical ». Thèse, France : Université de Toulouse, Université Toulouse III-Paul Sabatier.

(Zouaghi et al., 2012) Zouaghi, Anis, Laroussi Merhbene, et Mounir Zrigui. 2012. « Combination of Information Retrieval Methods with LESK Algorithm for Arabic Word Sense Disambiguation ». *Artificial Intelligence Review* 38 (4) : 257-69.

(Zulaini et al., 2013) Zulaini, Yahya, Muhamad Taufik Abdullah, Azreen Azman, et Rabiah Abdul Kadir. 2013. « Query Translation using concepts similarity based on Quran Ontology for cross-language information retrieval ». *Journal of Computer Science* 9 (7) : 889-97.