Recherche d'Information Monolingue et Translinguistique : de la Désambiguïsation vers l'Expansion Sémantique de Requêtes

Oussama Ben Khiroun

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

Recherche d'Information Monolingue & Translinguistique : de la Désambiguïsation vers l'Expansion Sémantique de Requêtes

---

par :

Oussama Ben Khiroun

Soutenue le 08/03/2018 devant le jury composé de

Président : Pr. Henda Hajjami Ben Ghezala, ENSI, Univ. Manouba, Tunisie
Rapporteur : Pr. Chiraz Latiri, ISAMM, Univ. Manouba, Tunisie
Rapporteur : Pr. Mathieu Roche, CIRAD & Univ. Montpellier, France
Examinateur : Pr. Rim Faiz, IHEC, Univ. Carthage, Tunisie
Directeur de thèse : Pr. Narjès Bellamine Ben Saoud, ENSI, Univ. Manouba, Tunisie

A ma mère et mon père,

A ma femme et ma petite Chaïma,

A mes frères et mes proches,

A tous ceux que j'aime.

Remerciements

Mes remerciements s'adressent à ma directrice de thèse Professeur Narjès Bellamine Ben Saoud et mon encadrant Docteur Bilel Elayeb pour leur disponibilité, leur soutien perpétuel, leurs précieuses directives et leurs idées scientifiques. Qu'ils trouvent ici le fruit de nos efforts comme témoignage de ma gratitude et de mon respect.

Je remercie également Docteur Ibrahim Bounhas pour ses idées enrichissantes et son aide précieuse.

Je tiens à exprimer ma profonde gratitude aux membres du jury qui m'ont honoré d'avoir accepté d'évaluer ce travail. En particulier, je remercie :

Professeur Henda Hajjami Ben Ghezala d'avoir accepté de présider le jury de ma thèse.

Professeur Chiraz Latiri et Professeur Mathieu Roche pour l'honneur qu'ils m'ont fait en acceptant d'être les rapporteurs de cette thèse.

Professeur Rim Faiz pour avoir accepté d'être l'examinatrice de ma thèse.

Je tiens à remercier aussi tous les membres du laboratoire RIADI et particulièrement les membres de notre équipe pour leurs encouragements persistants.

Je remercie aussi tous mes enseignants de l'ENSI qui ont contribué à ma formation. Qu'ils trouvent ici le résultat de leurs efforts.

Je n'oublie pas de saluer fortement tous mes amis et les membres de ma grande famille notamment ma mère, mon père, mes frères et ma femme pour leur patience et encouragement. Un grand merci à la famille Boughzala, je pense particulièrement à ma belle-mère Saloua Kort, pour leurs aides et encouragements. Qu'ils trouvent dans cette thèse une récompense de leurs sacrifices.

Table des matières

Table des figures

Liste des tableaux

Liste des Algorithmes

Liste des symboles

CLIR	Cross-Lingual Information Retrieval (recherche d'information translinguistique)
DPP	Degré de Pertinence Possibiliste
DSC	Dictionnaire Sémantique de Contextes
KWIC	Key Words In Context
LSA	Latent Semantic Analysis (analyse sémantique latente)
NER	Name Entity Recognition (reconnaissance d'entités nommées)
PRF	Pseudo Relevance Feedback (pseudo-rétroaction de pertinence)
QE	Query Expansion (expansion de requêtes)
RI	Recherche d'Information
RIT	Recherche d'Information Translinguistique
SRI	Système de Recherche d'Information
TALN	Traitement Automatique du Langage Naturel
WSD	Word Sense Disambiguation (désambiguïsation sémantique)

Introduction Générale

Le développement quasi-exponentiel des connaissances réparties sur des domaines d'intérêt variés a conduit à la génération d'une masse importante d'information de plus en plus difficile à gérer et à maintenir. Au sein de cet environnement à grande échelle caractérisé à la fois par le grand nombre d'utilisateurs et l'immense masse de données, il devient essentiel de concevoir et de développer des outils permettant un accès efficace et organisé à l'information.

Dans ce contexte, nous nous intéressons au domaine de la recherche d'information (RI) qui se focalise sur l'acquisition, l'organisation, le stockage et la recherche des données. Nous avons besoin de systèmes de recherche d'informations (SRI) qui constituent des outils informatiques visant à capitaliser l'information et à localiser les documents pertinents. Compte tenu d'une exigence d'information exprimée sous forme de requête, la pertinence est quantifiée selon un modèle d'appariement entre les termes de la requête et les documents.

Quelle que soit la sémantique donnée à la représentation des objets (document ou requête) ou à la définition de pertinence ; ces modèles ont un comportement général identique. La majorité d'entre eux représentent les documents et les requêtes par des listes de mots-clés pondérés. Par conséquent, à partir du concept de requête/réponse, la pertinence du résultat donné par un SRI dépend principalement de la requête. Cependant, l'utilisateur est parfois incapable de proposer des mots clés qui décrivent explicitement et clairement son besoin intentionnel ; ce qui peut détériorer la qualité des résultats attendus. Par conséquent, il devient crucial de développer des outils automatisés qui permettent de formuler et de satisfaire les besoins informationnels des utilisateurs à travers des interfaces d'interaction Homme/Machine.

La plupart des SRIs actuels sont axés sur le traitement des documents et des requêtes dans une seule langue. Cependant, devant la prolifération des corpus documentaires écrits en différentes langues, de nouveaux défis apparaissent pour passer d'un cadre de recherche d'information (RI) qualifié de « *monolingue* » vers un cadre de RI « *multilingue* ».

Cette thèse s'intègre dans le cadre de RI monolingue et translinguistique (appelée en anglais « *Cross Lingual Information Retrieval* »). Le but est de faciliter l'accès aux ressources d'information dans des langues différentes tout en étudiant la contribution des méthodes d'expansion et de désambiguïsation de requêtes sur le processus de recherche d'information.

1. Problématique de la thèse

La technique de reformulation de requêtes est une des stratégies mises en place dans les SRIs pour améliorer leurs performances et satisfaire au mieux les utilisateurs. Elle consiste de manière générale à enrichir la requête de l'utilisateur en ajoutant de nouveaux termes pour mieux exprimer son besoin. Ceci peut être déterminé en exploitant diverses sources de connaissance comme des dictionnaires, des relations sémantiques dans un thésaurus, des études du profil utilisateur, de classification des documents, voire du jugement de pertinence des résultats de recherche par l'utilisateur lui-même, appelé *rétroaction de pertinence*.

Néanmoins, l'utilisateur d'un SRI se contente souvent de donner quelques mots clés qui peuvent ne pas décrire explicitement et clairement son besoin intentionnel ; ce qui peut altérer la qualité des résultats attendus du SRI. En effet, parmi les problèmes rencontrés dans la phase de recherche d'information est l'ambiguïté sémantique des termes des requêtes. Ce phénomène d'ambiguïté est souvent renforcé dans le cas des requêtes courtes qui n'explicitent pas nécessairement le contexte de recherche. Par conséquent, la tâche de désambiguïsation sémantique s'avère importante lors de la phase de reformulation ainsi que pour la phase de traduction des requêtes dans le cadre de RI translinguistique.

Afin de résoudre ce problème d'ambiguïté sémantique conjointement au contexte incertain et imprécis de recherche, la théorie des possibilités s'apprête naturellement à ce genre d'application [Dubois et Prade, 2011]. Dans un premier lieu, nous menons des études et expérimentations sur la désambiguïsation sémantique monolingue des textes en étendant le modèle qualitatif de RI proposé par [Elayeb, 2009]. Ce modèle de base a été adapté dans nos anciens travaux sur la tâche d'expansion de requêtes monolingue [Ben Khiroun *et al.*, 2011] et a prouvé son efficacité par rapport au modèle à base de dénombrement de circuits [Elayeb *et al.*, 2011]. Dans ces travaux, l'approche proposée se base sur l'utilisation du dictionnaire *Le Grand Robert* organisé en Réseau de Petits Mondes Hiérarchiques (RPMH). Néanmoins, l'utilisation de telles ressources s'affronte au problème de couverture de lexique face aux nouvelles terminologies qui ne sont pas nécessairement présentes dans les dictionnaires classiques.

Dans le but de résoudre les limites de l'utilisation des dictionnaires classiques dans la tâche de désambiguïsation monolingue, nous proposons une approche qui combine ce type de données structurées avec des connaissances extraites à partir des corpus documentaires. Ainsi, les définitions contenues dans les dictionnaires sont enrichies par des connaissances contextuelles liant les mots avec leurs contextes. Nous proposons dans ces travaux un « *Dictionnaire Sémantique de Contextes* » pour mettre en place l'ensemble des relations sémantiques entre les mots [Ben Khiroun *et al.*, 2012, Elayeb *et al.*, 2015b].

Nous nous focalisons en deuxième lieu sur l'analyse et l'expérimentation de l'effet de la désambiguïsation sémantique des requêtes combinée avec l'expansion des requêtes sur la recherche d'information. Une représentation de connaissance en graphe de co-occurrence a été utilisée pour calculer la similarité entre les termes de requêtes (dans le cas de l'expansion) ou entre les termes et les sens (dans le cas de désambiguïsation).

L'approche proposée est basée sur les réseaux possibilistes pour la désambiguïsation et l'expansion des requêtes en considérant, pour la modélisation du graphe de co-occurrence, que deux nœuds sont liés s'ils existent dans la même phrase. Les arêtes sont non orientées et pondérées par la fréquence normalisée de co-occurrence des termes connexes. D'autre part, les mots ambigus sont liés avec leurs sens appropriés dans le dictionnaire [Ben Khiroun *et al.*, 2014, Elayeb *et al.*, 2015a].

Nous nous intéressons dans la dernière partie de la thèse à la RI translinguistique dans laquelle la requête est représentée dans une langue source et la collection des documents est représentée dans une autre langue cible. La problématique principale dans la RI translinguistique est de faire l'appariement adéquat entre les requêtes et les documents écrits dans deux langues différentes ; ceci passe obligatoirement par une phase de traduction. La traduction des requêtes vers la langue cible, dans laquelle sont écrits les documents, représente l'approche la plus courante dans les travaux de RI translinguistique [Nie, 2010]. En effet, cette approche est réalisée avec un coût plus optimisé du fait que les requêtes sont généralement limités à quelques termes. Nous étendons le cadre d'étude des approches de désambiguïsation et d'expansion de requêtes possibilistes appliquées sur la RI monolingue vers un cadre de RI translinguistique [Ben Khiroun *et al.*, 2018].

2. Organisation de la thèse

Le manuscrit de thèse de doctorat est composé de deux parties organisées en 6 chapitres.

La **première partie** intitulée « état de l'art sur la recherche d'information » est structurée comme suit :

Nous commençons, dans le premier chapitre, par définir les concepts fondamentaux de RI, les phases de prétraitement ainsi que les modèles de RI. Cette section sera suivie d'un aperçu sur l'application de la RI dans un cadre translinguistique. En fin du chapitre, nous présentons le cadre d'évaluation des SRIs en énumérant les mesures et les collections standards de test.

Ensuite, nous présentons dans le deuxième chapitre les domaines d'application de la désambiguïsation sémantique ainsi que les différentes ressources et mesures utilisées dans cette tâche. Nous détaillons dans le même chapitre une étude des approches de désambiguïsation sémantique des textes dans le cadre monolingue ; qui sera suivi d'un état de l'art sur l'application de la désambiguïsation dans le cadre translinguistique.

Le troisième chapitre s'intéresse à la reformulation de requêtes, ses techniques et les défis qu'elle confronte. Dans ce chapitre, nous présentons dans un premier lieu la corrélation existante entre les tâches d'expansion de requête et la désambiguïsation sémantique. Nous nous focalisons, en termes d'approches, sur la rétroaction de pertinence et les méthodes globales d'expansion de requêtes. Dans la dernière section du chapitre, nous présentons une vue sur l'application de l'expansion de requêtes dans la RI translinguistique.

Dans la **deuxième partie** du manuscrit, nous détaillons l'ensemble des « contributions » (théoriques, techniques et expérimentales) en décrivent les composantes de la méthodologie proposée à travers les trois chapitres restants :

Dans le quatrième chapitre, nous introduisons les fondements de base de la théorie des possibilités, ses applications dans la RI ainsi que ses adaptations dans notre nouveau Système Possibiliste d'Expansion Et de Désambiguïsation SEmantique de Requêtes (SPEEDSER). Le modèle conceptuel du système SPEEDSER ainsi que ses différents modules présentés via des interfaces graphiques sont décrits tout au long de ce chapitre.

Le cinquième chapitre présente un cadre applicatif et expérimental du système proposé dans la désambiguïsation sémantique des textes. Nous modélisons la structure du dictionnaire sémantique de contexte qui servira dans l'approche possibiliste proposée pour la désambiguïsation sémantique. La dernière section se focalise sur l'étude expérimentale de notre approche proposée en utilisant le corpus ROMANSEVAL comme standard de test.

Dans le sixième et dernier chapitre, nous projetons cette étude sur la RI en étudiant à la fois la contribution de la désambiguïsation sémantique combinée avec l'expansion de requêtes. Ce chapitre se divise en deux grandes parties : la première partie sera dédiée au cadre de RI monolingue et qui sera étendue par la suite vers un cadre translinguistique dans la deuxième partie du chapitre. Nous tirons profit, dans l'approche générique proposée, des relations sémantiques entre les termes en se basant sur des ressources de

graphes de co-occurrences dans les deux cadres de RI (monolingue et translinguistique).

En conclusion, nous dressons un bilan de nos travaux, en mettant en exergue nos approches proposées et en rappelant les motivations liées à la problématique traitée dans cette thèse. Nous concluons par la proposition des perspectives de recherche possibles à ces travaux.

Première partie

ÉTAT DE L'ART SUR LA RECHERCHE D'INFORMATION

Chapitre

1 De la RI Monolingue vers la RI Translinguistique

Sommaire

" Information is not knowledge. "

— Albert Einstein

Introduction

La Recherche d'Information (RI) est un domaine qui s'attache à définir des modèles et des systèmes dont le but est de faciliter l'accès à un ensemble de documents sous forme électronique (ou corpus), afin de permettre à l'utilisateur de retrouver les documents dont le contenu correspond le mieux à son besoin d'information. Le degré d'adéquation entre le contenu d'un document et l'information recherchée définit la notion de pertinence.

Un besoin en information est une représentation mentale de ce que l'utilisateur souhaite rechercher. Ce besoin est représenté sous forme d'une requête. Par conséquent, la requête n'est donc qu'une représentation possible d'un besoin en information. Les deux concepts de « requête » et « besoin » sont souvent confondus.

La notion de pertinence est souvent subjective vu qu'elle dépend du domaine du sujet recherché, le champ disciplinaire auquel il se rapporte et la tâche, c'est-à-dire l'activité que l'utilisateur vise réaliser avec les documents retrouvés. A titre d'exemple, la pertinence des documents retournés pour le terme « *cancer* » dépendra du besoin de l'utilisateur en information de nature médicale (la maladie), astrologique (l'horoscope) ou lié aux sciences de la vie (l'animal), etc.

On distingue principalement deux types de pertinence : (i) la *pertinence système*, c'est-à-dire l'évaluation par un système de l'adéquation entre des documents et une requête, et (ii) la *pertinence utilisateur* qui se traduit par des jugements de pertinence sur les documents fournis en réponse à une requête. Les travaux de recherche récents s'accordent sur la difficulté de la définition de cette notion et qu'il n'existe pas de consensus pour la pertinence [Borlund, 2003].

Dans ce chapitre, nous nous intéressons à présenter le processus de RI dans la première section, en détaillant les composants d'un système de recherche d'information (SRI), les phases de prétraitement ainsi que les modèles de RI. Cette section sera suivie d'un aperçu sur l'application de la RI dans un cadre translinguistique. En fin du chapitre, nous présentons le cadre d'évaluation des SRI.

1.1 Processus de recherche d'information

Le processus de RI commence par traiter d'abord tous les documents de la collection pour créer un index, afin de faciliter la recherche et la rendre plus rapide. Cet index est formé par des descripteurs obtenus après un processus assez complexe, utilisant divers outils, techniques et normes [Manning *et al.*, 2008]. Parmi ceux là, on note :

– les traitements linguistiques : qui décomposent le texte en un ensemble de mots, les normalisent et exploitent des dictionnaires, des thésaurus et des outils d'analyse pour définir ces descripteurs selon le sens des mots, des phrases et les rapports y figurant ;

– les traitements statistiques : plusieurs méthodes sont disponibles dans la littérature, dont celle basée sur la mesure de fréquence qui fixe deux seuils de fréquence maximal et minimal et ne garde que les termes qui figurent entre ces derniers, et celle basée sur la mesure de la pondération TF-IDF qui attribue un poids à chaque terme selon sa fréquence dans le document et le corpus entier et garde ceux avec les poids les plus élevés.

Cette phase, appelée *phase d'indexation*, est effectuée au préalable sur les documents et lors de la saisie de la requête.

Suite à cela vient la *phase d'appariement* (ou *matching* en anglais), qui consiste à appliquer l'un des modèles de pondération et d'appariement pour déterminer les documents que le système considère comme totalement ou partiellement pertinents à la requête. La pertinence d'un document est le degré d'adéquation entre son contenu et l'information recherchée. Parmi les méthodes d'appariement nous citons essentiellement le modèle booléen, les modèles vectoriels et les modèles probabilistes [Manning *et al.*, 2008, Elayeb, 2009]. Le but de la RI est de satisfaire le besoin initial en information de l'utilisateur ; d'où, une phase d'évaluation est engagée. Cette phase consiste à retourner à l'utilisateur les résultats de la phase précédente d'appariement, pour qu'il puisse évaluer leurs pertinences selon son jugement (voir figure 1.1).

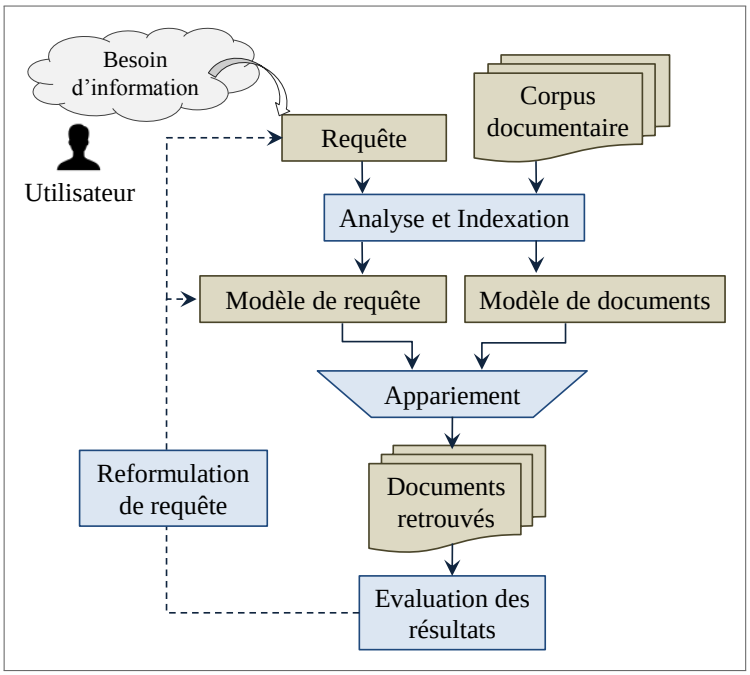


Figure 1.1 – Architecture générale d'un SRI

1.1.1 Composants d'un SRI

L'objectif d'un système de recherche d'information (SRI) est d'aiguiller la recherche dans le fond documentaire, en direction de l'information relativement pertinente à un besoin d'information exprimé par une requête. A cet effet, le SRI assure les fonctionnalités de communication, stockage, organisation et recherche d'information [Tamine, 2000].

Pour réaliser cet objectif énoncé, un SRI construit deux représentations internes : l'une correspondant à l'ensemble des documents et l'autre à la requête ou plus généralement au besoin d'information. Cette phase est désignée par *phase de représentation*. Pour bien élaborer ces représentations, le système intègre dans cette dernière phase une phase supplémentaire d'*analyse et d'indexation*. L'opération clé qui permet de retrouver les documents pertinents est la *mise en correspondance* entre les deux représentations (*matching* ou appariement).

La formulation du besoin et la restitution du résultat se font via une interface qui assure l'interaction entre les utilisateurs et le système de recherche proprement dit. Cette interface permet l'acquisition et l'interprétation des requêtes ainsi que la présentation et la visualisation des résultats. Cette même interface servira à la récupération de l'évaluation de l'utilisateur en terme de pertinence sur les documents retrouvés pour une éventuelle *reformulation de requête*. Il s'agit de la phase d'évaluation du résultat, appelée aussi phase de *rétroaction de pertinence* (ou *relevance feedback* en anglais). Pour juger l'efficacité du SRI de manière absolue et non par rapport à une recherche particulière, une *phase d'évaluation* globale est également nécessaire.

La figure 1.1 illustre l'architecture générale d'un SRI regroupant plusieurs éléments à savoir :

- la collection de documents qui constitue l'ensemble des informations exploitables et accessibles par l'utilisateur.
- le besoin en termes d'information d'un utilisateur exprimé par une requête.
- la représentation des documents et des requêtes (indexation ou analyse) supportée par un ensemble de règles et notations permettant la transformation d'une requête ou d'un document d'une description brute vers une description structurée.
- l'appariement requête-document se charge de faire la correspondance entre requête et documents.

1.1.2 Analyse et Indexation

Le document est l'élément central dans l'architecture d'un SRI. Il est considéré comme un objet complexe en état d'évolution constante grâce au développement des technolo-

gies de l'information et de la communication. L'émergence des documents électroniques et l'évolution des techniques de la communication ont favorisé la dématérialisation du document. Ainsi, le document électronique remplace de plus en plus le document en format papier.

D'une manière intuitive, la recherche peut s'effectuer directement dans le texte brut : il s'agit de la recherche plein texte ou recherche en texte intégral, sans effectuer de prétraitements. Par contre, dans une recherche élaborée, il s'agit d'introduire une phase d'analyse ou/et d'indexation qui permet de ressortir les sujets spécifiques des documents et des représentants ou termes/candidats descripteurs. L'analyse consiste à extraire l'information utile, qui sert à la fois à représenter le contenu des documents et à retrouver ceux qui correspondent à un sujet ou une question.

Ainsi en pratique, nous cherchons plutôt des *représentations* (instances ou extensions) des concepts. Ces représentants peuvent être de forme différentes : des mots simples, des termes (éventuellement composés), ou des doublets de mots (groupes de deux mots). En fait, le choix de représentants dépend de deux critères essentiellement : la facilité de traitement et la précision de représentation de sens. Deux techniques d'indexation peuvent être appliquées au contenu des documents : l'approche *linguistique* et l'approche *statistique*.

1.1.2.1 Traitement linguistique

Les techniques d'indexation à base de traitement linguistique présentent l'intérêt de traiter le contenu en considérant de manière plus ou moins riche sa nature linguistique. Dans ce cas, le contenu est considéré comme un ensemble de mots qui entretiennent différents rapports entre eux et donc non comme de simples chaînes de caractères.

L'analyse linguistique peut être décomposée en cinq niveaux [Kammoun Bouzaiene, 2006] :

1. Un *niveau morphologique* pour identifier les mots en tant que chaînes de caractères. Pour cela, il faut d'abord commencer par exploiter les caractères dits « séparateurs de mots » ; par la suite, neutraliser les caractères sans pertinence (ignorer par exemple les retours à la ligne et les signes de ponctuation) et normaliser le texte (en transformant tous les caractères sous forme minuscule par exemple).
2. Un *niveau lexical* qui permet de reconnaître des formes diverses des mots (conjuguées, dérivées, etc.). Il s'agit de vérifier l'existence des unités lexicales dans un dictionnaire et d'engager des procédures de manipulation de ces mots s'ils ne sont pas identifiés, telles que des fonctions morphologiques et des fonctions de redressage orthographiques.

3. Un *niveau syntaxique* qui tient compte de l'agencement des mots dans une phrase et indique leur structure. Il s'agit donc de dégager les rapports entre les unités lexicales pour élaborer une représentation syntaxique (temps des verbes conjugés, structures de groupes nominaux, etc.) qui vont jouer le rôle de règles de manipulation et de décomposition des objets linguistiques.
4. Un *niveau sémantique* (ou *syntaxico-sémantique*) qui identifie le sens du mot selon son emplacement dans la phrase. Cette phase utilise généralement une « base de connaissances » comprenant des relations sémantiques entre les mots.
5. Un *niveau pragmatique* qui identifie le contexte des phrases.

Tous ces traitements exigent une bonne connaissance de la langue et exige l'intervention d'outils supplémentaires tels que :

- Des dictionnaires où toute l'information linguistique attachée aux unités morphologiques, syntaxiques et sémantiques est codée de manière à pouvoir être exploitée de façon automatisée.
- Des thésaurus et/ou des terminologies.
- Des outils d'analyse de corpus (traitement lexical, syntaxique, etc.).

Il s'avère que l'utilisation du TALN peut permettre d'augmenter la performance et d'affiner la recherche d'un SRI mais elle augmente considérablement sa complexité. L'un des problèmes à définir ici est la profondeur de l'analyse syntaxique mise en œuvre, pour la recherche des sous-unités sémantiquement pertinentes pour représenter le document et l'expansion du vocabulaire d'indexation.

1.1.2.2 Traitement statistique

Dans ce type d'approche, l'idée est d'utiliser des mots comme des représentants de concepts. L'hypothèse adoptée par le traitement statistique d'analyse et d'indexation est que la fréquence d'apparition d'un terme dans un document est intimement liée à son importance pour ce même document. Différentes manières d'exploiter cette fréquence d'apparition ont été proposées en considérant le document d'une façon isolée ou par rapport au corpus auquel il appartient.

La mesure TF-IDF (de l'anglais « *Term Frequency-Inverse Document Frequency* ») est une mesure statistique très répandue dans la RI. En effet, selon les travaux de Sparck Jones [Sparck Jones, 1972] et Salton [Salton et Buckley, 1988], les notions de spécificité et d'exhaustivité peuvent être corrélées. Les auteurs proposent dans ce sens des interprétations statistiques qui sont fonction de la distribution des termes dans les documents qu'ils indexent et dans la collection. D'une part, l'exhaustivité est liée au nombre de termes d'indexation affectés à un document. D'autre part, la spécificité est

liée au nombre de documents indexés par un terme donné. Le poids final TF-IDF d'un terme est obtenu en multipliant la fréquence des termes dans le document (*tf*) avec la fréquence inverse du terme dans les documents de la collection (*idf*). La mesure *idf* est vu comme un indicateur de la spécificité et elle est dotée d'un pouvoir discriminant pour distinguer entre les documents de la collection. Par conséquent, un terme qui a une valeur de TF-IDF élevée doit être à la fois important dans ce document, et aussi il doit apparaitre peu dans les autres documents.

1.1.2.3 Filtrage, racinisation et lemmatisation

En plus du filtrage effectué sur la base du calcul de fréquences, on procède souvent dans les SRI à une phase d'élimination de certains mots qualifiés de « mots vides » ou « *stop words* ». Pour ce faire, on utilise un « anti-dictionnaire » (appelé en anglais « *stop list* »). Ces mots sont souvent des prépositions, des prénoms, certains adverbes, des adjectifs, etc. La liste peut varier selon le domaine d'application.

Par ailleurs, on recense beaucoup de mots qui ont des formes légèrement différentes, par contre leur sens est similaire, notamment pour les mots conjugués. L'objectif est donc d'éliminer ces différences non significatives pour les réduire à une même forme. On distingue deux principaux types de transformation des mots à savoir la « lemmatisation » et la « racinisation ». Ces problèmes à résoudre entrent principalement dans le cadre du TALN.

L'objectif de la lemmatisation est de réduire chaque mots en une entité appelée *lemme* (ou *forme canonique*). Ainsi, la lemmatisation regroupe les différentes formes que peut revêtir un mot (le nom, le pluriel, le verbe à l'infinitif, etc.). La lemmatisation nécessite souvent des recours à des connaissances linguistiques tels que les dictionnaires ou des corpus annotés, nécessaires pour la phase d'apprentissage afin de vérifier les mots transformés [Savoy, 1997]. Ceci rend le traitement linguistique plus difficile et limite le nombre des outils de lemmatisation disponibles pour une langue donnée. Pour la langue française, nous citons, à titre d'exemple, l'outil LIA Tagg [1] et l'outil TreeTagger [2] qui supporte plusieurs autres langues. Ces deux outils sont également des étiqueteurs morphosyntaxiques (en anglais « *Part-of-Speach (POS) Tagger* ») qui attribuent automatiquement des catégories grammaticales pour chaque mot traité.

A la différence de la lemmatisation, la racinisation (appelé aussi *troncature* et en anglais *« stemming »*) opère un mot simple sans connaissance préalable du contexte. En effet, la racinisation permet de transformer des flexions en leur radical ou racine (en anglais « *stem* »). La racine d'un mot correspond à la partie du mot restante après suppression

1. http://lia.univ-avignon.fr/chercheurs/bechet/download_fred.html
2. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

des préfixes et des suffixes. Par exemple, le mot « *meilleur* » a « *bon* » comme lemme ; cependant, cette transformation n'est pas identifiée dans le processus de racinisation qui laisse le mot tel qu'il est et le considère comme forme racine. La mise en place d'outil de racinisation consiste en des algorithmes de déduction de la racine (radical) en éliminant un ensemble de préfixes et suffixes. Nous citons à titre d'exemple les algorithmes les plus connus de [Lovins, 1968], de [Porter, 1997] et de [Paternostre *et al.*, 2002].

Les algorithmes de racinisation sont répandus dans la mise en place des SRI grâce à leur rapidité par rapport aux outils de lemmatisation qui sont plutôt utilisés dans les processus d'étiquetage morphosyntaxique. En effet, les mots-clés d'une requête ou d'un document sont représentés par leurs racines plutôt que par les mots d'origine. Plusieurs variantes d'un terme peuvent ainsi être groupées dans une seule forme représentative, ce qui réduit la taille du dictionnaire, c'est-à-dire le nombre de termes distincts nécessaires pour représenter un ensemble de documents. Un dictionnaire de taille réduite permet de gagner à la fois de l'espace et du temps d'exécution. Cependant, l'usage de la racinisation peut présenter un risque de baisse de la précision.

1.1.3 Modèles de recherche d'information

Un modèle de recherche d'information spécifie, d'une part, les représentations utilisées pour les requêtes des utilisateurs et les documents du corpus et, d'autre part, la façon dont ces représentations sont comparées. Les opérations de comparaison entre documents et requêtes suivent des modèles de mise en correspondance (*matching* ou *appariement*). Ceci se fait par le calcul d'une mesure appelée *pertinence système* supposée représenter la pertinence du document vis-à-vis de la requête. Le calcul de la pertinence système tient en compte généralement les poids des termes dans les documents et éventuellement leurs poids dans les requêtes.

La variété de méthodes de représentation des documents et des requêtes, ainsi que la diversité des méthodes de pondération des termes (que se soit dans les documents ou dans les requêtes) a donné naissance à une panoplie de modèle de RI dans la littérature. La différence d'un modèle à un autre se manifeste dans l'ensemble des documents retournés par le SRI et dans l'ordre de leur apparition. Ce dernier paramètre influence le jugement et le degré de satisfaction de l'utilisateur qui souhaite toujours avoir les documents pertinents à son sens (on parle ici de la *pertinence utilisateur*) en tête des résultats retournés.

La figure 1.2 présente une classification des différents modèles de RI qui se déclinent en trois grandes familles : les modèles booléens, les modèles vectoriels et les modèles probabilistes [Baeza-Yates et Ribeiro-Neto, 1999, Manning *et al.*, 2008].

Les modèles booléen sont basés sur des approches ensemblistes de représentation des

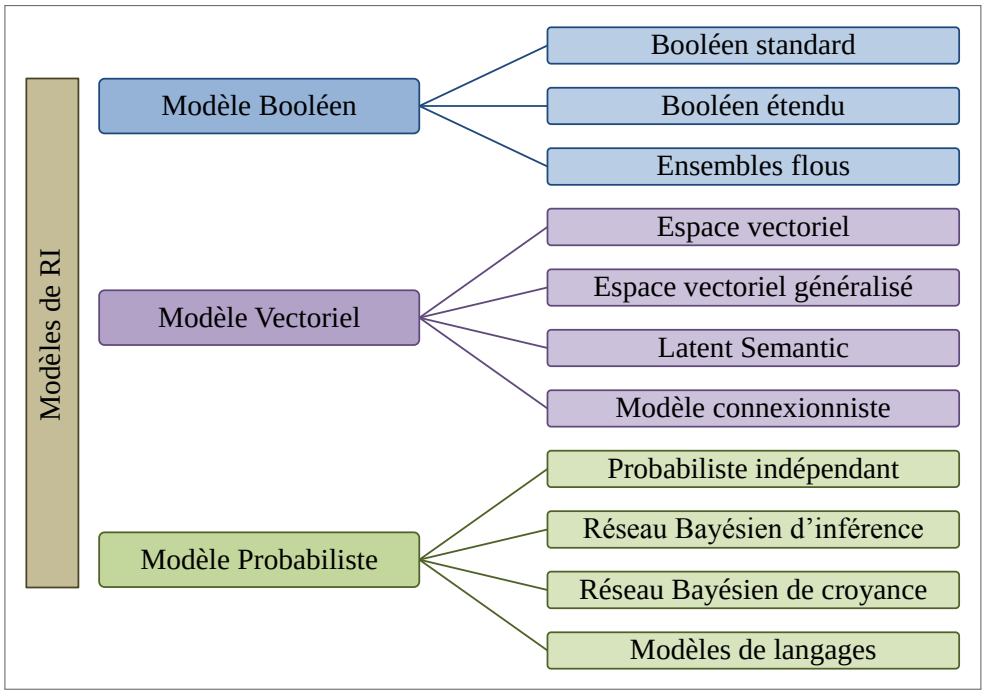


Figure 1.2 – Taxonomie des modèles de RI

documents. Ils constituent les premiers modèles utilisés en RI. Dans les modèles vectoriels, les documents et les requêtes sont représentés par des vecteurs de poids dans un espace vectoriel composé des termes d'indexation. La mesure de distances entre vecteurs permet de quantifier la pertinence d'un document vis-à-vis une requête. Les modèles probabilistes s'appuient sur la théorie des probabilités. Ainsi, la pertinence d'un document vis-à-vis d'une requête est vue comme une probabilité de pertinence document/requête [Manning *et al.*, 2008].

1.2 Recherche d'information translinguistique

Le rôle de la Recherche d'Information Translinguistique (RIT) est d'identifier les documents qui répondent aux besoins de l'utilisateur, d'après le(s) langage(s) de la requête et des documents recherchés. Il s'agit d'un sous-domaine actif de la RI avec laquelle il se rejoint étant centré sur la recherche de documents en partant d'un besoin en information posé par l'utilisateur. Contrairement à la RI classique, la RI translinguistique doit solliciter les requêtes et les documents qui sont écrits dans des langues différentes.

En effet, la RI translinguistique tente de surmonter la barrière de langue entre les requêtes et les documents : dans la vie réelle, un utilisateur d'un SRI qui soumet une requête en français pourrait aussi être intéressé par des documents en anglais, en allemand ou en arabe. Comment éliminer les barrières de la langue en permettant à un SRI de sélectionner des documents exprimés dans une langues différentes de celle de la

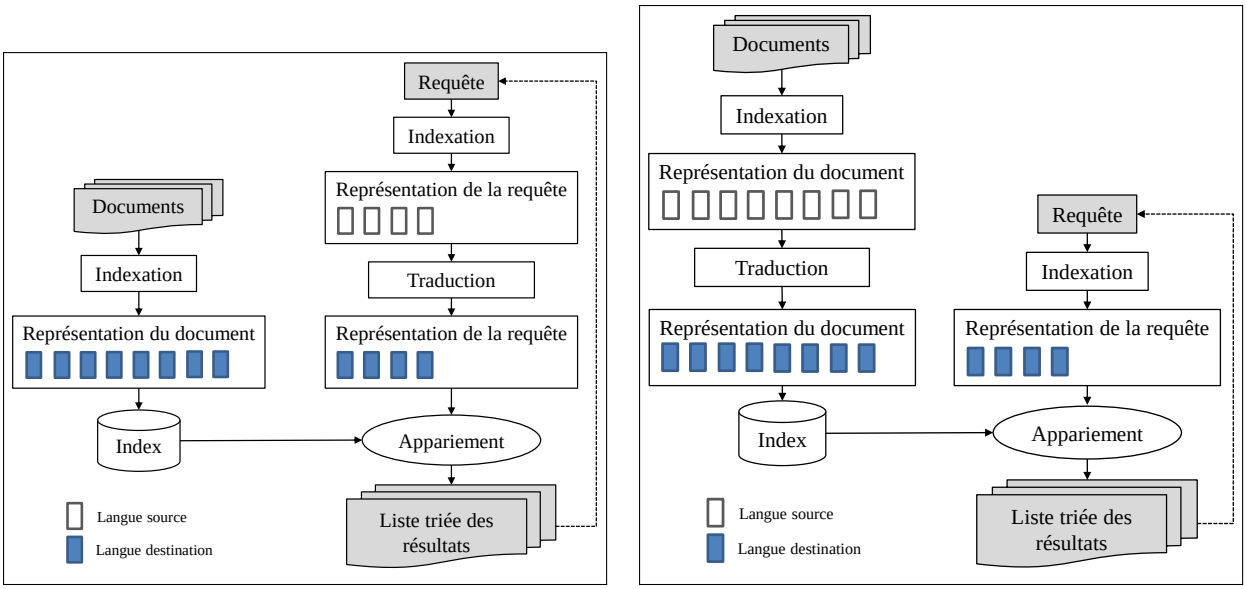
requête ?

Afin de résoudre le problème d'hétérogénéité linguistique, la solution intuitive consiste à traduire la requête et / ou les documents avant d'effectuer la recherche. Nous distinguons ainsi trois approches générales de traduction qui peuvent être utilisées dans la conception d'un SRI translinguistique [Zhou *et al.*, 2012] à savoir :

- Traduire la requête pour correspondre à la représentation du document (figure 1.3.a) ;
- Traduire le document pour correspondre à la représentation de la requête (figure 1.3.b) ;
- Traduire la requête et le document vers une troisième langue dite *pivot* (figure 1.4).

La première méthode est la plus répandue dans les SRI translinguistiques vu que la longueur de la requête est généralement courte ce qui rend sa traduction plus rapide et facile. Cependant, la longueur réduite de la requête peut présenter éventuellement des ambiguïtés en présentant des informations contextuelles limitées pour la phase de traduction.

En contre partie, la deuxième méthode de traduction des documents garde l'avantage, théoriquement, d'avoir plus d'informations contextuelles pour déterminer la traduction correcte. Cependant, étant donnée le volume des documents, cette traduction devient assez lente tout en ignorant également la langue dans laquelle l'utilisateur souhaite avoir des résultats. Ainsi, il faudra traduire les documents dans toutes les langues possibles.



(a) traduction des requêtes **(b)** traduction des documents

Figure 1.3 – RI translinguistique basée sur la traduction de la requête ou des documents [Zhou *et al.*, 2012]

1.2.1 Architectures des SRI translinguistiques

L'architecture d'un SRI translinguistique présentée dans la figure 1.3.(a) résume l'approche la plus répandue basée sur la traduction de la requête de la langue source vers la langue cible (ou de destination) dans laquelle les documents sont écrits. Après traduction de la requête, cette dernière subit une phase d'indexation pour extraire les termes discriminants. Par la suite, un appariement est appliqué entre cette nouvelle représentation de la requête dans la langue cible et les documents.

L'application de cette phase d'appariement attribue des scores de pertinence aux documents. Le résultat final de la recherche est fourni à l'utilisateur sous forme d'une liste de documents triés par ordre décroissant selon les score de pertinence précédemment calculés.

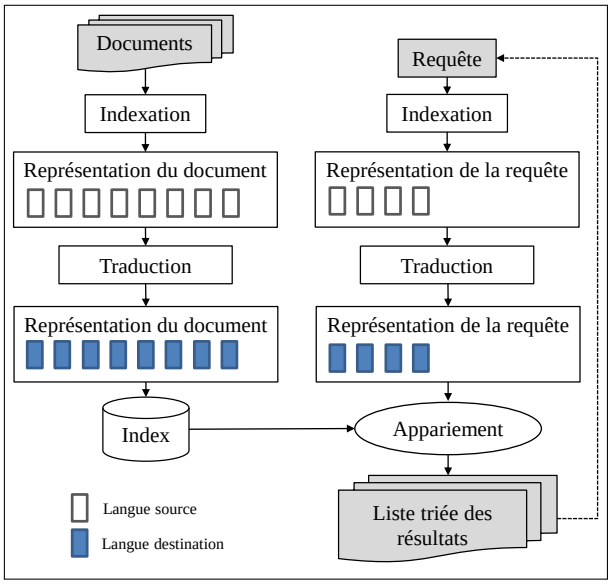


Figure 1.4 – RI translinguistique basée sur la traduction des documents et de la requête en utilisant une langue pivot [Zhou *et al.*, 2012]

Un SRI translinguistique peut être constitué des modules suivants : un module de pré-traduction, un module de traduction, un module de post-traduction et un module de RI comme présenté dans la figure 1.5.

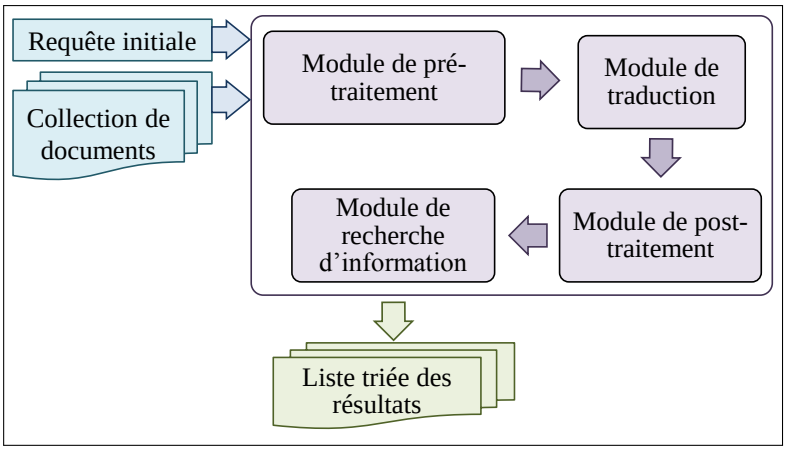


Figure 1.5 – Modules d'un SRI translinguistique

En amont, le module de pré-traduction regroupe l'ensemble des tâches de *tokenisation*, l'élimination des mots vides, la *lemmatisation* et éventuellement l'expansion des termes. Il s'agit d'un module de pré-traitement.

En aval, le module de post-traduction agit sur les termes dans la langue cible qui ont été traduits de la requête, des documents ou les deux à la fois selon l'architecture de traduction. Le traitement final regroupe les deux aspects de pondération des termes avec des poids et l'expansion des termes après traduction.

1.2.2 Approches de traduction de requêtes

La traduction est une activité clé pour les moteurs de RI translinguistique. Au cours des dernières années, la communauté de RI translinguistique a développé une large gamme de techniques et de modèles de traduction de texte.

1.2.2.1 Approches basées sur les dictionnaires bilingues

Les dictionnaires bilingues disponibles sous format numérique compréhensible par une machine (*Machine Readable Dictionary*) deviennent de plus en plus disponibles et utilisables à la fois dans les modules de traduction des SRI translinguistiques.

En dépit de la simplicité de cette approche de traduction, elle souffre de deux inconvénients majeurs [Zhou *et al.*, 2012] : (1) l'*ambiguïté* en proposant de multiples traductions et (2) la *faible couverture* du vocabulaire (surtout face aux termes modernes ou techniques que les dictionnaires n'inclus pas nécessairement ; ces termes non couverts sont appelés *termes hors vocabulaire* (ou en anglais *OOV : Out Of Vocabulary*)) .

1.2.2.2 Approches basées sur les corpus parallèles

Afin de résoudre le problème de l'incomplétude des dictionnaires de traduction, l'utilisation de corpus alignés a été développée afin d'extraire l'information manquante des dictionnaires. Un corpus aligné est constitué d'un ensemble de documents alignés par phrase avec leur équivalent dans une autre langue représentant ainsi des traductions mutuelles les unes des autres [Zhou *et al.*, 2012].

L'exploitation des corpus parallèles se fonde sur les relations de co-occurrence : plus un mot en langue cible co-occurre avec un mot source dans les textes parallèles, plus il y a une forte relation de traduction entre les deux mots.

L'application de cette approche dans le cadre de traduction des requêtes dépend de la disponibilité de corpus parallèles dont l'acquisition est souvent coûteuse.

1.2.2.3 Traduction automatique

Ces approches nécessitent un système de traduction automatique pour traduire la requête ou le corpus documentaire de sorte que tous les deux soient dans la même langue. Les systèmes de traduction automatique sont devenus populaires ces dernières années en raison de la disponibilité des ressources linguistiques nécessaires dans le processus d'apprentissage. Néanmoins, parmi les inconvénients que présente la traduction automatique est que les systèmes de traduction automatiques existants ne supportent qu'un nombre réduits de paires de langues cible/source [Zhou *et al.*, 2012].

1.2.3 Discussion et résumé des approches de traduction dans la RI translinguistique

Chacune des approches, décrites dans les sous-sections précédentes, présente des défis dans le cadre de RI translinguistique [Atkins et Rundell, 2008, Nie, 2010] à savoir :

- La désambiguïsation de traduction, qui est souvent dérivée des problèmes d'homonymie et de polysémie [3] ;
- Les langues flexionnelles : un mot peut avoir des formes différentes pour une même catégorie grammaticale [4]. Ce problème peut être résolu par la lemmatisation ou

3. ***Homonymie*** : réfère à un mot qui a au moins deux sens complètement différents ;
Polysémie : réfère un mot qui a deux sens différents mais qui sont en relation.
ex. d'homonyme : « *avocat* » peut signifier le fruit ou le métier de droit.
ex. de mot polysémique : le sens du mot « *étoile* » change selon le contexte : « *une étoile de mer* », « *une étoile filante* », « *une chanteuse étoile* », etc.
Les phrases suivantes présentent ainsi une forte ambiguïté : « *Un avocat mange de l'avocat* », « *La mode est un mode de vie* ».

4. ex : *cheval* (singulier), *chevaux* (pluriel)

la racinisation (appelée en anglais *stemming*) [5] ;

– La détection des entités nommées (appelée en anglais *Named Entity Recognition*). L'extraction et la traduction des entités nommées sont essentielles dans le domaine du traitement automatique du langage naturel, la RI translinguistique, la construction de lexique bilingue, etc ;

– Le manque de couverture (*Out-of-Vocabulary (OOV) problem*) lors de l'utilisation des dictionnaires dans les approches de traduction. En effet les requêtes saisies par l'utilisateur sont généralement courtes et même l'expansion de la requête ne peut pas aider à récupérer les mots manquants faute d'insuffisance de contexte ;

– Dans de nombreux documents, les termes techniques et les noms propres sont des éléments importants du texte. Cependant, les dictionnaires ne contiennent que les noms propres les plus couramment utilisés et les termes techniques les plus connus tels que les grandes villes et les pays. Leur traduction est cruciale pour augmenter la précision d'un système de RI translinguistique. Une méthode commune utilisée pour traiter les mots-clés intraduisibles consiste à inclure le mot non traduit dans la requête de langue cible. Néanmoins, si ce mot n'existe pas dans la langue cible, la requête sera moins susceptible de récupérer les documents pertinents.

Dans le tableau 1.2, nous présentons un résumé des différentes approches de traduction utilisées dans la RI translinguistique.

D'autres travaux récents se sont focalisés sur la combinaison de deux ou plusieurs approches de traduction. Nous citons les travaux de [Kim *et al.*, 2015] qui ont exploré la combinaison des ressources de traduction lexicales et statistiques. En effet, les auteurs ont utilisé Wikipedia et un dictionnaire électronique comme connaissance de traduction lexicale. En deuxième lieu, ils ont exploré des corpus parallèles pour extraire statistiquement les candidats de traduction. [Kim *et al.*, 2015] ont prouvé que l'utilisation conjointe des trois ressources de traduction (c-à-d un dictionnaire, un corpus parallèle et des connaissances de Wikipedia) donne de meilleurs résultats en comparaison avec l'utilisation d'une seule ressource.

5. La lemmatisation a pour objectif de retrouver le lemme d'un mot, par exemple l'infinitif pour les verbes. La racinisation consiste à supprimer les préfixes/suffixes des mots, ce qui peut résulter en un mot qui n'existe pas dans la langue.

Approche de traduction	**Description de l'approche**
Basée sur dictionnaire bilingue	Se base sur un ensemble de mots dans une lange source avec leurs traductions dans une langue de destination. Elle présente un problème de couverture des mots (noms propres, termes modernes/techniques, etc.)
Basée sur un corpus (parallèle/comparable)	**Corpus parallèle :** Chaque phrase est alignée avec sa traduction **Corpus comparable :** les traductions ne correspondent pas exactement aux phrases ; mais elles couvrent le même sujet. --- (+) Donne de meilleures performances que les dictionnaires bilingues. (-) Les corpus ne sont pas disponibles dans toutes les langues ; dans ce cas la construction des corpus parallèles demande des efforts considérables de collecte, traitement et validation des données (cette validation est souvent manuelle).
Traduction Automatique (TA) (*Machine translation*)	**Approche 1 :** Traduire les requêtes dans la langue destination des documents (-) Les requêtes sont souvent courtes et ne sont pas traduite correctement faute de non significativité du contexte des mots **Approche 2 :** Traduire les documents dans la langue source des requêtes (-) Les documents courts manquent de structuration sémantique suffisante pour identifier les contextes des termes ambigus. (-) La traduction automatique des documents longs est coûteuse en terme de traitement informatisé. --- La TA est souvent inefficace dû au coût de traitement et sa non disponibilité pour certains couples de langues.
A base de co-occurrence	Utilisée souvent pour la désambiguïsation des traductions. Nécessite l'utilisation d'un dictionnaire bilingue et d'un corpus monolingue pour assurer la tâche de désambiguïsation en calculant le degré de similarité sémantique entre les traductions possibles et les termes qui co-occurrent (c-à-d représentants le même contexte) dans le corpus.
A base d'ontologie	Basée sur une spécification explicite de conceptualisation. Exploite la combinaison de connaissances ontologiques et leurs liaisons avec des dictionnaires [Monti *et al.*, 2013]. (-) Manque de couverture dans les ontologies pré-construites (-) Complexité d'appariement entre concepts d'ontologie et définitions du dictionnaire.
A base de *Wikipedia*	*Wikipedia* constitue une ressource multilingue riche et disponible librement (environ 6 millions d'articles dans 250 langues en évolution continue) [Sorg et Cimiano, 2012] Les articles traitant le même sujet dans plusieurs langues sont inter-liés. Ces liens peuvent être exploitées dans la désambiguïsation des traductions.

Table 1.2 – Comparaison des approches de traduction dans la RI translinguistique

L'utilisation des méthodes de plongements de mots (ou de « *word embeddings* ») semble

prometteuse selon des travaux plus récents [Vulić et Moens, 2015, Mitra et Craswell, 2017]. Ces méthodes se focalisant sur l'apprentissage d'une représentation de mots d'un dictionnaire par un vecteur de nombres réels correspondant. Ceci facilite notamment l'analyse sémantique des mots. Selon cette nouvelle représentation, les mots apparaissant dans des contextes similaires possèdent des vecteurs correspondants qui sont relativement proches.

En conclusion, la RI translinguistique fournit une nouvelle technique pour la recherche de documents à travers différentes langues. En utilisant les différentes techniques de traduction, la RI translinguistique permet de fournir le résultat de recherche dans une langue autre que la langue de requête.

Le travaux de l'état de l'art ont montré qu'il est plus pratique de traduire uniquement la requête que tous les documents [Nie, 2010, Zhou *et al.*, 2012]. La traduction de documents en utilisant la traduction automatique est coûteuse en termes de calcul et en terme de taille de la collection de documents qui est importante.

Afin d'aboutir à une bonne performance, les systèmes de RI translinguistiques doivent défier principalement les points suivants :

- Les requêtes et les documents sont rédigés dans des langues différentes et doivent donc être traduits ;
- Un document peut être rédigé en plusieurs langues. Une phase d'identification automatique de la langue est primordiale dans ce cas de figure ;
- Les mots d'une requête peuvent être ambigus. Donc, une phase de désambiguïsation sémantique doit être effectuée ;
- Les requêtes sont généralement courtes. L'expansion peut enrichir éventuellement le contexte de la requête initiale ; ce qui facilite la traduction et la désambiguïsation.

1.3 Évaluation des systèmes de recherche d'information

L'évaluation des SRI constitue une étape importante dans l'élaboration d'un modèle de recherche d'information. Elle permet de caractériser chaque modèle et de fournir des éléments de comparaison entre les modèles existants.

Ainsi une telle évaluation intervient lors de la mise en œuvre d'un SRI puisqu'elle permet de paramétrer son modèle, d'estimer l'impact de chacune de ses caractéristiques et enfin de fournir des éléments de comparaison avec les autres SRI.

Un SRI idéal a deux objectifs : retrouver tous les documents pertinents et rejeter tous les documents non pertinents. Ces deux objectifs sont évalués principalement par des mesures de précision et de rappel que nous définissons ci-après. Les critères d'évaluation reposant sur le temps de réponse et l'espace utilisé pour le stockage d'information semblent être plus ou moins importants lors de l'évaluation des SRI [Baccini *et al.*, 2012].

1.3.1 Les mesures de RI

Plusieurs mesures ont été citées dans la littérature pour l'évaluation des SRI. Nous présentons ci-après celles les plus importantes.

Le rappel

La mesure de *rappel* est définie par le nombre de documents pertinents retrouvés au regard du nombre de documents pertinents que possède la base de données. Cela signifie que lorsque l'utilisateur interroge un système de RI, il souhaite voir apparaitre tous les documents qui pourraient répondre à son besoin d'information. Si cette adéquation entre le questionnement de l'utilisateur et le nombre de documents présentés est importante alors le taux de rappel est élevé. A l'inverse si le système possède de nombreux documents intéressants mais que ceux-ci n'apparaissent pas, on parle alors de *silence*. Le *silence* correspond à la proportion complémentaire du rappel représentant les documents pertinents non retrouvés.

Le rappel et le silence se calculent par les formules suivantes [Manning *et al.*, 2008] :

$$Rappel = \frac{|P \cap R|}{|P|} \in [0,1] \quad et \quad Silence = 1 - Rappel \tag{1.1}$$

Avec :

- P : représente l'ensemble de documents pertinents dans tout le corpus.
- R : représente l'ensemble de documents retrouvés par le SRI.

La précision

La mesure de *précision* représente le nombre de documents pertinents retrouvés rapporté au nombre de documents total retournés par le SRI pour une requête donnée. Le principe est le suivant : quand un utilisateur interroge le SRI, il souhaite que les documents proposées en réponse à son interrogation correspondent à son attente. Tous les

documents retournés superflus ou non pertinents constituent du *bruit*. Le *bruit* correspond à la proportion de documents retrouvés qui ne sont pas pertinents. Si la précision est élevée, cela signifie que peu de documents inutiles sont proposés par le SRI et que ce dernier peut être considéré comme « précis ».

On calcule la précision et le bruit selon les formules suivantes [Manning *et al.*, 2008] :

$$Précision = \frac{|P \cap R|}{|R|} \in [0,1] \quad et \quad Bruit = 1 - Précision \tag{1.2}$$

Par effectuer ces mesures, il faut disposer des réponses idéales aux requêtes en question (dites aussi « pertinence utilisateur »). Il s'agit des jugements manuels de pertinence des documents par rapport à un ensemble de requêtes de test. La figure 1.6 résume les aspects d'évaluation décrits auparavant.

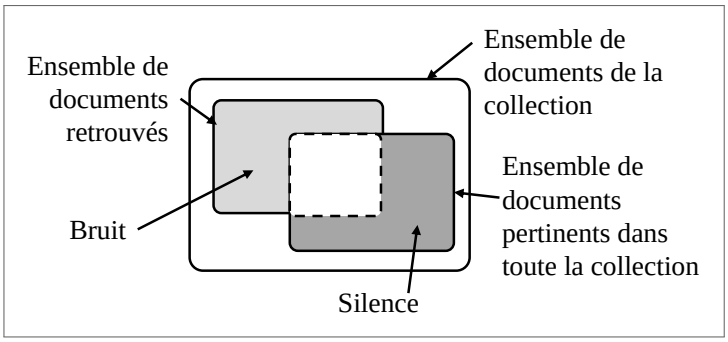


Figure 1.6 – Modélisation de la pertinence système vs pertinence utilisateur

La F-mesure

La F-mesure correspond à une moyenne harmonique de la précision et du rappel. Elle diminue lorsque l'un de ses paramètres est petit et vice versa selon la formule qui suit :

$$F - mesure = \frac{(1+\beta^2) précision \times rappel}{(\beta^2 \times précision) + rappel} \tag{1.3}$$

Le paramètre β permet de pondérer la précision ou le rappel, il est égal généralement à 1.

La précision moyenne (*MAP*) et la précision exacte (*R-précision*)

Les mesures de précision moyenne (*MAP*) et de précision exacte (*R-précision*) sont souvent utilisées dans les travaux de RI. Elle permettent de donner des mesures de

performances plus fines que les mesures de rappel et précision classiques [Manning *et al.*, 2008].

La précision moyenne est calculée comme suit :

$$MAP = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{|rel_j|} \sum_{r=1}^{N_j} P(r) \times isRel(r) \tag{1.4}$$

Avec :

$$P(r) = \frac{Nombre\ de\ documents\ pertinents\ trouvés\ au\ rang\ r\ ou\ moins}{r} \tag{1.5}$$

Où :

- $|Q|$: nombre total de requêtes ;
- $|rel_j|$: nombre de documents pertinents pour la requête j dans toute la collection ;
- N_j : nombre de documents retournés par la requête j ;
- $isRel(r)$: fonction binaire indiquant si le résultat au rang r est un document pertinent (1) ou pas (0).

La précision exacte (*R-précision*) correspond à :

$$R - précision = \frac{1}{|Q|} \sum_{j=1}^{|Q|} Précision(\{D_{kj}\}) \tag{1.6}$$

Avec : $Précision(\{D_{kj}\})$ correspond à la précision des k-premiers résultats de la requête j, où k désigne le nombre de documents pertinents associés à la requête j dans la collection.

Autres mesures de RI

Une autre mesure utilisée dans l'évaluation des SRIs est le critère *MRR* (*Mean Reciprocal Rank*) [Voorhees, 1999]. Le rang réciproque (en anglais « *reciprocal rank* ») d'une requête correspond à l'inverse du rang du premier document pertinent retourné ; c-à-d. 1 si le premier document retourné est pertinent, 1/2 si en deuxième rang, 1/3 en troisième rand, etc. Si aucun document pertinent n'est retourné, la valeur du rang réciproque est égale à 0. Le calcul de la moyenne des rangs réciproques *MRR* est ainsi proportionnelle à :

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rang_i} \tag{1.7}$$

Avec :

- $|Q|$: nombre total de requêtes ;
- $rang_i$: rang du premier document pertinent retourné pour la $i^{ème}$ requête.

La différence principale de la mesure *MRR* avec la mesure de précision moyenne *MAP* réside dans le fait que seul le premier document pertinent retourné compte dans l'évaluation. Autrement, les autres documents pertinents (ou non pertinents) situés après le rang réciproque n'influencent pas l'interprétation de la performance de RI selon le critère *MRR*.

Il existe d'autres mesures moins répandues dans la RI mais souvent utilisées dans d'autres domaines tel que la classification et l'extraction d'information. Nous citons, à titre d'exemple, la mesure *ROC* (*Receiver-Operating-Characteristic*) couramment utilisée dans le domaine de la médecine pour évaluer la validité des tests diagnostiques. Le traçage de la courbe *ROC* en RI consiste à présenter le rappel en fonction du taux de faux positifs [6]. L'aire *AUC* (désigné en anglais par « *Area under Curve* ») sous la courbe *ROC* peut être interprétée comme la probabilité, quand on prend deux échantillons (un positif et un négatif), que le système de RI identifie mieux le positif que le négatif.

1.3.2 Les collections standards de test

Une collection de tests comprend un ensemble de documents, un ensemble de requêtes et la liste des documents de la collection pertinents pour chaque requête. L'évaluation repose alors sur la comparaison de la liste de documents retrouvés par un système et celle des documents pertinents [Manning *et al.*, 2008]. Nous présentons dans ce qui suit un aperçu de quelques collections standards de test :

CRANFIELD : Il s'agit de l'une des premières collections de tests à donner des mesures quantitatives de l'efficacité de la recherche d'information. Néanmoins, de nos jours, elle devient trop petite et n'offre que les expérimentations les plus élémentaires. Recueillie au Royaume-Uni à partir de la fin des années 1950, elle contient 1398 résumés des articles de journaux de l'aérodynamique, une série de 225 requêtes, et des jugements de pertinence exhaustive de tout couple (requête, document) [Cleverdon, 1967].

REUTERS : Pour la classification de texte, la collection de test la plus utilisée est la collection nommée *Reuters-21578* contenant 21578 d'articles de presse. Plus récemment, l'agence *Reuters* a publié un corpus documentaire plus grand, nommé *Reuters Corpus Volume 1* (RCV1) et composé de 806791 documents. Son ampleur

6. le taux de Faux Positifs (*FP*) est quantifié par la fraction des documents non pertinents qui sont retournés sur le nombre total des documents non pertinents.

et ses riches annotations le favorisent pour être une bonne base des travaux touchant la classification dans la RI [Manning *et al.*, 2008] [7] .

20 NEWSGROUPS : Il s'agit d'une autre collection de test largement utilisée pour la classification. Elle est composée de 1000 articles pour chacun des 20 *« newsgroups »* utilisés de Usenet [8] (le nom du « *newsgroups* » étant considéré comme catégorie. ex : *rec.sport.hockey*, *talk.politics.mideast*, *sci.electronics*, etc.). Après la suppression des doublons d'articles, elle contient exactement 18941 articles [Manning *et al.*, 2008] [9].

TREC : *Text REtrieval Conference* (TREC). L'Institut National des Standards et de la Technologie (NIST) des États-Unis a organisé de grandes compagnes d'évaluation de RI depuis 1992. Les collections de tests les plus connues sont celles utilisées pendant les 8 premières évaluations TREC entre 1992 et 1999. Au total, elles comprennent 6 CD contenant 1,89 millions de documents (principalement, mais pas exclusivement, des articles de presse) et des jugements de pertinence pour 450 des besoins d'information (requêtes), qui sont appelés *« topic*s ». Au début, les collections TREC étaient composées chacune de 50 requêtes de test, évaluées sur différents ensembles de documents (même sur des documents des anciennes années). Les versions TREC 6-8 fournissent 150 requêtes sur environ 528.000 d'articles de presse et des services de diffusion d'information étrangères. Dans [Manning *et al.*, 2008], les auteurs considèrent que les collections TREC sont les meilleurs corpus de test à utiliser dans les travaux de RI [10].

CLEF : *Cross Language Evaluation Forum* (CLEF). Cette série d'évaluation s'intéresse aux langues européennes et à la recherche d'information monolingue, bilingue et multilingue [Braschler et Peters, 2004] [11].

Il faut noter que cette liste n'est pas exhaustive [12]. Pour qu'un corpus de tests soit significatif, il faut qu'il possède un nombre de documents assez élevé (un corpus de taille moyenne contient au moins 100.000 documents). Il faut avoir aussi quelques dizaines de requêtes, traitant des sujets variés, pour que l'évaluation soit la plus objective et significative.

7. http://about.reuters.com/researchandstandards/corpus/
8. Usenet (également connu sous le nom Netnews) est un système en réseau de forums, inventé en 1979 et basé sur le protocole NNTP. C'est un ensemble de protocoles servant à générer, stocker et récupérer des « articles » (des messages qui sont proches, dans leur structure, des courriels), et permet l'échange de ces articles entre les membres d'une communauté qui peut être répartie sur une zone potentiellement très étendue. Usenet est organisé autour du principe de groupes de discussion ou groupes de nouvelles (en anglais *« newsgroups »*), qui rassemblent chacun des articles (contributions) sur un sujet précis.
9. http://mlg.ucd.ie/datasets/20ng.html
10. http://trec.nist.gov
11. http://www.clef-campaign.org
12. voir site de l'European Language Ressources Association (ELRA) : http://catalog.elra.info

Le jugement de pertinence des documents vis-à-vis des requêtes constitue une phase critique dans la construction des corpus de tests. Ceci peut remettre la question sur la nécessité du recours à de tels standards de test (« *benchmark* » en anglais) et leurs influences sur la créativité des chercheurs du domaine de la RI [Voorhees et Harman, 2005, Daoud *et al.*, 2010].

Conclusion

Nous avons détaillé dans ce chapitre les différents acteurs qui interviennent dans un système de recherche d'information. Nous avons montré que le besoin de l'utilisateur ne doit pas se limiter uniquement à sa requête. Son interaction avec le système est aussi une composante essentielle pour améliorer la qualité de la recherche. En fait, pour satisfaire davantage le besoin d'information d'un utilisateur et pour l'accompagner dans le processus de recherche, d'autres techniques ont été introduites telles que l'expansion et la désambiguïsation de requêtes. Nous nous focalisons dans le chapitre suivant sur la désambiguïsation sémantique monolingue et translinguistique.

Chapitre

2 Désambiguïsation Sémantique Monolingue et Translinguistique

Sommaire

" We live in a world where there is more and more information, and less and less meaning."

— Jean Baudrillard

Introduction

La désambiguïsation sémantique des textes (WSD) consiste à identifier automatiquement le ou les sens possibles d'un mot polysémique dans un contexte donné. Elle constitue une tâche fondamentale pour la recherche d'information (RI) et le traitement automatique des langues (TAL).

Le développement et l'amélioration des techniques de désambiguïsation sémantique ouvrent de nombreuses perspectives prometteuses pour la RI. En effet, la désambiguïsation contribue à l'amélioration des performances des SRI en augmentant la pertinence des documents retournés par le SRI : Elle consiste à se focaliser sur le sens dominant de la requête et à se détacher de ses sens secondaires selon le contexte et filtrer ainsi les réponses retournées par le système en retournant celles qui sont sémantiquement pertinentes.

Nous présentons dans ce chapitre une étude des ressources utilisées lors de la mise en place d'un processus de désambigüisation, l'ensemble des mesures appliquées, ainsi que les différentes approches de WSD monolingue et translinguistique.

2.1 Historique et domaines d'application

Le problème de désambiguïsation sémantique était discuté à travers plusieurs travaux et compagnes de recherche sur la RI et WSD à savoir : TREC (Text Retrieval Evaluation Conference), TREC-QA [Voorhees et Harman, 2005], ACE (Automatic Content Extraction) [Doddington *et al.*, 2004] et Senseval [Kilgarriff, 1998].

Historiquement, l'intérêt sur le problème de désambiguïsation sémantique a commencé dans les années 1980s. En effet, de nombreuses ressources lexicales ont été développées tels que les dictionnaires électroniques, les glossaires, les thésaurus et les ontologies. Le travail sur les définitions (c-à-d les dictionnaires) a été initié par Lesk, en 1986, qui a proposé de relier des définitions de mots s'ils ont des mots communs [Lesk, 1986]. Ensuite, d'autres travaux ont étendu cette approche [Véronis et Ide, 1990, Wilks *et al.*, 1990, Basile *et al.*, 2014, Sawhney et Kaur, 2014].

Plus récemment, l'apparition des corpus documentaires et leur multitude ont inspiré les chercheurs à proposer des nouvelles approches en utilisant les méthodes statistiques basées sur la co-occurrence [Van Rijsbergen, 2004]. L'idée consiste à analyser les mots qui co-occurrent avec des mots polysémiques dans de grands corpus. Ces systèmes sont pertinents, car ils visent à modéliser le sens de chaque mot en fonction de son contexte, à partir de corpus sémantiquement étiqueté. L'adéquation entre un sens donné et le

mot à traiter est calculée en utilisant une mesure de similitude entre les caractéristiques des sens modélisées et celles du contexte considéré.

Les compagnes Senseval et SemEval ont présenté des études comparatives des problèmes de désambiguïsation en utilisant des corpus [Navigli, 2009, Erk et Strapparava, 2010]. Les résultats ont prouvé que les approches fondées sur des corpus d'apprentissage ont atteint des taux de réussite plus élevés que les autres. En effet, les meilleurs taux de réussite était de 80% pour les noms, 70% pour les verbes et 75% pour les adjectifs.

Il existe diverses applications de la désambiguïsation des sens de mots dont certaines sont les suivantes :

Recherche d'information (RI) : Résoudre l'ambiguïté dans une requête est un des problèmes les plus importants dans un SRI. A titre d'exemple, le mot « *dépression* » dans une requête peut avoir des significations différentes selon le contexte : (i) *maladie*, (ii) *dépression pneumatique* ou (iii) *économique* (de sens *crise économique*). Ainsi, trouver le sens exact d'un mot ambigu dans une requête particulière avant de retourner sa réponse est un problème à résoudre à cet égard.

Extraction de connaissance (EC) : La WSD joue un rôle important pour l'extraction de l'information dans différents travaux de recherche à savoir dans la recherche en bio-informatique, la reconnaissance d'entités nommées, la résolution de coréférence [1], l'expansion des acronymes (ex. *MG* désigne magnésium ou milligramme), etc. Tous ces domaines peuvent également être exprimés en tant que problèmes de WSD pour les noms appropriés .

Traduction automatique (TA) : La WSD est requise pour la TA car la traduction de mots d'une langue vers une autre peut avoir des traductions différentes basées sur les contextes de leur utilisation. Par exemple, les deux phrases « *Il a marqué un but* » et « *Ce fut son but dans la vie* », le mot « *but* » porte des significations différentes ; ce qui pose un problème d'ambiguïté lors de sa traduction [Xiong et Zhang, 2014, Neale *et al.*, 2016].

Analyse de contenu : Il existe différents sous-domaines d'application de l'analyse de contenu dont on distingue l'*analyse d'opinion* (en anglais *opinion mining*) et l'*analyse de sentiment* (connu en anglais par *sentiment analysis*). Ces applications peuvent bénéficier de la WSD. Par exemple, la classification des blogs a récemment suscité de plus en plus d'intérêt dans la communauté Internet. En effet, à mesure que le nombre de blogs augmente à un rythme exponentiel, il y a besoin d'un moyen simple et efficace en même temps de les classer, de déterminer leurs principaux sujets et d'identifier les connexions pertinentes entre les blogs et même entre des publications du blog unique. Un deuxième domaine de recherche

1. En linguistique, la coréférence est le phénomène qui consiste à désigner le même objet par plusieurs expressions différentes contenues dans une phrase ou dans un document, .

connexe est celui de l'analyse de réseaux sociaux, qui devient de plus en plus actif avec les évolutions récentes du Web [Singh *et al.*, 2013, Liu *et al.*, 2015].

Lexicographie : La lexicographie moderne est basée principalement sur l'analyse des corpus, de sorte que le WSD et la lexicographie peuvent fonctionner en boucle. En effet, la WSD fournit des groupes empiriques de sens approximatifs et des indicateurs contextuels statistiquement significatifs du sens pour les lexicographes, qui eux-même fournissent des inventaires de sens et des corpus annotés utiles pour la WSD. De plus, les dictionnaires intelligents et les thésaurus aident à fournir un dictionnaire à référencement sémantique ainsi que de meilleures fonctionnalités contextuelles de recherche de sens.

2.2 Ressources lexicales

Afin d'assurer la tâche de désambiguïsation sémantique, les sources de connaissance fournissent des données qui sont essentielles pour associer les sens aux mots ambigus. Ce type de ressources est utilisé principalement dans les approches de désambiguïsation *à base de connaissances*.

L'éventail de ces ressources couvre les corpus textuels (soit en format annoté avec des sens de mots ou en format brut non annoté), les dictionnaires, les thésaurus, les glossaires, les ontologies, etc. L'ensemble des ressources lexicales peut être classé en deux principales catégories comme suit :

A. Les ressources structurées :

Les dictionnaires sont devenus une source de connaissances utilisée pour le traitement du langage naturel depuis les années 1980s, lorsque les premiers dictionnaires ont été rendus disponibles en format électronique. D'autre part, il y a eu recours aux thésaurus qui présentent des relations entre les mots telles que la synonymie et l'antonymie [Kilgarriff, 2003].

L'apparition de WordNet, comme inventaire de sens et dictionnaire à la fois, a joué un rôle très important dans la recherche d'information, le traitement du langage naturel et également la désambiguïsation sémantique.
En effet, WordNet constitue la ressource la plus utilisée pour la désambiguïsation sémantique en anglais grâce à la richesse des relations modélisées entre ses entrées sous forme d'une hiérarchie de sens. WordNet est structuré autour d'entités, nommées *synsets*, qui représentent un ensemble de synonymes en formant ainsi un concept. Les *synsets* représentent également des sens de mots inter-liés entre eux par des relations sémantiques [Barque et Chaumartin, 2008].

B. Les ressources non structurées :

Ces types de ressources peuvent être de natures :

– **Corpus documentaires** : c'est-à-dire les recueils de textes utilisés pour l'apprentissage des modèles linguistiques. Les corpus peuvent être annotés par des sens ou en format brut (c'est-à-dire non annotés). Les deux types de ressources, respectivement avec ou sans annotation, sont utilisés dans la désambigüisation sémantique, et sont plus utiles dans les approches supervisées et non supervisées, respectivement. Les tableaux 2.1 et 2.2 présentent une comparaison de quelques corpus en format brut et en format annoté.

Nom	# de mots	Référence
Brown Corpus	1 million	[Francis, 1965]
British National Corpus (BNC)	100 millions	[Leech, 1993]
Wall Street Journal (WSJ)	30 millions	[Paul et Baker, 1992]
American National Corpus	22 millions	[Ide et Suderman, 2006]
Gigaword Corpus	2 milliards	[Graff *et al.*, 2007]
Europarl	~54 millions	[Koehn, 2005]

Table 2.1 – Quelques corpus en format brut utilisés dans la désambiguïsation sémantique

Nom	# de contextes annotés	Référence
SemCor	234 000	[Miller *et al.*, 1994]
Line-hard-serve	4 000	[Leacock *et al.*, 1998]
Interest	2 369	[Bruce et Wiebe, 1994]
Defence Science Organisation	192 800	[Ng et Lee, 1996]
Open Mind Word Expert	288	[Chklovski et Mihalcea, 2002]
Groningen Meaning Bank (GMB)	666 562	[Basile *et al.*, 2012]
ROMANSEVAL	3 624	[Segond, 2000]

Table 2.2 – Aperçu de quelques corpus annotés. Tous ces corpus sont annotés avec différentes versions de WordNet, à l'exception du corpus Interest (étiqueté avec les sens LDOCE), GMB et ROMANSEVAL. Ce dernier supporte les langues française et italienne.

– **Ressources de collocation** : ce type de ressources présente la tendance de co-occurrence d'un mot avec d'autres à partir de l'analyse des corpus. Parmi les exemples de ces ressources, nous citons : *Word Sketch Engine*, *JustTheWord*, *The British National Corpus collocations*, *The Collins Cobuild Corpus Concordance*, etc. [Navigli, 2009]

Le corpus Web1T [2] constitue un des plus grand jeu de données contenant les co-occurrences des textes [Islam et Inkpen, 2009]. En effet, il regroupe les fréquences d'apparition des groupes de mots (allant à 5 termes) dans un corpus de plus de 1 trillion de mots issues du Web.

Les Voisins De Le Monde [3] est une base lexicale distributionnelle en français, construite automatiquement à partir d'un corpus d'articles du journal *Le Monde* composé de 200 millions de mots. Cette ressource présente les co-occurrences syntaxiques [4] d'un terme à partir du corpus ainsi que ses voisins distributionnels [5].

– **Autres types de ressources :** tels que les lexiques, les listes des fréquences de mots, les listes de mots vides, etc.

2.3 Mesures de similarité pour la désambiguïsation sémantique

Le calcul de la mesure de similarité sémantique est très utilisé dans plusieurs domaines de l'informatique. La mise en place de telles métriques répond à des problèmes posés dans le domaine de TAL, la RI et la désambiguïsation sémantique.

Dans les travaux de [Harispe *et al.*, 2015], une étude approfondie des mesures de similarité sémantique a été établi de façon exhaustive. Ces mesures permettent de déterminer la similarité entre des termes ou concepts sans qu'il y est nécessairement une ressemblance syntaxique. Leurs utilisations reposent généralement sur une bonne organisation des documents en structure de type linéaire (ex. inventaire de sens) ou hiérarchique grâce à l'utilisation de bases de connaissances comme les ontologies. Nous présentons dans ce qui suit un résumé des mesures les plus utilisées.

2.3.1 Mesure de Lesk

L'algorithme de *Lesk* évalue la proximité sémantique entre deux sens (S_1, S_2) comme le nombre de mots communs dans les définitions correspondantes ($\mathcal{D}(S_1), \mathcal{D}(S_2)$) issues d'un dictionnaire [Lesk, 1986] :

$$Sim_{Lesk}(S_1, S_2) = |\mathcal{D}(S_1) \cap \mathcal{D}(S_2)| \tag{2.1}$$

2. corpus contribué par Google Inc.
3. http://redac.univ-tlse2.fr/voisinsdelemonde/ [dernière consultation 10-04-2016]
4. par exemple, le nom "peur" apparaît de façon très régulière dans les contextes syntaxiques "trembler de ~", "frissonner de ~", "~ tenailler", "exorciser ~", etc.
5. par exemple, le nom "traité" a pour voisins "convention", "accord", "constitution" car tous ces noms ont comme cooccurrents syntaxiques : "stipulation de ~", "ratifier ~", "renégociation de ~", "ratification de ~", "signataire de ~", "signature de ~", etc.

Une autre variante de l'algorithme de *Lesk* consiste à comparer chaque sens candidat avec le contexte du mot w à désambiguïser. Le contexte est représenté par la phrase dans laquelle le mot polysémique apparaît. Étant donné un mot cible w, le score suivant est calculé pour chaque sens S :

$$Sim_{LeskVar}(S) = |contexte(w) \cap \mathcal{D}(S)| \tag{2.2}$$

Avec : $contexte(w)$ est l'ensemble des mots contenu dans une fenêtre de contexte autour du mot cible w.

Cependant, l'algorithme de *Lesk* est très sensible à la présence des mots. En effet, une absence des mots discriminants, c'est-à-dire représentant fortement les sens dans les définitions, retourne des résultats non significatifs. En outre, l'algorithme détermine les chevauchements seulement parmi les gloses des sens considérés.

L'extension la plus classiquement utilisée est celle proposé par [Banerjee et Pedersen, 2002] sous le nom de « *Lesk étendu* » . Cette mesure nécessite une ressource composée de définitions pour les sens de mots qui doivent être inter-liés. Il s'agit ainsi d'enrichir la définition initiale du sens par les mots des définition des sens (S') qui lui sont liés, soit :

$$Sim_{LeskEtendu}(S) = \sum_{S' : S \underrightarrow{rel} S' ou\ S = S'} |contexte(w) \cap \mathcal{D}(S')| \tag{2.3}$$

Avec : $\mathcal{D}(S')$ est l'ensemble de mots appartenant à la définition textuelle d'un sens S' qui est soit S lui-même, soit relié à S par une relation rel. [Banerjee et Pedersen, 2002] ont montré que la désambiguïsation profite amplement des informations issus de concepts connexes par rapport à la version d'origine de l'algorithme *Lesk*.

2.3.2 Mesures de similarité basées sur les corpus

Les mesures de similarité sémantique peuvent également être obtenues en appliquant une analyse statistique sur les documents d'un corpus et en utilisant les techniques de traitement du langage naturel. L'avantage est que les mesures axées sur le corpus sont indépendantes. En effet, elles n'ont pas besoin de ressources de connaissances externes, ce qui peut surmonter le problème de couverture dans les taxonomies. Trois orientations principales ont été poursuivies dans cette catégorie d'approches :

Similarité basée sur la co-occurrence : Cette approche étudie les mots qui co-occurrent dans les textes avec l'hypothèse que les couples de mots fréquents révèlent

l'existence d'une certaine dépendance entre ces mots. La première mesure a été introduite par [Church et Hanks, 1990] et désigné par *Mutual Information* (MI).
Les travaux de [Turney, 2001] ont montré que la mesure *Pointwise Mutual Information* (PMI) calculée sur un très grand corpus (issu du web) et en utilisant une fenêtre de co-occurrence de taille moyenne peut être efficacement utilisé pour extraire des synonymes. D'autres travaux plus récentes en linguistique ont utilisé la mesure (PMI) pour extraire des collocations et des associations entre mots [Role et Nadif, 2011, Damani, 2013].

Similarité basée sur le contexte : Cette approche est basée sur l'intuition que des mots semblables tendent à se produire dans des contextes similaires [Manning *et al.*, 2008]. Le modèle d'espace vectoriel est ici utilisé comme dispositif de mesure sémantique. En effet, cette méthode utilise l'espace vectoriel comme espace de mot et ses dimensions ne sont que des mots. En outre, un mot qui se trouve dans un corpus sera désigné comme vecteur et sa fréquence d'apparition sera comptée dans son contexte. Ensuite, une matrice de co-occurrence est créée et des mesures de similarité issus des techniques de *clustering* sont appliquées tels que l'algorithme *k-nearest neighbor (k-NN)* [Niu *et al.*, 2004, Anaya-Sánchez *et al.*, 2006, Chen *et al.*, 2014].
[Reisinger et Mooney, 2010] ont proposé un modèle multi-prototype d'espace vectoriel, où les contextes de chaque mot sont regroupés en *cluster* (ou groupe), puis chaque *cluster* génère un vecteur distinct pour un mot en faisant la moyenne sur tous les vecteurs contextuels dans le *cluster*. Dans [Huang *et al.*, 2012], les auteurs ont introduit des vecteurs distribués continus, basés sur des modèles probabilistes en utilisant les réseaux de neurones, pour les représentations des mots.

Latent Semantic Analysis (LSA) : La LSA est une méthode de représentation du sens contextuel des mots à l'aide des calculs statistiques sur un large corpus, sous la forme d'un espace sémantique vectoriel de grande dimension [Landauer et Dooley, 2002]. Ces calculs permettent d'inférer des relations profondes entre mots ou ensembles de mots en analysant la distribution des mots dans la somme de leurs contextes.
Dans le domaine de WSD, l'intérêt de cette technique est de permettre la construction automatique de connaissances sémantiques génériques (indépendantes du domaine) [Bestgen, 2006]. La méthode LSA peut présenter des limites si on réduit la WSD à un problème de classification de sens [Roche et Chauché, 2006]. Cependant, les approches de modélisation en espace vectoriel (appelé en anglais *Vector Space Model*) de façon générale ont prouvées leurs utilités dans le traitement sémantique des textes [Turney et Pantel, 2010].

Autres mesures de similarité distributionnelle : Les mesures de similarité géométriques *Cosinus, Jaccard* et *Dice* sont largement utilisées dans le domaine de la RI et la WSD [Bannour *et al.*, 2011, Tyar et Win, 2015]. En partant de la représentation vectorielle des contextes, ces mesures peuvent être exprimés ainsi :

$$Cosinus(\vec{v},\vec{w}) = \frac{\vec{v}\cdot\vec{w}}{\|\vec{v}\|\|\vec{w}\|} \tag{2.4}$$

$$Jaccard(\vec{v},\vec{w}) = \frac{\vec{v}\cdot\vec{w}}{\|\vec{v}\|^2+\|\vec{w}\|^2-\vec{v}\cdot\vec{w}} \tag{2.5}$$

$$Dice(\vec{v},\vec{w}) = \frac{2\vec{v}\cdot\vec{w}}{\|\vec{v}\|^2+\|\vec{w}\|^2} \tag{2.6}$$

Avec : $\vec{v}\cdot\vec{w} = \sum v_i w_i$ et $\|\vec{v}\| = \sqrt{\sum(v_i)^2}$.

Dans les travaux de [Bannour *et al.*, 2011], les expérimentations ont donnée faveur à la mesure *Cosinus* qui présente de meilleurs résultats par rapport aux mesures *Jaccard* et *Dice*. En effet, ces deux dernières mesures sont incompatibles avec l'ensemble des expérimentations menées vu qu'elles font remonter des termes non pertinents.

2.3.3 Mesures de similarité basées sur les ressources lexicales structurées

Les mesures basées sur la distance taxonomique dans un thésaurus représentent une parmi les mesures couramment utilisées dans les ressources lexicales structurées. En effet, plusieurs travaux utilisent cette propriété de proximité entre les concepts tels ceux existants dans WordNet. Le principe général de ces mesures est de compter le nombre d'arcs qui séparent deux sens dans WordNet.

Z. Wu et M. Palmer [Wu et Palmer, 1994] définissent la similarité selon la distance qui sépare deux concepts par rapport à leur sens commun le plus spécifique (S_3) que la racine de la taxonomie (voir figure 2.1). Ainsi, la similarité entre deux sens S_1 et S_2 est représentée par :

$$Sim_{wup}(S_1,S_2) = \frac{2\times depth(S_3)}{depth(S_1)+depth(S_2)} \tag{2.7}$$

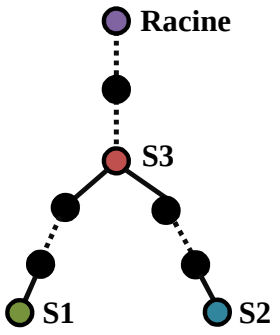


Figure 2.1 – Deux sens S_1 et S_2 et leur sens commun (S_3) le plus spécifique dans une taxonomie.

Où $depth(S_3)$ est le nombre d'arcs qui séparent S_3 de la racine et $depth(S_i)$ le nombre d'arcs qui séparent S_i de la racine en passant par S_3. Cette mesure a l'avantage d'avoir de bonnes performances

par rapport aux autres mesures de similarité [Lin, 1998, Sharma *et al.*, 2016, Jayakody, 2016].

La mesure de [Wu et Palmer, 1994] a été généralisée par [Stojanovic *et al.*, 2001] qui utilisent la profondeur des concepts dans une hiérarchie pour prendre en compte l'héritage multiple comme suit :

$$Sim_{Sto}(S_1, S_2) = \frac{depth(S_3) + 1}{(depth(S_1) + 1) + (depth(S_2) + 1) - (depth(S_3) + 1)} \tag{2.8}$$

Plusieurs autres mesures de similarité basées sur les arcs sont présentés dans la littérature [Tchechmedjiev, 2012, Batchkarov *et al.*, 2016]. Dans [Ngom, 2015], les expérimentations sur les coefficients de corrélation avec le jugement humain de [Miller et Charles, 1991] prouvent l'utilité de ces mesures en dépit de leurs dépendances de la structure hiérarchique de la taxonomie utilisée.

2.4 Approches de désambiguïsation sémantique

Dans ce qui suit, nous présentons les différentes approches répondant à la tâche de WSD. Ces approches sont classées en fonction de la source de connaissance de WSD, dans un premier lieu, et la façon avec laquelle elles sont structurées dans un deuxième lieu. Ces deux critères influencent fortement la performance de WSD [Vidhu Bhala et Abirami, 2014]. Les approches de WSD peuvent être classées dans quatre principales familles : (i) les approches de modélisation des connaissances ou de raisonnement, (ii) les approches supervisées ; (iii) les approches non supervisées ; et (iv) les approches hybrides.

2.4.1 Approches à base de connaissance

Ce type d'approche tente de modéliser la compréhension humaine de la langue à travers des modèles connexionnistes (basés sur les réseaux de neurones) ou symboliques (appelés aussi cognitifs) qui sont répandus dans le domaine de l'intelligence artificielle (IA) [Audibert, 2003]. Ces techniques conduisent à plusieurs développements dans le domaine de l'IA en partant de la notion des *réseaux sémantiques* [Masterman, 1961].

Toutefois, le grand volume des connaissances, nécessaires dans ce type d'approche pour résoudre l'ambiguïté des mots, est collecté et traité manuellement. D'autre part, ces approches utilisent des bases de connaissances spécifiques qui ne garantissent pas une couverture suffisante de la langue.

Afin de résoudre les problèmes liés aux méthodes supervisés de WSD (voir section 2.4.2), la WSD à base de connaissance semble prometteuse comme choix [Ponzetto et

Navigli, 2010]. En effet, sans avoir recours aux données issues de corpus, les systèmes à base de connaissances se basent uniquement sur l'information issue d'une base de connaissances lexicales (désignée en anglais par *Lexical Knowledge Base*) pour répondre au besoin de WSD .

Les approches à base de graphe, utilisant WordNet comme ressource, se distinguent dans cette catégorie en profitant des algorithmes de la théorie des graphes. Dans [Sinha et Mihalcea, 2007], les auteurs ont proposé des mesures de centralité de graphe [6], où les sens des mots dans le contexte sont liés avec des arêtes pondérées par des scores de similarité. [Navigli et Lapata, 2010] ont exploité un sous-graphe en utilisant les parcours en profondeur sur l'ensemble du graphe de WordNet, puis ont appliqué une série de mesures de centralité de graphe. En utilisant des jeux de données disponibles librement, [Agirre *et al.*, 2014] ont personnalisé l'algorithme *PageRank* de [Brin et Page, 1998] pour supporter l'exploration des connaissances en graphes.

Plus récemment, [Faralli et Navigli, 2012] ont proposé un nouveau cadre pour la WSD de domaine. En amont, une méthode d'apprentissage (dite également d'amorçage ou *bootstrapping* en anglais) a été développée dans le but d'obtenir des glossaires pour plusieurs domaines à partir du Web de façon itérative. En aval, les glossaires acquis comme inventaire de sens ont été utilisés dans la tâche de WSD.

[Pilehvar *et al.*, 2013] ont proposé une approche unifiée pour mesurer la similarité sémantique des sens des mots à différents niveaux lexicaux. Pour toutes sortes de données linguistiques, l'approche proposée tire parti d'une représentation probabiliste. Par ailleurs, et en utilisant la similarité sémantique à divers niveaux lexicaux dans trois expériences (sens général ou grossier du mot, similarité de mot et similarité sémantique des textes), la nouvelle représentation sémantique fournie par l'approche unifiée donne de meilleurs résultats que d'autres travaux de similarité qui sont souvent particulièrement conçu pour chaque niveaux à part (sens, mot et texte).

Ponzetto et Navigli ont proposé une approche basée sur l'extension de WordNet avec des millions de relations sémantiques générés à partir des articles de Wikipedia [Ponzetto et Navigli, 2010]. En effet, les sens dans WordNet sont automatiquement connectés avec des pages Wikipedia, et les relations sémantiques associatives pertinentes de Wikipedia sont reliées à WordNet. Ainsi, la ressource collectée devient plus riche. Les expérimentations, menées en utilisant cette ressource étendue de WordNet, ont consolidé l'importance de la prise en compte d'un grand nombre de relations sémantiques dans les systèmes à base de connaissances. En outre, Ponzetto et Navigli ont confirmé que les systèmes à base de connaissances peuvent donner de meilleures performances que les systèmes supervisés dans un scénario de WSD de domaine spécifique tel que

6. En théorie des graphes, les mesures de centralité servent à identifier les sommets les plus significatifs.

conclu par [Agirre *et al.*, 2010].

En étudiant l'ensemble de ces travaux, on remarque que WordNet était souvent la ressource linguistique de référence (ou *de facto*) la plus utilisée. Néanmoins, les travaux les plus récentes encouragent le développement et la collecte de ressources, autres que WordNet. Ces ressources sont par la suite utilisées dans la validation des approches de WSD sur des standard de test.

L'apparition des dictionnaires informatisés (ou électroniques) a contribué à l'automatisation de la tâche de WSD. Les approches basées sur les dictionnaires, consistent principalement à chercher parmi les mots qui co-occurrent leurs sens qui maximisent la similarité dans ce contexte de co-occurrence.

[Lesk, 1986] a proposé de relier chaque sens à la liste de mots apparaissant dans sa définition. D'autres mesures de similarité sémantique pour la WSD ont été développées (revenir en détails, dans ce chapitre, vers la section 2.3.1 page 34).

Une approche plus sophistiquée a été proposée par [Véronis et Ide, 1990] qui a généré des réseaux de neurones à partir des définitions du dictionnaire anglais CED (*Collins English Dictionary*). D'autres chercheurs ont essayé d'utiliser de plus amples informations dans la WSD tels que les liens sémantiques de LDOCE [Audibert, 2004].

[Brun *et al.*, 2001] ont proposé un système de WSD basée sur l'utilisation d'un dictionnaire électronique. Ce système a d'abord été conçu pour la WSD en anglais [Brun, 2000] et plus tard adapté au français. Ce système semble particulièrement prometteur dans de nombreuses applications : (i) il peut être intégré dans un système d'aide à la compréhension pour les langues étrangères ; (ii) il peut être intégré dans le processus d'indexation sémantique et plus généralement dans toute application adopté pour l'extraction et la compréhension de la connaissance contenue dans les documents électroniques.

2.4.2 Approches supervisées

L'utilisation des approches supervisées (appelées également approches d'apprentissage automatique) se base principalement sur des ressources annotées manuellement. En effet, cet ensemble de données annotées constitue des connaissances d'apprentissage qui permettent de ramener le problème de désambiguïsation vers la classification en déterminant le sens le plus approprié d'un mot selon son contexte donné.

Les méthodes supervisées les plus utilisées, dans la désambiguïsation sémantique, se basent sur l'utilisation des arbres de décision, les réseaux de neurones, les classifications naïves bayésiennes, les approches de type plus proches voisins et les machines à vecteurs de support (en anglais *Support Vector Machine* (SVM)) [Navigli, 2009].

Les approches supervisées tendent selon la littérature à donner de meilleurs résultats que les approches non supervisées, à la fois en termes de vitesse et de qualité. Cependant leur principal désavantage est qu'elles dépendent d'un grand volume de données textuelles qui doivent être annotées manuellement.

Autrement, les approches supervisées pour les WSD sont très dépendantes de la quantité de données d'apprentissage. Une façon d'augmenter la quantité de ces données est d'utiliser Wikipedia comme une source de données étiquetées sémantiquement [Mihalcea, 2007]. Lorsqu'un concept est mentionné dans un article de Wikipedia, le texte de l'article peut contenir un lien explicite vers la page Wikipedia des concepts, désignée par un identifiant unique.

Par conséquent, ce lien peut être utilisé en tant qu'annotation de sens. Par exemple, les mots ambigües « *avocat* » et « *fruit* » sont liés à différents articles de Wikipedia en fonction de leurs significations dans le contexte, y compris la page *AVOCAT (fruit)*, la page *AVOCAT (métier)*, et *FRUIT (alimentation humaine) ;* ainsi de suite, comme dans les exemples suivants de Wikipedia :

> ▶ L'avocatier (*Persea americana*) est une espèce d'arbre fruitier de la famille des Lauracées, originaire du Mexique et d'Amérique centrale. Il est notamment cultivé pour ses [[**Fruit (alimentation humaine)|fruit**]]s, les [[**Avocat (fruit)|avocats**]], ...
>
> ▶ Jacques Vergès, né le 20 avril 1924 au Laos, officiellement le 5 mars 1925 à Ubon Ratchathani au Siam (actuelle Thaïlande) et mort le 15 août 2013 à Paris, est un [[**avocat (métier)|avocat**]] franco-algérien.
>
> ▶ Cédric Anger (né en 1975 à Caen) est un réalisateur et scénariste français... Réalisateur : 2002 : Novela (court métrage) ; 2007 : Le Tueur ; 2011 : [[**L'Avocat (film, 2011)|L'Avocat**]]...

Ces phrases peuvent alors être ajoutées aux données d'apprentissage d'un système supervisé. Pour utiliser Wikipedia de cette façon, il est cependant nécessaire de *mapper* (faire la correspondance) des concepts Wikipedia à tout inventaire de sens pertinent pour l'application de WSD. Les algorithmes automatiques qui « *mappent* » (ou apparient) entre Wikipedia et WordNet, par exemple, s'intéressent à trouver le sens de WordNet qui a le plus grand chevauchement lexical avec le sens de Wikipedia. Ceci se fait en comparant le vecteur de mots dans le *synset* de WordNet, et les sens associés au vecteur de mots dans le titre de la page Wikipedia, ainsi que les liens sortants et la catégorie de la page [Ponzetto et Navigli, 2010].

2.4.3 Approches non supervisées

Les méthodes de WSD non supervisées ne dépendent pas de sources de connaissances externes, d'inventaires de sens, de dictionnaires électronique ou d'un ensemble de don-

nées annotées par des sens [Agirre et Edmonds, 2007]. Ces algorithmes n'attribuent généralement pas de sens aux mots, mais ils discriminaient la sémantique des mots en fonction de l'information, trouvée dans les corpus non annotés. Ainsi, ces approches utilisent principalement les textes des corpus dans la phase d'apprentissage des modèles de WSD.

Les corpus ont été tout d'abord utilisés dans les années soixante-dix par [Weiss, 1973] qui a proposé d'apprendre des règles de désambiguïsation à partir de corpus étiquetés. Ensuite, les méthodes d'apprentissage supervisé automatique ont été largement utilisées [Audibert, 2003]. Ces approches fondées sur des corpus, visant à éviter les limites de dictionnaires électroniques traditionnels, proposent de construire des dictionnaires modélisant des connaissances contextuelles.

En effet, [Véronis, 2001] affirme qu'il n'est pas possible de progresser dans la WSD tandis que les dictionnaires ne comprennent pas dans leurs définitions des critères de répartition ou des indices de surface (syntaxe, collocations, etc.). Travaillant dans sa même équipe, [Reymond, 2002] a proposé un dictionnaire de « répartition » pour assurer la WSD automatique. L'idée est d'organiser les mots en unités lexicales ayant des propriétés de répartition cohérentes. Ce dictionnaire contenait initialement la description détaillée de 20 noms communs, 20 verbes et 20 adjectifs. Cette ressource lui a permis d'étiqueter manuellement chacune des 53000 occurrences des 60 termes dans le corpus dans le cadre du projet *SyntSem* (le corpus contient environ 5,5 millions mots). Ce corpus constitue une ressource initiale pour étudier les critères de désambiguïsation sémantique automatique permettant de mettre en œuvre et évaluer les algorithmes de WSD. Audibert a travaillé sur ce dictionnaire pour étudier les différents critères de désambiguïsation (à savoir la co-occurrence, l'information du domaine et les synonymes de mots co-occurrents) [Audibert, 2003].

Dans les dernières années, la communauté de WSD a confirmé que la tâche de désambiguïsation doit être intégrée dans des applications réelles de traitement automatique de la langue comme la traduction automatique ou la RI multilingue. En effet, il serait difficile de parvenir à une amélioration concrète si la WSD reste considérée comme une tâche de recherche isolée.

Nous décrivons, dans les sous-sections qui suivent, les principales techniques utilisées dans les approches de WSD non supervisées :

2.4.3.1 Induction des sens par clustering

En partant de l'idée proposée par [Harris, 1970], le sens d'un mot peut être dérivé (ou *induit*) du contexte en découvrant les sens à partir du texte (on parle ici de *Word Sense Induction* (WSI)). L'hypothèse sous-jacente de cette idée est que les mots sont

sémantiquement proches s'ils apparaissent dans des documents similaires, dans des fenêtres contextuelles similaires ou dans des contextes syntaxiques similaires [Cruys, 2010]. L'algorithme de Lin [Lin, 1998] se base sur le *clustering* de mots. En effet, il se repose sur des statistiques de dépendance syntaxique entre des mots co-occurrents dans le corpus pour produire des ensembles pour chaque sens découvert d'un mot cible [Cruys et Apidianaki, 2011].

La méthode de regroupement (ou de *clustering*) contextuel est basée sur des vecteurs de contextes qui servent à identifier les sens des mots [Chen *et al.*, 2015]. Cette méthode utilise l'espace vectoriel comme espace de mots dans lequel chaque mot est représenté par un « vecteur de contexte ». Ensuite, une matrice de co-occurrence est créée, en regroupant l'ensemble des vecteurs de contextes, pour servir au calcul des mesures de similarité.

Comme un grand nombre de dimensions serait traité, l'analyse sémantique latente (LSA) peut être appliquée pour réduire l'espace multidimensionnel résultant par décomposition de valeur singulière (SVD ou *singular value decomposition*) [Golub et Van Loan, 1989]. L'application de SVD permet de trouver les principaux axes de variation dans l'espace des mots. La réduction de la dimensionnalité a pour effet de prendre l'ensemble des vecteurs de mots dans l'espace de grande dimension et de les représenter dans un espace de dimension inférieure. En conséquence, on s'attend à ce que les dimensions associées à des termes, ayant des significations similaires, soient fusionnées. Après la réduction, la similarité contextuelle entre deux mots peut être mesurée de nouveau en termes de mesures géométriques comme cosinus (voir autres mesures dans la section 2.3.2 page 35) entre les vecteurs correspondants [Turney et Pantel, 2010].

En utilisant une fonction de similarité, les algorithmes de *clustering* sont appliqués à un ensemble de vecteurs caractéristiques de mots tels que l'algorithme *K-means*, *Bisecting K-means* [Steinbach *et al.*, 2000], *Average-link* [Sokal et Michener, 1958] et *Unicon* [Lin et Pantel, 2001]. L'algorithme *Clustering by Committee* (CBC) de [Pantel et Lin, 2002] utilise également des contextes syntaxiques destinés à la tâche de l'induction des sens en exploitant une matrice de similarité pour quantifier les similarités entre les mots. Il s'appuie sur la notion de comités pour produire les différents sens du mot.

Cependant, ce type d'approche reste difficile à appliquer à grande échelle pour de nombreux domaines et langues.

2.4.3.2 Graphe de co-occurrence

Cette méthode se base sur la construction d'un graphe non orienté de co-occurrence ayant l'ensemble des sommets $\mathcal{S}$ et l'ensemble d'arêtes $\mathcal{A}$; où $\mathcal{S}$ représente les mots dans le texte et une arête $a \in \mathcal{A}$ est ajoutée si les mots co-occurrent dans le même

paragraphe ou texte. Le graphe est représenté par une matrice d'adjacence. Ensuite, la méthode de *clustering* de Markov est appliquée pour trouver le sens du mot [Dongen, 2000].

A chaque arête du graphe est associée un poids qui modélise la fréquence de co-occurrence de ces mots connectés. Le poids w_{ij} pour une arête (i, j) est donné par la formule :

$$w_{ij} = 1 - \max\{P(w_i|w_j), P(w_j|w_i)\} \tag{2.9}$$

avec $P(w_i|w_j) = \frac{f_{ij}}{f_j}$; et f_{ij} est la fréquence de co-occurrence des mots w_i et w_j ; et f_j est la fréquence d'apparition de w_j dans le texte.

En conséquence, les mots avec une fréquence élevée de co-occurrence sont pondérés d'un poids proche de zéro, tandis que les mots qui se produisent rarement ensemble ont des poids proches de 1. Les arêtes avec un poids supérieur à un certain seuil sont ignorés.

Ensuite, un algorithme itératif est appliqué sur le graphe et le nœud ayant le degré [7] relatif le plus élevé, est sélectionné comme racine ou *hub*. L'algorithme prend fin, lorsque la fréquence d'un mot pour son *hub* est inférieure à un seuil minimal. Enfin, tout le *hub* est désigné comme le sens du mot cible donné. Les *hubs* du mot cible qui ont un poids nul sont reliés et l'arbre couvrant minimal (ou *Minimum Spanning Tree : MST*) est créé à partir du graphe. Cet arbre est utilisé pour désambigüiser le sens réel du mot cible, racine de l'arbre (voir figure 2.2).

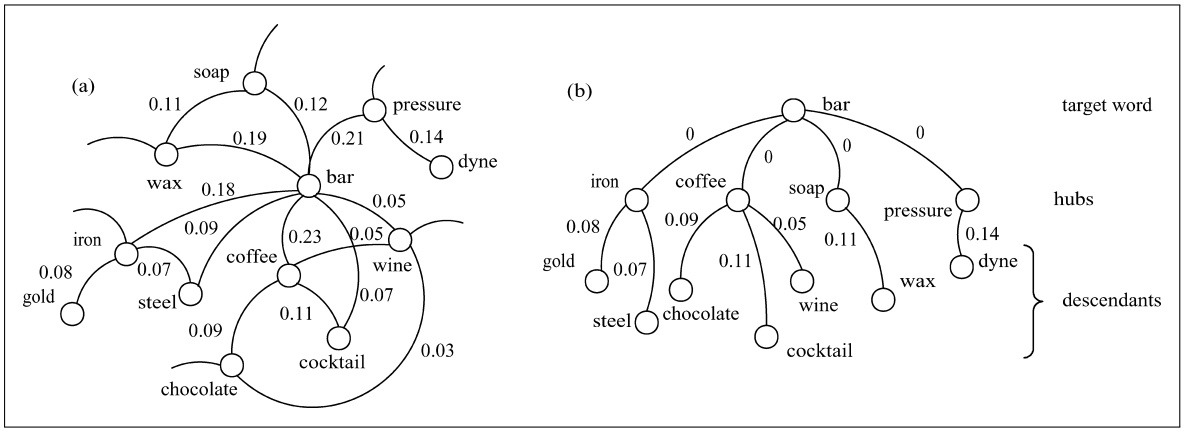


Figure 2.2 – (a) Exemple d'un graphe de co-occurrence. (b) L'arbre couvrant minimal (MST) pour le mot « *bar* ». [Navigli, 2009]

[Véronis, 2003, Véronis, 2004] propose un algorithme, nommé *HyperLex*, de recherche des zones à haute densité dans le graphe de co-occurrence et permet, contrairement aux

7. Le degré d'un nœud en théorie de graphe est le nombre d'arêtes sortantes du nœud. Il a été prouvé que le degré d'un nœud et sa fréquence dans le texte original sont fortement corrélés [Dongen, 2000].

méthodes traditionnelles d'analyse textuelle (comme les vecteurs de mots), d'ignorer les usages non fréquents.

Les méthodes de résolution de WSD basées sur les graphes de co-occurrence ont été largement utilisés dans d'autres travaux [Duque *et al.*, 2015, Koppula *et al.*, 2017].

2.4.4 Approches hybrides

Une approche hybride utilise en même temps des connaissances extraites de ressources lexicales (i.e. dictionnaires traditionnels) et des informations contextuelles ou distributionnelles apprises à partir de l'analyse de corpus. Ces approches ont atteint des taux de réussite très encourageants en termes de précision de désambiguïsation [Yarowsky, 2000].

Dans le système de [McRoy, 1992], l'auteur utilisait 13000 sens organisées de manière conceptuelle tout en mettant en place un système complexe de pondération à l'aide de plusieurs ressources. Comme autre exemple de l'approche combinant un thésaurus et un corpus, [Yuret et Yatbaz, 2010] ont proposé de combiner WordNet avec un grand volume de textes non annotés. Dans [Jimeno-Yepes *et al.*, 2011], les auteurs ont mis en place une approche d'apprentissage des classifieurs en se basant sur le corpus MEDLINE et le thésaurus MeSH.

[Stevenson et Wilks, 2001] ont présenté un système combinant plusieurs sources d'information : un filtrage basé sur l'étiquetage morphosyntaxique, les collocations extraites à partir du corpus, un chevauchement avec les définitions du dictionnaire LDOCE, des catégories de sujets et des restrictions de sélection. Les résultats du système proposé par Stevenson et Wilks avaient une précision de désambiguïsation de 95% au niveau des homographes de LDOCE et de 90% au niveau des sens.

[Mihalcea et Moldovan, 1998] ont présenté une approche basée sur l'idée de la densité sémantique. Ils utilisent la taxonomie et les définitions de WordNet en liaison avec les statistiques qui découlent directement de corpus récupérés sur Internet.

Dans les travaux de [Lafourcade et Brun, 2017], trois approches d'extraction des relations sémantiques ont été combinées. La première approche se base sur le réseau sémantique JeuxDeMots [Lafourcade, 2007] en tant que base de connaissance ; tandis que les deux autres approches se basent sur l'extraction des liens sémantiques à partir des textes. La combinaison de ses approches améliore la performance en terme de précision par rapport aux expérimentations faites sur chaque approche de façon indépendante.

Enfin, les ontologies sont de plus en plus utilisé récemment dans les SRI à base de WSD. Par exemple, [Barathi et Valli, 2010] ont combiné WordNet et les connaissances de domaine, modélisées comme ontologie, pour améliorer les résultats de RI.

2.4.5 Synthèse des approches de désambiguïsation sémantique

L'étude présentée sur les différentes approches de WSD nous conduit à conclure que l'adoption d'une approche en dépit d'autres dépend fortement des ressources disponibles lors de la phase d'apprentissage. Ainsi, afin de palier aux inconvénients de chaque approche, il est recommandé d'intégrer plusieurs ressources lors de la tâche de WSD.

Afin d'établir une synthèse générale des différentes approches discutées, nous présentons une comparaison de ces approches dans le tableau 2.4.

Approche de WSD	**Description de l'approche**
A base de connaissance	⊕ Ces algorithmes donnent de bonne précision. ⊖ Présence de divergences structurelles et de contenu entre les ressources utilisées qui ne sont pas souvent disponibles dans toutes les langues. ⊖ Dépendance forte de la couverture des ressources utilisées.
Supervisée	⊕ Elles donnent les meilleurs résultats dans les évaluations des systèmes de désambiguïsation sémantique. ⊖ Élaboration coûteuse des données d'entrainement et dispersion des données (il est difficile d'avoir un ensemble d'entrainement annoté qui couvre tout le lexique d'une langue).
Non supervisée	⊕ Pas besoin de corpus sémantiquement annotés ou de sources externes de connaissances (dictionnaires, thésaurus, etc.). ⊖ Les algorithmes de cette méthode sont difficiles à implémenter et leurs performances sont inférieures aux deux précédentes méthodes.
Hybride	⊕ Profite des avantages des approches supervisées et non supervisées. ⊖ Plus difficile à mettre en place avec un choix de pondération aléatoire (ou à la limite heuristique) pour les approches combinées lors de l'hybridation.

Table 2.4 – Comparaison des approches de désambiguïsation sémantique

Malgré la bonne performance des algorithmes supervisés de WSD, ils ont perdu progressivement du terrain pour les autres méthodes. En outre, ce type d'approche ne peut pas être facilement adapté à d'autres langues sans réapprentissage ; ce qui nécessite des données annotées dans cette langue. D'autre part, la réutilisation des modèles d'une langue à une autre conduit souvent à une mauvaise performance de classification [Khapra *et al.*, 2009].

D'autre part, l'apprentissage non supervisé est le plus grand défi pour les chercheurs de WSD [Panchenko *et al.*, 2017]. Les approches de WSD non supervisées sont composées de techniques d'induction de sens ou de discrimination visant à découvrir des

sens automatiquement basés sur des corpus non étiquetés. Par opposition à la WSD supervisée, ces approches utilisent des techniques d'apprentissage automatique sur des corpus non annotés, sans connaissance au préalable [Nasiruddin, 2013].

Nous étudions, dans la dernière section du chapitre, l'application des approches de WSD déjà étudiées dans un contexte translinguistique.

2.5 Désambiguïsation sémantique translinguistique

L'ambigüité sémantique constitue l'un des défis à résoudre lors de la traduction automatique. En effet, ceci se manifeste dans l'abondance des équivalences et les nuances liées à la complexité des structurations syntaxiques ; et par conséquent, il devient difficile de trouver des solutions automatiques efficaces lors de la sélection des traductions possibles. De nombreuses recherches sont entreprises pour résoudre ces problèmes d'ambiguïté et aboutir à une désambiguïsation qui distingue les significations souhaitées pour la traduction.

Parmi les expériences de désambiguïsation tentées depuis quelques années, les chercheurs ont eu recours à de nombreuses techniques dont (i) la sélection "statistique" des mots traduits de la requête, (ii) le calcul basé sur la moyenne relative de la fréquence des termes, (iii) l'utilisation de plusieurs critères afin de déterminer le sens d'un mot dans un contexte, y compris les valeurs syntaxique, sémantique et pragmatique, les relations de co-occurrences syntaxiques, (iv) le développement des requêtes utilisant la mise en graphes des termes et des documents, etc.

Nous prenons l'exemple du mot ambigu « *avocat* » ayant les deux sens traduits en anglais comme suit : « *lawyer* » (métier d'avocat), « *advocado* » (fruit). Dans l'exemple PH1 présenté ci-après, la proposition de traduction n'est pas adéquate à la différence de celle de l'exemple PH2 :

Phrase PH1 : *L'**avocat**, riche en lipides, apporte très majoritairement des acides gras insaturés bénéfiques pour la santé cardiovasculaire.*

Traduction PH1 : *The **~~lawyer~~** ✗, rich in fat, provides mostly beneficial unsaturated fatty acids for cardiovascular health.*

Phrase PH2 : *J'ai embauché un **avocat** qui a lutté pour mes prestations.*

Tradution PH2 : *I hired a **lawyer** ✓ who struggled for my benefits.*

En effet, le sens d'un mot dans son contexte est facilement identifié grâce aux connaissances extralinguistiques des interlocuteurs dans une communication orale (inférences, connaissances du monde, etc.). Dans la phase de traduction humaine, les ambigüités sont résolues de manière implicite grâce aux riches connaissances extralinguistiques du

traducteur. En contre partie, afin de rendre la traduction automatisée, il y aura besoin de modèles sophistiqués capables de considérer tous les sens candidats d'un mot et d'en sélectionner le plus approprié au contexte.

La première tâche de WSD translinguistique (appelée en anglais *Cross-Lingual WSD*) a été organisée à la compagne SemEval-2007 puis réintroduite dans SemEval-2010, comportant 16 contributions de 5 différentes équipes de recherche [Lefever et Hoste, 2010]. La troisième édition a été proposée à SemEval-2013 pour lequel de nouvelles données de test sont annotées en vue de souligner la faisabilité et la difficulté de WSD translinguistique [Lefever et Hoste, 2013]. Dans cette édition, les résultats de 12 soumissions officielles ont été rapportés pour 5 équipes de recherche différentes, en plus du système *Parasense* qui a été développé par l'équipe d'organisation. Cinq traductions de l'anglais vers les différentes langues cibles (français, italien, espagnol, néerlandais et allemand) ont été utilisés pour générer les étiquettes de sens.

Le corpus parallèle Europarl a été utilisé pour construire l'inventaire de sens. Pour un mot polysémique donné, différents sens sont regroupés en groupe (ou *cluster*) de traductions possibles. Par la suite, les évaluateurs (ou juges) choisissent le *cluster* le plus pertinent pour chaque phrase de test et proposent les trois premières traductions de la liste prédéfinie des traductions dans Europarl.

Il y avait deux types d'évaluations dans les systèmes participants : (i) une évaluation multilingue dans laquelle les traductions dans les cinq langues cibles sont évaluées ; et (ii) une évaluation bilingue où les traductions dans une langue cible sont évaluées. Les évaluations réalisées sont : l'évaluation « *meilleur résultat* », dans laquelle seule la première traduction donnée par un système a été considérée, et l'évaluation des résultats « *top cinq* » dans laquelle les cinq premières traductions fournies par un système ont été considérées. Les résultats ont été effectuées à l'aide des métriques de rappel et de précision .

Plus récemment, [Vulic et Moens, 2014] proposent une approche probabiliste de modélisation de la similarité sémantique inter-linguistique dans un contexte qui ne nécessite que des données comparables. L'approche s'appuie sur une idée de projeter des mots et des ensembles de mots dans un espace sémantique partagé par des concepts sémantiques latents indépendants. Ces concepts multilingues latents sont induits ainsi à partir d'un corpus comparable sans recours à des ressources lexicales supplémentaires. Le sens des mots est représenté comme une distribution de probabilité qui modélise les représentations de mots isolées avec des connaissances contextuelles. Les résultats sur la suggestion des traductions de mots dans 3 paires de langues [8] révèlent l'utilité des modèles contextuels proposés de similarité sémantique translinguistique.

8. paires de langues ES→EN , IT→EN et NL→EN
(ES : espagnol ; IT : italien ; NL : néerlandais ; EN : anglais)

Les corpus parallèles sont considérés comme une source commune de connaissances pour effectuer la tâche de désambiguïsation dans un contexte multilingue. Partageant une signification cachée qui peut être utile pour extraire des connaissances sur une langue, ces corpus sont de bonnes ressources pour assurer la désambiguïsation translinguistique [Resnik, 2004].

2.5.1 Approches de désambiguïsation translinguistique à base des graphes

[Duque *et al.*, 2015] a conclu que les systèmes basés sur l'utilisation des algorithmes de graphes sont les plus performants dans les systèmes qui ont participés aux compétitions SemEval 2010 et 2013. Certains de ces algorithmes ont été largement utilisés dans la littérature [Mihalcea, 2005, Navigli et Lapata, 2010, Agirre *et al.*, 2014]. Ainsi, nous nous focalisons, dans cette section, sur les approches utilisant les structures en graphes et exploitant les relations de co-occurrence extraites de l'analyse de corpus.

[Véronis, 2004] propose l'algorithme HyperLex qui est une approche basée sur un corpus en construisant un graphe de co-occurrence pour toutes les paires de mots qui co-occurrent dans le contexte du mot cible. Ce type de graphe a les propriétés des *réseaux de petits mondes hiérarchiques* (RPMH). Par conséquent, le graphe possède des composants fortement connectés (ou *hubs*) qui identifient les utilisations principales (ou sens) du mot cible, et peuvent ainsi être utilisés pour exécuter une tâche de WSD.

Les auteurs présentent dans [Agirre *et al.*, 2006] une étude comparative entre l'algorithme HyperLex de Véronis et un algorithme adapté de PageRank [Brin et Page, 1998] pour la désambiguïsation sémantique. Ainsi, ils ont exploré l'utilisation de deux algorithmes de graphes pour la désambiguïsation des sens nominaux. La performance de PageRank était presque la même que celle d'HyperLex, avec l'avantage de PageRank d'utiliser moins de paramètres d'optimisation.

Silberer et Ponzetto dans [Silberer et Ponzetto, 2010] se sont inspirés des travaux de [Véronis, 2004] et [Agirre *et al.*, 2006]. En fait, ils ont présenté un système basé sur un graphe de co-occurrence. Ce dernier est construit à partir de corpus parallèles multilingues avec l'application des algorithmes de graphes développés précédemment pour la WSD monolingue. Par la suite, l'algorithme d'arbre couvrant minimal (en anglais *Minimum Spanning Tree*) est appliqué sur le graphe pour effectuer la tâche de WSD.

[Duque *et al.*, 2015] présentent une approche qui comprend la génération automatique de dictionnaires bilingues et la construction d'un graphe de co-occurrence utilisé pour sélectionner les traductions les plus appropriées du dictionnaire. Ils ont mis en œuvre

des algorithmes qui combinent à la fois un dictionnaire et un graphe de co-occurrence pour effectuer la sélection des traductions finales. Ces algorithmes sont basés sur (i) des sous-graphes (ou *communautés*) contenant des groupes de mots avec des significations connexes, (ii) des distances entre les nœuds représentant les mots, et (iii) l'importance relative de chaque nœud dans le graphe. En utilisant les standards de test SemEval-2010 et SemEval-2013 pour évaluer leur système, ils ont prouvé la validité de l'approche de graphe non supervisée. Dans cette approche, le document entier est considéré comme une information cohérente, tandis que d'autres travaux considèrent des fenêtres de taille spécifique pour construire le contexte et le calcul des co-occurrences.

2.5.2 Combinaison des ressources lexicales et statistiques pour la RI translinguistique

Comme il existe une diversité de techniques de traduction de requêtes, l'idée de combiner ces techniques a été étudiée dans des travaux récents afin d'examiner si une approche est complémentaire à une autre [Nie, 2010, Azarbonyad *et al.*, 2013, Schamoni *et al.*, 2014].

Par exemple, [Herbert *et al.*, 2011] ont introduit un modèle translinguistique utilisant Wikipedia afin d'apparier les concepts dans une langue à leurs équivalents dans une autre langue. Cet appariement (ou *mapping*) est assuré grâce aux liens de redirection et inter-langues dans les versions multilingues de Wikipedia. Dans ce travail, les auteurs ont montré que les traductions de Wikipedia peuvent améliorer les performances des SRI translinguistiques basés sur la traduction automatique statistique. En fait, les requêtes sont traduites avec le service en ligne *Google Translate* et étendues avec de nouvelles traductions. Ces traductions sont obtenues en mappant des syntagmes nominaux dans la requête vers des concepts dans la langue cible en utilisant Wikipedia.

[Türe et Boschee, 2014] ont introduit une nouvelle méthode pour construire une formule unique pour chaque requête. Ils ont conçu cette idée comme un ensemble de problèmes de classification binaires. Les résultats montrent que l'apprentissage des classifieurs peut être utilisé pour produire des combinaisons des poids de manière efficace.

[Kim *et al.*, 2015] ont exploré la combinaison des ressources de traduction lexicales et statistiques en utilisant à la fois Wikipedia et un dictionnaire électronique comme connaissance de traduction lexicale. De plus, ils ont exploré des corpus parallèles pour extraire statistiquement les candidats de traduction. [Kim *et al.*, 2015] ont prouvé que l'utilisation conjointe des trois ressources de traduction (c-à-d un dictionnaire, un corpus parallèle et des connaissances de Wikipedia) donne de meilleurs résultats en comparaison avec l'utilisation d'une seule ressource. Ils ont proposé également une approche d'expansion des requêtes en post-traduction en utilisant une marche aléatoire

sur le graphe des liens des concepts dans Wikipedia. Cette approche apporte d'autres améliorations par rapport aux techniques alternatives en l'évaluant sur la collection de test anglais-coréen NTCIR-5.

Conclusion

La désambiguïsation sémantique est considérée comme tâche fondamentale, principalement pour les domaines de recherche d'information (RI) et de traitement automatique des langues (TAL). Le développement et l'amélioration de ces techniques et ces approches présentées tout au long de ce chapitre ouvrent de nombreuses perspectives prometteuses pour la RI dont l'objectif est de répondre au besoin de l'utilisateur.

En partant d'un contexte réduit à quelques mots-clés, les requêtes formulées par l'utilisateur du SRI peuvent avoir un taux d'ambiguïté important. Afin d'enrichir ce contexte, les techniques d'expansion de requêtes accompagnent l'utilisateur dans le but de mieux exprimer son besoin en information. Nous présentons ainsi ces différentes techniques et approches d'expansion des requêtes dans le prochain chapitre.

Chapitre

3 Expansion de Requêtes dans la RI Monolingue et Translinguistique

Sommaire

" To know what you know and what you do not know, that is true knowledge. "

— Confucius

Introduction

Une recherche complète se présente comme un processus itératif mettant en œuvre plusieurs requêtes qui se succèdent et qui permettent d'affiner progressivement les réponses du système [Manning *et al.*, 2008]. C'est dans cette perspective que les techniques de reformulation de requêtes ont été utilisées. En effet, un concept peut être défini par des termes différents qui ne sont pas nécessairement tous utilisés dans la formulation originelle de la requête. Par ailleurs, le besoin peut s'exprimer de manière peu claire au départ pour l'utilisateur, mais pourrait s'éclaircir et se préciser au fur et à mesure de la recherche des résultats.

D'autre part, les requêtes des utilisateurs sont généralement trop courtes pour décrire précisément les besoins en information. Par conséquent, des termes importants peuvent manquer dans la requête ; ce qui conduit à une mauvaise couverture des documents pertinents, d'où un taux faible de rappel. Pour résoudre ce problème, il existe des techniques d'expansion des requêtes, utilisant une variété d'approches, exploitant plusieurs sources de données et employant différentes méthodes pour choisir les termes les plus appropriés pour l'expansion [Carpineto et Romano, 2012].

Nous présentons dans un premier lieu la corrélation existante entre les tâches d'expansion de requête et la désambiguïsation sémantique dans la section 3.1. Les sections 3.2 et 3.3 traitent les principales approches d'expansion de requêtes. Dans la dernière section du chapitre nous présentons une vue sur l'application de l'expansion de requêtes dans la RI translinguistique.

3.1 L'expansion des requêtes en résolution de l'ambiguïté de contexte

Les travaux cherchant à désambiguïser sémantiquement les requêtes se basent essentiellement sur l'analyse du contexte, autrement les co-occurrences qui existent dans la requête. Cependant, l'expansion de la requête peut présenter une solution au problème de désambigüisation. En effet, étendre une requête permet de reconstituer un contexte linguistique absent face à une longueur courte de requête [Joho *et al.*, 2004].

Le problème de requête ambigüe peut être résolu en la décomposant en plusieurs « sous-requêtes » [Santos *et al.*, 2010]. Ces derniers sont des reformulations provenant par exemple de moteurs de recherche (voir l'exemple de proposition des *recherches associés* dans le moteur de recherche Google dans la figure 3.1).

Cependant, l'extraction du contexte de RI reste une tâche difficile. En effet, les requêtes

Figure 3.1 – Proposition de requêtes alternatives (ou recherches associées) pour enrichir le contexte de la requête utilisateur « *jaguar* » dans le moteur de recherche Google (résultats de avril 2017)

ne sont pas les seules expressions des besoins d'informations formulées par l'utilisateur d'un SRI, même si elles restent les principales traces exploitables du contexte. A titre d'exemples, les historiques de recherche contiennent de nombreuses informations concernant le contexte de la RI : informations sur la tâche de recherche, sur l'utilisateur ou encore sur le contexte de l'information [Allan *et al.*, 2003].

Une autre approche d'enrichissement de contexte consiste dans la *diversification des résultats de recherche* (DRR). Il s'agit de sélectionner divers documents à partir des résultats de recherche afin de couvrir autant d'intentions informationnels que possible [Santos *et al.*, 2015, Jiang *et al.*, 2017]. En effet, dans les approches existantes, les résultats initiaux sont supposés être suffisamment diversifiés et couvrent bien les aspects de la requête. Cependant, les résultats initiaux n'arrivent pas souvent à couvrir certains aspects. Les travaux de [Bouchoucha *et al.*, 2014, Bouchoucha *et al.*, 2015, Liu *et al.*, 2014] proposent une nouvelle approche de DRR qui consiste à *diversifier l'expansion de requête* (DER) afin d'avoir une meilleure couverture des aspects. Les termes d'expansion sont sélectionnés à partir d'une ou de plusieurs ressources suivant le principe de pertinence marginale maximale [Carbonell et Goldstein, 1998].

Selon [Adam *et al.*, 2013], les enrichissements et les modifications du contexte exprimé dans une requête peuvent être classées en 4 catégories :

- la *généralisation* : consiste à supprimer un ou plusieurs mots. Son objectif est d'élargir le champ de recherche visant ainsi à réduire le silence ;
- la *spécification* : c'est l'action inverse à la généralisation. Cette action vise à rétrécir le champ de recherche pour réduire le bruit ;
- la *reformulation* : il s'agit de paraphrase. Elle consiste à remplacer des mots de la requête par leurs synonymes ;
- le *mouvement parallèle* : ce mouvement engendre une modification importante de la requête créant une alternative, par le remplacement par exemple d'un produit ou d'une marque.

Le tableau 3.1 présente des exemples de chacune des 4 catégories de reformulation des requêtes.

Type	**Action**	**Exemple**
Généralisation	Suppression de mots	inégalités sociales de santé → inégalités ~~sociales~~ de santé
	Remplacement par un hyperonyme	sociologie arts martiaux → sociologie du sport
Spécification	Ajout de mots	faillite → théories de la faillite
	Remplacement par un hyponyme	activités dans la montagne → escalades dans la montagne
Reformulation	Remplacement par un synonyme	territoire voiture → territoire automobile
Mouvement parallèle	Remplacement par un co-hyponyme	siège romain → siège gaulois

Table 3.1 – Quatre types d'actions effectuées lors de la reformulation de requêtes [Adam *et al.*, 2013]

Nous présentons dans la figure 3.2 un scénario de recherche du mot « *jaguar* » sur le moteur de recherche Bing [1]. La requête est désambigüisée dans une première phase en proposant deux sens possibles du mot « *jaguar* » ainsi qu'un ensemble de suggestion d'auto-complétion (phase 1).

Si l'utilisateur clique sur le premier sens, la requête sera étendue en « *Jaguar Cars* » (scénario 2). Dans le cas où l'utilisateur sélectionne l'autre sens, la requête sera totalement substituée en « *Panthera onca* », qui représente le nom scientifique de l'animal (scénario 3). D'autres recherches similaires sont affichées (scénario 4) : Il s'agit d'une sorte de mouvement parallèle de la requête d'origine en proposant un ensemble de co-hyponymes (dans ce cas des félins tels que « *léopard* », « *puma* », « *tigre* », « *lion* », etc.).

Les approches pour s'attaquer au problème de reformulation de requête se divisent principalement en deux classes : les méthodes dites locales et les méthodes globales.

1. https ://www.bing.com

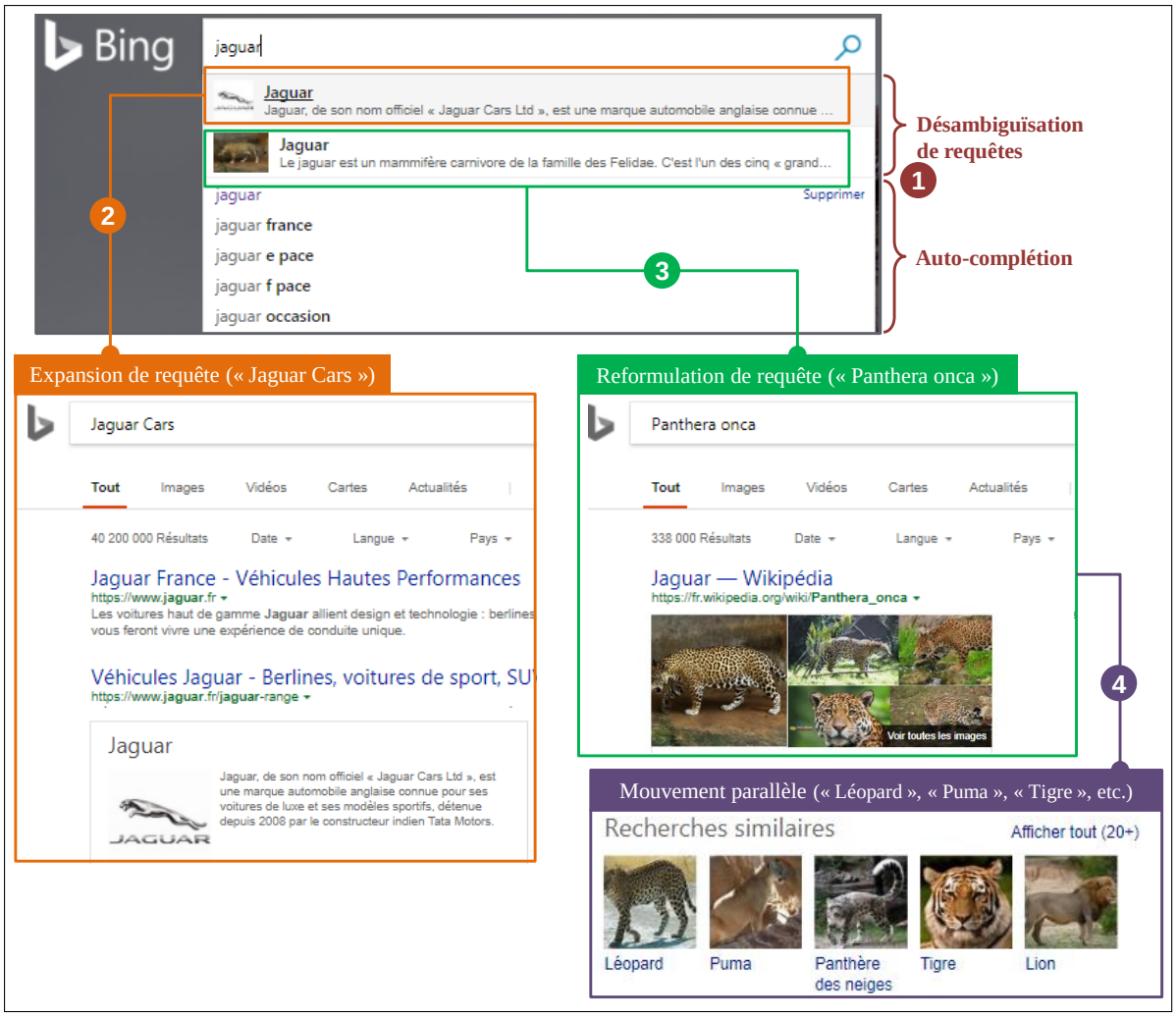


Figure 3.2 – Différents scénarios de recherche et de reformulation du mot « *jaguar* » sur le moteur de recherche Bing (résultats de mai 2017)

Les méthodes locales ajustent une requête relative aux documents qui apparaissent initialement pour correspondre à la requête. Nous distinguons principalement les méthodes locales de rétroaction de pertinence et de pseudo-rétroaction de pertinence qui seront détaillées dans la section 3.2.

Les méthodes globales sont des techniques permettant d'étendre ou de reformuler des termes de requête indépendamment de la requête et des résultats renvoyés. Les modifications apportées à la requête entrainent la correspondance de la nouvelle requête avec d'autres termes sémantiquement similaires. Ces méthodes seront traitées dans la section 3.3.

3.2 La rétroaction de pertinence

Ce processus, nommé aussi par *réinjection de pertinence* ou *retour de pertinence*, a été introduit par [Rocchio, 1971]. Son principe consiste à prendre en considération l'évaluation de l'utilisateur (dite *pertinence utilisateur*) des documents jugés pertinents par le système. Son évaluation consiste à indiquer parmi les documents proposés ceux qui sont jugés pertinents et ceux qui ne le sont pas. La reformulation de la requête se traduit donc par une repondération des termes ou par l'ajout ou le retrait de termes contenus dans les documents pertinents/non pertinents à la requête.

3.2.1 Algorithme de Rocchio

Rocchio décrit une stratégie permettant de dériver itérativement le vecteur requête optimal à partir d'opérations sur les vecteurs documents pertinents et les vecteurs documents non pertinents [Rocchio, 1971]. En partant d'une requête initiale, l'utilisateur fournit des informations de jugement de pertinence au système. Ces retours (ou *feedbacks*) permettent de reformuler automatiquement le profil de la requête. Ceci permet aux documents, au fur et à mesure des itérations, de se rapprocher de plus en plus des intentions de l'utilisateur (voir figure 3.3).

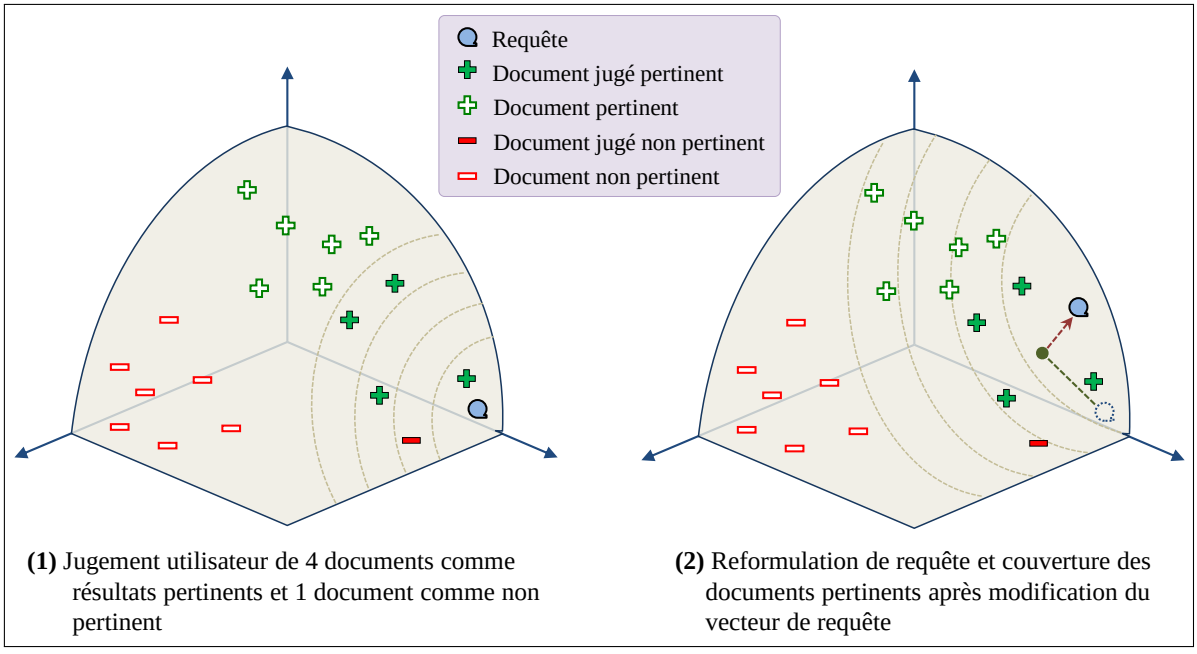


Figure 3.3 – Modélisation de l'algorithme de Rocchio en prenant en compte le jugement de pertinence de l'utilisateur [Lavrenko, 2008]

La formule posée par Rocchio est la suivante (équation 3.1) :

$$Q^{new} = \alpha Q^{old} + \beta \frac{1}{|reldocs|} \sum_{reldocs} w_{t_i} - \gamma \frac{1}{|nonreldocs|} \sum_{nonreldocs} w_{t_i} \qquad (3.1)$$

avec :

- α permet de moduler l'importance de la requête précédente Q^{old} ;
- β permet de moduler le vecteur profil moyen des documents choisis ;
- γ permet de moduler le vecteur profil des documents rejetés ;
- α, β, γ représentent des paramètres positifs. Leurs valeurs sont à fixer dans l'intervalle [0,1] ;
- w_{t_i} identifie un terme des documents ;
- $|reldocs|$ représente le nombre des documents pertinents ;
- $|nonreldocs|$ représente le nombre des documents non pertinents.

Le modèle introduit ainsi des poids négatifs pour les termes que l'utilisateur ne désire pas retrouver dans les documents recherchés ce qui modélise le *feedback négatif*.

[Ide, 1971] propose une formule dérivée de celle de Rocchio en éliminant les facteurs de normalisation exprimés respectivement par les nombres de documents pertinents et non pertinents et en limitant le nombre de documents non pertinents comme suit (équation 3.2) :

$$Q^{new} = \alpha.Q^{old} + \beta. \sum_{reldocs} w_{t_i} - \gamma T_{nonreldocs} \tag{3.2}$$

Avec : $T_{nonreldocs}$ correspond au vecteur des documents qui sont classés les moins pertinents.

[Miao *et al.*, 2012] incorporent des informations de proximité dans le modèle de Rocchio de base en proposant un modèle de proximité, appelé PRoc, avec trois variantes. Dans ce modèle, le concept de proximité des fréquences de termes est introduit pour modéliser les informations de proximité dans les documents pseudo-pertinents, qui sont ensuite utilisés dans trois types de mesures de proximité. Les résultats expérimentaux sur les collections TREC montrent que les modèles PRoc proposés sont efficaces et généralement supérieurs aux modèles de rétroaction de pertinence de l'état de l'art avec des paramètres optimaux.

Dans les travaux de [Ksentini *et al.*, 2016], les auteurs revisitent les paramètres du modèle de Rocchio afin de prendre en compte les relations entre les termes. Ces dernières sont définies par une méthode statistique basée sur l'optimisation des moindres carrés. Le processus d'évaluation a été réalisé sur la base de données CLEF-eHealth-2014.

Il a été prouvé par différents travaux de recherche que la rétroaction de pertinence améliorait les résultats de la recherche selon les taux de rappel et de précision [Wang *et al.*, 2008, Yan *et al.*, 2003]. Cependant, la mention des documents non pertinents, désignée aussi par le *feedback négatif*, ne donne pas de résultats aussi satisfaisants que le *feedback positif* [Manning *et al.*, 2008].

3.2.2 Rétroaction de pertinence probabiliste

Sur la base du modèle probabiliste, [Harman, 1992] a développé des formules de pondération de requête en utilisant le jugement de l'utilisateur sur la pertinence des documents restitués par le système. L'équation 3.3, dérivée des travaux de [Robertson et Jones, 1976], détermine une pondération des termes évaluant la distribution des termes de la requête dans les documents jugés pertinents et les documents jugés non pertinents par l'utilisateur.

$$w_i = log\frac{p_i(1-q_i)}{q_i(1-p_i)} = log\frac{(r_i+0,5)/(n-r_i+0,5)}{(R-r_i+0,5)/(N-R-n_i+r_i+0,5)} \tag{3.3}$$

Avec :

$$p_i = \frac{r_i+0,5}{R+1} \quad et \quad q_i = \frac{n_i-r_i+0,5}{N-R+1} \tag{3.4}$$

- r_i correspond au nombre de documents pertinents qui sont indexés par le terme t_i ;
- n_i correspond au nombre de documents qui sont indexés par le terme t_i ;
- R correspond au nombre de documents pertinents ;
- N correspond au nombre de tous les documents dans la collection ;
- 0,5 est un facteur d'ajustement.

L'utilisation du coefficient 0,5 comme facteur d'ajustement permet d'augmenter la précision de l'ordre de 25% sur la base CRANFIELD selon [Harman, 1992].

Haines et Croft ont défini dans [Haines et Croft, 1993] une méthode de repondération en utilisant une version révisée de la formule de pondération de [Robertson et Jones, 1976]. La recherche initiale suit la fonction de pondération des termes suivante (équations 3.5, 3.6) :

Recherche initiale :

$$w_{ijk} = (C + idf_i) \times f_{ik} \tag{3.5}$$

Avec :

- C : constante ;
- idf_i : fréquence absolue du terme t_i dans la collection ;
- f_{ik} : la fréquence du terme t_i dans le document k.

Pour repondérer des termes par réinjection de pertinence, Haines et Croft se basent sur la formule de [Robertson et Jones, 1976] (équation 3.3) comme suit :

Rétroaction (*feedback*) :

$$w_{ijk} = [C + log\frac{p_{ij}(1 - q_{ij})}{q_{ij}(1 - p_{ij})}] \times f_{ik} \tag{3.6}$$

Avec :

- w_{ijk} : le poids du terme t_i dans la requête j et le document k ;
- p_{ij} : probabilité que le terme t_i soit assigné à un ensemble de documents pertinents pour une requête j.
  $p_{ij} = (r_i + 0,5)/(r_i + 1)$ si $r_i > 0$, $p_{ij} = 0{,}01$ si $r_i = 0$;
- q_{ij} : probabilité que le terme t_i apparaisse dans un ensemble de documents non pertinents pour une requête j.
  $q_{ij} = (n_i - r_i + 0,5)/(N - R + 1)$ si $r_i > 0$, $q_{ij} = 0{,}01$ si $r_i = 0$;
- $f_{ik} = K + (1 - K) \times freq_{ik}/max(freq_k)$; où :
  $freq_{ik}$: la fréquence du terme t_i dans le document k ;
  $freq_k$: la fréquence maximale d'un terme dans le document k ;
  K : constantes.

3.2.3 Pseudo-rétroaction de pertinence

Une approche alternative, connue sous le nom de *pseudo-rétroaction de pertinence* (en anglais *pseudo relevance feedback* ou *blind relevance feedback*), utilise des techniques de réinjection automatique de termes à l'aveugle pour construire une nouvelle requête. Plus précisément, le système de recherche restitue un ensemble de documents répondant à la requête initiale. Ainsi au lieu de juger explicitement les documents, les k premiers documents sont supposés comme étant pertinents (ou documents *pseudo-pertinents*) (voir figure 3.4). Par analogie, les documents qui sont restitués en fin des résultats retournés peuvent être considérés comme non pertinents [Buckley *et al.*, 1995].

L'idée de base derrière la pseudo-rétroaction de pertinence est qu'une itération de réinjection basée sur les documents les plus similaires à la requête initiale de l'utilisateur pourrait donner une meilleure restitution des documents. La pseudo-réinjection de pertinence peut être bénéfique si les requêtes initiales permettent de retrouver des documents pertinents en tête de liste. Dans le cas contraire, elle provoque une dégradation des performances en introduisant plus de bruit dans les résultats restitués. Pour éviter cette dégradation, les termes d'expansion candidats pourraient être affichés à l'utilisateur, mais des études ont montré que cela n'est pas particulièrement efficace. Suggérer des requêtes alternatives basées sur une analyse des journaux de requêtes (fichiers *logs*) est une alternative plus fiable pour l'expansion de requête semi-automatique [Croft *et al.*, 2009].

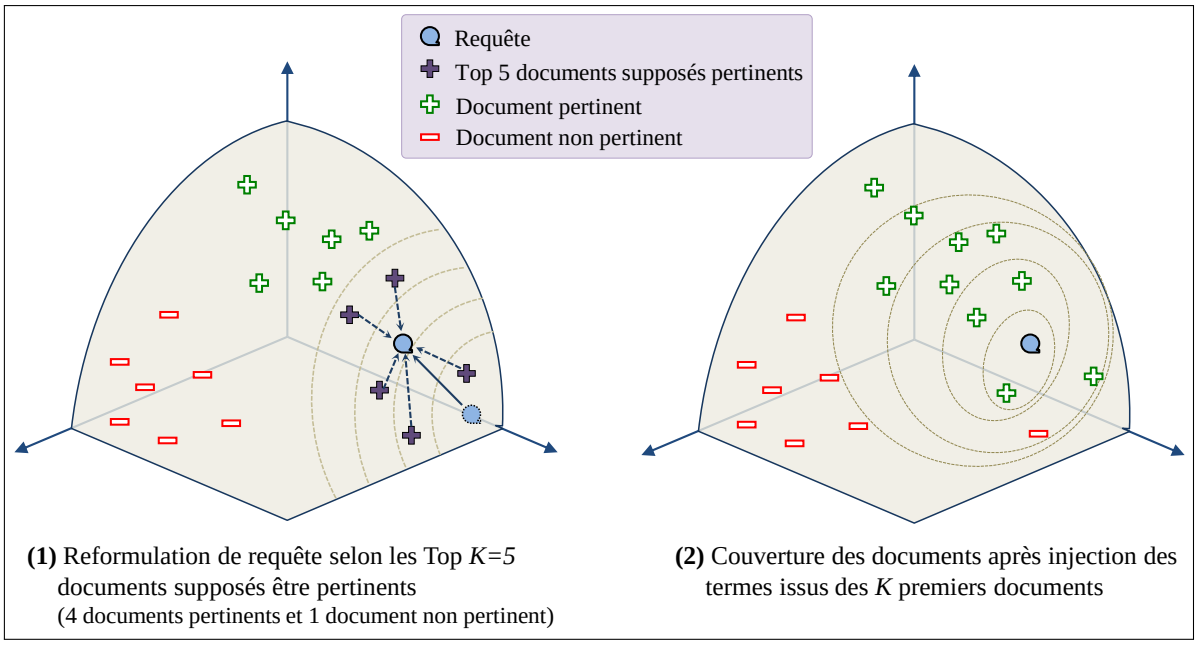


Figure 3.4 – Exemple de modélisation de la pseudo-rétroaction de pertinence avec K=5 top documents [Lavrenko, 2008]

Dans [Yoo et Choi, 2010], Yoo et Choi critiquent l'apport de la pseudo-réinjection de pertinence dans l'amélioration de la RI en menant une étude sur les documents de la base MEDLINE [2]. Cette base bibliographique présente un grand défis causé par la faible précision lors de la recherche de données. Le problème d'injection de bruit s'aggrave en appliquant la technique de pseudo-réinjection de pertinence qui extrait des termes des premiers résultats retournés. Dans leurs expérimentations, Yoo et Choi ont établi une étude comparative de 6 méthodes de pondération (Rocchio, LCA, CHI2, EMIM, KLD et RSV) [3] pour la sélection des termes avec une faveur donnée à la méthode LCA de [Xu et Croft, 2000]. Les résultats ont prouvé également la forte dépendance de la performance de RI avec le nombre de termes à choisir ainsi que le nombre de documents sur lesquels agit la pseudo-réinjection de pertinence.

Dans [Williams et Giles, 2016], les auteurs proposent une stratégie récursive de pseudo-rétroaction de pertinence. Cette approche récursive permet de générer une arborescence utilisée pour le classement des résultats en appliquant un modèle mathématique approprié. Les expériences sur les ensembles de données Reuters-21578 et WebKB montrent une amélioration des résultats de la stratégie récursive en terme de mesure MAP.

[Keikha *et al.*, 2017] proposent deux méthodes d'expansion de requêtes supervisées et non supervisées qui sont inspirées de l'approche de pseudo-rétroaction de pertinence. En

2. MEDLINE est une base de données bibliographiques regroupant la littérature relative aux sciences biologiques et biomédicales. Elle regroupe plus que 26 millions d'enregistrements. Lien : www.nlm.nih.gov/pubs/factsheets/medline.html

3. K*ullback–Leibler Divergence* (KLD), *Robertson Selection Value* (RSV), *CHI-squared* (CHI2), *Expected Mutual Information Measure* (EMIM) et *Local Context Analysis* (LCA).

considérant les articles extraits de Wikipedia comme des documents de rétroaction, les termes dans ces articles sont sélectionnés pour l'expansion. Dans la méthode de pseudo-rétroaction de pertinence, il est possible que les premiers résultats, considérés comme pertinents pour la requête, contiennent des documents non pertinents ; pouvant ainsi avoir un impact négatif sur les résultats de l'expansion. Dans l'approche proposée par [Keikha *et al.*, 2017], les auteurs extraient les articles de Wikipedia qui sont susceptibles d'être liés à la requête ; et donc de diminuer la probabilité que des documents non pertinents soient inclus dans la collection de documents de rétroaction. Ils utilisent également les articles de redirection et de désambiguïsation de Wikipedia pour aider à surmonter le problème de non-concordance de vocabulaire.

[Mbarek *et al.*, 2017] proposent une approche pour l'expansion de requêtes en utilisant un document dit *absorbant* qui est le produit croisé de documents non pertinents et qui sera orthogonal aux documents non pertinents. Les auteurs ont montré que ce document absorbant pourrait extraire de meilleurs termes d'expansion à partir des top k documents les mieux classés. Les expériences montrent des améliorations pour les deux collections, TREC-7 et TREC-8, sur le modèle d'appariement BM25.

3.3 Méthodes globales d'expansion de requêtes

Dans l'état de l'art établi par [Carpineto et Romano, 2012], les auteurs proposent une taxonomie des principales techniques d'expansion selon la figure 3.5. Dans les sous-sections qui suivent, nous présentons les méthodes, dites globales, qui sont les plus utilisées dans la littérature ; à savoir les méthodes basées sur des ressources linguistiques ainsi que les méthodes basées sur l'analyse de corpus.

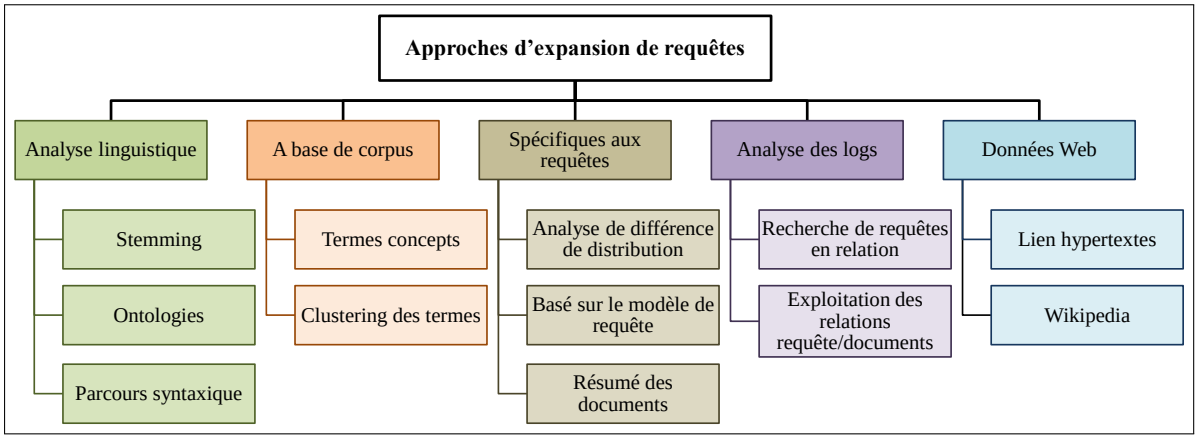


Figure 3.5 – Taxonomie des approches d'expansion automatique de requêtes [Carpineto et Romano, 2012]

3.3.1 Approches basées sur l'exploitation de ressources linguistiques

Afin de trouver des termes d'expansion sémantiquement proches des termes de la requête initiale, l'utilisation de ressources linguistiques externes de type dictionnaire ou thésaurus, s'avère intéressante [Picton *et al.*, 2008].

Plusieurs travaux ont été menés en utilisant la base lexicale WordNet pour réaliser l'expansion de requêtes [Voorhees, 1994, Fang, 2008, Agirre *et al.*, 2010]. Le principe général est le suivant : aux mots d'origine de la requête sont ajoutés ceux qui figurent dans les mêmes classes synonymiques qu'eux (*synsets*), éventuellement dans les classes rattachées par une relation d'hyperonymie [4]. Cependant, il est en particulier difficile de contenir les risques de propagation de l'ambiguïté des termes polysémiques [5], qui a pour effet de dégrader la précision de la recherche.

L'exploitation de WordNet dans l'expansion des requêtes nécessitent la réponse aux questions suivantes :

- Si un mot de requête apparait dans plusieurs *synsets*, lequel choisir ?
- Une fois un *synset* est choisi, quels sont les mots à ajouter à la requête ? est-ce que les synonymes contenus dans ce *synset* seront les seuls à être ajoutés ? ou il faut considérer les autres types de relations sémantiques telle que l'hyperonymie ?

[Zhang *et al.*, 2009] ont utilisé WordNet pour désambiguïser les termes de requête, puis ont ajouté des synonymes de mots de requête pour l'étendre. Les expérimentations sur la collection CACM ont amélioré la précision P@10 d'environ 7% par rapport aux requêtes originales non étendues. La même approche d'expansion a été utilisé par [Liu *et al.*, 2004] en rajoutant les hyponymes ainsi que les synonymes depuis WordNet. Ils ont testé leur méthode sur TREC9, TREC10 et TREC12 et obtenu de meilleurs résultats.

[Fang, 2008] rapporte des résultats positifs pour l'expansion de requêtes basée sur WordNet dans un cadre de recherche *axiomatique*. Dans la méthode décrite par Fang, l'ensemble des termes d'expansion candidats se compose de tous les mots de tous les synsets dans lesquels les termes de requête occurrent. Les termes candidats d'expansion sont sélectionnés sur la base du chevauchement de vocabulaire entre ses glossaires et les *glosses* WordNet des termes de requête.

4. L'*hyperonymie* est la relation sémantique hiérarchique d'un terme à un autre selon laquelle l'extension du premier terme, plus général, englobe l'extension du second, plus spécifique. Le premier terme est dit *hyperonyme* de l'autre, ou *superordonné* par rapport à l'autre. Dans l'exemple : « *Parmi les félins nous pouvons distinguer les tigres et les lions* » ; « *félin* » est un hyperonyme de « *tigre* » et de « *lion* ».

5. La *polysémie* est la qualité d'un mot ou d'une expression qui a deux voire plusieurs sens différents (on le qualifie de *polysémique*). Exemple : le mot « *souris* » peut désigner souris d'ordinateur, l'animal, le sourire, etc.

Plus récemment, [Pal *et al.*, 2014] proposent une méthode combinée à base de WordNet qui considère sa distribution, son association statistique avec les termes de requête, ainsi que sa relation sémantique avec la requête. Cette méthode surpasse les méthodes basées sur WordNet existantes telles que KLD [Carpineto *et al.*, 2001] et RM3 [Abdul-Jaleel *et al.*, 2004]. La combinaison de diverses sources d'information semble bien fonctionner et donne des résultats qui, dans l'ensemble, sont meilleurs que les méthodes individuelles impliquées dans la combinaison selon [Pal *et al.*, 2014].

La multiplication des expériences menées sur des bases textuelles diverses et à l'aide de ressources lexicales variées ont démontré, par la diversité de leurs résultats et de leurs conclusions, que la nature des ressources exploitées est cruciale [Picton *et al.*, 2008]. L'usage des ontologies et bases lexicales indépendantes du domaine telle que WordNet pose un problème dû à leurs larges couvertures ; et par la suite la présence des termes ambigus au sein de l'ontologie. Pour des SRIs particuliers, les ontologies spécifiques au domaine représente un choix plus approprié. Une ontologie spécifique au domaine modélise les termes et les concepts spécifiquement utilisés dans un domaine donné tel que le droit, la médecine, l'agriculture, la géographie, l'histoire, etc. [Bhogal *et al.*, 2007].

A titre de travaux sur des ontologies de domaine, [Fu *et al.*, 2005] présentent des techniques d'expansion de requête basées à la fois sur une ontologie géographique et sur le domaine de tourisme. Dans ce travail, une requête est développée par dérivation de son empreinte géographique. Les termes spatiaux tels que les noms de lieux sont modélisés dans l'ontologie géographique et les termes non spatiaux sont codés dans une ontologie de domaine touristique.

Dans la compagne TREC Genomics 2003, [Hersh *et al.*, 2003] testent une approche d'expansion en utilisant des phrases basées sur des synonymes de noms de gènes et une autre approche basée des ressources de connaissances externes. Les résultats de la première expérience étaient meilleurs que les résultats de la deuxième expérience. Ainsi, ils concluent que les résultats de l'expansion de la requête pourraient s'améliorer si la requête concerne une tâche spécifique.

[Nilsson *et al.*, 2006] utilisent une ontologie de domaine spécifique basée sur le système d'information de l'Université de Stockholm (intitulé SUiS) pour effectuer l'expansion de la requête. Les types de questions dans le système SUiS sont limités à *qui*, *quoi*, *quand* et *où*. Au lieu d'étendre les requêtes avec toutes les relations sémantiques fournies par une ontologie telle que WordNet, seuls les synonymes et les hyponymes sont utilisés pour augmenter la précision. Les expériences ont montré une amélioration des résultats.

[Dramé *et al.*, 2014] présentent la participation de l'équipe ERIAS à la 3ième tâche du ShARe/CLEF eHealth Evaluation Lab 2014. Le but de cette tâche est d'évaluer l'efficacité des SRIs pour aider les patients à accéder facilement aux informations médicales

pertinentes. L'approche proposée est basée sur le modèle d'espace vectoriel (SVM) et utilise deux extensions de ce modèle pour améliorer ses performances. Plus précisément, le thésaurus MeSH est utilisé pour l'expansion de requêtes avec différentes configurations. Des expériences sur une grande collection de documents ont montré que l'utilisation de ces ressources externes peut améliorer les performances dans la récupération des informations médicales.

[Lv *et al.*, 2015] proposent une nouvelle approche de modélisation linguistique pour la recherche de *microblog* dans Twitter en inférant différents types d'informations contextuelles. Les requêtes sont étendues en utilisant des termes de connaissances dérivés de Freebase [6]. En outre, afin de répondre aux besoins d'information en temps réel des utilisateurs, des informations temporelles sont incorporées dans les méthodes d'expansion afin que l'approche proposée puisse favoriser les *tweets* récents dans les résultats de recherche par rapport à un sujet donné. Les résultats expérimentaux sur deux corpus TREC-Twitter (2012 et 2013) démontrent une amélioration par rapport à d'autres méthodes de l'état de l'art.

En conclusion, les ontologies ont été utilisées pour un large éventail de tâches de RI. Les ontologies spécifiques au domaine conviennent mieux grâce à leurs terminologies moins ambigües. Les ontologies générales conviendraient aux requêtes générales ; mais le processus d'expansion de la requête peut nécessiter une tâche intermédiare de désambiguïsation ou une interaction de la part de l'utilisateur [Bhogal *et al.*, 2007].

3.3.2 Approches basées sur l'analyse de corpus

Dans ces approches, l'expansion de requête se base sur une analyse statistique de la collection de documents qu'interroge le SRI. Il s'agit de chercher des associations de termes dans la collection afin d'ajouter des termes voisins à la requête. Ceci se fait généralement de façon automatique par le calcul des liens contextuels entre termes [Kuzi *et al.*, 2016, Diaz *et al.*, 2016, Ermakova et Mothe, 2016] ou le recours à des méthodes de classification automatique de documents [Bellot *et al.*, 1998, Chifu et Mothe, 2014].

Les relations entre les mots peuvent être modélisées en comparant les distributions de mots trouvées dans les flux de langage naturel [Schütze, 1993]. Ces modèles construisent des distributions de mots en identifiant fréquemment des mots cooccurrents dans le langage naturel. Une approche intéressante de stockage de ces distributions consiste à les représenter dans des espaces vectoriels de grande dimension [Turney et Pantel, 2010].

6. Freebase constituait un projet de rassemblement et de connexion des connaissances du web, sous forme sémantique. Après son rachat, Google annonce la fermeture de Freebase en juin 2015 et le transfert de son contenu à Wikidata [Wikipedia].

La création de représentations vectorielles de mots permet d'utiliser des techniques d'algèbre linéaire pour modéliser des relations entre objets, y compris des associations syntagmatiques et paradigmatiques. Ces approches sont souvent appelées modèles d'espace sémantique, car la distance entre les mots dans l'espace reflète souvent divers groupements sémantiques, c'est-à-dire des mots liés par une association sémantique [Turney et Pantel, 2010].

Il y a eu un certain nombre de travaux au tour d'espaces sémantiques basés sur des corpus. La modélisation de ces espaces intègrent des phases d'apprentissage des associations sémantiques directement à partir du texte, à savoir HAL (*Hyperspace Analogue to Language*) [Lund et Burgess, 1996] et l'analyse sémantique latente (LSA) [Landauer et Dooley, 2002]. Des modèles plus récents ont incorporé des avancées ayant abordé des problèmes des modèles antérieur, notamment le manque d'informations structurelles stockées dans les représentations et la capacité à supporter des représentations de tenseur [7] d'ordre supérieur. Le modèle TE (*Tensor Encoding*), proposé par [Symonds *et al.*, 2011], a démontré une efficacité dans une gamme de tâches sémantiques qui prennent en compte la modélisation des associations syntagmatiques et paradigmatiques [Symonds *et al.*, 2014].

[Galeas et Freisleben, 2008] ont proposé une analyse de la distribution des mots comme méthode statistique descriptive pour calculer une représentation des positions de mots dans un document. Sur la base de cette approximation statistique, deux méthodes sont proposées pour améliorer l'évaluation de la pertinence des documents : (1) une procédure de classement de pertinence basée sur la façon dont les termes de requête sont distribués dans les documents initialement récupérés, et (2) une technique d'expansion de requête basée sur le chevauchement des distributions de termes dans les documents les mieux classés. Les résultats expérimentaux obtenus pour la collection de documents TREC-8 démontrent que l'approche proposée conduit à une amélioration d'environ 6,6% par rapport à TF-IDF [Buckley *et al.*, 1995] sans reformulation de la requête ni application de techniques de rétroaction de pertinence.

[Kuzi *et al.*, 2016] proposent une méthode à base de Word2Vec [Mikolov *et al.*, 2013] qui, une fois appliquée sur l'ensemble du corpus, elle sélectionne des termes sémantiquement liés à la requête. [Diaz *et al.*, 2016] démontrent que les modèles de plongements de mots (en anglais *word embedding*) tels que Word2Vec et GloVe [Pennington *et al.*, 2014] utilisés dans l'apprentissage global, donnent des performances inférieures aux méthodes spécifiques aux requêtes.

7. En mathématiques, plus précisément en algèbre multilinéaire et en géométrie différentielle, un tenseur désigne un objet très général, dont la valeur s'exprime dans un espace vectoriel.

3.3.3 Autres méthodes d'expansion de requêtes

L'exploitation des fichiers de *log* (ou de journal) constitue une piste intéressante de recherche [Agosti *et al.*, 2012]. Il existe deux techniques principales d'expansion de requêtes exploitant les journaux de recherche. La première est de traiter les requêtes individuelles comme des documents et d'extraire des caractéristiques de celles liées à la requête d'origine, avec ou sans utilisation des résultats de recherche associés (par exemple, [Jones *et al.*, 2006, Yin *et al.*, 2009]). La deuxième technique, plus largement utilisée, consiste à exploiter la relation entre les requêtes et les résultats de recherche pour fournir un contexte supplémentaire. Des exemples de cette dernière approche comprennent l'utilisation des meilleurs résultats de requêtes passées [Fitzpatrick et Dent, 1997], la recherche de requêtes associées aux mêmes documents [Billerbeck *et al.*, 2003] ou des clics de l'utilisateur [Beeferman et Berger, 2000], et l'extraction des termes directement à partir des résultats cliqués [Cui *et al.*, 2003].

Dans le travail de [Saneifar *et al.*, 2010], les auteurs extraient des passages pertinents à partir des fichiers logs tout en mettent en place une méthode d'enrichissement de requêtes. En effet, la méthode d'apprentissage utilisée se base sur la notion de *monde lexical*, des connaissances morpho-syntaxiques et une fonction de pondération des termes. Ce travail a été étendu dans [Saneifar *et al.*, 2014] en se basant sur deux étapes de rétroaction de pertinence. Les résultats, basé sur une nouvelle fonction de pondération appelée TRQ (*Term Relatedness to Query*), donnent une valeur MRR [8] de 0,87 ; alors que la valeur MRR était de 0,71 en utilisant les requêtes non étendues.

De nombreuses études sur l'expansion de requêtes à partir d'un corpus local souffrent de deux problèmes entraînant des performances faibles de RI : (1) les relations entre les termes sont limitées ; (2) et les termes de requête non répertoriés dans le corpus n'ont pas de termes d'expansion. Afin de résoudre ces deux problèmes, des travaux récents ont bénéficié de l'analyse des articles Wikipedia qui constitue un corpus riche en information sémantique [Milne et Witten, 2008, Almasri *et al.*, 2014, Gan et Hong, 2015].

A titre d'exemple, dans [Gan et Hong, 2015], des relations sémantiques entre les termes ont été extraites à partir de Wikipedia. Ces relations sont superposées sur un réseau de Markov de base qui a été pré-construit en utilisant un corpus local. Par conséquent, un nouveau réseau de Markov est formé avec des relations plus nombreuses et plus riches pour chaque terme. L'évaluation est effectuée sur trois corpus standard de RI (ADI, CISI et CACM). Les résultats expérimentaux montrent que la technique proposée du réseau de Markov superposé est efficace pour sélectionner plus de candidats pertinents pour l'expansion de requêtes.

8. Moyenne des Réciproques du Rang (MRR ; en anglais *Mean Reciprocal answer Rank*) [Voorhees, 1999].

3.3.4 Synthèse et discussion

Les différentes approches d'expansion de requêtes incluent principalement la rétroaction de pertinence, l'exploitation de connaissances issues des documents ou à partir de ressources linguistiques externes. Sur la base de cette étude, les ressources externes telles que les dictionnaires, les thésaurus et les ontologies ont une couverture linguistique plus élevée par rapport aux corpus textuels.

Cependant, construire des ontologies prend beaucoup de temps et nécessite des outils sophistiqués pour extraire et organiser les connaissances. En outre, les ontologies ne sont pas disponibles pour toutes les langues et tous les domaines. Les thésaurus sont construits à partir des index des documents. Par conséquent, ils souffrent des problèmes d'ambiguïté et de couverture car tous les sens des mots ne sont pas représentés et clairement distingués.

La performance d'expansion de requêtes dans les approches LSA ou basées sur l'analyse distributionnelle dépend de la qualité du corpus qui ne peut pas nécessairement couvrir toute la langue. Par conséquent, il est difficile de gérer tous les sens d'un mot donné. De plus, il est difficile de prouver que le corpus contient les relations nécessaires entre les mots qui sont utiles pour l'étude quantitative. Même si toutes ces relations existent, il n'est pas garanti qu'elles sont traitées avec LSA ou l'analyse distributionnelle ou toute autre méthode basées sur l'analyse de corpus.

Néanmoins, quelque soit l'approche utilisée pour l'expansion, l'introduction automatisée de nouveaux termes aux requêtes peut induire des problèmes et lever d'autres défis, comme synthétisé par [Audeh, 2014], à savoir :

- Risque de dérive de la requête : Ce phénomène se produit lors de l'expansion de la requête en altérant le but initial du besoin en information formulé par l'utilisateur. Dans ce cas de figure, les résultats retournés par les nouveaux termes risquent d'être plutôt pertinents par rapport à un autre sens différent de celui cherché à la base.
- Paramètres des techniques de rétroaction de pertinence : Le choix des paramètres, principalement le nombre de documents de retour de pertinence ainsi que le nombre de terme à extraire, représente le plus grand défi des approches de rétroaction de pertinence. Plusieurs études ont confirmé qu'un bon réglage des paramètres ne marchera pas forcément sur d'autres collections de tests, et même pas sur deux requêtes différentes sur la même collection de documents [Montgomery *et al.*, 2004, Ksentini *et al.*, 2016, Romberg, 2017].
- Adéquation de l'approche choisie d'expansion des requêtes : Plusieurs critères peuvent favoriser le choix d'une approche d'expansion donné tels que la disponibilité des ressources, le temps de calcul et la nature de stockage nécessaire.

Cependant, d'autres paramètres de fond doivent être considérés en fonction des priorités de l'utilisateur dans un contexte donné. Par exemple, une bonne approche pour l'expansion dans la recherche Web est celle qui améliore la précision, alors que c'est le rappel qu'il faut observer pour une approche d'expansion des requêtes dans un domaine médical. Un autre critère important lors du choix de la méthode d'expansion est la nature de la collection de documents. Par exemple, une collection dynamique, susceptible d'être modifiée fréquemment, est probablement moins adaptée aux méthodes d'expansion dépendantes de statistique sur la collection qu'une collection stable avec des documents inchangés [Audeh, 2014].

3.4 L'expansion de requête en RI translinguistique

Les systèmes de RI translinguistique cherchent à identifier les informations pertinentes dans une collection de documents dans des langues autres que celle dans laquelle l'utilisateur a formulé sa requête. Le besoin de transformer la requête, le document, ou les deux, en une représentation commune est nécessaire et ceci dépend des ressources de traduction disponibles. Ainsi, la performance du système est limitée par la qualité des traductions tout en considérant que les ressources avec une couverture plus large sont préférables. Cependant, les ressources linguistiques de haute qualité sont généralement difficiles à obtenir et à exploiter [McNamee et Mayfield, 2002]. Afin de résoudre les problèmes de couverture de termes lors du processus de traduction, l'expansion de requêtes a été utilisée dans la littérature en l'appliquant avant/après la traduction (voire avant et après la traduction).

3.4.1 L'expansion des requêtes en pré- et post-traduction

L'une des principales différences entre la traduction de requêtes et la traduction de texte est que la traduction de requêtes ne se limite pas à une traduction mot par mot, ni à des traductions littérales [Nie, 2010]. En effet, les traductions pour une requête peuvent être des termes qui y sont liés. Par exemple, même si le mot anglais « *hospital* » (*hôpital*) n'est pas une traduction d'une requête sur « *soins de santé* », ajouter ce terme à la traduction anglaise de la requête peut être bénéfique. La question est de savoir comment sélectionner ces mots de langue cible qui sont fortement liés à la requête dans la langue source.

Dans un contexte multilingue, il semble plausible que l'expansion préalable à la traduction soit effectivement utile. Si une ressource contient un nombre restreint de termes de recherche traduisibles, la dégradation résultant du processus de traduction entraînera

l'indisponibilité de nombreux mots de requête importants pour retourner des documents pertinents. Dans le cas contraire, si de nombreux mots liés à la requête sont traduits, alors le nombre final de termes disponibles pour rechercher la langue cible est plus grand. Cette méthode suppose que l'ensemble des termes traduits représente toujours la sémantique de la requête (c'est-à-dire que la demande d'information de l'utilisateur n'est pas significativement modifiée par l'expansion et la traduction). Si la traduction de la requête ne produit pas une requête comportant de nombreux termes, une expansion supplémentaire via la rétroaction de pertinence peut probablement améliorer la précision ainsi que le rappel.

L'idée exploitée dans l'expansion avant et après la traduction est similaire à la pseudo-rétroaction de pertinence selon [Nie, 2010]. L'expansion en pré-traduction utilise la requête d'origine pour récupérer un ensemble de documents à partir d'une collection dans la langue source. Un ensemble de termes est extrait et ajouté dans la requête avant la traduction. Par analogie, dans l'expansion en post-traduction, un ensemble de documents de langue cible est récupéré à l'aide de la requête traduite, et un ensemble de termes y est extrait et utilisé pour reformuler la requête.

[Adriani et Rijsbergen, 1999] proposent une technique d'expansion de requête basée sur une mesure de similarité statistique. Ils utilisent également une technique de désambiguïsation des sens fondée sur la similarité des termes. La technique d'expansion de requête est ensuite appliquée à la traduction. Adriani et Rijsbergen démontrent l'efficacité de combiner les deux techniques en utilisant des requêtes dans trois langues (à savoir l'allemand, l'espagnol et l'indonésien) pour récupérer les documents anglais de la collection standard TREC.

La pseudo-rétroaction de pertinence a été étudié dans le contexte de RI translinguistique. Par exemple, [Lee et Croft, 2014] revisitent le problème de CLIR en mettant l'accent sur les problèmes qui se posent avec les textes informels, tels que les blogs et les forums. Pour traiter le problème de *bruit*, causé par la traduction et le caractère informel des documents, les auteurs proposent de choisir entre la pseudo-rétroaction de pertinence (PRF) inter- et intra-langue, en fonction des propriétés du langage de la requête et des documents. Les expérimentations montrent que la PRF inter-langues est particulièrement utile pour les requêtes avec une mauvaise qualité de traduction. En contre partie, la PRF intra-langue est plus utile pour les requêtes bien traduites car elle réduit l'impact de toute erreur de traduction potentielle dans les documents.

Dans [Wang *et al.*, 2015], les auteurs utilisent une méthode de PRF basée sur l'alignement des sujets de faible pertinence (WRTA : *Weak Relevant Topic Alignment*) pour l'expansion de requête translinguistique sur des pages Web non alignées. Les sujets, dans différentes langues, sont alignés sur la base de leurs traductions. Des termes d'expansion utiles sont extraits en fonction de la similitude des termes bilingues. Les

résultats sur des données non parallèles dérivées du Web montrent la contribution du modèle de PRF basé sur WRTA dans la RI translinguistique.

Plusieurs autres études ont expérimenté et comparé les techniques d'expansion avant et après la traduction. [Ballesteros et Croft, 1997] ont constaté que les expansions avant et après la traduction peuvent améliorer la précision et le rappel. [McNamee et Mayfield, 2002] ont montré que l'expansion en pré-traduction est plus efficace que l'expansion après traduction, tandis que [Levow *et al.*, 2005] ont prouvé l'inverse.

3.4.2 Analogie entre expansion de requête et traduction de requête

Certaines études ont examiné à part les deux problèmes liés à l'expansion des requêtes dans la RI monolingue et à la traduction de requêtes pour la RI translinguistique. Cependant, [Nie, 2003] évoque la tâche de traduction de requêtes comme cas particulier de l'expansion de requêtes.

L'idée de base est d'améliorer la requête d'origine en y ajoutant certains termes reliés sémantiquement. Le succès d'une approche d'expansion des requêtes dépend principalement de deux facteurs : (1) la détermination de termes fortement liés à la requête et (2) la pondération des termes d'expansion par rapport aux termes d'origine. Nous avons présenté, dans les sections 3.2 et 3.3, trois principales méthodes pour identifier les termes d'expansion en se basant sur :

- La rétroaction de pertinence : les termes sont extraits à partir des premiers documents retournés en réponse à la requête ;
- Une ressource linguistique : tel qu'une base lexicale comme WordNet ;
- L'analyse de corpus : les termes cooccurrents dans le document sont utilisés pour enrichir la requête d'origine.

En comparant ces approches à celles de la RI translinguistique, une similarité entre elles peut être dégagée comme suit :

Traduction basée sur les dictionnaires vs. expansion de requêtes basée sur un thésaurus :

Un dictionnaire bilingue ou un thésaurus multilingue joue un rôle similaire dans la RI translinguistique à celui joué par un thésaurus monolingue en RI monolingue [Hedlund *et al.*, 2004]. Dans les deux cas, les relations stockées dans des ressources lexicales, construites manuellement, sont étudiées pour déterminer des termes connexes dans la même langue ou dans une langue différente (c'est-à-dire les traductions possibles).

Traduction basée sur corpus parallèle/comparable vs. analyse de co-occurrence :

L'extraction des relations de similarité inter-langues entre les termes à partir d'un corpus parallèle ou d'un corpus comparable peut être considérée comme une approche similaire à l'analyse de co-occurrence en RI monolingue [Braschler et Schäuble, 2000]. L'analyse de co-occurrence utilisée dans la RI monolingue a été largement étendue à des textes comparables et parallèles. La seule différence réside dans le fait que les co-occurrences dans ce dernier cas sont dans les textes correspondants en deux langues. Les modèles de traduction statistique peuvent également être considérés comme une analyse inter-langues de co-occurrence plus sophistiquée utilisant des phrases alignées [Nie, 2003].

3.4.3 Synthèse et discussion

L'étude établie dans la sous-section 3.4.1 sur l'expansion de requêtes avant ou après traduction, montre que les deux processus d'expansion sont capables d'améliorer l'efficacité de la RI dans une certaine mesure. Cependant, il n'existe aucune conclusion claire selon laquelle une méthode d'expansion est jugée meilleure que l'autre. En effet, l'impact de ces processus d'expansion dépend fortement des collections de documents à partir desquelles les termes d'expansion sont extraits.

L'expansion après traduction est généralement basée sur un ensemble de documents de langue cible extraits de la même collection sur laquelle la recherche finale est effectuée. Néanmoins, il est souvent courant dans la littérature d'utiliser une collection de documents différente pour l'expansion en pré-traduction. La différence entre cette collection et celle sur laquelle la recherche est effectuée peut expliquer les grandes différences observées sur l'impact de l'expansion de la pré-traduction : lorsque ces collections couvrent des sujets similaires, on pourrait s'attendre à ce que l'expansion de la traduction puisse trouver des termes supplémentaires fortement liés à la requête avant la traduction. Si, au contraire, les collections couvrent des sujets très différents, l'expansion en pré-traduction pourrait dégrader la performance de recherche.

Par exemple, si la requête en anglais « *drug traffic* » (en français « *trafic de drogue* ») est étendue avant la traduction en utilisant une collection de documents médicaux en anglais ; puis traduite et utilisé dans la recherche dans une collection générale en français, on peut s'attendre à une dérive de la requête vers le sens de « *médicament légal* », ce qui est nocif. Ceci est dû à la polysémie du mot « *drug* » qui peut être traduit par « *médicament* » ou « *drogue* ». Ainsi, la collection utilisée pour l'expansion de la traduction devrait être fortement liée aux sujets de la requête.

La comparaison établie dans la sous-section 3.4.2 met en relief le fait que de nombreuses approches de RI translinguistique ont leurs homologues en RI monolingue. À savoir, la plupart des approches de traduction de requêtes peuvent être exprimées comme un

cas spécial d'expansion de requêtes en RI monolingue. Cela suggère que les méthodes développées pour l'expansion des requêtes en RI monolingue peuvent généralement être adaptées pour la traduction dans la RI translinguistique. Cette vue unifiée a été adoptée par Nie [Nie, 2003, Nie, 2010], en montrant une forte analogie entre l'expansion de requêtes, d'une part, et la traduction de requêtes en RI translinguistique, d'autre part.

Néanmoins, cette comparaison ne signifie pas que la différence de la langue entre l'expansion de requêtes en RI monolingue et la traduction de requêtes dans RI translinguistique n'est pas importante. Au contraire, la plupart des approches pour la RI translinguistique visent spécifiquement les problèmes posés par les différences linguistiques. En dépit de ces différences langagières, les deux problèmes partagent de nombreux points communs et des approches similaires.

En fait, dans les deux cas, l'objectif est de déterminer les termes sémantiquement en relation avec la requête initiale. Considérons par exemple la requête Q : « *prévention de propagation du virus Zika* » en français. On peut identifier, en utilisant différentes ressources, les termes français suivants représentent les significations impliquées dans la requête : *« fièvre »*, « *transmission du virus* », « *pandémie* », « *vaccin* », « *campagne de vaccination* », etc.

Dans un contexte de RI translinguistique, nous pouvons identifier les traductions suivantes dans d'autres langues en relation avec la requête Q : { *prevention*, *Zika*, *spreading* } (en anglais) et { زيكَا, فيروس, انتشَار, من, الوقَاية } (en arabe).

Si notre objectif est de traduire uniquement la requête terme à terme, il s'agit d'une traduction stricte. Cependant, il y a généralement une tendance à induire les termes en relation avec la requête après traduction comme par exemple : « *virus transmission* », « *pandemic* », « *vaccine* », « *vaccination campaign* » (en anglais). Ces termes sont en effet reliés à la requête d'origine sans pour autant être des traductions directes.

Conclusion

La reformulation de requêtes représente une méthode d'amélioration des performances d'un SRI en terme de rappel et précision. Nous nous intéressons principalement à la

tâche d'expansion de requêtes qui consiste à proposer de nouveaux termes à l'utilisateur afin de mieux exprimer son besoin selon des approches variées. L'application de ces approches à travers les modèles de RI est très variée dans la littérature et tourne principalement autour de la rétroaction de pertinence et les méthodes dites globales d'expansion de requête. Ces dernières se déclinent selon les ressources exploitées : des ressources linguistiques externes (dictionnaires, ontologies), l'analyse de corpus (co-occurrence dans les documents, Wikipedia, etc.) ou d'autres ressources telles que l'analyse des fichiers *logs*. Les méthodes d'expansion, appliquées souvent dans un cadre monolingue, peuvent être projetés sur un cadre de RI translinguistique. La combinaison de la tâche d'expansion en parallèle avec la tâche de désambiguïsation sémantique est également prometteuse en présence de polysémie dans les requêtes.

Partant de cette étude de l'état de l'art, nous définissions dans le chapitre suivant une architecture d'un système semi-automatisé incluant les tâches désambiguïsation sémantique et d'expansion de requêtes. L'objectif d'un tel système est d'accompagner l'utilisateur dans les phases de RI afin de lever l'ambiguïté des termes polysémiques. La proposition de tels systèmes contribue dans la facilitation de la RI via des interfaces d'interaction Homme-Machine.

Deuxième partie

CONTRIBUTIONS

Chapitre

4 Modèle d'un Système Possibiliste d'Expansion Et de Désambiguïsation SEmantique de Requêtes (SPEEDSER)

Sommaire

" Accepting that the world is full of uncertainty and ambiguity does not and should not stop people from being pretty sure about a lot of things. "

— Julian Baggini

Introduction

Au terme de l'état de l'art fait dans la première partie de ce rapport, nous concluons que la performance d'un SRI en terme de pertinence dépend fortement de la manière selon laquelle l'utilisateur exprime son besoin d'information. La formulation de ce besoin en exprimant un ensemble réduits de mots-clés dans la requête rend le contexte pauvre, incertain et imprécis. Ce problème s'amplifie davantage en présence de mots ambigus qui portent plusieurs sens. Afin de résoudre ce problème d'ambiguïté sémantique (posé principalement par la polysémie) conjointement au contexte incertain et imprécis de recherche, la théorie des possibilités s'apprête naturellement à ce genre d'application [Dubois et Prade, 2011].

Suite à l'évolution scientifique et l'augmentation du volume des connaissances, l'utilisateur d'un SRI se confronte, de nos jours, à deux problèmes informationnels majeurs : La variété des données d'une part et l'ampleur de l'information d'autre part. En d'autres termes comment arriver à l'information pertinente et exacte, autrement augmenter la précision et le rappel du SRI, à la fois à partir d'un grand corpus documentaire ?

Étant confronté à cette grande masse d'information, il s'avère indispensable de développer des outils facilitant l'acquisition et l'interprétation de l'information textuelle à travers des interfaces automatisées [Oard *et al.*, 2008]. Ceci s'applique pour les différentes tâches qui accompagnent le processus de RI, à savoir : l'expansion et la désambiguïsation sémantique des requêtes ainsi que le choix des traductions adéquates dans le cas d'une RI translinguistique.

Dans ce chapitre, nous introduisons les fondements de base de la théorie des possibilités, ses applications dans la RI ainsi que ses adaptations dans notre nouveau Système Possibiliste d'Expansion Et de Désambiguïsation SEmantique de Requêtes (SPEEDSER). Dans la deuxième partie, nous détaillons le modèle conceptuel du système SPEEDSER ainsi que ses différents modules présentés via des interfaces graphiques. En fin de chapitre, nous établissons un tableau comparatif de notre système SPEEDSER avec d'autres systèmes de l'état de l'art pour la RI monolingue et translinguistique.

4.1 Théorie des possibilités et applications à la RI

La théorie des possibilités introduite par [Zadeh, 1978] et développée par [Dubois et Prade, 2012] traite l'incertitude sur l'intervalle [0,1], appelé *échelle possibiliste*, d'une manière qualitative ou quantitative. En effet, Zadeh a formalisé la théorie des possibilités pour traiter l'incertitude permettant ainsi de traiter l'ignorance et de prendre en compte la pertinence d'une information incertaine.

Dans cette théorie, l'information fournie par une source sur la valeur réelle d'une variable x est codée sous forme d'une distribution de possibilités dont les valeurs sont supposées être mutuellement exclusives, puisque x prend en définitive une seule valeur (sa vraie valeur), qui appartient à un ensemble Ω donné [Sandri, 1991]. La théorie des possibilités se base sur deux mesures de confiance : la mesure de *possibilité* et la mesure de *nécessité* [Fabiani, 1996].

4.1.1 Distribution de possibilité

La théorie des possibilités est basée sur les distributions de possibilité. Une distribution de possibilité, notée par π, est une application de Ω (l'univers de discours) vers l'échelle [0,1] traduisant une connaissance partielle sur le monde, noté ω. L'échelle possibiliste est définie de deux manières.

Dans le cadre numérique les valeurs des possibilités traduisent souvent les bornes supérieures des probabilités. Dans le cadre qualitatif, les valeurs de possibilité peuvent être considérées comme un ordre de classement des états possibles. La combinaison des distributions de possibilité, exprimée à l'aide des normes triangulaires (*t-normes*) dépend du cadre. Les opérateurs « produit » et « minimum » peuvent être utilisés pour combiner des distributions de possibilité indépendantes dans les cadres quantitatif et qualitatif respectivement.

Normalisation : Une distribution de possibilité est dite $\alpha-$normalisée, si son degré de normalisation, noté $\alpha(\pi)$, est égal à α. Ainsi :

$$\alpha = \alpha(\pi) = max_{\omega}\pi(\omega) \tag{4.1}$$

Lorsque $\alpha = 1$, π est dite normalisée.

Marginalisation : Soit une distribution de possibilité jointe, π sur Ω , une distribution marginale relative aux sous ensembles de variables peut être dérivée en utilisant l'opérateur maximum. Ainsi, $\forall$ X $\subseteq$ V, $\forall$ x $\subseteq$ dom(X) :

$$\pi(x) = max_{\omega \in \Omega}\{\pi(\omega) : \omega[X] = x\} \tag{4.2}$$

Où :

- V : ensemble de variables $\{A_1, A_2, ..., A_N\}$;
- X : sous ensemble de V ;
- $dom(X)$: domaine de X, produit cartésien des domaines des variables de X ;
- x : une instance de X, si $X = \{A_1, A_2, ..., A_j\}$, alors $x = (\alpha_1, \alpha_2, ..., \alpha_j)$;

– $\omega[X] = x$: configuration de X dans ω.

Une distribution de possibilité π sur ω permet de qualifier les évènements en terme de mesure de *plausibilité* et de *certitude* respectivement.

4.1.2 Mesures de possibilité et de nécessité

Dire qu'un évènement est non possible n'implique pas seulement que son évènement contraire est possible mais qu'il est certain. Deux mesures duales sont utilisées : la mesure de possibilité $\Pi(\phi)$, et la mesure de nécessité $N(\phi)$.

– La possibilité d'un évènement A, notée $\Pi(A)$ est obtenue par $\Pi(A) = max_{x \in A}\pi(x)$ et décrit la situation la plus normale dans laquelle A est vraie ;

– La nécessité $N(A) = min_{x \notin A}(1 - \Pi(\bar{A}))$d'un évènement A reflète la situation la plus normale dans laquelle A est faux.

La distance entre $N(A)$ et $\Pi(A)$ évalue le niveau d'ignorance sur A. Rappelons que $N(A) > 0$ impliqueΠ$(A) = 1$. Lorsque A est un ensemble flou, cette propriété n'est plus vérifiée et dans ce cas l'inégalité $N(A) \leq \Pi(A)$ est vérifiée.

4.1.3 Réseaux Possibilistes (RP)

Les travaux existant sur les réseaux possibilistes sont soit des adaptations directes de l'approche probabiliste [Benferhat *et al.*, 1999], ou des méthodes d'apprentissage à partir de données imprécises [Borgelt *et al.*, 2000]. La théorie des possibilités offre deux définitions du conditionnement, ce qui conduit à deux définitions des réseaux causaux possibilistes. Les réseaux possibilistes basés sur le produit sont très similaires aux réseaux probabilistes.

Définitions

Un graphe possibiliste orienté sur un ensemble de variables $V = \{A_1, A_2, ..., A_N\}$ est caractérisé par une composante qualitative et une composante numérique. La première est un graphe acyclique orienté. La structure du graphe représente l'ensemble des variables ainsi que l'ensemble des relations d'indépendance. La seconde composante quantifie les liens du graphe en utilisant des distributions de possibilité conditionnelles de chaque nœud dans le contexte de ses parents. Ces distributions de possibilité doivent vérifier la contrainte de normalisation. Pour chaque variable A_i :

– Si A_i est un nœud racine et dom_{A_i} le domaine de A_i, la possibilité à priori de A_i doit satisfaire :

$$max_{a_i}\Pi(a_i) = 1, \forall a_i \in dom_{A_i} \tag{4.3}$$

– Si A_i n'est pas un nœud racine, la distribution conditionnelle de A_i dans le contexte de ses parents doit satisfaire :

$$max_{a_i}\Pi(a_i|\theta_{A_i}) = 1, \forall a_i \in dom_{A_i} \tag{4.4}$$

Avec :

– dom_{A_i} : Le domaine de A_i ;

– θ_{A_i} : L'ensemble des configurations possibles des parents de A_i ;

Réseaux possibilistes basés sur le produit

Un graphe possibiliste basé sur le produit est un graphe possibiliste où les possibilités conditionnelles sont obtenues par le conditionnement produit. La distribution de possibilité des réseaux possibilistes basés sur le produit, notée par π_p, est obtenue par la règle de chainage :

$$\pi_p(A_1, A_2, ..., A_N) = PROD_{i=1..N}\Pi(A_i|\theta_{A_i}) \tag{4.5}$$

– Avec : $PROD$ est l'opérateur produit.

4.1.4 Modèle possibiliste quantitatif de RI

Pour évaluer les valeurs de nécessité et de possibilité , appliquées dans un cadre de RI, Brini propose dans [Brini *et al.*, 2004a, Brini *et al.*, 2004b] un modèle de base quantitatif de pondération des différents nœuds dans un réseau possibiliste (RP). L'approche proposée tente de distinguer entre les termes possiblement représentatifs des documents (ceux qui sont absents sont écartés) et ceux nécessairement représentatifs, c'est-à-dire les termes qui suffisent à caractériser les documents.

Le modèle de base proposé part des quatre hypothèses suivantes :

Hypothèse 1 : Un terme est d'autant moins représentatif d'un document qu'il apparaît peu fréquemment dans ce document.

Hypothèse 2 : Un terme est d'autant plus nécessairement représentatif du document qu'il apparaît fréquemment dans ce document et peu fréquemment dans les autres documents de la collection.

Hypothèse 3 : A priori, un document possède une égale possibilité d'être pertinent ou non pour un utilisateur potentiel, soit :

$$\Pi(d_j) = \Pi(\neg d_j) = 1, \forall j \tag{4.6}$$

Hypothèse 4 : Les termes de chaque document de la collection sont conditionnellement indépendants de ce document.

D'après l'hypothèse 1, $\Pi(t_i|d_j)$ peut être estimée par la fréquence tf_{ij} de t_i dans d_j :

$$\Pi(t_i|d_j) = nft_{ij} = tf_{ij}/max(tf_{kj}) \tag{4.7}$$

Où nft_{ij} correspond à la fréquence normalisée.
Avec l'hypothèse 3, on déduit :

$$\Pi(t_i \wedge d_j) = \Pi(t_i|d_j) \tag{4.8}$$

Le degré de possibilité évalue à quel point un terme est « typique » (ou discriminant) du document et donc à quel point il est possible qu'il contribue à sa restitution. S'il apparaît avec une fréquence maximale ($ntf_{ij} = 1$), alors il est considéré comme le meilleur candidat potentiel pour sa représentation.

Pour le modèle de base de Brini, la requête est composée d'une simple liste de mots-clés. Lorsque la requête est connue, un processus de propagation est déclenché à travers le réseau, modifiant les valeurs des possibilités à priori des documents en vertu de leurs liens avec les termes d'indexation.

Soit une requête $Q = (t_i, ..., t_T)$ (interprétée conjointement [1]), alors :

$$\Pi(d_j|Q) = (\Pi(Q|d_j) * \Pi(d_j))/\Pi(Q) \tag{4.9}$$

Si un document D_j est composé des termes T, l'hypothèse 4 jointe à l'hypothèse 3 simplifie la formule 4.9 ; $\Pi(d_j|Q)$ est alors proportionnel à :

$$\Pi'(d_j|Q) = \Pi(t_1|d_j) * ... * \Pi(t_T|d_j) = nft_{1j} * ... * nft_{Tj} \tag{4.10}$$

La nécessité représente la mesure duale de la possibilité dans la logique possibiliste. Elle exprime l'idée qu'un document soit certainement pertinent ou pas. Cette certitude, notée $N(d_j|Q)$ est donnée par :

$$N(dj|Q) = 1 - \Pi(\neg d_j|Q) \tag{4.11}$$

Où $\Pi(\neg d_j|Q)$ est proportionnel d'après les hypothèses 3 et 4, à :

$$\Pi'(\neg d_j|Q) = \Pi(t_1|\neg d_j) * ... * \Pi(t_T|\neg d_j) = (1 - \phi_{1j}) * ... * (1 - \phi_{Tj}) \tag{4.12}$$

1. Le réseau possibiliste est basé sur le produit

ϕ_{ij} est le degré de nécessaire pertinence donné (conformément à l'hypothèse 2) par :

$$\phi_{ij} = \mu_1(nC/nd_i) * \mu_2(nft_{ij}) \tag{4.13}$$

Avec :

- nC : nombre de documents de la collection ;
- nd_i : nombre de documents de la collection contenant le terme t_i ;
- μ_1 et μ_2 : fonctions de normalisation.
  Typiquement μ_1 : fonction croissante de type logarithmique, μ_2 : la fonction identité.

Les documents préférés sont ceux qui ont des valeurs $N(d_j|Q)$ et $\Pi(d_j|Q)$ élevées. Ils sont ainsi qualifiés de documents possiblement et nécessairement pertinents pour la requête Q.

Brini a proposé également dans [Brini *et al.*, 2007] un modèle plus sophistiqué que le premier modèle quantitatif de base et y intègre d'autres facteurs lors du calcul de score de pertinence tel que la longueur des documents (ainsi, les documents courts sont favorisés par rapport aux documents longs ou inversement). Elle a comparé aussi les différentes stratégies d'agrégation des termes de la requête (par conjonction, disjonction, etc) pour aboutir à un modèle possibiliste optimal de RI.

4.1.5 Modèle possibiliste qualitatif de RI

Elayeb a proposé dans le système SARIPOD [Elayeb, 2009] une extension du modèle possibiliste quantitatif de base de [Brini *et al.*, 2004a] vers un cadre possibiliste qualitatif. Travaillant sur des documents semi-structurés, il applique le modèle de Brini, non pas à la totalité d'un document, mais à ses entités logiques. Ces derniers sont obtenues suite à une analyse des documents permettant de générer les fragments logiques de chaque page Web retrouvée [Elayeb *et al.*, 2009][Elayeb, 2009].

Il définit le degré de pertinence possibiliste mixte de chaque entité logique (ELd_i) d'un document d_i par :

$$DPMEL(d_i) = \Pi(ELd_i|Q) + N(ELd_i|Q) \tag{4.14}$$

Avec :

- $\Pi(ELd_j|Q) = \Pi(t_1|ELd_j) * ... * \Pi(t_T|ELd_j) = nft_{1j} * ... * nft_{Tj}$; nft_{ij} : fréquence normalisée des termes de la requête dans l'entité logique ELd_i ;
- $N(ELd_i|Q) = 1 - \Pi(\neg ELd_j|Q)$; avec :
  - $\Pi(\neg d_j|Q) = (1 - \phi_{EL1j}) * ... * (1 - \phi_{ELTj})$ et
  - $\phi_{ELij} = Log_{10}(nCEL/nELd_i) * (nft_{ij})$ où

- $nCEL$: nombre d'entités logiques des documents de la collection ;
- $nELd_i$: nombre d'entités logiques des documents de la collection contenant le terme t_i.

Le calcul du score final de pertinence d'un document se fait en impliquant les scores de chaque entité logique. Il a également proposé une sorte de pondération des fragments logiques favorisant ainsi un élément à un autre comme suit :

$$DPM(d_i) = \sum_j (\alpha_j * DPMEL_j(d_i)) \tag{4.15}$$

Où α_j : coefficient de préférence pour l'entité logique j.

L'adoption du cadre possibiliste permet d'éclaircir les définitions de la pertinence ainsi que la représentativité d'un terme dans un document. La notion de pertinence d'un document, étant donnée une requête, est modélisée par une double mesure.Ces degrés mesurent d'une manière générale le degré de possibilité et de nécessité de l'information véhiculée par les arcs du réseau possibiliste. Cette information concerne la représentativité d'un terme dans un document et permet de quantifier la pertinence d'un document étant donnée une requête.

4.1.6 Vers une généralisation du modèle possibiliste

En s'inspirant des travaux antérieurs de [Brini *et al.*, 2007] et de [Elayeb *et al.*, 2009] qui ont proposés respectivement un modèle quantitatif et qualitatif de RI, nous proposons d'étendre ces travaux vers un cadre général d'utilisation (figure 4.1).

La théorie des possibilités à été appliquée dans d'autres travaux en relation avec le domaine de RI. Dans [Bounhas *et al.*, 2015], les auteurs proposent un classifieur possibiliste pour la désambiguïsation morphologique des textes arabes. Le modèle proposé utilise un analyseur morphologique pour faire l'apprentissage du classifieur à partir de documents non voyellés et l'appliquer ensuite sur un corpus non voyellé.

Dans [Chebil *et al.*, 2016], les auteurs proposent une approche d'indexation des documents biomédicaux à partir d'un réseau possibiliste. Cette approche permet d'extraire les concepts biomédicaux en effectuant une correspondance partielle entre les documents et le vocabulaire biomédical.

[Garrouch et Omri, 2015] proposent un modèle de RI basé sur les réseaux possibilistes en le comparant à un modèle bayésien de RI. La structure de ce modèle réduit les dépendances entre les termes d'indexation à ceux qui sont les plus pertinents. L'approche utilisée pour extraire l'ensemble de ces dépendances se concentre sur les dépendances locales entre les termes dans chaque document.

Nous notons également que des travaux plus récents d'extension du modèle possibiliste ont été faits dans un cadre de RI purement translinguistique. En effet, dans [Elayeb *et al.*, 2018], les auteurs ont proposé une transformation *probabilité*→*possibilité* comme moyen d'introduire une tolérance supplémentaire dans le processus de traduction de requête. À partir de l'identification des phrases nominales, la requête dans le langue source est traduite en unités à l'aide de patrons de traduction.

Ce travail a été étendu dans [Ben Romdhane *et al.*, 2017] vers une approche de traduction de requête possibiliste *discriminative* en utilisant à la fois un dictionnaire bilingue et un corpus parallèle. L'objectif principal est de surmonter certains inconvénients des techniques basées sur les dictionnaires.

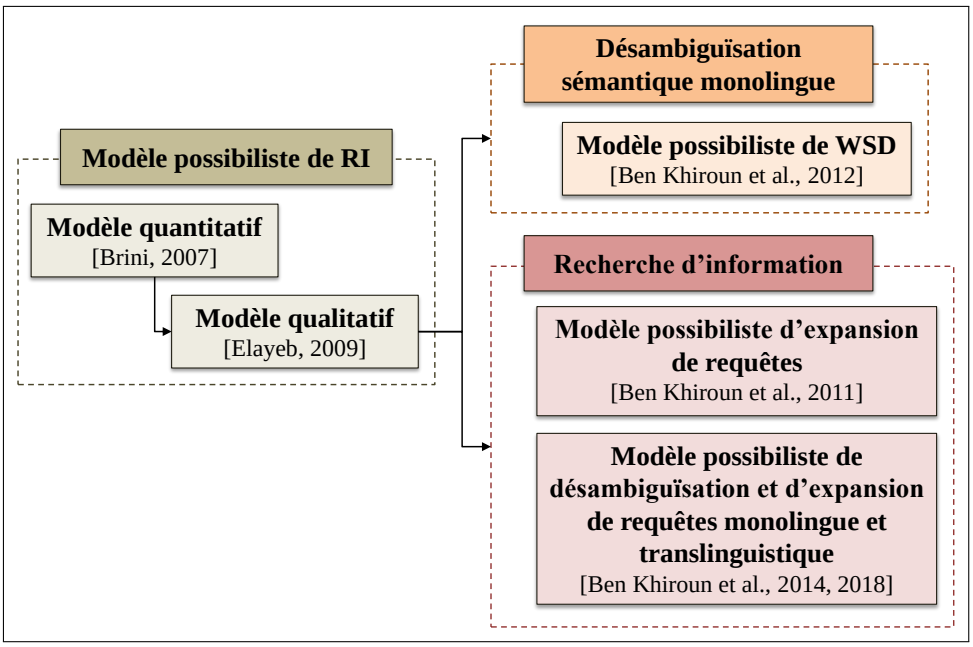


Figure 4.1 – Extension du modèle possibiliste qualitatif

Dans notre travail de thèse, nous traitons l'incertitude posée par la polysémie en ayant recours aux réseaux possibilistes. Ce type de réseaux offre un cadre de modélisation des dépendances entre les termes ambigus d'une part et les mots avec lesquels ils ont une relation sémantique d'autre part. Nous désignons par relation sémantique l'appartenance à la définition dans un dictionnaire, la co-occurrence dans un contexte de phrase, ou tout autre lien aidant à résoudre cette ambiguïté par un calcul possibiliste. Nous étudions ainsi, dans le chapitre 5, la contribution de la théorie des possibilités dans la désambiguïsation monolingue des textes (WSD) [Ben Khiroun *et al.*, 2012].

Ces relations sémantiques peuvent également contribuer dans la phase d'expansion des requêtes dans la RI en injectant les termes liés sémantiquement avec la requête d'origine [Ben Khiroun *et al.*, 2011]. Par conséquent, nous étudions, dans le chapitre 6, l'apport de l'expansion et la désambiguïsation des requêtes dans le processus de la RI

monolingue [Ben Khiroun *et al.*, 2014], dans un premier lieu, et la RI translinguistique, en deuxième lieu [Ben Khiroun *et al.*, 2018].

Dans ces axes d'extension introduits, la représentation des nœuds dans le réseau possibiliste dépendent de la nature de la tâche à réaliser (désambiguïsation des textes, expansion ou/et désambiguïsation des requêtes, désambiguïsation des traductions). Les arcs reliant chaque couple de nœuds décrivent une relation de dépendance et sont quantifiés par deux mesures : la possibilité et la nécessité.

Ainsi, quel que soit le type de la relation décrite par un arc entre deux nœuds, sa quantification est engendrée par ces deux mesures. Alors que la première est utile pour écarter certaines informations, la seconde mesure renforce les informations restantes.

Nous présentons dans ce qui suit le modèle conceptuel de notre système intitulé SPEEDSER tout en détaillant les différents modules et outils qu'il intègre.

4.2 Modèle conceptuel du système SPEEDSER

Cette section décrit l'approche globale de conception d'un système de recherche d'information incluant des outils d'expansion et de désambiguïsation de contexte. Nous distinguons principalement 5 modules complémentaires dans l'architecture globale de notre système SPEEDSER présenté dans la figure 4.2.

La première composante du système est le *module d'expansion* [1] qui est fondé sur l'utilisation de ressource linguistique de type dictionnaire [2].

Afin de résoudre l'ambiguïté pour les mots ayant plusieurs sens dans un texte, le module de désambiguïsation sémantique joue un rôle important dans notre système.

Dans un premier lieu, nous nous attaquons uniquement au volet monolingue pour les textes écrits dans une seule langue. Le module de *désambiguïsation monolingue des sens* [2] favorise le sens le plus approprié en se référant à des ressources linguistiques [B] à savoir : (i) un dictionnaire de sens, désigné aussi par « *Dictionnaire Sémantique de Contexte* » [B.1] et (ii) un graphe de co-occurrences [B.2] construit à partir de corpus documentaire.

Une phase d'expansion de requêtes peut-être appliquée en fin du processus de désambiguïsation monolingue [1'].

Dans un second lieu, nous modélisons la tâche de désambiguïsation dans un contexte translinguistique où les requêtes et les documents à rechercher ne sont pas dans la même langue.

2. Les nomenclatures [x] font référence aux modules de SPEEDSER comme schématisés dans la figure 4.2

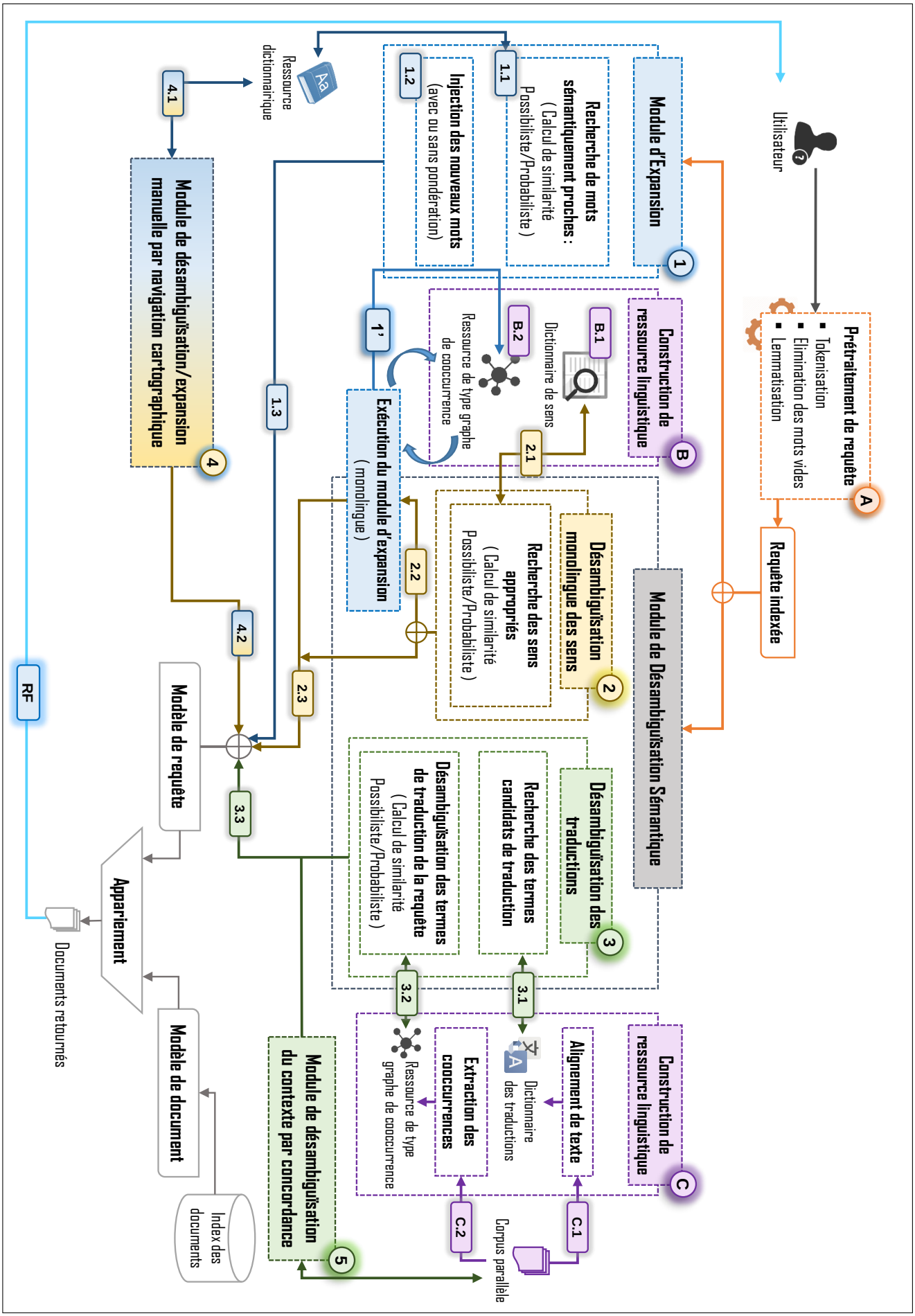


Figure 4.2 – Architecture générale du système SPEEDSER

Le module de *désambiguïsation des traductions* (ou *module de désambiguïsation translinguistique*) [3] favorise les meilleurs traductions selon un calcul de score de similarité sémantique. En effet, deux types de ressources sont utilisés dans cette phase à savoir : (i) un dictionnaire des traductions [C.1] extrait depuis l'alignement de textes parallèles écrits dans deux langues et (ii) un graphe de co-occurrences [C.2] construit à partir de l'analyse des contextes dans le corpus documentaire.

Afin de proposer une meilleure interaction avec l'utilisateur, le système SPEEDSER propose deux modules interactifs d'expansion et de désambiguïsation sémantique :

- Un *module de désambiguïsation / expansion manuelle par navigation cartographique* [4] : ce module propose une interface de visualisation des relations entre les mots et leurs sens issus à partir d'un dictionnaire électronique.
  L'exploration des liens sémantiques, schématisés sous forme de graphe, offre un cadre de navigation semi-automatisé d'expansion et de désambiguïsation des contextes.
- Un *module de désambiguïsation du contexte par concordance* [5] : la proposition des différents contextes, dans lesquels apparaît un mot, constitue une approche intéressante pour l'utilisateur. En effet, l'examen des mots voisins à un mot ambigu facilite la distinction du vrai sens approprié.
  Ce module inclue un concordancier ainsi qu'une recherche dans un dictionnaire bilingue qui servent dans la tâche de recherche translinguistique.

Les principaux modules du système SPEEDSER seront détaillés dans les sous-sections ci-après.

4.2.1 Prétraitement de requête

Dans ce module [A], la requête subit en premier lieu une phase d'analyse, appelée *tokenisation*, afin de dégager les différents termes (ou *token*) et leurs poids respectifs.

Une fois la requête analysée et parcourue, il y aura une phase d'élimination des mots vides. Ceci est réalisé en se référant à un fichier texte, souvent appelé *stop word list*, qui regroupe une liste « *noire* » de mots à ignorer. Ces mots sont considérés comme non discriminants (non significatifs) pour les documents écrits en langue française, par exemple, comme les pronoms, les prépositions et les articles.

La dernière étape (racinisation ou *stemming* en anglais) consiste à supprimer les suffixes et les terminaisons connus pour réduire les différentes formes d'un mot à leur racine. Par exemple, l'adjectif « *petit* » existe sous quatre formes : « *petit* », « *petite* », « *petits* » et « *petit*es ». La forme réduite (dite aussi *canonique*) de tous ces mots est « *petit* ». Ce mécanisme de réduction donne de meilleurs résultats pour la phase d'appariement puisque les entrées de l'indexe de la collection sont aussi sous forme réduite.

4.2.2 Appariement requêtes/documents

Afin d'assurer l'appariement entre requête/documents, les documents de la collection de recherche doivent subir une phase d'indexation au préalable assuré par la plate-forme d'expérimentation Terrier [Ounis *et al.*, 2006, Macdonald *et al.*, 2012] que nous avons utilisée pour la réalisation de notre système (voir figure 4.3).

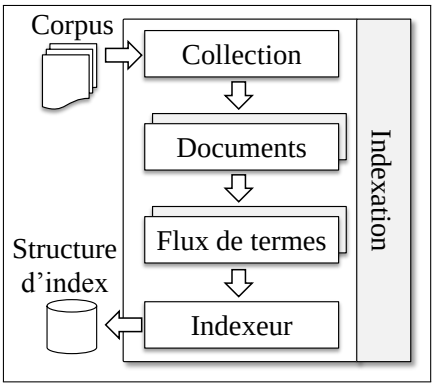


Figure 4.3 – Processus d'indexation dans Terrier

Chaque terme extrait d'un document possède principalement trois propriétés :

- La chaîne de caractères représentant le terme ;
- La position dans laquelle apparaît le terme dans le document ;
- Le champs dans lequel apparaît le terme [3].

Les termes des documents subissent un pré-traitement en passant par une chaîne de traitement (*term pipeline*). Nous nous intéressons uniquement à l'élimination de mots vides et à la réduction des termes sous forme canonique (*stemming*) spécifique à la langue française.

Ces deux opérations contribuent à la compatibilité avec le modèle de requêtes dans le processus d'appariement puisque les requêtes suivent la même logique de pré-traitement décrit auparavant dans la section 4.2.1.

En dernière étape, l'indexeur stocke des structures adéquates de représentation des documents dans divers fichiers. Ces structures se résument dans :

Le lexique (ou *Lexicon*) : Le lexique stocke le terme et son identificateur (un numéro unique pour chaque terme), ainsi que les statistiques globales du terme (fréquence du terme globale dans la collection et le nombre de documents contenant ce terme) et la liste des entrées dans l'index inversé ;

L'index inversé (ou *Inverted Index*) : Pour chaque terme, l'index inversé stocke : l'identificateur du document correspondant et la fréquence de ce terme dans ce document ;

L'index de documents (ou *Document Index*) : L'index de documents sauvegarde le numéro du document (un identificateur externe unique du document), l'*id* du document (une identification interne unique du document) ; la longueur du document en termes de jetons (*tokens*) et l'offset du document dans l'index direct ;

3. cette propriété est spécifique aux documents écrits dans un langage de balisage comme XML ou HTML ; on désigne par champs ici les balises.

L'index direct (ou *Direct Index*) : Il stocke pour chaque document les identificateurs des termes présents et leurs fréquences. Il s'agit d'une représentation orthogonale de l'index inversé et peut-être considéré comme la représentation intuitive des documents de la collection.

Une fois le contenu des documents représenté sous forme d'index, l'utilisateur exprime son besoin en information sous la forme d'une requête, qui est interprétée selon le modèle de requête et le système évalue la pertinence des documents par rapport à cette requête par l'intermédiaire de la fonction de correspondance.

Le processus d'appariement requêtes/documents permet de mesurer la pertinence d'un document vis-à-vis d'une requête. Ainsi à chaque réception d'une requête, notre système crée une représentation similaire à celle des documents, puis calcule un score de correspondance entre la représentation de chaque document et celle de la requête en suivant un modèle d'appariement.

Ce score traduit un degré de pertinence du système qui est supposé représenter le jugement de pertinence de l'utilisateur vis-à-vis du document. Notre système retourne par suite une liste de documents classés par ordre décroissant selon le score de pertinence.

4.2.3 Module d'expansion de requêtes

Une des techniques de reformulation de requêtes proposées dans les SRI repose sur l'utilisation de références lexicographiques comme un thésaurus, une ontologie ou un simple dictionnaire pour dégager de nouveaux termes. La recherche de ces nouveaux termes se fait par un jugement de proximité sémantique en se basant sur une méthode particulière.

Au sein du système SPEEDSER, nous implémentons deux formes d'expansion sémantique de requêtes (module 1) : l'approche d'expansion à base de dénombrement de circuits et l'approche d'expansion possibiliste [Ben Khiroun *et al.*, 2011]. Nous avons également mis en place une troisième méthode de reformulation interactive manuelle (module 4).

Le module d'expansion accepte en entrée la requête de l'utilisateur après la phase de prétraitement. L'utilisateur choisit un nombre de termes sémantiquement proches que le système doit ajouter à la requête à partir d'un dictionnaire. Cette ressource linguistique est représentée par une base de données construite à partir du dictionnaire français « Le Grand Robert ».

Gaume a montré dans [Gaume *et al.*, 2004, Gaume, 2004] que les graphes d'origine linguistique et notamment ceux qui sont construits à partir de dictionnaires sont de type

(RPMH) « Réseau de Petits Mondes Hiérarchiques » [4] (appelé en anglais *small-words networks*). Elayeb, dans [Elayeb, 2009], a prouvé expérimentalement que le dictionnaire « Le Grand Robert » est bel et bien un RPMH.

4.2.4 Module de désambiguïsation et d'expansion manuelle par navigation cartographique dans le dictionnaire

Notre système SPEEDSER facilite l'interaction avec l'utilisateur en lui offrant la visualisation des données dictionnairiques sous forme de graphe [5]. Ce dernier est construit en récupérant la liste des synonymes relatifs aux termes de la requête initiale et en considérant l'ensemble de ces mots (termes et synonymes) comme nœuds du graphe.

Ainsi, un arc orienté est présent entre deux mots (autrement deux nœuds) si l'un est présent dans la définition de l'autre. Ce mode de visualisation de données offre ainsi un cadre de navigation cartographique dans le Réseau de Petits Mondes Hiérarchiques (RPMH) du dictionnaire « Le Grand Robert » (figure 4.4).

Cette interface qui a été développée avec l'outil de visualisation des données Prefuse [Heer *et al.*, 2005] est zoomable. Ce mode de navigation offre une meilleure interaction Homme Machine face à des graphes de grandes tailles qui sont issus du dictionnaire. Un aperçu de la vue globale permet également d'afficher la concentration des relations sémantiques entre les mots et d'avoir un aperçu sur les composantes connexes [6] du graphe du dictionnaire (voir menu à droite en bas dans la figure 4.4). Autrement, la présence des composantes connexes indique une présence implicite de corrélation de sens.

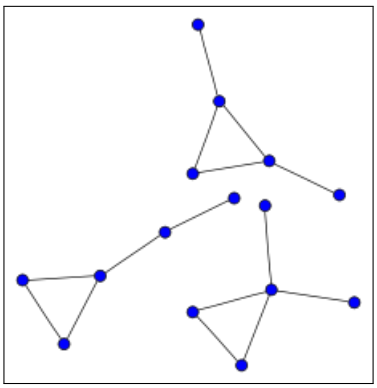


Figure 4.5 – Exemple de graphe non connexe, avec trois composantes connexes

4. Les RPMH sont des graphes peu denses, c'est à dire qu'ils ont relativement peu d'arcs au regard du nombre de leurs sommets. Ils se caractérisent principalement, selon Gaume [Gaume, 2004], par deux propriétés : (1) la moyenne des plus courts chemins entre les sommets L est petite ; (2) le taux de *clustering* ou d'agrégation C est grand.

5. Cette composante a été initiée dans notre premier système baptisé SPORSER [Ben Khiroun *et al.*, 2011, Elayeb *et al.*, 2011]

6. Un graphe non orienté $G = (S, A)$ est dit connexe si quels que soient les sommets u et v de S, il existe une chaîne de u vers v. Un sous-graphe connexe maximal d'un graphe non orienté quelconque est une composante connexe de ce graphe (voir exemple dans la figure 4.5).

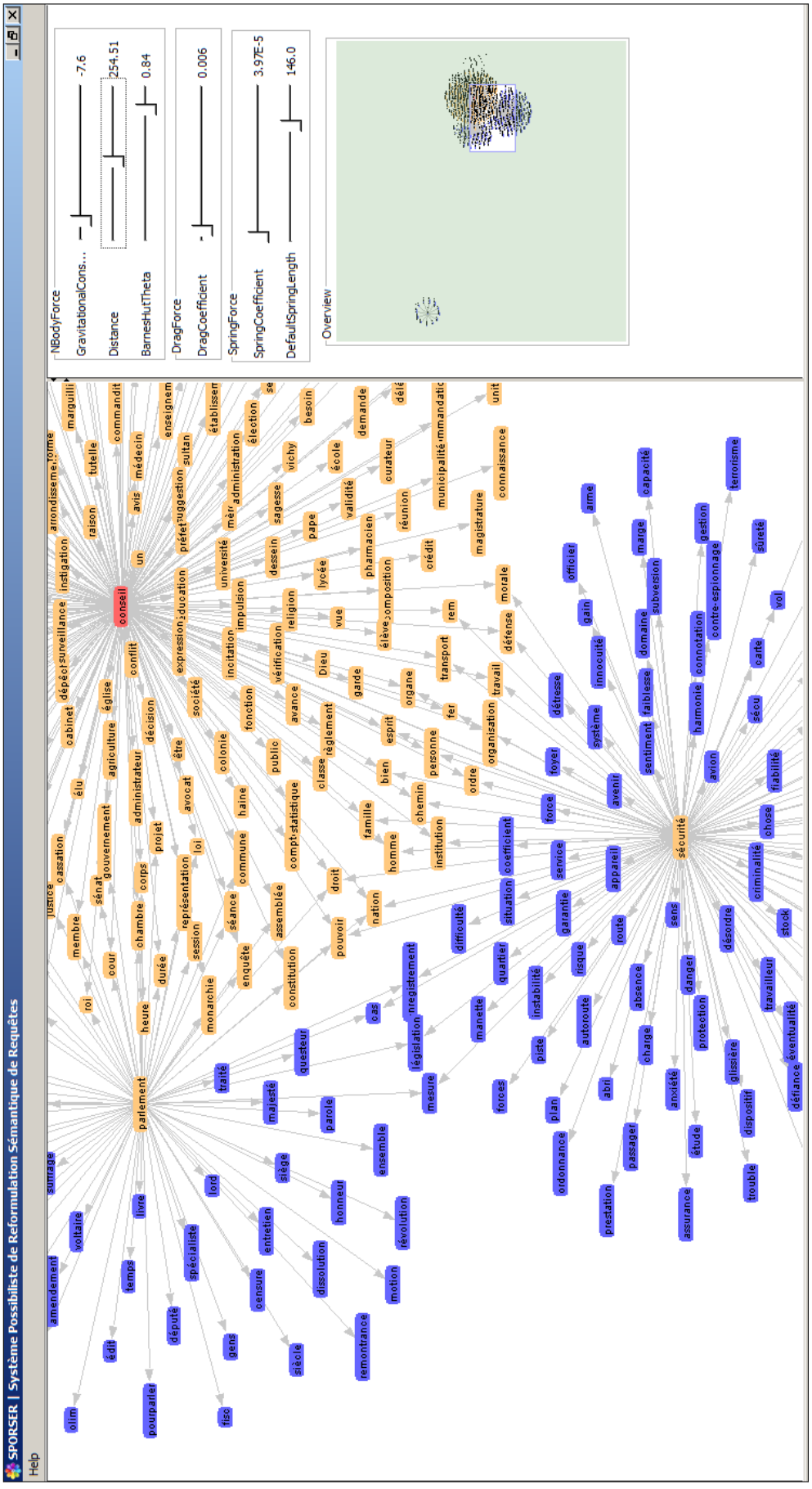


Figure 4.4 – Outil SPORSER : Navigation cartographique dans le graphe des synonymes

4.2.5 Désambiguïsation du contexte par concordance

Les concordanciers sont des logiciels qui construisent des concordances, c'est-à-dire, dans le plus simple des cas, une liste de contextes d'occurrence pour un terme de requête dans un corpus de texte.

Lorsque l'utilisateur soumet une requête, le système SPEEDSER fouille dans sa base de données et affiche toutes les occurrences trouvées dans leurs contextes. La forme la plus commune des concordances est l'indexation KWIC (*Key Word In Context*) où le terme de requête est centré dans une fenêtre de taille fixe [Manning et Schütze, 1999].

Les concordances peuvent être utilisées pour plusieurs finalités à savoir : comparer les divers emplois / sens d'un même terme, observer la fréquence des mots, identifier des collocation, observer des propriétés distributionnelles de certains mots, etc. Un concordancier se définit selon [Pincemin *et al.*, 2006, Pincemin, 2006] par trois paramètres :

1. l'expression d'un pivot (généralement un mot ou un groupe de mots) ;
2. la taille du contexte (il s'agit de la délimitation du contexte donné pour chaque occurrence relevée du pivot) ;
3. l'ordre de présentation des contextes sélectionnés (typiquement l'ordre de présence dans le corpus ou le tri alphabétique sur le mot qui précède le pivot (tri gauche) ou sur celui qui le suit (tri droit)).

Afin d'assister l'utilisateur dans la phase de désambiguïsation sémantique, nous avons mis en place un sous-système, baptisé TransKWIC (*Translation Key Word In Context*) et présenté dans la figure 4.6.

Sur l'interface présentée dans la figure 4.6, l'utilisateur sélectionne l'ensemble des documents à analyser depuis la zone 1. Il saisit par la suite l'ensemble des termes d'analyse (appelés aussi *pivots*) dans la zone 2. En cliquant sur le bouton d'analyse dans cette même zone, l'ensemble des concordances est affiché sous format KWIC (*Key Word In Context*). Ce format permet d'afficher les occurrences des mots recherchés de façon centrée et étiquetée par des couleurs différentes (zone 4).

Un affichage intégral des concordances peut être choisi comme présenté dans la figure 4.7. Dans ce cas de figure, chaque phrase contenant au moins un mot de la requête constitue un contexte de ce mot.

Le système TransKWIC assiste également l'utilisateur en proposant un ensemble de traductions possibles extraites à partir d'un dictionnaire bilingue organisé sous format CSV (zone 3).

Il est également possible de lancer l'outil SPORSER depuis TransKWIC (bouton 5) pour des fins de désambiguïsation en exploitant la fonctionnalité de navigation dans le graphe de dictionnaire.

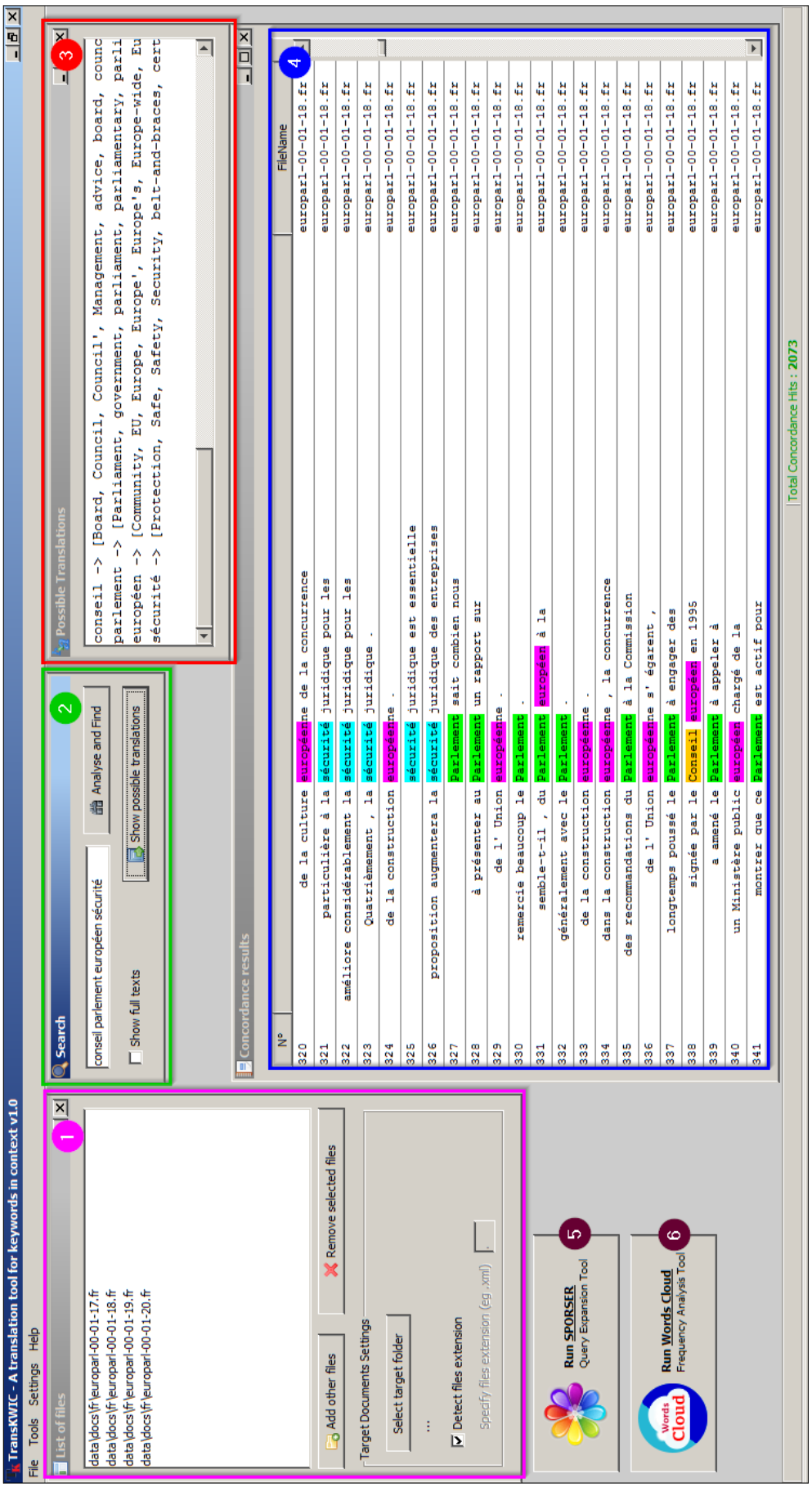


Figure 4.6 – Outil TransKWIC : Recherche de concordances et des traductions possibles

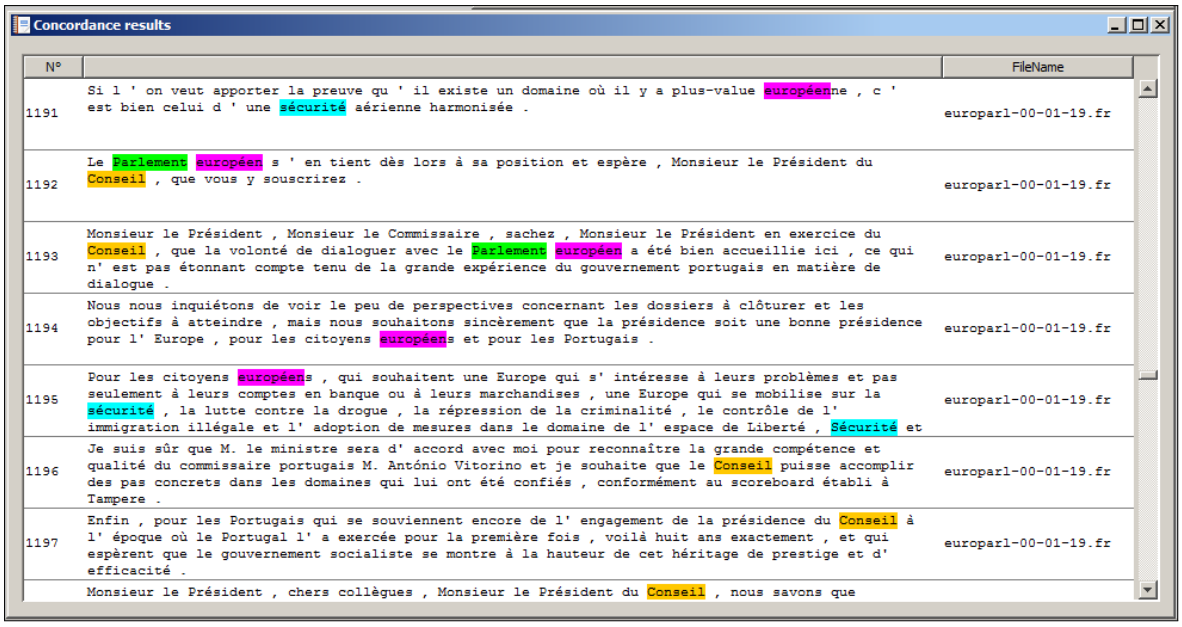


Figure 4.7 – Outil TransKWIC : Affichage intégral des concordances de la requête « *conseil parlement européen sécurité* »

De plus, l'outil TransKWIC propose une vue statistique de distribution des mots dans les documents à analyser (figure 4.8).

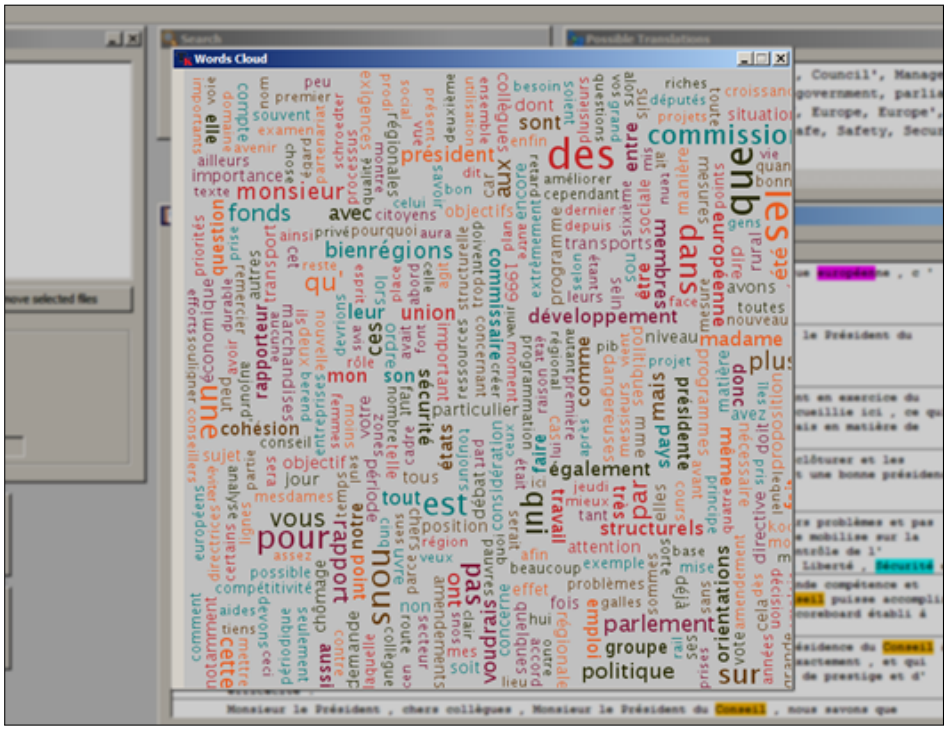


Figure 4.8 – Outil TransKWIC : Analyse de corpus et affichage de nuage de mots

En effet, il est possible d'afficher un nuage de tous les mots existants en cliquant sur le bouton 6 . Dans ce type de visualisation, la taille du mot varie en fonction de sa fréquence d'apparition. Cette fonctionnalité donne une vue globale sur le contexte des documents analysés en exposant en relief les mots récurrents.

4.3 Comparaison avec d'autres systèmes de l'état de l'art

Nous présentons dans le tableau 4.1 (page 96) une étude comparative du système SPEEDSER avec d'autres systèmes de l'état de l'art. L'objectif de l'utilisation de tels outils est d'assister l'utilisateur dans la recherche d'information monolingue et/ou translinguistique. Dans l'ensemble des outils présentés dans le tableau comparatif, l'utilisateur est considéré comme composante principale dans le processus de recherche. En effet, l'intervention manuelle de l'utilisateur peut contribuer à l'amélioration des résultats retournés par le système utilisé [Carpineto et Romano, 2012, Baeza-Yates et Ribeiro-Neto, 1999].

A notre connaissance et en se référant au tableau comparatif, l'outil *Sketch Engine* se distingue en proposant des fonctionnalités riches pour l'analyse des textes tels que l'extraction des mots clés, l'extraction des collocations, ainsi que l'intégration des outils de visualisation des distributions des mots (désignés par *word sketch*). L'outil *Sketch Engine* supporte également des corpus documentaires de l'état de l'art qui sont indexés par défaut. Cependant, son utilisation est commerciale et il est plus adapté pour les lexicographes.

Un des apports de notre système SPEEDSER, par rapport aux autres systèmes, consiste à modéliser d'une nouvelle manière la pertinence en se basant sur la théorie des possibilités. En fait, nous avons défini la pertinence possible d'un sens/traduction vis-à-vis des termes de la requête et sa pertinence nécessaire. La pertinence possible vise à éliminer les sens/traductions non pertinents, la pertinence nécessaire vise à renforcer la pertinence des sens/traductions non éliminés par la possibilité.

Conclusion

La spécification et la conception du système SPEEDSER présenté dans ce chapitre répondent bien à notre problématique de départ ; à savoir la combinaison des méthodes de désambiguïsation et d'expansion de requêtes via des procédures automatisées et semi-automatisées. Le système SPEEDSER offre ainsi des interfaces utiles à l'utilisateur pour lui faciliter la RI dans un cadre monolingue ou multilingue. En adoptant une approche top/down (passant du général au spécifique), nous présentons, dans les chapitres qui suivent, l'aspect validation et expérimentation des modules SPEEDSER. Nous commençons par l'évaluation du modèle possibiliste dans le cadre de désambiguïsation sémantique monolingue des textes dans le prochain chapitre.

Outil	Langues *	Ressources	Adaptation avec d'autres langues	Outil de visualisation	Expansion de requêtes	Concordancier	Traduction automatique
Mulinex [Capstick *et al.*, 2000]	fr, en, de	6 dictionnaires bilingues (avec entre 100.000 et 200.000 entrées)	-	-	✓	-	✓
Keizai [Ogden *et al.*, 1999]	en/ja, en/ko	Dictionnaire bilingue	-	-	-	-	-
WORDS [López-Ostenero *et al.*, 2002]	es/en	Systran	-	-	✓	-	✓
UCLIR [Abdelali *et al.*, 2003]	requêtes multilingues→en	Dictionnaires bilingues et traduction automatique	✓	-	-	-	-
UTACLIR [Hedlund *et al.*, 2004]	en→fr, de, es, it, nl, fi	Dictionnaires bilingues et multilingues	✓	-	-	-	✓
MultiLexExplorer [Luca *et al.*, 2006]	RI multilingue (de, fr, en, es, it)	EuroWordNet	-	✓	✓	-	-
Patent CLIR [Bian et Teng, 2005]	ja/en	Systèmes de traduction en ligne	-	-	-	-	✓
MIRACLE [Oard *et al.*, 2008]	en/es	Simple liste bilingue de termes	✓	-	✓	-	✓
KADRI CLIR Tool [Kadri et Nie, 2008]	en/ar CLIR - ar RI monolingue	Dictionnaires bilingues et des corpus parallèles extraits à partir de pages Web	-	-	✓	-	-
Multi Searcher [Ahmed et Nürnberger, 2010]	ar/en	Dictionnaire et corpus parallèle	-	-	-	-	✓
Sketch Engine [Kilgarriff *et al.*, 2014]	plus de 15 langues (ar, fr, en, de, it, ja, ko, es, etc.)	Corpus monolingues et bilingues	✓	✓	✓	✓	✓
SARIPOD [Elayeb, 2009]	fr monolingue	Dictionnaire monolingue	✓	✓	✓	-	-
SPORT [Elayeb *et al.*, 2018]	fr/en	Dictionnaire bilingue et corpus parallèle (Europarl)	✓	✓	✓	-	✓
SPEEDSER (TransK-WIC+SPORSER)	fr/en	Liste bilingue de termes, dictionnaire monolingue et corpus parallèle (Europarl)	✓	✓	✓	✓	-

Table 4.1 – Tableau comparatif du système SPEEDSER avec d'autres outils de RI monolingue et translinguistique

* Langues : *ar* (arabe), *en* (anglais), *fr* (français), *de* (allemand), *it* (italien), *es* (espagnol), *ja* (japonnais), *ko* (coréen), *nl* (néerlandais), *fi* (finnois)

Chapitre

5 Application et Expérimentations sur la Désambiguïsation Sémantique

Sommaire

" It is beyond a doubt that all our knowledge begins with experience. "

— Immanuel Kant

Introduction

Les exigences de la traduction automatique (ou assistée par ordinateur) et de la nature des SRI basés sur des requêtes composées de mots-clés ont encouragé le développement des outils pour la compréhension du langage naturel. Une des caractéristiques principales de la langue est qu'un mot, une expression ou une phrase peut avoir plusieurs sens différents.

Certains auteurs distinguent plusieurs types d'ambiguïté comme la polysémie et l'homonymie, mais nous considérons généralement par « ambigu » tout mot ayant la même orthographe et des sens différents, quel que soit le degré de proximité de ses sens [Nguyen et Ock, 2013, Vidhu Bhala et Abirami, 2014]. L'ambiguïté des mots introduit le phénomène de bruit dans les résultats retournés ce qui peut biaiser la pertinence de recherche de tout système basé sur la langue naturelle. Par conséquent, il est nécessaire d'identifier, dans une étape préliminaire, le sens exact des mots polysémiques à l'aide des techniques de désambiguïsation sémantique.

La tâche de WSD est définie par la capacité d'identifier la signification des mots dans le contexte d'une manière automatique [Navigli, 2009]. Cette tâche est importante dans de nombreux domaines tels que la reconnaissance optique de caractères (OCR), la lexicographie, la reconnaissance de la parole, la compréhension du langage naturel, la restauration d'accent [1], l'analyse de contenu, la classification de contenu, la recherche d'information et la traduction assistée par ordinateur [Ide et Wilks, 2007, Yarowsky, 2000].

Le problème de la désambiguïsation sémantique a été traité dans de nombreux travaux de recherche de différentes manières. Cependant, il a toujours été considéré comme une tâche difficile dans le domaine du traitement automatique du langage naturel (TALN). En effet, la tâche de désambiguïsation sémantique nécessite d'énormes ressources lexicales telles que les corpus annotés, les dictionnaires, les réseaux sémantiques et / ou les ontologies [Navigli, 2009]. Néanmoins, il n'existe pas encore de solutions efficaces pour répondre au problème d'ambiguïté dans les tâches de RI ou de traduction automatique. L'idée principale sur laquelle se fondaient de nombreuses recherches dans ce domaine est que les relations fines entre un mot et son contexte seront maximisées par le sens le plus probable de cette occurrence [Navigli, 2009, Zhou et Han, 2005].

Dans ce chapitre, nous proposons d'utiliser les réseaux possibilistes d'une part et des graphes sémantiques d'autre part comme moyen de représenter le sens pour la désambiguïsation automatique. En fait, de nombreux types d'informations peuvent être représentés grâce aux graphes dans lesquels les arcs modélisent les relations de synonymie, d'antonymie et d'hyperonyme. Par conséquent, l'étude des relations existantes

1. ou la restauration des voyellations pour la langue arabe par exemple.

entre les entrées d'un dictionnaire peut être réduite à l'étude d'un graphe visant à exploiter des réseaux de mots.

Nous présentons dans la première section de ce chapitre la structure d'un dictionnaire sémantique de contexte. Ce dernier servira dans l'approche possibiliste de désambiguïsation sémantique que nous détaillons dans la deuxième section. La dernière section se focalise sur l'étude expérimentale de notre approche proposée en utilisant le corpus ROMANSEVAL comme standard de test.

5.1 Proposition d'un dictionnaire sémantique de contextes

Les dictionnaires numériques compréhensibles par machine (dits en anglais *MRD : Machine Readable Dictionary*) présentent des réseaux sémantiques riches reliant l'ensemble des mots avec leurs définitions. Ces relations constituent des informations structurées et exploitables dans la phase de désambiguïsation sémantique des textes.

Cependant, l'évolution de la langue via de nouveaux termes (noms propres, termes techniques, nouvelles technologies, etc.) constitue un problème lors de l'utilisation des *MRDs*. En effet, les dictionnaires sont construits manuellement et ils sont liés à une période donnée ; ce qui rend la couverture des mots un problème récurrent. Ce problème d'inconsistance est connu depuis longtemps par les lexicographes [Ide et Wilks, 2007, Atkins et Rundell, 2008].

La grande majorité des travaux en WSD est basée sur des dictionnaires traditionnels ou d'autres ressources lexicales comme WordNet, ce qui n'est pas très différent en termes d'organisation des sens [Barque et Chaumartin, 2008]. Le problème est que les dictionnaires traditionnels ont été conçus pour l'usage humain plutôt que le traitement automatique. Ils manquent d'informations précises utiles pour la désambiguïsation. En conséquence, l'une des principales difficultés dans WSD est l'insuffisance des dictionnaires traditionnels. L'autre difficulté réside dans l'absence de corpus étiquetés sémantiquement et qui sont utiles pour la phase d'apprentissage [Audibert, 2004]. Même si ces corpus sont disponibles, l'existence du bruit et la dispersion des connaissances nécessaires à la désambiguïsation rendent cette tâche difficile.

Pour ces raisons, il est nécessaire de définir de nouveaux types de structures qui peuvent être formés et ensuite utilisés pour représenter des connaissances utiles pour la WSD. Un graphe contextuel sémantique est appris et mis à jour pendant le processus de désambiguïsation. Le mécanisme d'apprentissage devrait pouvoir acquérir de nombreux types de liens sémantiques entre un mot polysémique et les définitions d'un dictionnaire traditionnel.

Afin de résoudre les limites de l'utilisation des dictionnaires classiques dans la tâche de désambiguïsation monolingue, nous proposons une approche qui combine ce type de données structurées avec des connaissances extraites à partir des corpus documentaire. Ainsi, les définitions contenues dans les dictionnaires sont enrichis par des connaissances contextuelles liant les mots avec leurs contextes.

Le « Dictionnaire Sémantique de Contextes » ($\mathcal{DSC}$), que nous proposons dans cette section, est basé sur cette idée en assurant l'apprentissage automatique dans une plateforme sémantique de WSD. Ainsi, nous combinons les connaissances extraites des dictionnaires traditionnels avec les dépendances contextuelles tirées d'un corpus.

Par conséquent, nous modélisons le $\mathcal{DSC}$ sous forme d'un graphe pour mettre en place l'ensemble des relations sémantiques entre les mots. Pour construire et représenter de façon automatique le graphe $\mathcal{G} = (\mathcal{S}; \mathcal{A})$ associé à un mot, il faut définir l'ensemble des sommets $\mathcal{S}$ (les mots de textes et leurs définitions) ainsi que les arêtes $\mathcal{A}$ (les relations entre les mots et leurs définitions).

5.1.1 Ensemble des sommets

Les sommets du graphe sont constitués des entrées du dictionnaire sémantique de contextes (définition sémantique des mots de contextes), les entrées du dictionnaire et le contexte du mot polysémie. Nous notons :

- $Polysémie(ph)$: Ensemble des mots polysémiques dans la phrase ph : $p_1, p_2, \ldots, p_k$
- $Significatif(ph)$: Ensemble des mots significatifs et non polysémique constituants la phrase ph : $m_1, m_2, \ldots, m_i$.
- $Contexte(p, ph) = \{Significatif(ph) \cup Polysémie(ph) \backslash \{p\}\}$; avec p est un mot significatif $\in ph$.
- $Contexte(ph) = \{Significatif(ph) \cup Polysémie(ph)\}$.
- $Définition(m_i)$: Ensemble des termes de la définition du mot significatif m_i dans le dictionnaire si le mot n'est pas polysémique. Si le mot m_i est polysémique, cet ensemble regroupe les différents sens p_i^1, p_i^2 , ..., p_i^a

5.1.2 Ensemble des arêtes

Dans la littérature, il existe plusieurs types de réseaux lexicaux, suivant la nature de la relation sémantique qui définit les arcs du graphe. Les trois principaux types de relations utilisées dans notre $\mathcal{DSC}$ sont les suivants :

- *Relations syntagmatiques*, ou de co-occurrence ; nous construisons une arête entre deux mots s'ils co-occurrent dans le même contexte dans le corpus [Yuret et

Yatbaz, 2010]. Cette relation est formalisée comme suit :
$\forall m_i, m_j \in Contexte(ph)$ si $i \neq j$ alors $< m_j, m_i > \in \mathcal{A}$

– *Relations paradigmatiques*, notamment de synonymie ; nous construisons un graphe dans lequel deux sommets sont reliés par une arête si les mots correspondants entretiennent une relation synonymique [Ploux et Victorri, 1998]. En effet, si deux mots partagent des termes en communs dans leurs définitions du dictionnaire, alors :
$\forall m_i, m_j$ si $\{Définition(m_i) \bigcap Définition(m_j)\} \neq \emptyset$ alors $< m_j, m_i > \in \mathcal{A}$

– *Relations de proximité sémantique* ; il s'agit de relations moins spécifiques qui peuvent prendre en compte à la fois l'axe paradigmatique et l'axe syntagmatique. Ainsi, nous définissons une arête entre deux mots m_i et m_j si m_j apparait dans la définition de m_i [Véronis et Ide, 1990]. Ce type de relation peut être formalisé comme suit :
$\forall m_i \in Polysémie(ph)$, $\forall m_j \in Définition(m_i)$ alors $< m_j, m_i > \in \mathcal{A}$

Afin de structurer les contextes des mots polysémiques d'un corpus de test [2], nous avons adopté le format XML qui a une syntaxe générique et extensible (figure 5.1, figure 5.2 et tableau 5.1).

Balise	Attribut	Description
<sdc>	-	balise racine du document XML
<word>	id	mot de test
<context>	noccur	nombre d'occurrence dans le standard ROMANSEVAL
<sentence>	-	phrase polysémique (ambigüe)
<sens>	-	identifiant du sens sélectionné

Table 5.1 – Signification des balises et des attributs XML pour la représentation du $\mathcal{DSC}$

5.2 Proposition d'une approche possibiliste de désambiguïsation sémantique

Les approches de WSD ont besoin des modèles d'apprentissage, qui calculent les similarités (ou la pertinence) entre les sens et les contextes. Les modèles existants pour les WSD sont basés sur des données incertaines et imprécises et utilisent des modèles probabilistes d'apprentissage et d'appariement [Nguyen et Ock, 2013, Yuret et Yatbaz, 2010]. En revanche, la théorie des possibilités est naturellement conçue pour ce type d'application, car elle permet d'exprimer l'ignorance et de tenir compte de l'imprécision et de l'incertitude en même temps. Nous citons à titre d'exemple les travaux

2. Nous avons utilisé le standard ROMANSEVAL (voir détails Annexe A) pour la validation de la tâche de désambiguïsation sémantique.

```
<?xml version="1.0" encoding="UTF-8"?>
<sdc>
  <word id="concentration">
    <context noccur="216">
      <sentence>commission déjà annoncer février 1991 peut-elle temps
         prendre réflexion étude problème concentration médias Europe
      </sentence>
      <sens>1</sens>
    </context>
    <context noccur="211">
      <sentence>commission pouvoir intervenir égard opération
         concentration demande état membre condition définir articler
         paragraphe dudit règlement</sentence>
      <sens>1</sens>
    </context>
    <context noccur="255">
      <sentence>guerre civil opposer Nord Sud servir alibi déplacement
         masse population sud établissement camp concentration torture
         viol chose quotidien conversion forcé chrétien animiste
         apparaître organisation humanitaire obédience islamique utiliser
          menace assurer besoin élémentaire contraindre conversion
         refuser abjurer foi</sentence>
      <sens>3</sens>
    </context>
  ...
  </word>
  <word id="détention"> ... </word> ...
</sdc>
```

Figure 5.1 – Extrait de construction du $\mathcal{DSC}$ dans le format XML

de [Ayed *et al.*, 2014] dans lesquels les auteurs ont proposé une approche possibiliste pour la désambiguïsation morphologique des textes arabes en comparant des modèles possibilistes et probabilistes.

D'autre part, la désambiguïsation sémantique peut être considérée comme une tâche de classification dans laquelle une phase d'apprentissage et une deuxième phase de test se présentent dans le processus de désambiguïsation [Sugawara *et al.*, 2015].

En amont, durant la phase d'apprentissage, nous avons besoin d'étudier les dépendances entre les sens des mots et les différents contextes. Ceci peut être mis en place :

- en utilisant des corpus étiquetés : on parle ici d'un *apprentissage basé sur les jugements* que nous pouvons qualifier en tant qu'approche de désambiguïsation semi-automatisée ;
- en attribuant des poids aux différentes relations issues d'un dictionnaire traditionnel : on parle ici d'un *apprentissage basé sur dictionnaire* qui peut être considéré comme une approche automatisée (voir les détails dans la section 5.3.2). Dans ce cas de figure, nous devons trier l'ensemble de sens/contextes candidats selon un taux d'ambiguïté (voir section 5.2.2).

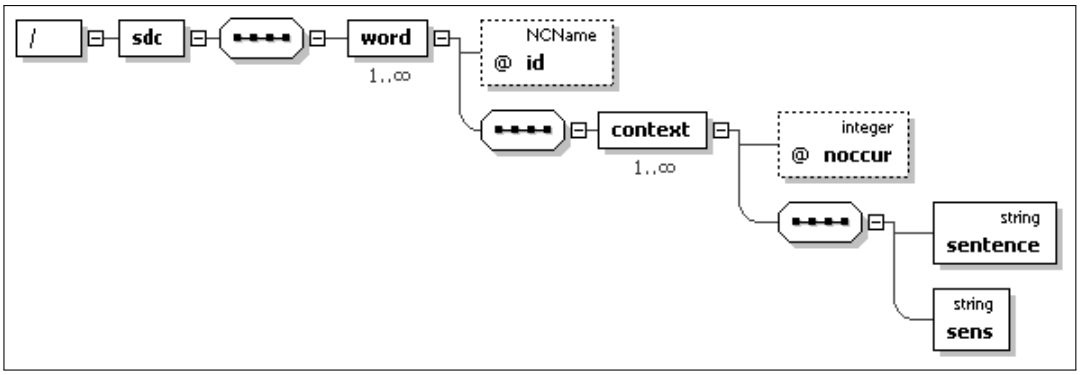


Figure 5.2 – Structure des fichiers XML du $\mathcal{DSC}$

En aval, durant la phase de test, nous procédons à une phase d'appariement entre le contexte dans lequel apparait un mot et les sens qui lui sont attribués dans le $\mathcal{DSC}$ afin de choisir le meilleur sens. Dans ce qui suit, nous présentons les détails des formules avec des exemples illustratifs pour le calcul des deux mesures de degré de pertinence possibiliste (DPP) et le taux d'ambiguïté.

5.2.1 Degré de pertinence possibiliste

Ayant une phrase polysémique (ou ambigüe) ph, nous notons $DPP(S_i|ph)$ le degré de pertinence possibiliste du sens S_i sachant ph. Nous supposons que la phrase ph est composée de T termes ($ph = \{t_1, t_2, \ldots, t_T\}$). Nous évaluons la pertinence d'un sens de mot S_i en ayant la phrase ph en appliquant le modèle d'appariement possibiliste de RI utilisé dans [Elayeb *et al.*, 2009, Ben Khiroun *et al.*, 2011].

Nous adaptons ainsi le modèle d'appariement entre requête/document au calcul de pertinence entre un sens et une phrase polysémique en utilisant une double mesure de pertinence : (i) la *pertinence possible* permet de rejeter les sens non pertinents à une phrase donnée et (ii) la *pertinence nécessaire* permet de se focaliser sur les sens à restituer ainsi que de renforcer la nécessité de faire figurer parmi les premiers de la liste des résultats en réponse à une phrase polysémique [Ben Khiroun *et al.*, 2012, Elayeb *et al.*, 2015b].

Le réseau possibiliste relie l'ensemble des sens des mots (S_i) avec les mots d'une phrase polysémique ($ph_i = \{t_1, t_2, \ldots, t_T\}$) comme le montre la figure 5.3.

La pertinence de chaque sens (S_j), sachant la phrase polysémique (ph_i), est calculée comme suit [Elayeb *et al.*, 2015b] :

La possibilité $\Pi(Sj|ph)$ est proportionnelle à :

$$\Pi'(S_j|Q) = \Pi(t_1|S_j) \times \ldots \times \Pi(t_T|S_j) = nft_{1j} \times \ldots \times nft_{Tj} \tag{5.1}$$

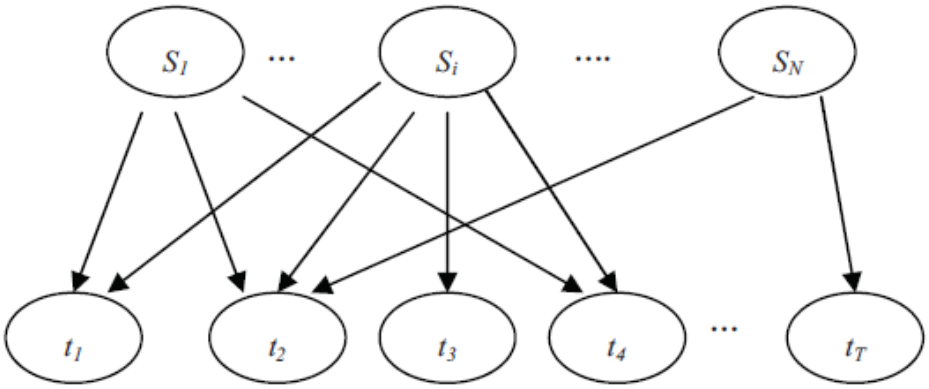


Figure 5.3 – Réseau possibiliste de l'approche de désambiguïsation

- avec : $nft_{ij} = tf_{ij}/max(tf_{kj})$ représente la fréquence normalisée du terme t_i dans le sens S_j ;
- et $tf_{ij} = \frac{nombre\ d'occurrence\ du\ terme\ t_i dans\ S_j}{nombre\ de\ termes\ dans\ S_j}$.

La certitude (ou nécessité) consiste à restituer un sens pertinent S_j pour une phrase. Notée par $N(S_j|ph)$, cette mesure est calculée comme suit :

$$N(S_j|ph) = 1 - \Pi(\neg S_j|ph) \tag{5.2}$$

Où :

$$\Pi(\neg S_j|ph) = (\Pi(ph|\neg S_j) \times \Pi(\neg S_j))/\Pi(ph) \tag{5.3}$$

De même $\Pi(\neg S_j|ph)$ est proportionnelle à :

$$\Pi'(\neg S_j|ph) = \Pi(t_1|\neg S_j) \times \ldots \times \Pi(t_T|\neg S_j) \tag{5.4}$$

Ce numérateur peut être exprimé par :

$$\Pi'(\neg S_j|ph) = (1 - \phi S_{1j}) \times \ldots \times (1 - \phi S_{Tj}) \tag{5.5}$$

Où :

$$\phi S_{ij} = Log_{10}(\frac{nCS}{nS_i}) \times nft_{ij} \tag{5.6}$$

- avec : nCS = nombre des sens du mot ;
- et nS_i = nombre des sens du mot contenant le terme t_j.

Nous définissons le *Degré de Pertinence Possibiliste (DPP)* de chaque sens S_j étant donné une phrase ph par :

$$DPP(S_j|ph) = \Pi(S_j|ph) + N(S_j|ph) \tag{5.7}$$

Les sens préférés sont ceux qui ont une valeur de $DPP(S_j|ph)$ élevée.

5.2.2 Calcul des taux d'ambiguïté des phrases polysémiques

Une phrase est considérée avoir un taux d'ambiguïté élevé si les sens correspondant aux mots ambigus dans la phrase ont des significations proches et/ou ne correspondent pas au contexte proposé par la phrase.

Nous calculons le taux d'ambiguïté d'une phrase polysémique en utilisant les deux mesures de possibilité et de nécessité comme suit : (i) indexer tous les sens possibles du mot ambigus ; (ii) utiliser l'index de chaque sens comme une requête ; (iii) évaluer la pertinence du sens ayant cette requête ; et (iv) juger le taux d'ambiguïté d'une phrase en tant que élevé si la phrase est pertinente pour plusieurs sens ou si elle n'est pertinente pour aucun sens [Elayeb *et al.*, 2015b].

Par conséquent, le taux d'ambiguïté est inversement proportionnel à l'écart type :

$$Taux\_ambiguïté(ph) = 1 - \sigma(ph) \tag{5.8}$$

Où $\sigma(ph)$ représente l'écart type des degrés de pertinence possibiliste correspondant à chaque mot polysémique contenu dans la phrase ph comme suit :

$$\sigma(ph) = \sqrt{\left(\frac{1}{N}\right) \times \sum_j (DPP(S_j|ph) - S)^2} \tag{5.9}$$

Avec : S : moyenne des $DPP(S_j|ph)$ et N : nombre de sens possibles dans le dictionnaire.

5.2.3 Exemple illustratif

Considérons l'exemple de phrase suivante [Elayeb *et al.*, 2015b] :

« *Le noyau de l'implantation de l'avocat est le fruit des efforts juridiques.* »

Dans le but de simplifier les calculs dans cet exemple, nous considérons le mot « *avocat* » comme seul terme polysémique dans le contexte de cette phrase. Nous supposons que ce mot a deux sens *S1* et *S2* désignés par : « *avocat_1* » et « *avocat_2* » comme suit :

avocat_1 — *Praticien et professionnel du **droit** dont la fonction traditionnelle est de **conseiller** ses clients sur des questions juridiques, quelles soient relatives à leur vie **juridique** quotidienne ou plus spécialisées, [...]*

avocat_2 — ***Fruit** comestible de l'avocatier, à pulpe jaune, contenant un gros **noyau**, fortement **conseillé** dans plusieurs cocktails de fruits et des confitures [...]*

Nous commençons par une phase de pré-traitement incluant l'élimination des mots vides et la lemmatisation. Puis, nous supposons que le sens « *avocat_1* » est indexé par trois termes {*conseiller, juridique, droit*} et le sens « *avocat_2* » est indexé par {*conseiller, fruit, noyau*}.

Nous appliquons les mêmes pré-traitements sur la phrase initiale. On suppose le résultat d'indexation suivant de la phrase polysémique : *ph*={*avocat, juridique, fruit, noyau*}

La mesure Π est calculée comme suit :

$$\begin{aligned}\Pi(avocat\_1|ph) &= nf(avocat, avocat\_1) \times nf(juridique, avocat\_1) \\ &\times nf(fruit, avocat\_1) \times nf(noyau, avocat\_1) = 0 \times (1/3) \times 0 \times 0 = 0\end{aligned}$$

Avec $nf(avocat, avocat\_1)$ est la fréquence normalisée du terme « *avocat* » dans le premier sens.

$$\begin{aligned}\Pi(avocat\_2|ph) &= nf(avocat, avocat\_2) \times nf(juridique, avocat\_2) \\ &\times nf(fruit, avocat\_2) \times nf(noyau, avocat\_2) = 0 \times 0 \times (1/3) \times (1/3) = 0\end{aligned}$$

Nous aurons les mesures Π souvent égales à 0 sauf dans le cas où tous les termes de la phrase existent dans l'index du sens.

En contre partie, nous avons des valeurs non nulles de la mesure N comme suit :

$$\begin{aligned}N(avocat\_1|ph) &= 1 - [(1 - \phi(avocat\_1, avocat)) \times (1 - \phi(avocat\_1, juridique)) \\ &\times (1 - \phi(avocat\_1, fruit)) \times (1 - \phi(avocat\_1, noyau))]\end{aligned}$$

Nous avons : $nf(avocat, avocat\_1) = 0$; ainsi : $\phi(avocat\_1, avocat) = 0$;

De même, $\phi(avocat\_1, juridique) = Log_{10}(2/1) \times 1/3 = 0,1$; $\phi(avocat\_1, fruit) = Log_{10}(2/1) \times 0 = 0$ et $\phi(avocat\_1, noyau) = 0$

Ainsi : $N(avocat\_1|ph) = 1-[(1-0)\times(1-0,1)\times(1-0)\times(1-0)] = 1-[1\times 0,9\times 1\times 1] = 0,1$ et la mesure de degré de pertinence possibiliste est égale à : $DPP(avocat\_1|ph) = 0,1$.

De même, la mesure de nécessité du deuxième sens est calculée comme suit :

$$\begin{aligned}N(avocat\_2|ph) &= 1 - [(1 - \phi(avocat\_2, avocat)) \times (1 - \phi(avocat\_2, juridique)) \\ &\times (1 - \phi(avocat\_2, fruit)) \times (1 - \phi(avocat\_2, noyau))]\end{aligned}$$

Avec : $\phi(avocat\_2, avocat) = 0$; $\phi(avocat\_2, juridique) = 0$; $\phi(avocat\_2, fruit) = Log_{10}(2/1) \times 1/3 = 0,1$ et $\phi(avocat\_2, noyau) = 0,1$

Enfin : $N(avocat\_2|ph) = 1-[(1-0)\times(1-0)\times(1-0,1)\times(1-0,1)] = 1-[1\times0,9\times0,9\times1] = 0,19$

La mesure de pertinence possibiliste du deuxième sens est égale à $DPP(avocat\_2|ph) = 0,19 > DPP(avocat\_1|ph)$.

Nous concluons, après calcul possibiliste, que la phrase *ph* est plus pertinente pour le sens « *avocat_2* » que pour le sens « *avocat_1* ». Ceci s'explique par le fait que la phrase contient deux termes du sens « *avocat_2* » (*fruit*, *noyau*) et un seul terme du sens « *avocat_1* » (*juridique*).

Afin de mesurer le taux d'ambiguïté de la phrase *ph*, nous calculons la moyenne des *DPP* associés aux sens : $S = (0,1 + 0,19)/2 = 0,145$.

L'écart type : $\sigma(ph) = (\sqrt{1/2 \times ((0,1–0,145)^2 + (0,19–0,145)^2)} = 0,045$.

Ainsi, $Taux\_ambiguïté(ph) = 1 - \sigma(ph) = 0,955$.

Par conséquent, la phrase de cet exemple est considérée très ambigüe vu que le taux d'ambiguïté calculé est proche de 1. Ceci est dû aux mesures *DPP* calculées (respectivement 0,1 et 0,19) qui sont très proches.

5.3 Expérimentations et étude comparative

Pour notre approche possibiliste proposée, nous évaluons la pertinence d'un sens de mot étant donné une phrase polysémique. D'autre part, le problème de la désambiguïsation peut être modélisé dans une perspective dynamique. Le calcul dynamique du sens dans un espace sémantique consiste à spécifier des contraintes sur chaque point de cet espace. Il permet d'obtenir des relations sémantiques entre les mots. A partir de ces relations, nous calculons la distance sémantique entre un mot polysémique et ses définitions, mentionnées dans un dictionnaire traditionnel, en ayant des informations contextuelles [Ben Khiroun *et al.*, 2012, Elayeb *et al.*, 2015b].

Par conséquent, nous proposons, évaluons et comparons dans cette section deux approches pour le WSD automatique. Dans la première approche possibiliste, la pertinence d'un sens de mot (resp. document en RI), étant donnée une phrase polysémique (resp. requête en RI), est modélisée par deux mesures : (i) la pertinence possible permet de rejeter des sens non pertinents ; et (ii) la pertinence nécessaire permet de renforcer les sens éventuellement pertinents. Dans la seconde approche probabiliste, nous calculons la distance sémantique entre les sens du mot en utilisant une distance probabiliste

existante, comme proposé dans [Gaume *et al.*, 2004]. Dans cette étape, nous considérons la topologie complète d'un dictionnaire traditionnel vu comme un graphe sur ses entrées. Nous montrons comment l'utilisation du $\mathcal{DSC}$ améliore les résultats du processus WSD. Les différentes expérimentations sont réalisées par la collection de test ROMANSEVAL tout en comparant nos deux approches à certains systèmes de WSD existants.

5.3.1 La collection de test ROMANSEVAL

Nous utilisons dans nos expérimentations la collection de test ROMANSEVAL qui propose un cadre de validation de la tâche de désambiguïsation sémantique. Cette collection inclut (i) un ensemble de documents et (ii) une liste de phrases de test contenant des mots ambigus. L'ensemble des documents se compose de textes parallèles en 9 langues issus du *Journal Officiel de la Commission Européenne*. Ces textes représentent des questions posées par les membres du parlement européen autour d'un large éventail de sujets (politique, recherche, santé, environnement, économie, etc.) et les réponses correspondantes de la commission européenne. La taille totale du corpus est d'environ 10,2 millions de mots (environ 1,1 million de mots par langue), qui ont été recueillis et préparés dans les projets MULTEXT et MLCC [Segond, 2000].

Le corpus a été divisé en mots étiquetés par des catégories grammaticales pour distinguer les noms (N), les adjectifs (A), et les verbes (V). Ensuite, les 600 mots les plus fréquents (200 N, 200 A et 200 V) ont été extraits ainsi que leurs contextes d'apparition. Ces mots ont été annotés par six étudiants en linguistique en les faisant correspondre avec les sens du dictionnaire français *Le Petit Larousse*. Chaque occurrence de mot peut avoir une ou plusieurs étiquettes de sens. Après cette première étape, les 60 mots les plus polysémiques ont été conservés (20 N, 20 A et 20 V) et leurs occurrences ont été étiquetées dans 3624 contextes (voir plus de détail dans l'annexe A).

5.3.2 Comparaison des méthodes d'apprentissage du DSC

Pour chaque test, nous avons préparé un $\mathcal{DSC}$ (voir la représentation des $\mathcal{DSC}$ dans la section 5.1 page 99). Il est utilisé comme un sous-ensemble d'apprentissage des phrases à évaluer dans le corpus ROMANSEVAL.

Pour chaque phrase analysée ph contenant un mot polysémique w, nous relions les mots de ph avec le sens correct de w. Le « sens correct » peut être identifié à partir des annotations présentes dans le corpus (on parle d'*apprentissage à base des jugements*) ou en utilisant les connaissances contextuelles de co-occurrence issus du dictionnaire *Le*

Petit Larousse (on parle d'*apprentissage à base de dictionnaire*) [Ben Khiroun *et al.*, 2012].

Afin de réaliser la tâche d'apprentissage à base des jugements, nous avons construit les $\mathcal{DSC}$s sous le format XML en appliquant la méthode de validation croisée [Kohavi, 1995]. Dans cette méthode, 90% des phrases aléatoires de la collection ROMANSEVAL sont utilisées pour la formation du $\mathcal{DSC}$ et les 10% restantes sont utilisées pour les tests. Cette tâche est répétée 10 fois pour tous les 60 mots ambigus dans chacune des dix exécutions. Les phrases sont lemmatisées en utilisant l'outil TreeTagger [3] pour la langue française.

Le processus d'apprentissage à base de dictionnaire, décrit dans l'algorithme 1, est effectué en appliquant la méthode 80/20 comme suit : en ayant le résultat final des phrases annotées avec des sens et triées en ordre décroissant (respectivement croissant) par taux d'ambiguïté, nous construisons le $\mathcal{DSC}$ à partir des 80% phrases les plus (respectivement les moins) ambigües. Nous évaluons par la suite le reste des phrases ambigües (20%) en considérant le $\mathcal{DSC}$ construit comme ressource de désambiguïsation.

Algorithme 1 : Apprentissage à base de dictionnaire

Entrées : phrases ambigües
Sorties : phrases annotées avec des sens et triées selon taux d'ambiguïté
Variables : w_i :mot ; S_i,S_{max} :sens ;

```
1  début
2      pour chaque phrase ambigüe faire
3          pour chaque mot ambigu w_i faire
4              calculer DPP pour chaque sens S_i ∈ dictionnaire
5              associer au mot w_i le sens S_max ayant le plus grand DPP
6          fin
7      fin
8      pour tous les phrases ambigües annotées avec des sens faire
9          trier en ordre croissant/décroissant les phrases selon taux d'ambiguïté
10     fin
11 fin
```

Nous proposons ainsi trois façons de construction du $\mathcal{DSC}$:

- Apprentissage à base des jugements ;
- Apprentissage à base de dictionnaire par ambigüité décroissante ;
- Apprentissage à base de dictionnaire par ambigüité croissante.

Pour comparer ces méthodes d'apprentissage, nous utilisons le taux d'accord calculé comme suit [Segond, 2000] :

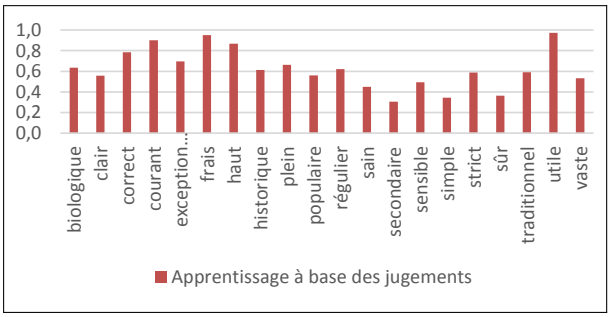
3. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

$$accord = \frac{|\left\{S_i \in \Delta,\ où\ S_i^{système} = S_i^{juges}\right\}|}{|\{S_i \in \Delta\}|} \tag{5.10}$$

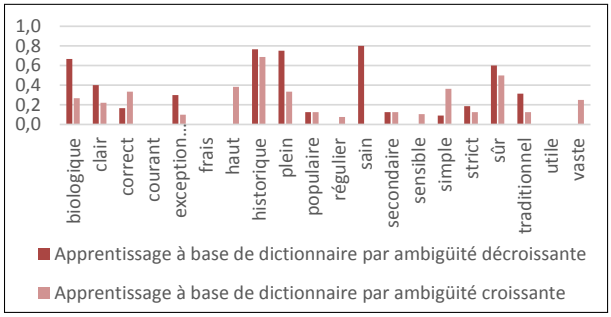
Avec : Δ : l'ensemble des sens (jugés par les annotateurs) correspondant aux phrases de test ; $S_i^{système}$: le sens sélectionné en calculant le *DPP* (calculé par le système) ; et S_i^{juges} : le sens attribué par les juges.

Résultats

Comme première interprétation des figures 5.4, 5.5 et 5.6, nous constatons que plus un mot est fréquent dans le corpus et a quelques sens, plus l'accord moyen est élevé. Ainsi, les verbes sont les plus ambigus dû au fait qu'ils sont moins fréquents dans le corpus. En contre partie, les noms (à l'exception de quelques uns) sont moins ambigus grâce à leurs fréquences dans les textes du corpus ROMANSEVAL [Ben Khiroun *et al.*, 2012].



(a) Apprentissage du $\mathcal{DSC}$ à base des jugements



(b) Apprentissage du $\mathcal{DSC}$ à base de dictionnaire

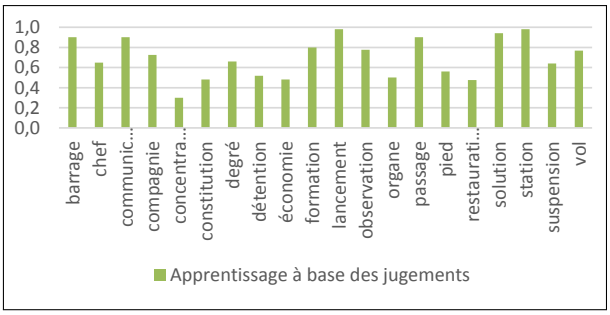
Figure 5.4 – Résultats d'accord moyen des méthodes d'apprentissage du $\mathcal{DSC}$ pour les adjectifs

La mesure d'accord moyen dépend également des caractéristiques du corpus utilisé. Par exemple, nous discutons le cas du terme « constitution » qui donne une valeur faible d'accord moyen par rapport aux autres noms. Ce terme possède plusieurs sens : (1) *constitution*, (2) *mise en place*, (3) *incorporation*, (4) *règle*, (5) *habitude* et (6) *code*.

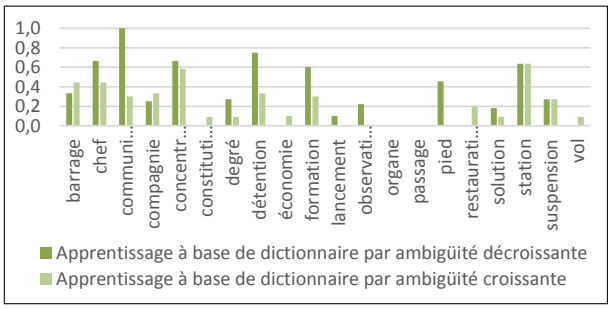
La nature des sujets traités dans les articles de ROMANSEVAL (discussions du parlement européen traitant principalement les domaines politiques et économiques) renforce cette ambigüité. Cette interprétation colle par exemple pour le mot « économie » ayant quatre sens différents : (1) *économie*, (2) *finances*, (3) *épargne* et (4) *élevage*.

Les différentes expérimentations (figures 5.4.b, 5.5.b et 5.6.b) confirment que l'apprentissage du $\mathcal{DSC}$ doit être fait en commençant par les phrases les plus ambigües dans la méthode d'apprentissage par dictionnaire.

La figure 5.7 présente les résultats des accords moyens pour les trois catégories grammaticales confondues. Nous remarquons que la méthode d'apprentissage à base des ju-

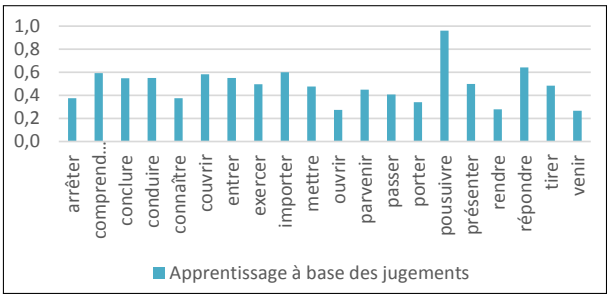


(a) Apprentissage du $\mathcal{DSC}$ à base des jugements

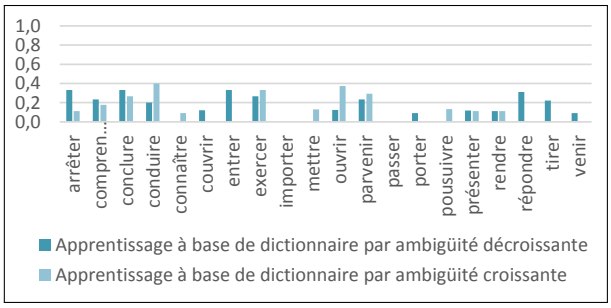


(b) Apprentissage du $\mathcal{DSC}$ à base de dictionnaire

Figure 5.5 – Résultats d'accord moyen des méthodes d'apprentissage du $\mathcal{DSC}$ pour les noms



(a) Apprentissage du $\mathcal{DSC}$ à base des jugements



(b) Apprentissage du $\mathcal{DSC}$ à base de dictionnaire

Figure 5.6 – Résultats d'accord moyen des méthodes d'apprentissage du $\mathcal{DSC}$ pour les verbes

gements donne des résultats meilleures vu qu'elle exploite des connaissances manuelles données par les évaluations des juges lors de l'étiquetage du standard ROMANSEVAL.

En contre partie, la méthode de construction du $\mathcal{DSC}$ en se basant sur le dictionnaire représente une méthode semi-automatisée qui peut s'avérer très utile dans le cas de non existence d'annotation sémantique du corpus utilisé. Dans ce cas de figure, il serait plus adéquat d'aborder les phrases les plus ambigües pour la phase d'apprentissage du $\mathcal{DSC}$.

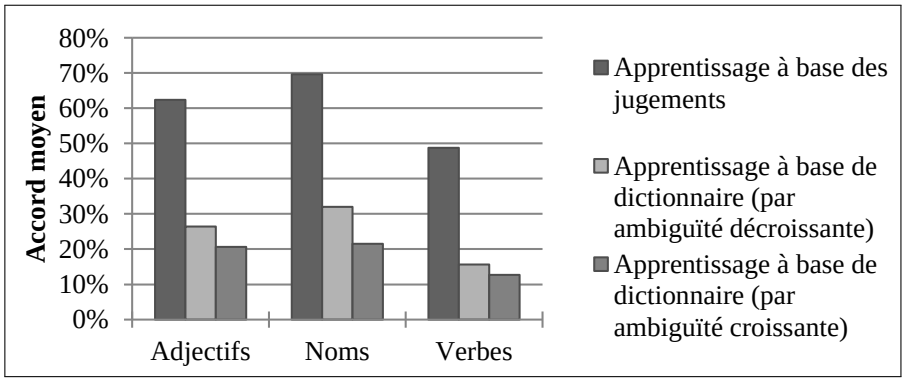


Figure 5.7 – Comparaison de l'accord moyen des méthodes de construction du $\mathcal{DSC}$

5.3.3 Approche probabiliste de désambiguïsation sémantique

Cette section présente une nouvelle approche probabiliste basée sur une version généralisée de la méthode PROX initialement proposée par [Gaume *et al.*, 2004] et ensuite étendue dans notre travail [Elayeb *et al.*, 2015b]. Nous présentons dans ce qui suit l'approche pour le calcul sémantique des sens.

La modélisation du problème en graphe ouvre la porte à des représentations mathématiques et informatiques permettant de mesurer l'acceptante d'un mot par rapport à ces définitions mentionnées dans le dictionnaire traditionnel. Pour ce faire, nous avons transformé le graphe en une matrice markovienne dont les états sont les sommets du graphe en question et ses arêtes les transitions possibles.

En amont, nous avons généré à partir du graphe issu du dictionnaire sémantique de contexte $\mathcal{DSC}$ une matrice de transition (section 5.3.3.1), qui sera transformée, en une matrice de Markov (section 5.3.3.2). En aval, nous appliquons l'algorithme de désambiguïsation basé sur la proximité sémantique (section 5.3.3.3).

5.3.3.1 Construction de la matrice de transition

Nous générons la matrice d'adjacence à partir du graphe $G =< S, A >$. Notons par $[G]$ la matrice carrée $n \times n$ telle que pour tout $r, s \in S, [G]_{r,s} = | < s, r > |$ si $(r, s) \in A$ et $[G]_{r,s} = 0$ si $(r, s) \notin A$. Nous appelons $[G]$ la matrice de transition de G. Puisque G n'est pas orienté donc $[G]$ est une matrice symétrique. En plus, G est un graphe réflexif ; d'où $\forall r \in S, [G]_{r,r} = 1$.

5.3.3.2 Construction de la matrice de Markov

A partir de la matrice de transition, nous construisons la matrice Markovienne. Notons $\left[\hat{G}\right]$ la matrice de Markov correspondante au graphe $G =< S, A >$ et définie par :

$$\forall r, s \in S, \left[\hat{G}\right]_{r,s} = \frac{[G]_{r,s}}{\sum_{x \in S} [G]_{r,x}} \tag{5.11}$$

5.3.3.3 Algorithme de désambiguïsation basé sur la proximité sémantique

Nous présentons en premier lieu l'approche PROX à partir de laquelle nous nous sommes inspirés. En second lieu, nous détaillons une version améliorée de cette méthode que nous avons adaptée dans le cadre de désambiguïsation basée sur la proximité sémantique.

Présentation de la méthode PROX La méthode PROX de [Gaume *et al.*, 2004] est une méthode stochastique pour l'étude de la structure des réseaux de petits mondes hiérarchiques. L'application des chaînes de Markov dans le cadre de désambiguïsation sémantique a apporté des résultats significatifs (voir, par exemple, [Loupy, 2000]).

Cette méthode consiste à transformer un graphe en une chaîne de Markov dont les états sont les sommets du graphe en question et ses arêtes les transitions possibles : une particule, partant à l'instant $t = 0$ d'un sommet s_0, se déplace en un pas sur s_1 l'un des voisins de s_0 sélectionné aléatoirement ; la particule se déplace alors à nouveau en un pas sur s_2, l'un des voisins de s_1 sélectionné aléatoirement, etc. Si au $t-^{ième}$ pas la particule est sur le sommet s_t elle se déplace alors en un pas sur le sommet s_{t+1} qui est sélectionné aléatoirement parmi les voisins de s_t tous équiprobables. Selon [Gaume, 2004], une trajectoire $(s_1, s_2, \cdots, s_t, \cdots)$ ainsi sélectionnée est une "balade" aléatoire sur le graphe, et ce sont les dynamiques de ces trajectoires qui donnent les propriétés structurelles des graphes étudiés. [Gaume *et al.*, 2004] définissent $PROX(G, i, r, s)$ comme la probabilité, qu'en partant à l'instant $t = 0$ du sommet r, la particule soit à l'instant $t = i$ sur le sommet s :

$PROX(G, i, r, s) = \left[\hat{G}^i\right]_{r,s}$où A^i est la matrice A multipliée i fois par elle-même.

Calcul dynamique des sens Nous proposons une méthode de calcul dynamique du sens d'un mot dans son contexte en exploitant le graphe sémantique et en calculant la distance sémantique. Notre approche est basée, essentiellement, sur le principe de PROX [Gaume *et al.*, 2004]. Cette méthode représente ainsi une mesure de similarité entre sommets d'un graphe en calculant la distance sémantique entre le mot et ses définitions, ce qui permet d'envisager une meilleure exploitation des graphes sémantiques.

On considère un lemme m_i qui constitue un mot polysémique. Nous notons :

- m_i est un nœud du graphe G.
- $Definition(m_i) = p_i^1, p_i^2, \cdots, p_i^a$
- $G^\infty = \lim_{t\to\infty} \left[\hat{G}^t\right]$, ainsi G^∞est un vecteur de $\mathbb{R}^a$
- $f_\infty(r, s) = \lim_{t\to\infty} PROX(G, t, s, r)$. La fonction f_∞représente la proximité sémantique entre les mots polysémiques et leurs définitions dans le $\mathcal{DSC}$.

Par conséquent, nous avons la propriété suivante :

Propriété : Puisque le graphe G est réflexif et fortement connexe alors [Elayeb *et al.*, 2015b] :

$\forall a \in S, \lim_{t\to\infty} PROX(G, t, s, r) = \lim_{t\to\infty} PROX(G, i, a, r)$

Cela signifie que la probabilité, pour un temps t assez long, d'atteindre un sommet s ne dépend pas du sommet de départ (s ou a).

On dit que α est fortement lié à α_i si est seulement si :

$\forall j \in \mathbb{N} \setminus \{i\}, f_\infty(\alpha, \alpha^i) > f_\infty(\alpha, \alpha^j)$.

Dans ce cas, le résultat de désambiguïsation du mot polysémique α est α_i.

Le mot β portant de l'information vérifie les propriétés suivantes :

- $\beta \in Contexte(\alpha, ph)$
- $\forall \delta \in Contexte(\alpha, ph) \setminus \{\beta\}, f_\infty(\alpha, \beta) = max_\delta(f_\infty(\alpha, \delta))$; avec β est la définition de α.

Dans ce cas, β est la définition sémantique de α dans le dictionnaire sémantique de contextes ($\mathcal{DSC}$).

5.3.3.4 Exemple illustratif

Considérons le premier exemple de phrase polysémique suivant [Elayeb *et al.*, 2015b] :

PH1 : « *François est un vrai avocat qui s'occupe des droits des étrangers sur Paris.* »

La phrase polysémique PH1 est représentée par les ensembles suivants :

- $Polysémie(PH1) = \{avocat,\ droit\}$
- $Significatif(PH1) = \{occuper,\ vrai,\ étranger\}$
- $Contexte(PH1) = \{vrai,\ avocat,\ occuper,\ droit,\ étranger\}$
- $Définition(avocat) =$

  avocat_1 *Praticien et professionnel du **droit** dont la fonction traditionnelle est de **conseiller** ses clients sur des questions juridiques, quelles soient relatives à leur vie **juridique** quotidienne ou plus spécialisées, [...]*

  avocat_2 ***Fruit** comestible de l'avocatier, à pulpe jaune, contenant un gros **noyau**, fortement **conseillé** dans plusieurs cocktails de fruits et des confitures [. . .]*

- $Définition(droit) =$

  droit_1 *Faculté reconnue de jouir d'une chose, d'accomplir une action.*
  droit_2 *Taxe, impôt*
  droit_3 ***Lois** et dispositions **juridiques** qui **règlent** les rapports entre les membres d'une société*

- $Définition(occuper) = \{Prendre\ possession\ d'un\ endroit.\}$

– $Définition(vrai)$ =$\{$*Qui présente un caractère de vérité.*$\}$

– $Définition(étranger)$ =$\{$*Qui est d'une autre nation ; qui est autre, en parlant d'une nation.*$\}$

Le graphe sémantique correspondant à l'exemple de la phrase PH1 est présenté dans la figure 5.8.

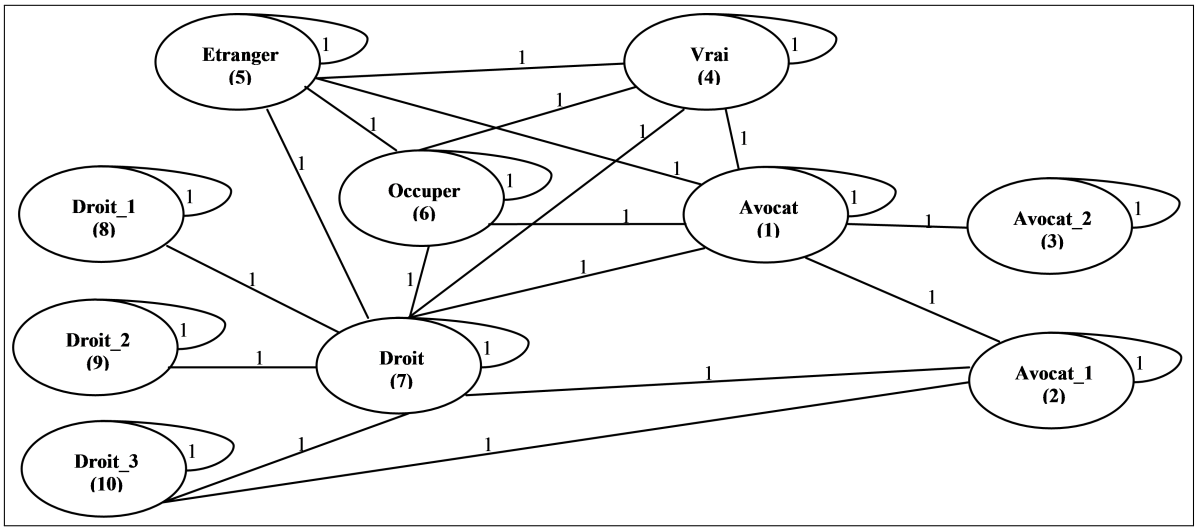


Figure 5.8 – Graphe sémantique de l'exemple PH1

Le calcul sémantique des sens est basé sur la matrice de Markov suivante de l'exemple PH1 :

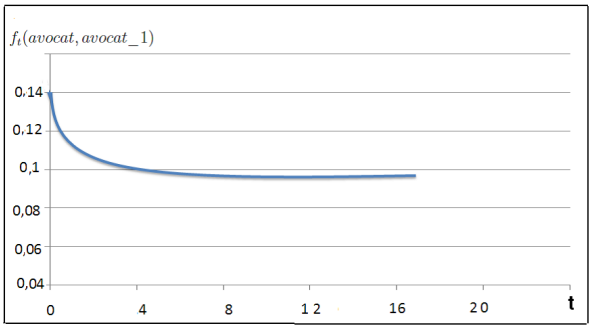
$$\left[\hat{G}\right] = \begin{pmatrix} (0) & (1) & (2) & (3) & (4) & (5) & (6) & (7) & (8) & (9) & (10) \\ (1) & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 & 0 & 0 \\ (2) & 1/4 & 1/4 & 0 & 0 & 0 & 0 & 1/4 & 0 & 0 & 1/4 \\ (3) & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ (4) & 1/5 & 0 & 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 0 & 0 \\ (5) & 1/5 & 0 & 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 0 & 0 \\ (6) & 1/5 & 0 & 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 0 & 0 \\ (7) & 1/9 & 1/9 & 0 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 & 1/9 \\ (8) & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 1/2 & 0 & 0 \\ (9) & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 \\ (10) & 0 & 1/3 & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 1/3 \end{pmatrix}$$

$$[\hat{G}]^{\infty} = \begin{pmatrix} (0) & (1) & (2) & (3) & (4) & (5) & (6) & (7) & (8) & (9) & (10) \\ (1) & 0,160 & 0,090 & 0,046 & 0,114 & 0,114 & 0,114 & 0,202 & 0,044 & 0,044 & 0,067 \\ (2) & 0,157 & 0,092 & 0,044 & 0,112 & 0,112 & 0,112 & 0,205 & 0,045 & 0,045 & 0,069 \\ (3) & 0,164 & 0,089 & 0,051 & 0,115 & 0,115 & 0,115 & 0,199 & 0,042 & 0,042 & 0,065 \\ (4) & 0,160 & 0,089 & 0,046 & 0,114 & 0,114 & 0,114 & 0,203 & 0,044 & 0,044 & 0,066 \\ (5) & 0,160 & 0,089 & 0,046 & 0,114 & 0,114 & 0,114 & 0,203 & 0,044 & 0,044 & 0,066 \\ (6) & 0,160 & 0,089 & 0,046 & 0,114 & 0,114 & 0,114 & 0,203 & 0,044 & 0,044 & 0,066 \\ (7) & 0,157 & 0,091 & 0,044 & 0,112 & 0,112 & 0,112 & 0,205 & 0,046 & 0,046 & 0,068 \\ (8) & 0,154 & 0,091 & 0,042 & 0,111 & 0,111 & 0,111 & 0,208 & 0,048 & 0,048 & 0,069 \\ (9) & 0,154 & 0,091 & 0,042 & 0,111 & 0,111 & 0,111 & 0,208 & 0,048 & 0,048 & 0,069 \\ (10) & 0,156 & 0,093 & 0,043 & 0,111 & 0,111 & 0,111 & 0,206 & 0,046 & 0,046 & 0,071 \end{pmatrix}$$

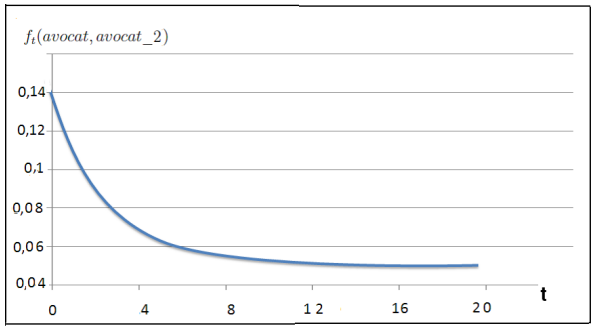
L'observation de la dynamique de la particule partant d'un nœud r vers un nœud s indique le rapport sémantique entre deux nœuds r et s. Les figures 5.9.a et 5.9.b représentent, respectivement, les courbes $f_t(avocat, avocat\_1) = \left[\hat{G}^t\right]_{avocat,avocat\_1}$ et $f_t(avocat, avocat\_2) = \left[\hat{G}^t\right]_{avocat,avocat\_2}$

Quand le temps t tend vers l'infini, chacune de ces courbes tend vers sa limite : $f_t(avocat, avocat\_1) = 0,091$ et $f_t(avocat, avocat\_2) = 0,045$.

D'après le tableau 5.2, nous remarquons dans $\left[\hat{G}\right]^{\infty}$ que « *droit* » puis « *avocat* » sont les mots les plus importants par rapport aux autres nœuds dans le graphe sémantique.



(a) Courbe de convergence de « *avocat_1* »



(b) Courbe de convergence de « *avocat_2* »

Figure 5.9 – Courbes de convergence des sens « *avocat_1* » et « *avocat_2* »

avocat (1)	avocat_1 (2)	avocat_2 (3)	vrai (4)	étranger (5)	occuper (6)	droit (7)	droit_1 (8)	droit_2 (9)	droit_3 (10)
0,160	0,090	0,046	0,114	0,114	0,114	0,202	0,044	0,044	0,067

Table 5.2 – Résultat de proximité sémantique de $\left[\hat{G}\right]^{\infty}$ de l'exemple PH1

Nous concluons que :

- « *avocat* » est plus lié à « *avocat_1* » qu'à « *avocat_2* » car $f_{\infty}(1,2) = 0,09 > f_{\infty}(1,3) = 0,046$.
- « *droit* » est plus lié à « *droit_3* » qu'à « *droit_1* » et « *droit_2* » car $f_{\infty}(7,10) = 0,068 > f_{\infty}(7,8) = f_{\infty}(7,9) = 0,046$.

Les deux définitions de « *avocat_1* » et « *droit_3* » sont ajoutées dans le dictionnaire sémantique de contextes. Ce résultat d'apprentissage est pris en compte dans la suite des analyses.

Considérons maintenant l'exemple d'une deuxième phrase polysémique :

PH2 : « *L'avocat pratique la loi.* »

La phrase polysémique PH2 est représentée par les ensembles suivants :

- $Polysémie(PH2) = \{loi\}$
- $Contexte(avocat, PH2) = \{pratiquer,\ loi\}$
- $Définition(pratiquer) = \{$*Applications des principes d'un art, d'une science ou d'une technique.*$\}$
- $Définition(loi) =$
  - *loi_1* *L'ensemble des* ***règles***, ***droits*** *et devoirs, édictée par une autorité, que toute personne doit suivre [Droit]..*
  - *loi_2* *Convention.*
  - *loi_3* *énoncé de phénomènes dans un domaine particulier.*
- « *avocat* » existe dans le dictionnaire sémantique de contextes
- Enrichissement du contexte de « *avocat* » : $Contexte(avocat) = \{pratiquer,\ loi,\ droit(droit\_3)\}$.
- $EspaceSémantique = \{avocat,\ avocat\_1,\ avocat\_2,\ droit\_3,\ pratiquer,\ loi,\ loi\_1,\ loi\_2,\ loi\_3\}$

Le graphe sémantique correspondant à l'exemple PH2 est présenté dans la figure 5.10.

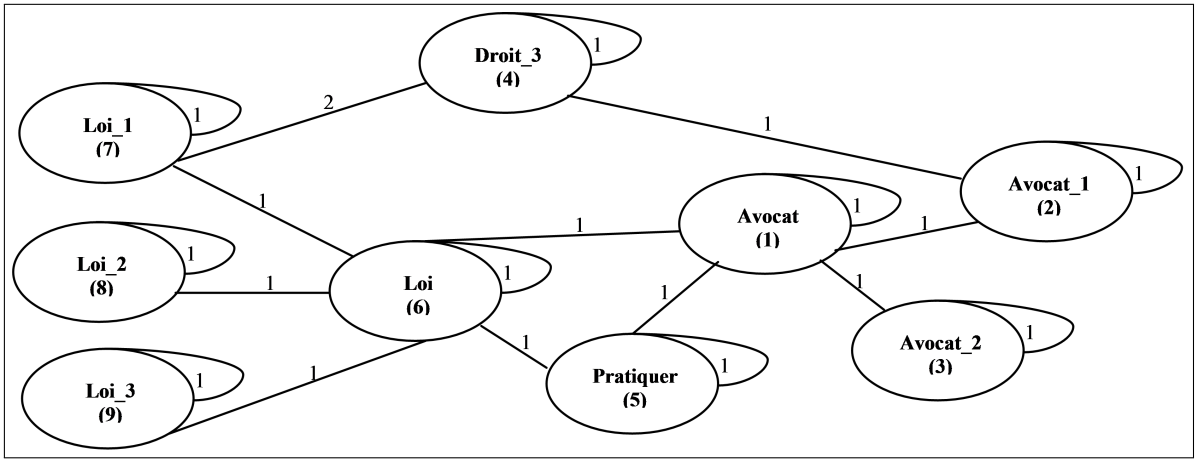


Figure 5.10 – Graphe sémantique de la phrase PH2

Remarquons que $|loi\_1, droit\_3| = 2$ car $loi\_1 \bigcap droit\_3 = \{loi, règle\}$.

avocat (1)	avocat_1 (2)	avocat_2 (3)	droit_3 (4)	pratiquer (5)	loi (6)	loi_1 (7)	loi_2 (8)	loi_3 (9)
0,035	0,021	0,014	0,028	0,014	0,036	0,028	0,014	0,014

Table 5.3 – Résultat de proximité sémantique de $\left[\hat{G}\right]^{\infty}$ de l'exemple PH2

La première ligne de $\left[\hat{G}\right]^{\infty}$ est donnée dans le tableau 5.3. Nous remarquons dans $\left[\hat{G}\right]^{\infty}$ que « *loi* » puis « *avocat* » sont plus importants que tous les autres sommets du graphe sémantique. On distingue que :

- « *avocat* » est plus fortement lié à « *avocat_1* » qu'à « *avocat_2* »
- « *loi* » est davantage liée à « *loi_1* » qu'à « *loi_2* » et « *loi_3* »

Nous remarquons aussi que « *loi* » , « *avocat* » , « *droit_3* » , « *loi_1* » et « *avocat_1* » sont fortement liées. Les modifications apportées au dictionnaire sémantique de contextes sont :

- Ajouter à l'entrée de « *avocat* » le concept « *loi_1* ». D'où « *avocat* » = « *avocat_1* » (« *droit* » = « *droit_3* » et « *loi* » = « *loi_1* »).
- Ajouter une nouvelle entrée « *loi* » = « *loi_1* » et « *avocat* » = « *avocat_1* ».

Les exemples rapportés pour notre approche ont été effectués à partir du cadre d'évaluation de l'importance de dictionnaire sémantique de contextes dans la désambiguïsation sémantique.

5.3.4 Étude comparative des approches possibiliste et probabiliste de WSD

Nous avons organisé le $\mathcal{DSC}$ dans trois fichiers par catégories (adjectifs, noms et verbes) qui ont été utilisés dans les deux approches de désambiguïsation (possibiliste et probabiliste) [Ben Khiroun *et al.*, 2012].

Pour évaluer notre système, nous calculons le taux d'exactitude pour chaque mot en utilisant la métrique *Kappa* de Cohen [Cohen, 1968, Eugenio, 2000]. Cette mesure est basée sur la différence entre la quantité de l'accord qui est effectivement présente (accord « observé ») et l'accord qui serait présent par le seul hasard (accord « attendue ») comme suit [Viera et Garrett, 2005] :

$$Kappa = \frac{Pobservée - Pattendue}{1 - Pattendue} \tag{5.12}$$

Avec :

- $Pobservée$: la proportion d'accord observée.
- $Pattendue$: la proportion d'accord aléatoire (ou attendue).

Pour calculer le degré *Kappa* d'accord entre deux jugements [4], nous représentons l'ensemble des jugements d'un mot ambigu par le tableau 5.4 suivant :

Jugement Système

Jugement Annotateur						
	Sens	1	2	...	m	Total
	1	n_{11}	n_{12}	...	n_{1m}	L_1
	2	n_{21}	n_{22}	...	n_{2m}	L_2
	.					
	.					
	.					
	m	n_{m1}	n_{m2}	...	n_{mm}	L_m
	Total	C_1	C_2	...	C_m	N

Table 5.4 – Représentation matricielle des jugements de sens pour deux jugements et m sens possibles

Avec :

- N : Nombre de cas à désambiguïser pour un mot donné.
- m : Nombre de sens possibles dans le dictionnaire.
- n_{ij} : Nombre de cas jugés en relation avec les sens i par les annotateurs et en relation avec le sens j par le système [5].

La concordance observée $Pobservée$ est représentée par la proportion des mots classés dans les cases diagonales du tableau de contingence, soit la somme de ces cases diagonales divisée par la taille de l'échantillon N des cas du mot ambigu comme suit :

$$Pobservée = \frac{1}{N} \sum_{i=1}^{m} n_{ii} \tag{5.13}$$

L'accord aléatoire attendu $Pattendue$ est proportionnel à :

$$Pattendue = \frac{1}{N^2} \sum_{i=1}^{m} L_i \times C_i \tag{5.14}$$

Le test *Kappa* prend ainsi en compte l'accord survenant par hasard et il est considéré comme une valeur raffinée. Landis et Koch [Landis et Koch, 1977] ont proposé une interprétation des valeurs de la mesure *Kappa* de Cohen comme l'indique le tableau *5.5*

Selon [Edmonds et Hirst, 2002], un mot peut avoir plusieurs sens sans pouvoir pour autant les différencier. Ce phénomène qualifié d'*indétermination* est fortement lié à l'expressivité de la langue en question.

4. nous prenons le cas de jugement par le système et le jugement convenu par les annotateurs.
5. les deux jugements sont en désaccord si $i \neq j$ et en accord total si $i = j$.

Kappa	Interprétation
< 0	Désaccord
0,0 à 0,20	Accord très faible
0,21 à 0,40	Accord faible
0,41 à 0,60	Accord modéré
0,61 à 0,80	Accord fort
0,81 à 1,00	Accord presque parfait

Table 5.5 – Interprétation des valeurs *Kappa* de Cohen [Landis et Koch, 1977]

Afin de renforcer notre évaluation de désambiguïsation, nous avons utilisé le *rappel*, la *précision*, et la *F-mesure* recommandés pour évaluer les SRI [Segond, 2000]. Dans notre cas, ces métriques sont calculées comme suit :

$$rappel = \frac{sens\ corrects\ retrouvés}{nombre\ total\ de\ sens\ dans\ le\ standard} \tag{5.15}$$

$$précision = \frac{sens\ corrects\ retrouvés}{nombre\ total\ de\ sens\ proposés} \tag{5.16}$$

$$F - Mesure = \frac{2 \times rappel \times précision}{rappel + précision} \tag{5.17}$$

Résultats

Nous présentons dans les figures 5.11, 5.12 et 5.13 une étude comparative détaillée des mesures de *Kappa* moyenne pour les 3 catégories grammaticales (nom, verbe, adjectif). Les résultats présentés dans ces figures comparent les performances obtenues de nos deux approches de WSD (possibiliste et à base de la méthode PROX) avec les résultats du système Xerox, utilisant la même collection de test ROMANSEVAL dans les expérimentations [Elayeb *et al.*, 2015b].

Les figures 5.14 et 5.15 présentent une comparaison de la mesure de *Kappa* des approches possibiliste (notée *POSS*), probabiliste (notée *PROBA* et détaillée dans la section 5.3.3) avec les résultats de cinq autres systèmes de WSD monolingue. Le choix de ces systèmes dans l'étude comparative de nos approches se base sur deux critères ; à savoir l'utilisation du corpus de test ROMANSEVAL dans les expérimentations et l'application à la langue française.

Ces systèmes ont été développés respectivement par EPFL (Ecole Polytechnique Fédérale de Lausanne), IRISA (Institut de recherche en informatique et Systèmes Aléatoire, Rennes), LIA-BERTIN (Laboratoire d'informatique, Université d'Avignon, et BERTIN, Paris), et XEROX (Xerox Research Centre Europe, Grenoble). Une étude des 5

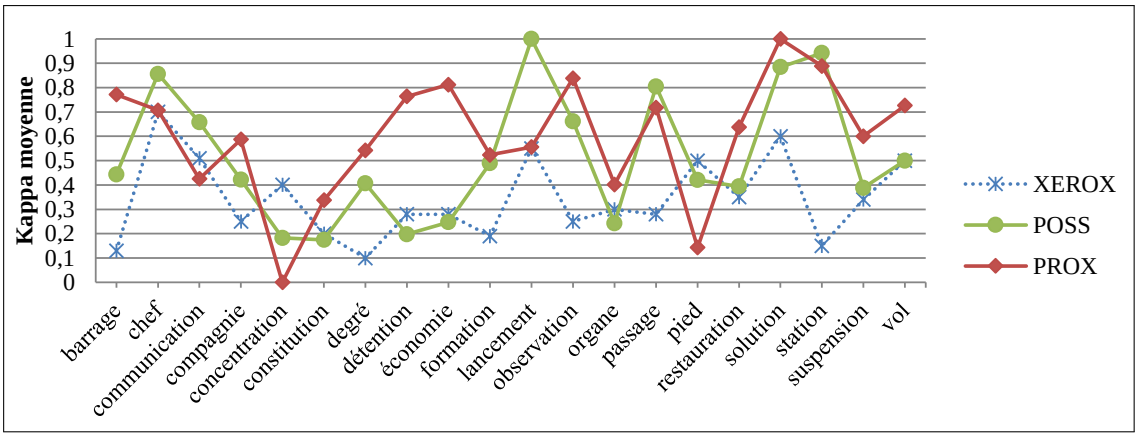


Figure 5.11 – Résultats détaillés des mesures de *Kappa* moyennes pour les noms

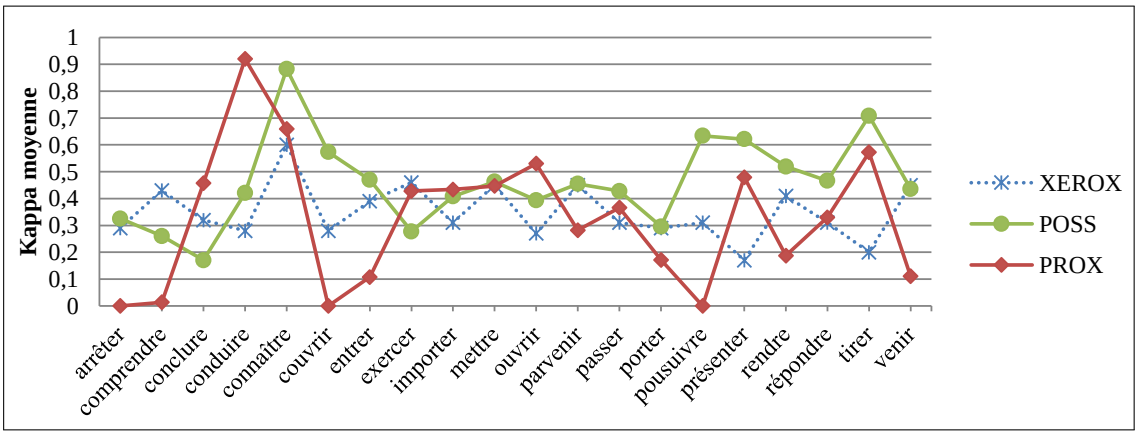


Figure 5.12 – Résultats détaillés des mesures de *Kappa* moyennes pour les verbes

systèmes a été établie dans [Segond, 2000].

Les résultats expérimentaux pour les adjectifs, présentés dans la figure 5.14 montrent une bonne performance de l'approche possibiliste par rapport aux 5 autres systèmes ainsi que l'approche probabiliste. En contre partie, pour les noms, l'approche probabiliste donne de meilleurs résultats. Concernant les verbes, l'approche possibiliste est meilleure que les systèmes EPFL, IRISA, Xerox et l'approche probabiliste ; cependant, elle est moins performante que les approches LIA1 et LIA2.

En se focalisant sur les résultat globaux, présentés dans la figure 5.15, l'approche possibiliste se distingue en terme de mesure de *Kappa* moyenne sur les trois catégories grammaticales. Par conséquent, la valeur élevée de la mesure *Kappa* égale à 0,47 prouve un accord significatif mesuré de notre système avec les jugements des jurys ayant étiquetés les sens du corpus ROMANSEVAL.

Nous notons également que le désaccord mesuré entre les jurys humains est assez important selon [Véronis, 1998]. En effet, la mesure *Kappa* varie entre 0,92 (pour le nom « détention ») et 0,007 (pour l'adjectif « correct »). Ainsi, il y a plus de chance que l'accord soit aléatoire dans d'étiquetage des sens entre les jurys pour certains mots.

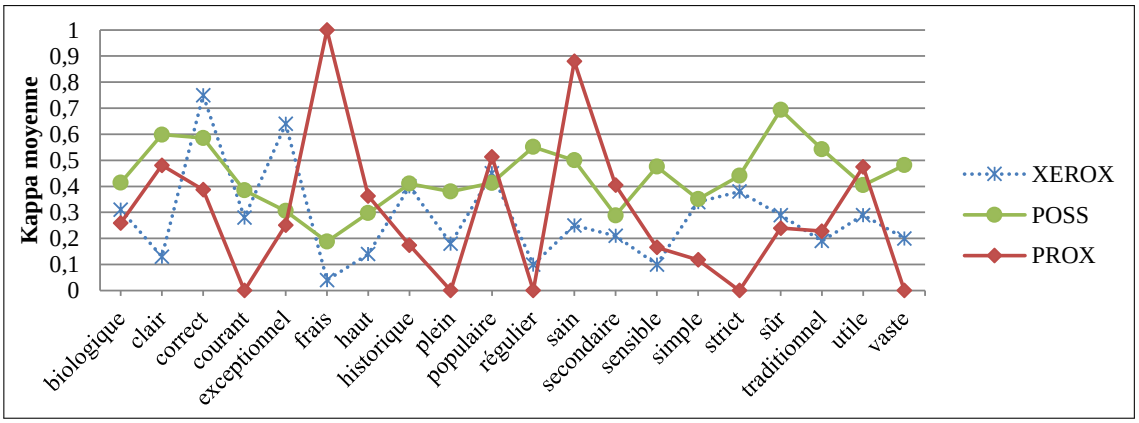


Figure 5.13 – Résultats détaillés des mesures de *Kappa* moyennes pour les adjectifs

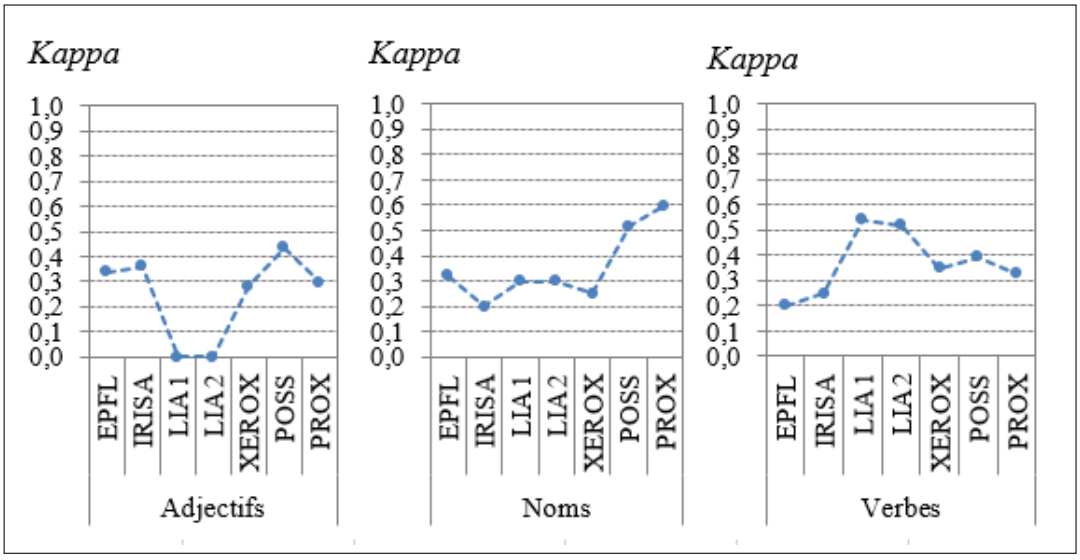


Figure 5.14 – Résultats de la mesure de *Kappa* moyenne par catégorie grammaticale (adjectifs, noms, verbes)

En conséquence, si les annotateurs humains ne sont pas parfaitement d'accord sur l'étiquetage des sens de plusieurs mots, les systèmes ayant produits des résultats aléatoires pour ces mots peuvent être considérés comme satisfaisants. Ce phénomène est bien connu, dans l'état de l'art sur la désambiguïsation sémantique des sens, reconnaissant le fait que les annotateurs humains tendent souvent au désaccord [Vidhu Bhala et Abirami, 2014].

Afin de comparer davantage la nouvelle approche de désambiguïsation sémantique à base de réseaux possibilistes avec l'approche probabiliste et celle de Xerox, nous avons établi dans le tableau 5.6 les mesures *p-valeur* associées au test des rangs signés de Wilcoxon pour échantillons appariés. Cette mesure a été introduite par Demšar dans [Demšar, 2006].

Les *p-valeur* sont calculées en comparant la moyenne *Kappa* de l'approche possibiliste à chacun des deux autres approches une par une. Les résultats donnés dans le

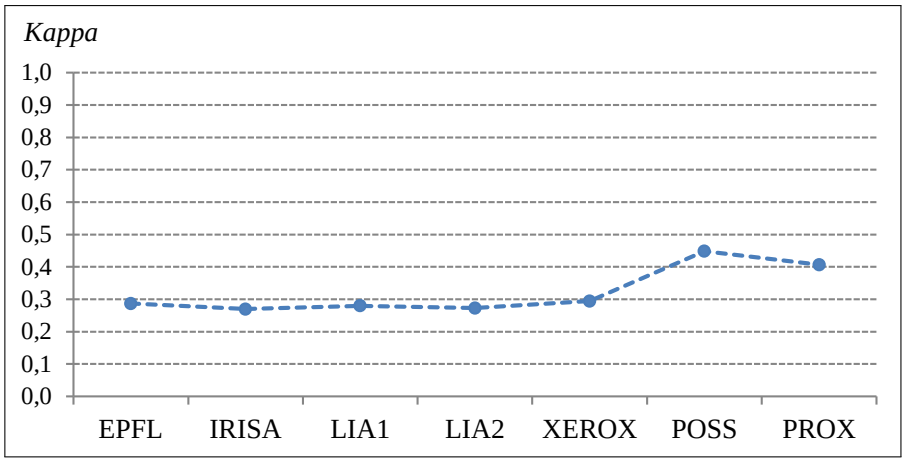


Figure 5.15 – Résultats de la mesure de *Kappa* moyenne pour les 3 catégories grammaticales combinées

tableau 5.6 montrent que notre approche possibiliste donne de meilleurs résultats que l'approche Xerox pour chaque catégorie grammaticale ($p\text{-}valeur < 0{,}05$)[6]. Le résultat est plus significatif pour toutes les catégories grammaticales confondues avec une $p\text{-}valeur = 0.000007$.

D'autre part, la méthode possibiliste dépasse l'approche probabiliste pour les adjectifs et les verbes ; cependant, les résultats ne sont pas assez significatifs pour les noms avec une *p-valeur* qui dépasse 0,05. Ceci est dû au fait que les noms présentent la catégorie grammaticale la plus ambigüe dans la langue française selon Veronis [Véronis, 1998].

	POSS vs Xerox	*POSS vs PROBA*
Adjectifs (A)	0,004550	0,033340
Noms (N)	0,011220	0,184992
Verbes (V)	0,016852	0,019569
Tous (A, N, V)	0,000007	0,052847

Table 5.6 – Résultats de la *p-valeur* associée au test des rangs signés de Wilcoxon pour échantillons appariés

Nous évaluons dans les tableaux 5.7 et 5.8 les différentes approches en terme de rappel, précision et F-mesure. Ces trois mesures ont constitué en effet les métriques principales de mesure de performance dans les compagnes SensEval/SemEval [Lefever et Hoste, 2013].

Les résultats du tableau 5.7 montrent que le système IRISA se classe en premier lieu juste avant l'approche possibiliste pour la catégorie grammaticale des adjectifs. Cependant, les systèmes LIA1 et LIA2 donnent des valeurs nulles pour la même catégorie grammaticale. En contre partie, ces deux systèmes, LIA1 et LIA2, se distinguent dans la désambiguïsation des noms et des verbes juste avant le système possibiliste.

6. 0,05 est la valeur seuil de significativité dans l'échelle de *Ronald Fisher* (voir détails dans [Biau *et al.*, 2010])

	Adjectifs (A)			*Noms (N)*			*Verbes (V)*		
	Rappel	*Précision*	*F-mesure*	*Rappel*	*Précision*	*F-mesure*	*Rappel*	*Précision*	*F-mesure*
EPFL	0,54	0,56	0,549	0,51	0,52	0,514	0,40	0,39	0,394
IRISA	**0,69**	**0,61**	**0,647**	0,55	0,48	0,512	0,29	0,28	0,284
LIA1	0,00	0,00	-	0,75	0,64	**0,690**	0,88	0,71	0,785
LIA2	0,00	0,00	-	**0,76**	0,63	0,688	**0,89**	**0,72**	**0,796**
XEROX	0,56	0,48	0,516	0,45	0,43	0,439	0,31	0,29	0,299
POSS	0,54	0,57	0,554	0,63	**0,66**	0,644	0,44	0,47	0,454
PROBA	0,49	0,53	0,509	0,59	0,63	0,609	0,40	0,44	0,419

Table 5.7 – Résultats des mesures *rappel*, *précision* et *F-mesure* par catégorie grammaticale (adjectifs, noms, verbes)

	(Adjectifs+Noms+Verbes)		
	Rappel	*Précision*	*F-mesure*
POSS	**0,543**	**0,570**	**0,556**
PROBA	0,507	0,546	0,526

Table 5.8 – Résultats des approches possibiliste et probabiliste en utilisant les mesures *rappel*, *précision* et *F-mesure* pour les 3 catégories grammaticales

En analysant les résultats globaux pour les trois catégories grammaticales confondues (tableau 5.8), l'approche possibiliste se distingue par rapport à l'approche probabiliste en terme de rappel, précision et F1 mesure. En effet, la mesure *Kappa* ne peut pas seule refléter la réelle performance des approches de désambiguïsation sémantique [Cohn, 2003]. Notre étude expérimentale, établie en utilisant des métriques variées, vise à mettre en place une évaluation objective de nos approches proposées et celles existantes.

Conclusion

Dans ce chapitre, nous avons proposé des contributions dans le domaine de la désambiguïsation sémantique monolingue qui est considérée comme l'une des tâches les plus difficiles dans le domaine du traitement sémantique [Navigli, 2009]. En effet, nous avons proposé, évalué et comparé deux approches possibiliste et probabiliste pour la WSD en les appliquant sur des textes français. Dans ces approches, nous avons modélisé et construit une ressource lexicale, appelée dictionnaire sémantique de contextes ($\mathcal{DSC}$) qui intègre des connaissances issues de dictionnaire traditionnel et de corpus étiqueté.

Tout d'abord, nous avons modélisé un réseau possibiliste afin de quantifier la pertinence d'un sens de mot polysémique par une double mesure : la pertinence possible permet de rejeter le sens non pertinent du mot, alors que la pertinence nécessaire permet de renforcer les sens non écarté par la possibilité. En deuxième lieu, nous avons adapté une distance existante pour proposer une nouvelle méthode probabiliste de WSD dans

laquelle les sens sont modélisés en topologie de graphe.

Afin d'évaluer et de comparer nos deux approches proposées avec des systèmes de désambiguïsation monolingues similaires, nous avons effectué des expérimentations sur la collection standard de test ROMANSEVAL. Les expériences utilisant la métrique *Kappa* ont montré une amélioration significative en terme de taux de désambiguïsation des mots français. La désambiguïsation possibiliste donne des résultats meilleures que le système Xerox pour les adjectifs, les noms et les verbes. Notons, cependant, que l'approche de WSD probabiliste dépasse les approches possibiliste et Xerox en l'appliquant sur les noms seuls.

En analysant les moyennes des résultats appliqués sur les trois catégories grammaticales ensembles, l'approche possibiliste est plus performante en termes des métriques rappel, précision et F-mesure en comparaison avec les autres systèmes. D'autre part, la modélisation et l'utilisation du $\mathcal{DSC}$, comme ressource lexicale de désambiguïsation, permettait d'améliorer les résultats de calcul des sens des mots grâce aux informations contextuelles représentées explicitement dans le $\mathcal{DSC}$.

Afin d'étudier l'apport des réseaux possibilistes ainsi que l'impact de la WSD sur la RI, nous étendons, dans le chapitre suivant, notre cadre d'étude en partant du contexte monolingue vers un cadre translinguistique.

Chapitre

6 L'Expansion de Requêtes et la Désambiguïsation Sémantique au Service de la RI

Sommaire

" A little knowledge that acts is worth infinitely more than much knowledge that is idle. "

— Gibran Khalil Gibran

Introduction

Le développement des systèmes de recherche d'information rencontre de nombreux défis, en particulier liés à la nature des requêtes qui sont formulées par l'utilisateur du SRI. En fait, l'utilisateur a tendance à formuler son besoin en information par des requêtes courtes (c-à-d avec un nombre réduit de mots clés) qui peuvent contenir également des termes ambigus. Par conséquent, le manque de contexte proposé dans la requête influence la qualité de recherche en proposant des documents non pertinents.

Le processus de désambiguïsation de requête est basé sur la tâche de WSD. En effet, la désambiguïsation du sens du mot consiste à sélectionner le sens approprié d'un mot en fonction de son contexte tel que présenté dans le chapitre précédant. Récemment, le domaine de WSD a été amélioré principalement grâce aux compétitions SensEval et SemEval. À titre d'exemple, certains travaux ont confirmé que l'efficacité des systèmes de traduction automatique a été considérablement améliorée, grâce à la WSD en accompagnant le processus de traduction [Chan et Ng, 2007, Sharma et Mittal, 2016].

Dans le domaine de RI, la tâche de désambigüisation sémantique s'impose en vue des deux angles suivants : (i) les termes de requête peuvent avoir des sens étroitement liés avec d'autres mots qui n'existent pas dans la requête. Par conséquent, le rappel des documents pertinents peut être amélioré si l'on tient compte de ces liens sémantiques entre les mots ; et (ii) les requêtes et les documents peuvent avoir plusieurs sens ; ce qui diminue la précision de la RI. Ainsi, la sélection du bon sens pour les termes des requêtes et des documents peut considérablement améliorer la précision en diminuant le bruit dans les résultats de recherche [Chifu et Ionescu, 2012].

Afin d'enrichir le contexte des requêtes, le recours aux méthodes d'expansion de requêtes peut être considéré comme solution partielle. Néanmoins, l'application d'expansion peut reformuler la requête d'origine en rajoutant des termes ambigus ; d'où la nécessité de recours de nouveau à la désambiguïsation sémantique afin de lever cette ambiguïté. Cette relation de dépendance entre la désambiguïsation et l'expansion de requêtes prouvent la nécessité de les combiner ensemble afin d'améliorer la RI.

Ainsi, nous étendons, dans ce chapitre, le cadre d'étude afin de répondre aux problématiques suivantes : Comment tirer profit de la désambiguïsation sémantique des textes dans le domaine de la recherche d'information en l'appliquant sur des requêtes ? Quelle est la contribution d'une approche combinée de désambiguïsation et d'expansion sémantique des requêtes sur la performance de RI ? Comment projeter cette approche sur un cadre translinguistique ?

Ce chapitre se divise en deux grandes parties : la première sera dédiée à une étude dans un cadre de RI monolingue et qui sera étendue vers un cadre translinguistique dans la

deuxième partie du chapitre.

6.1 Désambiguïsation et expansion des requêtes en RI monolingue

Plusieurs travaux dans la littérature ont étudié l'impact de l'expansion et de la désambiguïsation de requêtes dans la performance de RI. En effet, certaines méthodes basées sur des thésaurus ont amélioré l'efficacité de RI en élargissant les termes de requête lors de la désambiguïsation avec des synonymes issus de WordNet [Liu *et al.*, 2005, Voorhees, 1994, Fang, 2008]. En outre, l'utilisation de WordNet, comme ressource pour l'expansion des documents, a prouvé également des améliorations dans la performance des SRI selon [Cao *et al.*, 2005, Agirre *et al.*, 2010].

En partant de ces études, nous étudions, dans cette première section du chapitre, la contribution de la désambiguïsation sémantique avec l'expansion des requêtes sur la performance de RI. Pour assurer ces deux tâches, nous calculons la similarité entre les termes de requêtes (dans le cas de l'expansion) ou entre les termes et les sens (dans le cas de désambiguïsation).

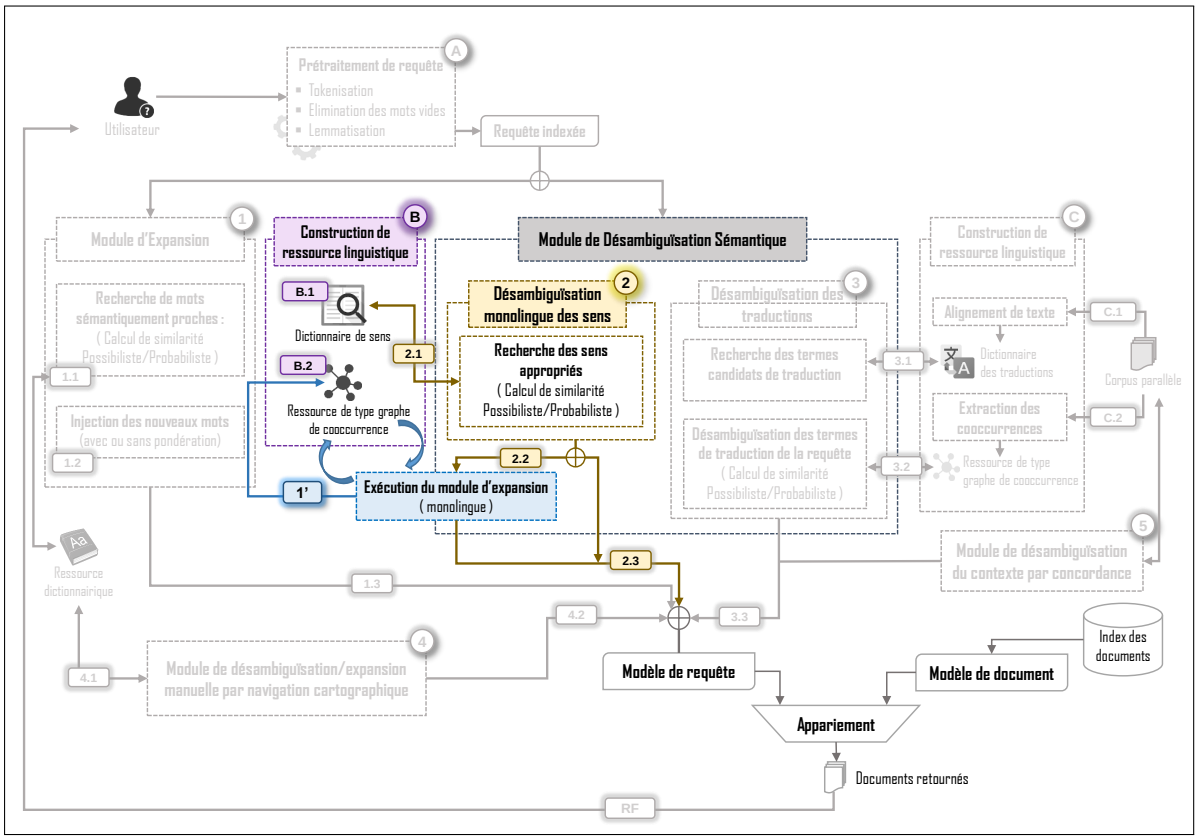


Figure 6.1 – Positionnement de l'approche de désambigüisation et d'expansion sémantique de requêtes dans l'architecture générale du système SPEEDSER

Le calcul de similarité se base sur une représentation de connaissance en graphe de co-occurrence qui sera l'objet de la sous-section 6.1.1. Ceci garantit l'adoption d'une démarche générique (en basculant entre termes et sens) tout en ayant recours à une ressource facile à construire et à exploiter à partir des textes ; à savoir le graphe de co-occurrence. (le positionnement de cette approche est mis en relief dans la figure 6.1 ; se référer au schéma général du système SPEEDSER page 86).

6.1.1 Représentation des connaissances et architecture du modèle

Afin d'avoir une représentation générique des données qui pourra être utilisée dans la désambiguïsation sémantique des requêtes, l'expansion des requêtes et la rétroaction de pertinence, nous optons pour un modèle de graphe basé sur la co-occurrence des termes. Les relations entre les termes sont extraites à partir de corpus textuels afin de modéliser les liens contextuels et sémantiques.

Notre approche est basée sur les réseaux possibilistes pour la désambiguïsation et l'expansion des requêtes. En effet, nous considérons, pour construire le graphe de co-occurrence, que deux nœuds sont liés s'ils existent dans la même phrase. Les arêtes sont non orientées et pondérées par la fréquence normalisée de co-occurrence des termes connexes. D'autre part, les mots ambigus sont liés avec leurs sens appropriés dans le dictionnaire.

Nous considérons les différentes composantes comme suit :

- T : l'ensemble de termes présents dans le corpus ;
- S : l'ensemble des sens dans le dictionnaire ;
- Un nœud t_i est lié à un nœud t_j si t_i et t_j co-occurrent dans la même phrase ; avec $\{t_i, t_j \in T\}$;
- Un nœud t_i est lié à un nœud s_j si t_i est un mot ambigu et s_j représente un sens de t_i ; avec $\{t_i \in T\}$ et $\{s_j \in S\}$.

Le schéma dans la figure 6.2 présente les différentes ressources utilisées ainsi que le processus de réalisation des tâches de désambiguïsation sémantique de requêtes, d'expansion et de rétroaction de pertinence. En partant d'une requête initiale, le module d'expansion est exécuté pour générer une requête étendue.

Si la requête contient des termes ambigus, le module de désambiguïsation est lancé avant l'application de l'expansion. Dans ce cas, le meilleur nœud représentant un sens et ayant le meilleur score possibiliste sera élu et les mots qui existent dans la définition de ce sens seront utilisés pour l'expansion de la requête de départ.

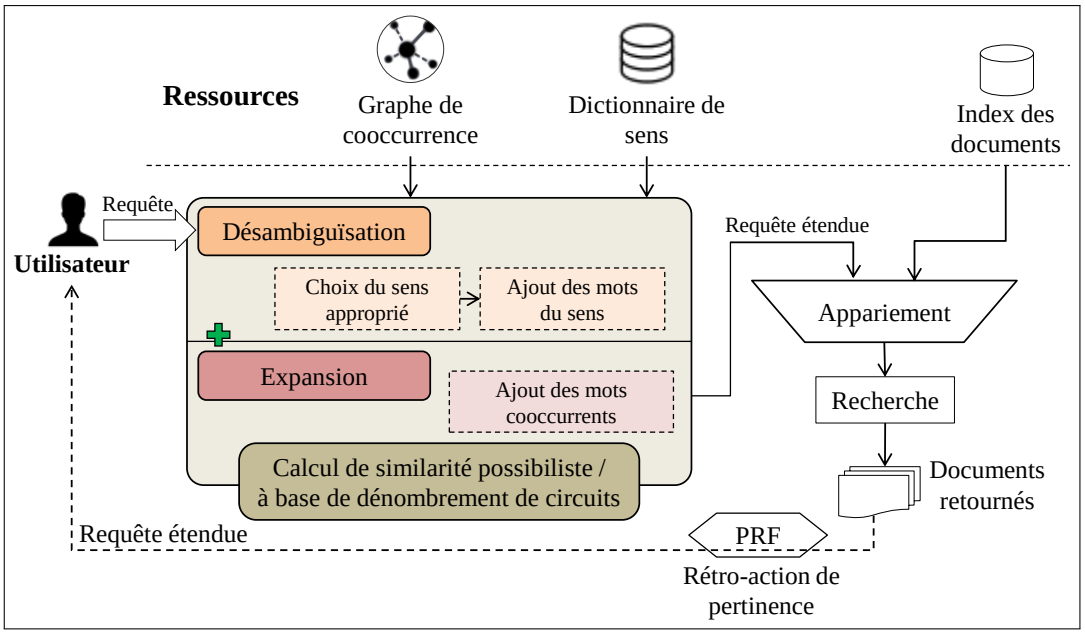


Figure 6.2 – Processus d'expansion de requête en utilisant la désambiguïsation sémantique.

Pour mettre en place le calcul possibiliste des scores de pertinence, nous utilisons le graphe de co-occurrence qui représente une structure adéquate pour modéliser les relations sémantiques entre les termes et les sens.

En fin du processus, une phase de rétroaction de pertinence est exécutée en extrayant les termes à partir des premiers documents retournés après appariement. L'ensemble du processus peut être réitéré.

6.1.2 Proposition d'une approche possibiliste appliquée à la désambiguïsation sémantique et l'expansion de requêtes

Pour calculer la similarité des termes dans les deux tâches d'expansion et de désambiguïsation sémantique de requêtes, nous avons basé notre approche sur la théorie des possibilités introduite par [Zadeh, 1978] et développée par plusieurs auteurs [Dubois et Prade, 2012]. Nous avons adapté l'architecture du modèle possibiliste de [Elayeb *et al.*, 2011] pour l'appliquer sur les graphes de co-occurrence. Nous définissons le degré de pertinence possibiliste (DPP) de chaque nœud n_j sachant une requête $Q = (t_1, t_2, \ldots, t_T)$ par :

$$DPP(n_j) = \Pi(n_j|Q) + N(n_j|Q) \tag{6.1}$$

Avec $\Pi(n_j|Q)$ et $N(n_j|Q)$ représentent respectivement les mesures de possibilité et de nécessité. La première mesure permet de rejeter les nœuds non pertinents (ceux qui ne sont pas près du contexte de la requête et ne peuvent pas être utilisés pour étendre

la requête ou lever l'ambiguïté qu'elle présente). La deuxième mesure est utilisée pour renforcer la pertinence des nœuds les plus importants. Les deux mesures sont calculées comme suit [Ben Khiroun *et al.*, 2014] :

$$\Pi(n_j|Q) = \Pi(t_1|n_j) \times \ldots \times \Pi(t_T|n_j) = nft_{1j} \times \ldots \times nft_{Tj} \tag{6.2}$$

$$N(n_j|Q) = 1 - \left[(1 - \phi n_{1j}) \times \ldots \times (1 - \phi n_{Tj})\right] \tag{6.3}$$

– Avec $nft_{ij} = tf_{ij}/max_k(tf_{kj})$ représente la fréquence normalisée des termes de la requête ; où tf_{ij} est le poids de l'arête qui relie les nœuds t_i et n_j dans le graphe ; i.e. le nombre de fois de co-occurrence des deux nœuds.

– Et :

$$\phi n_{ij} = Log_{10}(\frac{nCN}{nN_i}) \times nft_{ij} \tag{6.4}$$

Où :

- nCN : le nombre total des nœuds dans le graphe de co-occurrence qui sont reliés aux termes de la requête ;
- nN_i : le nombre de nœuds reliés au terme t_i.

L'utilisation de la fonction logarithmique permet de calculer la puissance discriminative des termes de la requête. Ainsi, nous sélectionnons les nœuds du graphe qui sont les plus proches des éléments les plus discriminants de l'information contextuelle représentée dans la requête.

6.1.3 Expérimentations et étude comparative

6.1.3.1 Scénarios des expérimentations

Afin d'étudier l'effet de la désambiguïsation sur l'expansion de requêtes en langue française, nous avons utilisé deux collections de test pour valider nos approches proposées à savoir : CLEF-2003 et ROMANSEVAL.

Le standard de test CLEF-2003, issu de la compagne CLEF [1], propose les outils nécessaires pour l'évaluation des SRI sur des gros corpus documentaires. La collection CLEF comporte :

– un ensemble de documents et un ensemble de requêtes ;

– une liste de documents pertinents pour chaque requête.

1. Cross-Language Evaluation Forum : http://www.clef-campaign.org

Suivant le modèle des compagnes TREC [2], chaque requête possède principalement trois champs, à savoir un titre bref (<title>), une phrase décrivant le besoin d'information (<desc>) et une partie narrative (<narr>) spécifiant plus précisément le contexte de la requête. Un identifiant est également attribué à chaque requête (balise <num>).

La collection CLEF-2003 monolingue pour la langue française est composée de documents extraits de « Le Monde 94 », « ATS94 » et « ATS95 » [3] formant ainsi 57 requêtes et plus de 300 Mo de données [Braschler et Peters, 2004].

Dans les expérimentations réalisées, nous nous sommes limités à un sous-ensemble de 15 requêtes issues de la collection CLEF-2003 et ayant la particularité d'intégrer des mots ambigus présents dans le standard ROMANSEVAL.

En effet, le standard ROMANSEVAL est utilisé pour évaluer les approches de désambiguïsation sémantique en proposant les ressources nécessaires pour la tâche de WSD à savoir : un ensemble de documents et un ensemble de phrase de test contenant des mots ambigus. L'ensemble de 60 mots ambigus dans ROMANSEVAL est distribué sur trois catégories grammaticales (adjectif, adverbe, nom) et les phrases ambigües ont été annotées par 6 experts en associant un ou plusieurs sens possibles pour chaque contexte (voir annexe A).

La plate-forme expérimentale Terrier [4] pour la recherche d'information a été utilisée pour évaluer notre système en appliquant le modèle d'appariement Okapi (BM25) et Snowball comme racinisateur (ou *stemmer* en anglais) [Ounis *et al.*, 2006, Macdonald *et al.*, 2012]. Différentes métriques ont été calculées à savoir le Rappel/Précision, la R-précision et la précision moyenne (MAP).

Afin de mettre en place la pseudo-rétroaction de pertinence basée sur les documents, nous avons utilisé le modèle *Bo1* (*Bose-Einstein1*) implémenté dans Terrier tout en choisissant la configuration par défaut suivante : le nombre de termes pour étendre les requêtes est égal à 10 et le nombre des top-premiers documents pertinents à partir desquels les termes sont extraits est égal à 3 documents.

6.1.3.2 Évaluation de l'approche possibiliste d'expansion de requêtes

Le choix du nombre de termes d'expansion à utiliser dans l'expansion automatique de requêtes a été étudié dans [Ogilvie *et al.*, 2009] à travers 8 SRI. Les résultats ont montré que le nombre de termes qui optimise la précision moyenne (MAP) varie selon le système étudié et le jeu de requêtes utilisé. Pour plusieurs requêtes, un nombre inférieur à 10

2. Text REtrieval Conference : http://trec.nist.gov
3. ATS : Agence Télégraphique Suisse
4. http://terrier.org

termes engendre les meilleures précisions selon les expérimentations réalisées [Ogilvie *et al.*, 2009].

Par conséquent, nous nous sommes basés sur cette hypothèse comme suit : nous considérons un ensemble de terme d'expansion réduit égal à *(N div 4)*, *(N div 2)* ou N ; avec N représente le nombre de termes dans la requête originale [5]. Ces nombres de termes ont été choisi en considérant que la partie narrative des requêtes dans CLEF-2003 est longue (plus de 10 termes). En conséquence, l'application de l'expansion sur de telles requêtes longues peut provoquer un *bruit* ; ce qui conduit à des résultats non interprétables comme détaillé dans [Pinto et Pérez-sanjulián, 2008].

Nous comparons dans le tableau 6.1 les différents scénarios d'expansion possibiliste (désignés par *QE*) basé sur le graphe de co-occurrence construit à partir de la collection ROMANSEVAL. Dans ces résultats, les deux dernières colonnes présentent respectivement la précision moyenne (MAP) et la précision exacte (R-précision) [Manning *et al.*, 2008]. Les résultats initiaux, avant application de l'expansion, sont présentés dans le scénario nommé *baseline*.

Méthode	**# de termes pour l'expansion**	**MAP**	**R-précision**
Baseline	-	0,5487	0,5174
QE	N	0,4180	0,4043
	N div 2	0,4700	0,4633
	N div 4	**0,5083**	**0,4742**

Table 6.1 – Résultats de l'expansion de requête possibiliste en utilisant le graphe de co-occurrence

Les résultats expérimentaux, présentés dans le tableau 6.1, montrent une perte de performance de RI proportionnellement au nombre des termes d'expansion pour les deux métriques MAP et R-précision.

En se référant aux résultats des courbes Rappel/Précision, présentés dans la figure 6.3, les 3 scénarios d'expansion de requêtes ne sont pas satisfaisants en les comparant à la courbe associée à la requête initiale (*baseline*). Cependant, nous pouvons affirmer que l'application de l'expansion (notamment pour le scénario *(N div 4)*) donne de meilleures performances pour les premiers documents retournés avec des rappels <=0,1 ; c-à-d. ces scénarios sont initialement meilleurs à trouver des documents pertinents.

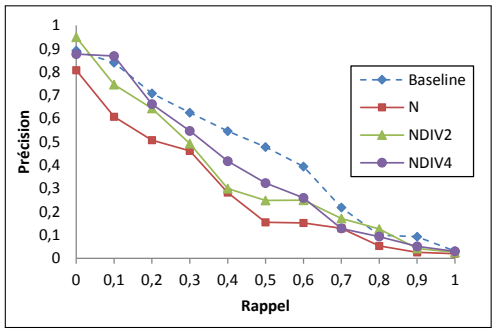


Figure 6.3 – Courbes Rappel/Précision pour la QE possibiliste en ajoutant N, *(N div 2)* et *(N div 4)* termes

5. L'opérateur *div* représente le résultat de la division euclidienne de deux entiers.

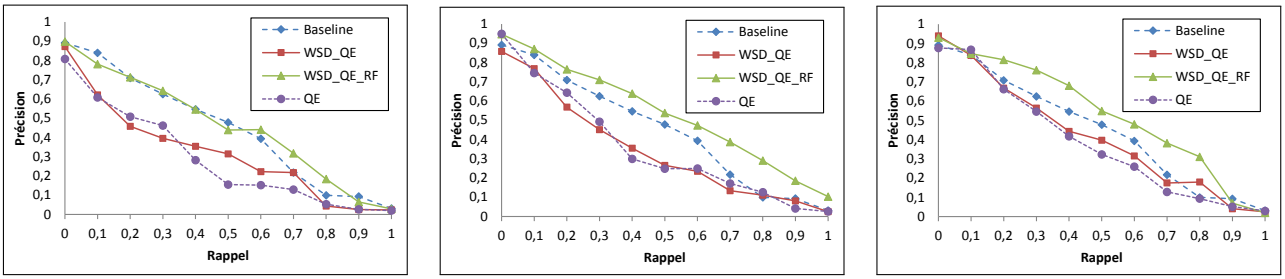
Ces résultats sont affectés par le degré d'ambiguïté des requêtes et la difficulté à pouvoir distinguer les sens corrects des mots ambigus. En effet, plus la requête est longue, plus les performances de RI sont faibles.

6.1.3.3 Combinaison des approches possibilistes de désambiguïsation et d'expansion de requête

Afin d'examiner l'effet de la tâche de désambiguïsation sur la RI, nous avons appliquée une phase de désambiguïsation sémantique avant l'expansion de requêtes. En effet, cette tâche permet de sélectionner les meilleurs sens des mots ambigus ; ce qui résulte dans la réduction de bruit dans les résultats retournés par le SRI.

Pour ce faire, les termes composant chaque meilleur sens, sélectionné par calcul possibiliste, sont injectés dans la requête avant l'application de l'expansion (scénario « *WSD_QE* »). Nous avons appliqué également la technique de pseudo-rétroaction de pertinence dans nos expérimentations en fin du processus de traitement de la requête (scénario « *WSD_QE_PRF* »).

Les résultats des différents scénarios sont présentés dans les figures 6.4.a, 6.4.b et 6.4.c.



(a) Ajout de N termes **(b)** Ajout de *(N div 2)* termes **(c)** Ajout de *(N div 4)* termes

Figure 6.4 – Courbes Rappel/Précision en ajoutant N, *(N div 2)* et *(N div 4)* termes avec application de l'expansion, la désambiguïsation et la rétroaction de pertinence.
Légende :
- *Baseline* : Scénario des requêtes initiales sans traitement.
- *QE* : Application de l'expansion de requêtes.
- *WSD_QE* : Application de la désambiguïsation avant expansion de requêtes.
- *WSD_QE_RF* : Application de la désambiguïsation avant expansion de requêtes et pseudo-rétroaction de pertinence.

L'étude des trois résultats en variant le nombre de termes ajoutés (N, *(N div 2)* et *(N div 4)*) montre que l'application de la désambiguïsation avec l'expansion a des améliorations mineures en les comparants aux résultats de l'expansion sans désambiguïsation. Néanmoins, ces scénarios affectent négativement les résultats des requêtes initiales (scénario « *Baseline* »).

Cependant, avec l'application de la pseudo-rétroaction de pertinence (PRF) après désambiguïsation sémantique et expansion de requêtes, les performances de RI s'améliorent

notamment pour un nombre limité de termes ajoutés (voir figures 6.4.b et 6.4.c).

Nous pouvons ainsi constater l'impact positif de l'application de la désambiguïsation sur l'expansion de requêtes notamment pour les premiers niveaux de rappel (<10%). Nous constatons également l'apport de l'application de *PRF* dans l'amélioration de performance de RI. Dans [Paskalis et Khodra, 2011], le même effet positif d'application de méthode de pseudo-rétroaction de pertinence a été discuté. Nous rejoignons également les interprétations de [Pinto et Pérez-sanjulián, 2008] sur l'utilisation des requêtes longues dans la RI qui génèrent du bruit lors de l'expansion.

6.1.3.4 Approche probabiliste à base de dénombrement de circuits pour la désambiguïsation et l'expansion de requêtes

Nous avons utilisé pour le calcul de proximité sémantique l'approche à base de dénombrement de circuits initié par [Elayeb *et al.*, 2009, Elayeb, 2009] dans le cadre d'étude de l'expansion de requêtes et son effet sur un SRI possibiliste. Les différents termes constituant le graphe construit [6] entretiennent des relations qui incluent parfois des circuits.

Ce facteur est assez important pour assurer l'existence de lien sémantique significatif entre deux mots en examinant la distance (en termes de nœuds du graphe) qui les sépare, et donc la longueur du circuit les reliant. En effet, la richesse d'une langue se traduit par la complexité du dictionnaire qui lui est associe. La multitude de sens associés à un mot donné dans un dictionnaire se traduit par une inter-connectivité accrue entre les nœuds du graphe associé au dictionnaire.

Ainsi, pour un terme donnée t_i d'une requête initiale Q, nous calculons le score de proximité sémantique de ce terme avec n'importe quel autre terme t_j selon la formule (6.5) [Elayeb *et al.*, 2015a] :

$$ProxSém(t_i, t_j) = \frac{NombreCircuits(t_i, t_j)}{Nombre\ maximum\ de\ circuits\ dans\ le\ graphe} \tag{6.5}$$

– Avec $NombreCircuits(t_i, t_j)$ représente le nombre de circuits commençant par le nœud t_i et retournant vers ce même nœud ; et passant par le nœud t_j (c-à-d $t_i \rightarrow ... \rightarrow t_j \rightarrow ... \rightarrow t_i$).

Pour assurer la tâche de désambiguïsation de requêtes, nous calculons la proximité sémantique d'un sens S_i par rapport à une requête Q comme suit :

$$ProxSém(S_i, Q) = \sum_{s_{ij} \in S_i} \sum_{t_k \in Q} ProxSém(s_{ij}, t_k) \tag{6.6}$$

6. décrit auparavant dans la section 6.1.1, page 129.

La longueur maximale de circuit est un des facteurs influant le calcul à base de dénombrement de circuit. En effet, plus le circuit est plus long, plus il y a des chances pour mélanger différentes composantes hétérogènes de sens. Cependant, la prise en compte de circuits trop courts uniquement aurait pour effet de scinder une même composante de sens en plusieurs. Une longueur maximale de quatre arcs représente le meilleur compromis selon [Elayeb, 2009].

6.1.3.5 Exemple illustratif

Considérons l'exemple de requête suivante tout en admettant qu'elle contient un terme ambigu [Ben Khiroun *et al.*, 2014] :

« *Les règles d'orthographe et de ponctuation pour la langue allemande ont été considérablement simplifiées.* »

La requête subit une phase de *tokenisation* et de *lemmatisation* en ignorant les mots vides (tels que les pronoms, les articles, etc.) comme suit :

[*règle, orthographe, ponctuation, langue, allemand, considérable, simple*]

Le résultat de requête après pré-traitement contient le terme ambigu « *simple* ». Par conséquent, le module de désambiguïsation est lancé et le sens ayant le meilleur score de proximité possibiliste parmi les sens issus du dictionnaire de ROMANSEVAL sera sélectionné (dans cet exemple, considérons le sens « AII1 ») :

AI2 *Qui n'est formé que* [...]

AI3 *Qui suffit à soi seul* [...]

AII1 *Qui est facile à comprendre* [...]

Ainsi, les termes existants dans la définition « AII1 » sont injectés dans la requête selon l'approche possibiliste.

En contre partie, considérons l'exemple résumé d'un graphe schématisé dans la figure 6.5 pour calculer la proximité sémantique en utilisant l'approche à base de dénombrement de circuits.

En énumérant le nombre de circuits des trois sens « AI2 », « AI3 » et « AII1 », le sens « AI3 » contenant les mots « *seul* », « *soi* » et « *suffisant* » possède le score de proximité sémantique le plus élevé. Ainsi, ce sens est considérée comme le plus adéquat et les termes qui le constituent sont ajoutés à la requête avant l'expansion.

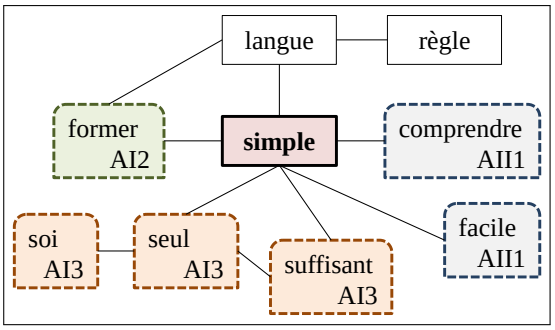


Figure 6.5 – Exemple de graphe de co-occurrence correspondant au terme « simple »

6.1.3.6 Comparaison des approches possibiliste et probabiliste de désambiguïsation et d'expansion de requêtes

Afin d'évaluer les deux approches possibiliste et à base de dénombrement de circuits, nous exposons les résultats globaux en terme de précision moyenne et de précision exacte dans le tableau 6.2. En effet, 4 scénarios ont été exécutés en appliquant l'expansion de requête (QE) seule et avec désambiguïsation (WSD) pour chaque approche. Le scénario *baseline* représente les résultats des requêtes initiales sans traitement.

Méthode		**MAP**	**R-précision**
Possibiliste	avec QE	0,5083	0,4742
	avec WSD & QE	**0,5124**	**0,4760**
À base de dénombrement de circuits	avec QE	0,4920	0,4633
	avec WSD & QE	0,5071	0,4642
Baseline	-	0,5487	0,5174

Table 6.2 – Résultats généraux de l'application de WSD et de QE pour les deux approches possibiliste et probabiliste (à base de dénombrement de circuits)

L'application de l'expansion de requêtes montre une dégradation des performances de RI en terme de précision moyenne et précision exacte pour tous les tests réalisés par rapport aux résultats de base. Néanmoins, les résultats de l'expansion possibiliste montrent une légère amélioration par rapport à ceux de l'approche à base de dénombrement de circuits. L'application d'une phase de désambiguïsation des requêtes contribue à l'amélioration des résultats dans les différents tests.

En fait, cette dégradation des performance de RI en appliquant l'expansion de requêtes (sans ou avec désambiguïsation) peut être justifiée par la génération de taux de bruits importants dans les documents retournés ; et par conséquent une baisse de précision.

[Ogilvie *et al.*, 2009] ont souligné que le nombre de termes d'expansion pour une précision optimale dépend fortement des SRI et des standards de test utilisés (et principalement l'ensemble de requêtes de test). Pinto et Pérez-sanjulián démontrent également que l'application d'expansion sur des requêtes longues, contenant plus de 10 mots, in-

duit à l'injection de taux de bruit important ; ce qui rend les résultats non interprétables [Pinto et Pérez-sanjulián, 2008].

Ainsi, nous avons limité le nombre de termes à ajouter au quart du nombre des termes qui composent la requête afin de réduire le phénomène de bruit en se référant aux premières expérimentations réalisées dans les sous-sections 6.1.3.2 et 6.1.3.3.

Dans le but de mener une étude plus détaillée, nous avons raffiné les expérimentations en traçant les courbes Rappel/Précision présentées dans la figure 6.6.

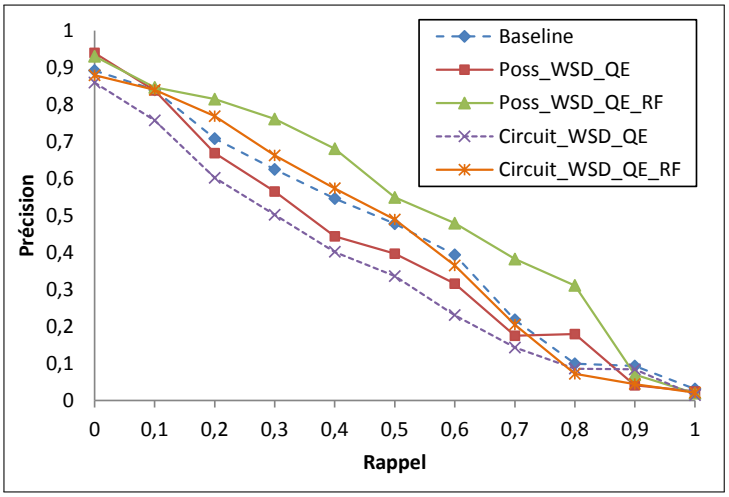


Figure 6.6 – Courbes Rappel/Précision entre approche possibiliste et à base de dénombrement de circuits avec application de la QE, la WSD et la rétroaction de pertinence.

Dans ces expérimentations, nous comparons l'effet de l'application de rétroaction de pertinence après la désambiguïsation et l'expansion de requêtes (voir les deux scénarios possibiliste *Poss_WSD_QE_RF* et probabiliste à base de dénombrement de circuits *Circuit_WSD_QE_RF*). Nous constatons l'apport positif de la rétroaction de pertinence dans les processus de RI surtout pour l'approche possibiliste. En effet, cette approche se distingue, par rapport à la méthode de dénombrement de circuits, par une double mesure de proximité sémantique lors de la recherche dans le graphe de co-occurrence des termes et des sens pour l'expansion sémantique et la désambiguïsation des requêtes respectivement.

6.1.4 Synthèse et discussion

Ce travail présente et compare des approches possibiliste et probabiliste basées sur une ressource de graphe de co-occurrence. Ainsi, nous avons comparé l'impact de la désambiguïsation des mots dans la performance de RI lors de l'application de l'expansion des requêtes et de la rétroaction de pertinence. Le graphe utilisé dans les différentes approches a été construit à partir des documents de la collection de test ROMANSEVAL.

Par ailleurs, Pinto et Pérez-sanjulián ont exploité WordNet comme ressource linguistique externe pour la désambiguïsation et l'expansion de requêtes [Pinto et Pérez-sanjulián, 2008] . Ils ont prouvé la nécessité d'incorporer une tâche de désambiguïsation dans le processus d'expansion afin d'augmenter la performance de RI. Les résultats expérimentaux d'application de désambiguïsation avec expansion de requêtes sont obtenus en utilisant des requêtes courtes et longues de la collection de test TREC-8 avec des valeurs de précisions moyennes (MAP) égales à 0,2030 et 0,1577 respectivement. Ces résultats ont confirmé que l'expansion appliquée seule sur des requêtes courtes (MAP=0,1655) et longues (MAP=0,0628) n'est pas suffisante pour améliorer la RI. Ces résultats montrent également que l'application d'expansion sur des requêtes longues dégrade la performance de RI faute de présence de l'effet de bruit.

D'autre part, Paskalis et Khodra ont proposé et évalué dans [Paskalis et Khodra, 2011] plusieurs scénarios de RI en utilisant la désambiguïsation (WSD), l'expansion de requête (QE), le stemming et une technique de rétroaction de pertinence. Pour la tâche de WSD, ils ont étudié une implémentation étendue de l'algorithme de *Lesk* [Banerjee et Pedersen, 2002] afin d'identifier les sens de chaque requête et les termes du document. Pour la tâche de QE, ils ont tout d'abord exploité un thésaurus basé sur la co-occurrence et construit automatiquement à partir de la collection de documents. En second lieu, ils ont profité d'une technique de PRF en utilisant un ensemble de documents pertinents de premier ordre afin d'en extraire certains termes représentatifs/pertinents. Ces termes sont finalement injectés dans la requête d'origine pour améliorer le processus d'expansion. En rejoignant la conclusion des travaux de [Paskalis et Khodra, 2011], nous avons montré la contribution importante de la rétroaction de pertinence en parallèle avec la désambiguïsation et l'expansion de requêtes dans l'amélioration de la recherche d'information.

Nous notons, dans ce cadre, que nos résultats ne peuvent pas cependant se comparer avec ces travaux cités vu la différence des ressources linguistiques utilisées. Afin d'étendre l'étude de la contribution des tâches de désambiguïsation et d'expansion de requête dans la RI, nous élargissons le cadre d'étude d'un contexte monolingue vers un cadre translinguistique. Cette étude fera l'objet de la deuxième partie de ce chapitre et sera détaillée dans la section 6.2.

6.2 Désambiguïsation et expansion des requêtes en RI translinguistique

On s'intéresse dans cette deuxième section du chapitre à la RI translinguistique (RIT) dans laquelle la requête est représentée dans une langue source et la collection des do-

cuments est représentée dans une autre langue cible. Ainsi, la problématique principale dans la RIT est de faire l'appariement adéquat entre les requêtes et les documents écrits dans deux langues différentes ; ceci passe obligatoirement par une phase de traduction.

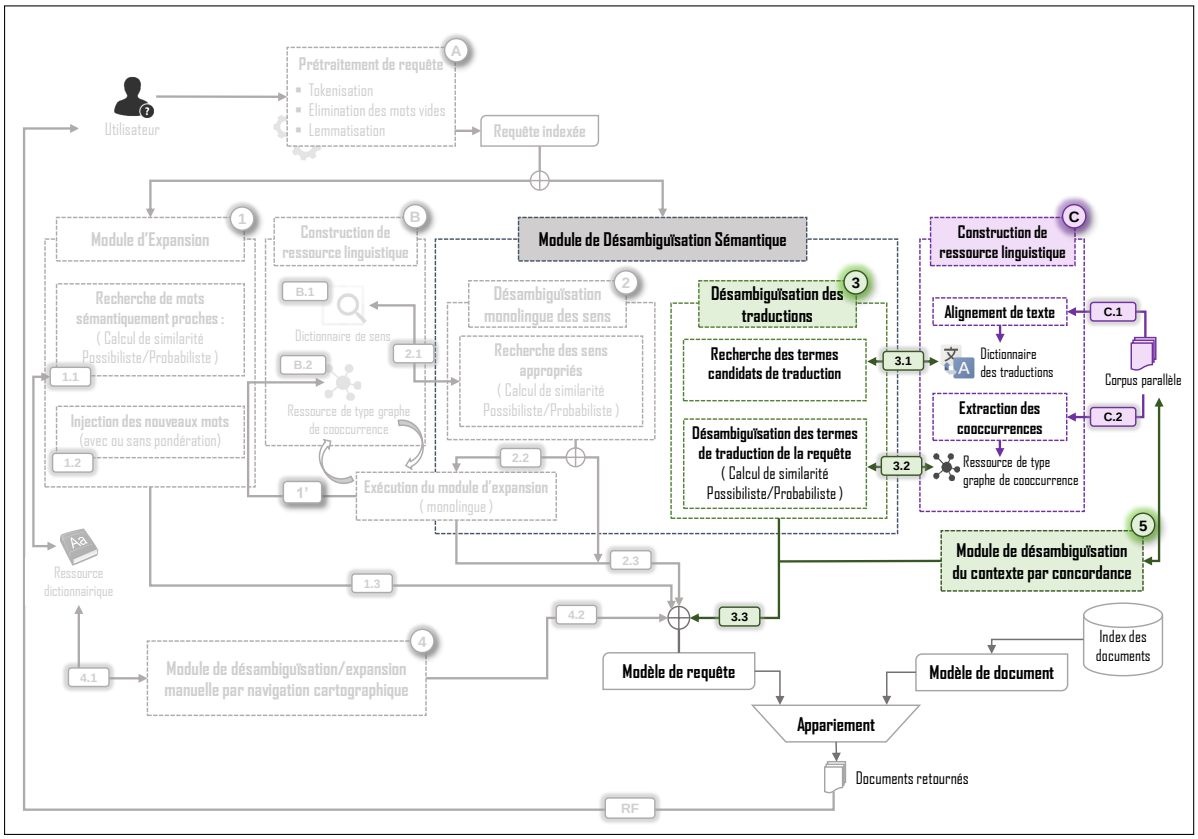


Figure 6.7 – Positionnement de l'approche de traduction dans l'architecture générale du système SPEEDSER

La traduction des requêtes vers la langue cible, dans laquelle sont écrits les documents, représente l'approche la plus courante dans les travaux de RIT. En effet, cette approche est réalisée avec un coût plus optimisé du fait que les requêtes sont généralement limités à quelques termes. Le choix des approches de traduction dépend fortement de la disponibilité des ressources utilisées dans la traduction.

Cependant, il est important à souligner que l'utilisation d'une seule ressource de traduction est souvent insuffisante [7]. En effet, les termes des requêtes ne sont pas nécessairement couverts par une ressource de traduction.

Afin d'étendre nos travaux réalisés en passant du contexte monolingue vers un cadre translinguistique de RI, nous étudions l'effet de la désambiguïsation des traductions dans la RIT tout en analysant l'apport de la rétroaction de pertinence sur la performance globale de RI.

Ainsi, nous essayerons de répondre aux deux points suivants :

7. voir discussion dans la section 1.2.3 page 19

– Comment extraire les termes candidats de traduction de la requête source ?
– Comment choisir la meilleure traduction si plusieurs termes candidats de traduction ont été identifiés ?

6.2.1 Architecture du modèle

Nous présentons dans cette section notre modèle de désambiguïsation des traductions et d'expansion des requêtes en RIT. Les différentes ressources et étapes de cette tâche sont schématisées dans la figure 6.8 [Ben Khiroun *et al.*, 2018].

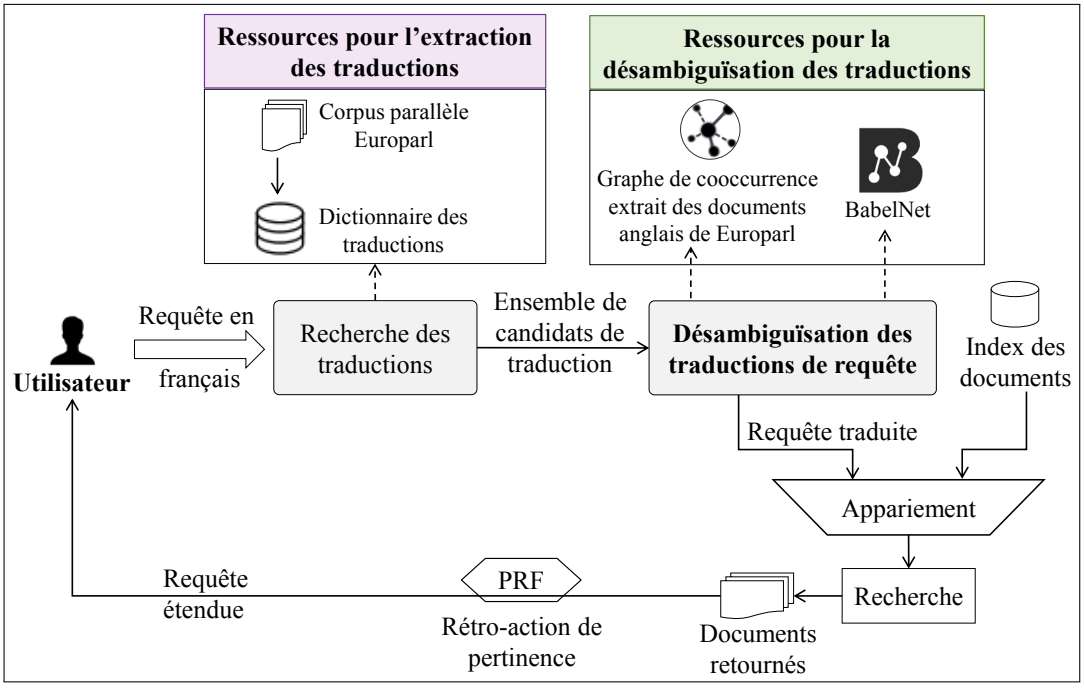


Figure 6.8 – Processus général de désambiguïsation et d'expansion des traductions pour la RIT

En partant d'une requête initiale écrite en français (qui représente la langue source), un ensemble de candidats de traduction en anglais est extrait à partir d'un dictionnaire personnalisé dont le processus de construction est décrit dans la section 6.2.1.1.

Si des mots présentent une ambiguïté lors de leurs traductions, le module de désambiguïsation traite ces mots afin de sélectionner le candidat de traduction le plus adéquat selon le contexte. Dans cette phase, deux ressources (une lexicale et l'autre statistique) sont sollicitées pour le choix des traductions pertinentes (voir plus de détails dans la section 6.2.1.2).

La rétroaction de pertinence est appliquée en fin du processus via l'extraction des termes les plus discriminants des premiers documents retournés et tout le processus peut être réitéré.

6.2.1.1 Extraction des candidats de traduction

Afin d'extraire les termes candidats de traduction associés aux mots d'une requête, nous construisons un dictionnaire bilingue par alignement des textes en français avec les textes correspondants en anglais du corpus Europarl. En effet, ce corpus contient des textes parallèles dans plus de 11 langues qui sont générés à partir des procès du parlement européen [Koehn, 2005]. Europarl a été conçu initialement pour la recherche en traduction automatique statistique de langue. Cependant, il est utilisé dans d'autre domaine d'application à savoir le traitement automatique de langue (TAL) et la désambiguïsation sémantique (WSD).

Pour mettre en place l'alignement des textes de Europarl au niveau mots, nous avons utilisé l'outil GIZA++ [8]. En effet, ce dernier est capable d'extraire des traductions possibles d'un mot cible et leurs probabilités d'occurrence correspondantes. Pour ce faire, il utilise un corpus parallèle comme base de connaissances (dans notre cas le corpus Europarl). Dans la première étape, l'outil GIZA++ effectue un alignement de mots sur le corpus initial, sans prétraitement. Une fois que l'alignement est fait, une table de probabilité est générée ; dans laquelle chaque mot dans la langue source (en l'occurrence le français) est relié à chacune de ses traductions possibles dans la langue cible (dans ce cas l'anglais) en attribuant une probabilité d'occurrence.

Dans notre dictionnaire de traduction proposé, les couples de termes, extraits dans les deux langues source et cible, sont structurés dans un fichier en format CSV qui constitue un format facile à exploiter et à enrichir. Néanmoins, le dictionnaire construit pose un problème de couverture. En effet, nous avons extrait un ensemble réduit de 717 mots en français qui sont inclus dans les requêtes du standard CLEF-2003. Le nombre final des termes candidats de traduction en anglais est égal à 2324 ; soit une moyenne de 3,2 traductions par mot.

6.2.1.2 Désambiguïsation des traductions de requête

Afin de procéder à la phase de désambiguïsation des candidats de traduction, nous avons utilisé deux types de ressources : une ressource statistique et une ressource lexicale.

Dans un premier lieu, nous avons extrait un graphe de co-occurrence issu des documents anglais du corpus Europarl. Chaque mot est relié avec les autres mots existant dans la même phrase [9]. Nous nous sommes basés sur l'hypothèse que si deux termes co-occurrent, alors ils tendent à être dépendants sémantiquement [Church et Hanks, 1990, Cao *et al.*, 2005].

8. http://www.statmt.org/moses/giza/GIZA++.html
9. nous considérons toute la phrase comme contexte de co-occurrence (appelé aussi *fenêtre*).

En deuxième lieu, nous avons utilisé BabelNet, qui est considérée comme une ressource lexicale riche intégrant des connaissances lexicographiques et encyclopédique issues respectivement à partir de WordNet et Wikipedia [Navigli et Ponzetto, 2012]. La construction de BabelNet a été réalisée en appliquant une correspondance (ou *mapping*) entre Wikipedia et WordNet dans une première phase. En deuxième phase, il y a eu recours aux systèmes de traduction automatique ; ce qui a permis de recueillir une grande quantité de concepts multilingues complétant ainsi les traductions manuellement éditées dans Wikipedia. Par conséquent, BabelNet peut être considérée comme un dictionnaire semi-automatique, puisque l'information multilingue comprend à la fois des traductions manuelles de Wikipedia et des traductions obtenues par la traduction automatique. De la même façon que WordNet, BabelNet regroupe les mots en différentes langues par groupes de synonymes appelés *Babel Synsets*. Pour chaque *Babel Synset*, BabelNet fournit des définitions textuelles (appelées *gloses*), obtenues à partir de WordNet et Wikipedia. Dans notre approche proposée, nous considérons que tous les mots composants un *Babel Synset* sont reliés sémantiquement.

Ces deux types de ressources, décrites dans cette section (graphe de co-occurrence et BabelNet), sont utilisées dans la représentation des vecteurs sémantiques qui sera l'objet de la section suivante.

6.2.2 Proposition d'une approche possibiliste de désambiguïsation des traductions

Considérons une requête dans une langue source noté $Q^{(src)} = \{T_1^{(src)}, T_2^{(src)}, \ldots, T_n^{(src)}\}$ et composé de n termes. Chaque terme $T_i^{(src)}$ de la requête peut avoir éventuellement une à plusieurs traductions possibles dans une autre langue cible.

Pour sélectionner les termes candidats de traduction dans l'approche que nous proposons, nous utilisons un dictionnaire bilingue construit à partir de l'alignement des textes en français (qui constitue la langue source) avec les textes anglais correspondant (langue cible) de la collection Europarl (comme nous l'avons décrit dans la section 6.2.1.1).

Notons $\Phi(T_i^{(src)}) = \{T_{ij}^{(cible)},\ j \in [1..m]\}$ l'ensemble des m traductions candidats pour le terme $T_i^{(src)}$.

Nous désignons par *vecteur de contexte,* pour un terme $T_i^{(src)}$, l'union des ensembles de termes candidats de traduction des termes $T_k^{(src)} \neq T_i^{(src)}$ formalisé comme suit [Ben Khiroun *et al.*, 2018] :

$$VC_i = \{\bigcup \Phi(T_k^{(src)}),\ k \neq i \text{ et } k \in [1..m]\} \tag{6.7}$$

Nous désignons par *vecteur sémantique* d'un candidat de traduction $T_{ij}^{(cible)}$ l'ensemble des termes extraits à partir du graphe de co-occurrence ou l'ensemble des *synsets* issus de BabelNet comme décrit auparavant dans la section 6.2.1.2.

Le vecteur sémantique est formulé par :

$$VS_{ij} =< s_{ij1}^{(cible)}, s_{ij2}^{(cible)}, ..., s_{ijk}^{(cible)} > \tag{6.8}$$

La pertinence d'un vecteur sémantique, désigné par VS_{ij} et relatif à une traduction $T_{ij}^{(cible)}$, par rapport au vecteur de contexte de la requête, est calculée en étendant le modèle possibiliste utilisé dans [Ben Khiroun *et al.*, 2012, Ben Khiroun *et al.*, 2014].

Nous adaptons ainsi le modèle d'appariement entre requête/document au calcul de pertinence entre un ensemble de termes représentant sémantiquement un candidat de traduction d'une part et un vecteur de contexte représentant les autres termes candidats de traduction, d'autre part, tout en utilisant une double mesure de pertinence :

La *pertinence possible* permet de négliger les traductions non pertinentes à une requête donnée. La *pertinence nécessaire* renforce, en contre partie, la nécessité de faire figurer les termes candidats pertinents de traduction dans la traduction finale de la requête.

Afin de désambiguïser les traductions possibles, nous calculons le score possibiliste de chaque vecteur sémantique, qui représente une traduction, par rapport au vecteur de contexte de la requête, qui inclue les traductions candidats associés aux autres termes, comme suit :

La mesure de possibilité $\Pi(VS_{ij}|VC_i)$ est proportionnelle à :

$$\Pi'(VS_{ij}|VC_i) = \Pi(w_1|VS_{ij}) \times \ldots \times \Pi(w_p|VS_{ij}) = nft_{1ij} \times \ldots \times nft_{pij} \tag{6.9}$$

- avec : $nft_{kij} = tf_{kij}/max(tf_{kij})$ représente la fréquence normalisée du terme de traduction $w_k \in VC_i$ dans le vecteur sémantique VS_{ij} associé au terme candidat de traduction $T_{ij}^{(cible)}$;
- et $tf_{kij} = \frac{nombre\ d'occurrence\ du\ terme\ w_k\ dans\ VS_{ij}}{nombre\ de\ termes\ dans\ VS_{ij}}$.

La certitude (ou nécessité) de restituer une traduction pertinente $T_{ij}^{(cible)}$ pour un contexte des termes traduits, notée $N(VS_{ij}|VC_i)$, est donnée par :

$$N(VS_{ij}|VC_i) = 1 - \Pi(\neg VS_{ij}|VC_i) \tag{6.10}$$

De même $\Pi(\neg VS_{ij}|VC_i)$ est proportionnelle à :

$$\Pi'(\neg VS_{ij}|VC_i) = \Pi(w_1|\neg VS_{ij}) \times \ldots \times \Pi(w_p|\neg VS_{ij}) \tag{6.11}$$

La mesure de nécessité peut être exprimée par :

$$\Pi'(\neg VS_{ij}|VC_i) = (1 - \phi_1(T_{ij}^{(cible)})) \times \ldots \times (1 - \phi_p(T_{ij}^{(cible)})) \tag{6.12}$$

Où :

$$\phi_k(T_{ij}^{(cible)}) = Log_{10}(\frac{nCT_i}{nT_{ik}}) \times nft_{kij} \tag{6.13}$$

- avec : nCT_i = nombre des candidats de traduction du terme $T_i^{(src)}$ de la requête initiale ;
- et nT_{ik} = nombre des candidats de traduction du terme $T_i^{(src)}$ contenant le terme $w_k \in VC_i$.

Nous définissons le *Degré de Pertinence Possibiliste (DPP)* de chaque traduction $T_{ij}^{(cible)}$ étant donné un contexte de termes de traduction VC_i par :

$$DPP(VS_{ij}|VC_i) = \Pi(VS_{ij}|VC_i) + N(VS_{ij}|VC_i) \tag{6.14}$$

Les traductions préférées sont celles qui ont les valeurs DPP les plus élevées.

Nous résumons les différentes étapes de notre approche proposée dans l'algorithme 2 [Ben Khiroun *et al.*, 2018].

Algorithme 2 : Désambiguïsation des traductions des requêtes

Entrées : requête $Q^{(src)}$ dans une langue source composé de n termes
Sorties : requête $Q^{(cible)}$ traduite dans une langue cible

```
début
    pour chaque terme T_i^(src) ∈ Q^(src) faire
        construire le vecteur de contexte VC_i
        pour chaque traduction possible T_ij^(cible) ∈ Φ(T_i^(src)) faire
            extraire le vecteur sémantique VS_ij depuis une ressource
            calculer le score possibiliste de VS_ij par rapport à VC_i
        fin
        ajouter la meilleure traduction à Q^(cible)
    fin
fin
```

6.2.3 Expérimentations et étude comparative

Dans cette section, nous évaluons et comparons la contribution de l'approche possibiliste de désambiguïsation des traductions en utilisant les deux ressources, à savoir de types statistique et lexical, présentées auparavant.

6.2.3.1 Collection de test

Afin d'évaluer l'approche de désambiguïsation des traduction, nous avons utilisé la collection standard de test CLEF-2003. Cette collection est adaptée pour la validation de recherche d'information monolingue tel que ce qui a été expérimenté dans la première partie de ce chapitre (voir section 6.1, page 128). Cette même collection CLEF-2003 offre également un cadre de validation translinguistique en proposant des documents dans différentes langues [10] [Braschler et Peters, 2004]. Le tableau 6.3, présente la répartition des articles en anglais et en français dans la collection CLEF-2003.

Collection		**Taille (Mo)**	**Nombre de documents**	**Nombre moyen de mots par document**
Langue	**Sources**			
Anglais	LA Times 94	425	113005	421
	Glasgow Herald 95	154	56472	343
Français	Le Monde 94	158	44013	361
	ATS 94	86	43178	227
	ATS 95	88	42615	234

Table 6.3 – Statistiques sur les documents anglais et français de CLEF-2003

Le standard de test CLEF-2003 fournit ainsi un environnement de validation pour la recherche d'information translinguistique en incluant un ensemble de documents, un ensemble de requêtes de test et une liste de jugement de pertinence des documents pour chaque requête. Chaque requête est représentée sous format XML par un titre (`<title>`), une description moyenne (`<description>`) et une partie narrative (`<narra>`) spécifiant plus précisément le contexte de la requête [11]. La figure 6.9 présente un exemple de requête (ou *topic*).

```
<top>
  <num> C148 </num>
  <EN-title> Damages in Ozone Layer </EN-title>
  <EN-desc> What holes in the ozone layer are not an effect of pollution?
      </EN-desc>
  <EN-narr> Not all damage to the ozone layer is caused by pollution.
     Relevant documents will  give information on other causes for holes
     in the ozone layer. </EN-narr>
</top>
```

Figure 6.9 – Exemple de requête (*topic*) en anglais du standard CLEF-2003

Les expérimentations, qui suivent, ont été réalisées en utilisant la plate-forme de RI Terrier avec OKAPI BM25 comme modèle d'appariement entre requête-document [Ounis

10. Ces documents existant dans plusieurs langues ont des propriétés communes : mêmes périodes, sujets et contenus des documents comparables. Les langues supportées dans CLEF-2003 : néerlandais, anglais, finnois, français, allemand, italien, russe, espagnol et suédois.

11. Nous nous avons utilisé uniquement la partie `<title>` pour la réalisation de nos expérimentations afin de minimiser l'effet de *bruit*

et al., 2007, Macdonald *et al.*, 2012].

6.2.3.2 Évaluation de l'approche possibiliste de traduction de requêtes

Nous comparons dans la figure 6.10 notre approche possibiliste pour la désambiguïsation des traductions de requête en se basant sur deux ressources (comme décrit dans la section 6.2.1.2 page 142). Les courbes désignées par « *Cooccurrence* » et « *Babelnet* » se réfèrent respectivement à la ressource statistique de type graphe de co-occurrence et à la ressource lexicale extraite à partir des synsets de BabelNet.

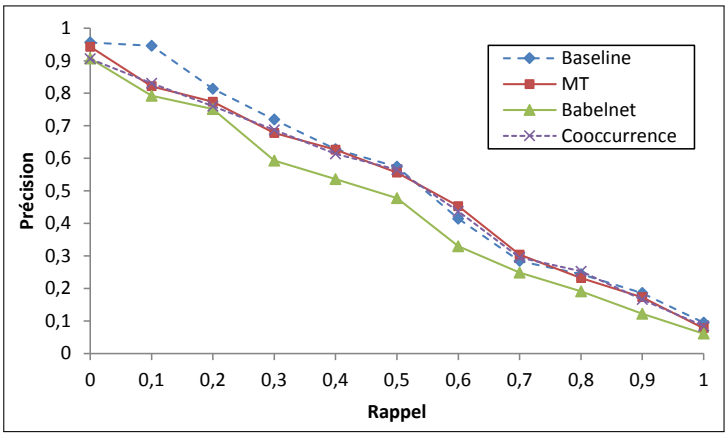


Figure 6.10 – Courbe Rappel/Précision des méthodes de désambiguïsation des traductions sans application de rétroaction de pertinence.
Légende :
- *Baseline* : Scénario des requêtes d'origine dans CLEF-2003 (requêtes écrites en anglais).
- *MT* : Application de traducteur automatique (*Google Translate*).
- *Babelnet* : Application de la désambiguïsation des traductions possibiliste à base de BabelNet.
- *Cooccurrence* : Application de la désambiguïsation des traductions possibiliste à base de graphe de co-occurrences.

La courbe désignée par « *baseline* » représente les valeurs de précision des requêtes d'origine écrite en anglais dans le standard de test CLEF-2003. Nous introduisons également les résultats de traduction en utilisant le traducteur automatique *Google Translate* [12] dans la courbe désignée par *MT*.

En analysant les courbes Rappel/Précision des 4 méthodes de traduction, nous remarquons que la courbe « *baseline* » reste en moyenne au-dessus des autres courbes. Ceci implique que les 3 autres méthodes ne proposent pas des traductions proches à celles proposées par les évaluateurs humains en anglais. Nous remarquons cependant que la performance de notre approche possibiliste à base de graphe de co-occurrence est presque identique à celle basée sur le traducteur automatique tout en restant au-dessus de la courbe de traduction possibiliste à base de sous-graphe de BabelNet.

12. https://translate.google.com

Comme nous avons distingué l'effet positif de l'application de la rétroaction de pertinence dans un cadre monolingue dans la première partie de ce chapitre, nous expérimentons cette technique en l'appliquant au cadre translinguistique. Nous présentons ainsi, dans la figure 6.11, le rôle de la rétroaction de pertinence dans l'approche possibiliste proposé après traduction des requêtes. Pour se faire, nous avons utilisé le modèle *Bo1* (*Bose-Einstein 1*) implémenté dans la plate-forme Terrier en appliquant la configuration par défaut à savoir : le nombre de termes d'expansion est égal à 10 et le nombre des tops documents en tête de liste, depuis lesquels les termes sont extraits, est limité à 3 documents.

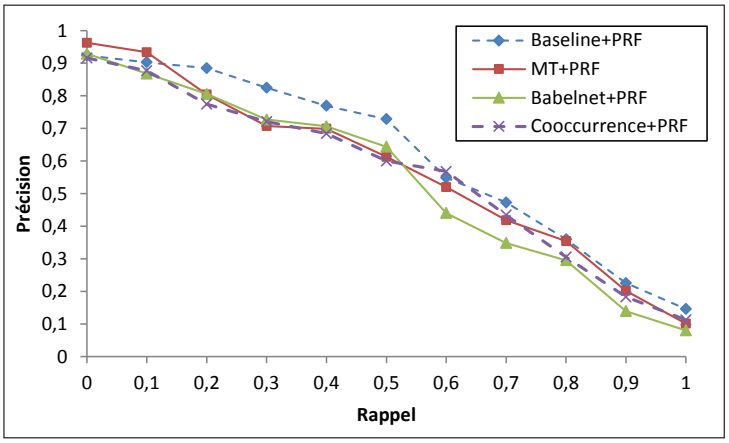


Figure 6.11 – Courbe Rappel/Précision des méthodes de désambiguïsation des traductions avec application de pseudo-rétroaction de pertinence (PRF).

A titre d'observation, nos remarquons que les courbes avec application de PRF sont légèrement au-dessus de celles sans application de PRF avec une petite amélioration en faveur de l'approche à base de BabelNet et celle du scénario « *baseline* ».

Le tableau 6.4 détaille les valeurs de précision aux @5, @10, @15... et @1000 top documents pour les 4 scénarios d'expérimentation en appliquant la rétroaction de pertinence. Les résultats montrent que l'utilisation de graphe de co-occurrence comme ressource de désambiguïsation en appliquant le modèle possibiliste donne de meilleures performances que les autres tests pour les premiers documents retournés. Néanmoins, la performance du traducteur automatique (MT) est meilleure en examinant le reste des derniers documents retournés en fin de la liste.

Dans le but de comparer davantage l'approche basée sur la co-occurrence avec l'approche basée sur BabelNet comme ressource de désambiguïsation et celle basée sur la traduction automatique, nous calculons les mesures *p-valeurs* associées au test des rangs signés de Wilcoxon pour échantillons appariés [Demšar, 2006]. Les *p-valeurs* sont calculées en comparant les couples de mesures de précision de l'approche fondée sur la ressource de co-occurrence vis-à-vis de chacun des deux autres scénarios.

Comme le montre le tableau 6.5, les résultats de la *p-valeur* prouvent que les améliora-

	Baseline	**MT**	**Babelnet**	**Cooccurrence**
P@5	0,4148	0,3741	0,3741	**0,3815**
P@10	0,3333	0,3259	0,3037	**0,3352**
P@15	0,2975	0,2864	0,2691	**0,3012**
P@20	0,2741	0,2657	0,2398	**0,2759**
P@30	0,2309	0,2259	0,2049	**0,2364**
P@50	0,1785	0,1789	0,1596	**0,1848**
P@100	0,118	**0,1185**	0,1056	0,1174
P@200	0,0705	**0,0694**	0,065	0,0683
P@500	0,0334	**0,0315**	0,0304	0,0313
P@1000	0,0175	**0,0164**	0,0162	0,0163

Table 6.4 – Valeurs détaillées des précisions pour les différentes approches de traduction avec application de PRF

tions observées de l'approche de co-occurrence, par rapport aux scénarios MT (*p-valeur* = 0,010301 <0,05) et BabelNet (*p-valeur* = 0,003509 <0,05), sont statistiquement significatives [Biau *et al.*, 2010].

	Co-occurrence vs MT	*Co-occurrence vs BabelNet*
p-valeur	0,010301	0,003509

Table 6.5 – Résultats de la *p-valeur* associée au test des rangs signés de Wilcoxon pour échantillons appariés sur les mesures de précision moyenne

6.2.3.3 Comparaison de l'approche possibiliste de traduction de requêtes avec l'approche à base de dénombrement de circuits

La mesure à base de dénombrement de circuits a été adaptée dans le cadre monolingue comme ça été présenté auparavant dans la section 6.1.3.4 page 135. En effet, dans un cadre monolingue, l'exploitation des relations à base de dénombrement de circuits se contente à mesurer le degré de similarité entre deux termes donnés.

En passant vers un cadre translinguistique, nous appliquons ce même modèle pour désambiguïser les termes de traduction des requêtes en calculant la similarité sémantique d'un terme donné t_i avec un candidat de traduction t_j selon la formule (6.5) page 135.

Afin d'optimiser la recherche de circuit dans un graphe, nous avons extrait un sous-ensemble de *synsets* de BabelNet qui incluent les candidats de traduction. Ainsi, ce sous-ensemble couvre uniquement les traductions qui correspondent aux requêtes de CLEF-2003. Nous avons considéré la longueur maximale de circuit de l'ordre de 4, déduite comme la longueur de circuit optimale par [Elayeb, 2009].

La figure 6.12 montre les valeurs de précision moyenne (MAP) et de R-précision utilisées dans l'évaluation des scénarios de test. La mesure MAP représente la précision moyenne

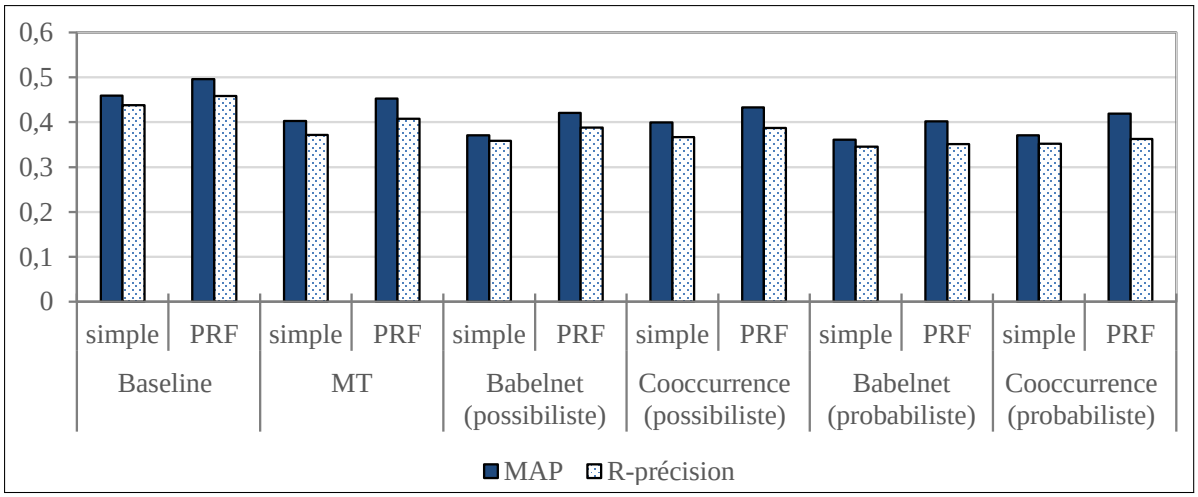


Figure 6.12 – Comparaison des résultats MAP et R-précision des approches de traduction (possibilistes et probabilistes à base de circuits) avec et sans application de PRF (scénarios « PRF » et « simple », respectivement)

des requêtes (*topics*) du standard CLEF-2003. La R-précision définie la précision au rang R [Manning *et al.*, 2008, Baccini *et al.*, 2012].

Comme première observation globale, nous remarquons que les scénarios de tests à base de co-occurrence surpassent légèrement les scénarios basés sur BabelNet. Toutefois, tous les scénarios sont au dessous des performances des scénarios des requêtes d'origine (*baseline*) et ceux de traduction automatique (MT) en considérant les métriques MAP et de R-précision. En effet, la performance des traducteurs automatiques tire son ampleur du fait qu'un grand volume de textes est utilisé dans la phase d'apprentissage de ces outils. Les jugements de pertinence des traductions, effectués par les utilisateurs, contribue également dans la qualité des traductions proposées [Hosseinzadeh Vahid *et al.*, 2015]. La performance des traducteurs automatiques dépend cependant des couples de langues sur lesquelles ils s'appliquent [Tobin, 2015]

En examinant nos résultats de point de vue comparaison possibiliste-probabiliste, les résultats montrent une avance en faveur des scénarios possibilistes par rapport aux scénarios basées sur le dénombrement des circuits.

La contribution des réseaux possibilistes dans la recherche translinguistique, telle que présentée dans d'autres travaux (principalement l'approche basée sur la transformation de la probabilité en possibilité dans [Elayeb *et al.*, 2018] et l'approche possibiliste discriminative de [Ben Romdhane *et al.*, 2017]), rejoint les interprétations faites sur nos résultats expérimentaux actuels.

6.2.4 Synthèse et discussion

Dans un cadre de validation de RI translinguistique, nous avons comparé des approches possibiliste et probabiliste (par dénombrement de circuits) basées sur deux types de ressources. Ces ressources, de nature statistique et lexicale respectivement, sont structurées en graphe via exploitation des relations de co-occurrence dans les textes du corpus Europarl, et respectivement, via extraction des relations sémantiques dans les *synsets* du réseau BabelNet.

L'étude expérimentale a prouvé, d'une part, que l'approche possibiliste proposée donne de meilleurs résultats par rapport à celle basée sur le dénombrement de circuits. D'autre part, l'utilisation de graphes de co-occurrences a permis d'améliorer légèrement les performances de RI par rapport à l'exploitation des sous-réseaux extraits de BabelNet. De plus, l'application de la technique de rétroaction de pertinence a considérablement amélioré les résultats des différents scénarios de test, ce qui rejoint nos travaux antérieurs faites dans un cadre de RI monolingue [Ben Khiroun *et al.*, 2014, Elayeb *et al.*, 2015a] ainsi que les travaux de [Paskalis et Khodra, 2011].

En vue de perspective, nous proposons de résoudre le problème de couverture des mots en raison de la nature du processus d'extraction de dictionnaire bilingue qui est proposé dans notre travail. En fait, les approches fondées sur la connaissance reposant sur des corpus alignés dépendent de la taille et du type de texte analysé. Cela pourrait être un grand défi face au manque de ressources parallèles pour certaines langues comme l'arabe tel que présenté dans [Elayeb et Bounhas, 2016]. Une autre direction potentiellement intéressante consisterait à étudier l'impact de l'application de l'expansion de requête avant et après le processus de traduction. De plus, nous pouvons étudier la contribution des techniques d'expansion de requêtes, autres que la rétroaction de pertinence, comme celles basées sur des dictionnaires ou qui exploitent des relations ontologiques par exemple.

Conclusion

Nous avons établi dans ce chapitre une étude approfondie sur l'apport l'application de la désambiguïsation sémantique et l'expansion de requêtes via la rétroaction de pertinence sur l'amélioration du processus de RI. Afin d'avoir des ressources génériques, nous avons modélisé les connaissances sous forme de graphes de co-occurrence pour extraire les relations sémantiques entre les termes et les sens. Nous avons également établi une étude comparative de notre approche possibiliste avec d'autres approches de l'état de l'art en passant d'un cadre monolingue vers un cadre translinguistique.

Conclusion Générale et Perspectives

Le développement des systèmes de recherche d'information (SRIs) rencontre de nombreux défis, en particulier liés à la nature des requêtes qui sont formulées par l'utilisateur du SRI. En fait, l'utilisateur a tendance à formuler son besoin en information par des requêtes courtes qui peuvent contenir également des termes ambigus. Par conséquent, le manque de contexte proposé dans la requête influence la qualité de recherche en renvoyant des documents non pertinents.

Nous nous sommes intéressés, dans le cadre de cette thèse, au domaine de recherche d'information (RI) et plus particulièrement à la désambiguïsation sémantique et l'expansion des requêtes. Nous avons exposé l'importance de la phase de désambiguïsation sémantique des textes face à la richesse des langues naturelles. En effet, la désambiguïsation est une tâche intermédiaire fondamentale pour la plupart des applications de RI ; ainsi que d'autres applications telles que le traitement automatique du langage naturel (TALN). La désambiguïsation sémantique a pour objectif d'améliorer la pertinence des documents sélectionnés par le SRI. Elle consiste à se focaliser sur le sens dominant de la requête et à se détacher de ses sens secondaires selon le contexte et filtrer ainsi les réponses retournées par le système en favorisant les documents pertinents.

Dans le but de renforcer la pertinence de recherche, le SRI ne doit pas se contenter d'une analyse simple de la collection de documents et d'une mise en correspondance directe, dite appariement, entre les requêtes et les documents. Les techniques d'expansion de requêtes, et en particulier celle de rétroaction de pertinence, sont introduites dans le processus global de RI afin d'améliorer la qualité de la recherche. Le recours aux méthodes d'expansion de requêtes peut être considéré comme une solution d'enrichissement du contexte des requêtes courtes. Néanmoins, l'application d'expansion peut reformuler la requête d'origine en rajoutant des termes ambigus ; d'où la nécessité de recours de nouveau à la désambiguïsation sémantique afin de résoudre cette ambiguïté. Cette relation de dépendance entre la désambiguïsation et l'expansion de requêtes prouvent la nécessité de les combiner ensemble afin d'améliorer la RI.

Principales contributions

Sur le plan théorique

Nous avons proposé une conceptualisation des relations sémantiques via une structure de type dictionnaire sémantique de contexte ($\mathcal{DSC}$) et une structure de type graphe de co-occurrence pour les tâches de désambiguïsation et d'expansion de requête. Ainsi, nous nous sommes basés sur l'extraction des liens sémantiques et l'exploration des relations de corrélations entre les termes à partir des textes analysés. Ceci rend l'approche proposée générique et indépendante des langues et des ressources utilisées.

Sur le volet modèle théorique, nous avons étendu le modèle possibiliste qualitatif de [Elayeb, 2009] pour l'application aux tâches de désambiguïsation et d'expansion de requête. Ce cadre possibiliste renforce la pertinence ainsi que la représentativité d'un terme dans un contexte en proposant une double mesure pertinence. La pertinence possible permet de rejeter les contextes (sens ou termes d'expansion) non pertinents à une requête donnée. La pertinence nécessaire permet de se focaliser sur les contextes à restituer ainsi que de renforcer la nécessité de faire figurer les documents pertinents parmi les premiers résultats en réponse à une requête.

Nous avons étudié la contribution de la désambiguïsation avec l'expansion de requête via rétroaction de pertinence pour l'amélioration du processus de RI. Pour assurer ces deux tâches, nous calculons la similarité entre les termes de requêtes (dans le cas de l'expansion) ou entre les termes et les sens (dans le cas de désambiguïsation) en se référant à une représentation de connaissance en graphe de co-occurrence.

En partant de l'analogie faite par [Nie, 2010] entre l'expansion de requête monolingue et la traduction de requête en RI translinguistique, nous avons projeté l'utilisation des graphes de co-occurrence dans le cadre de RI translinguistique. Ainsi, nous avons démontré l'extensibilité du modèle possibiliste de base vers différentes applications telle que la désambiguïsation des termes candidats de traduction.

Sur le plan pratique et technique

Nous avons proposé une architecture et une implémentation d'un Système Possibiliste d'Expansion Et de Désambiguïsation SEmantique de Requêtes (SPEEDSER) dédié à l'expansion et la désambiguïsation de requêtes en RI monolingue et translinguistique. Ce système intègre des interfaces Homme-Machine pour assister l'utilisateur dans la tâche de reformulation en exploitant le module de navigation dans le graphe de dictionnaire de type réseaux de petits mondes hiérarchiques (RPMH) pour un cadre de RI monolingue. Le système offre également une analyse de concordance des termes ainsi

qu'une recherche des traductions possibles pour assister l'utilisateur dans un cadre de RI et d'analyse linguistique bilingue en français et en anglais.

Sur le plan expérimental

Nous avons établi des études comparatives entre des approches possibilistes avec des approches probabilistes ainsi que d'autres approches de l'état de l'art pour la désambiguïsation des textes, l'expansion de requêtes et la désambiguïsation des traductions des requêtes.

Dans les différentes validations expérimentales, nous avons utilisé des corpus standards de test destinés à la tâche de WSD (ROMANSEVAL) et de RI monolingue et multilingue (CLEF-2003) en appliquant les métriques issus du domaine de RI, à savoir : rappel, précision, F-mesure, MAP, R-précision, Kappa, p-valeur, etc. Ces études nous ont permis d'évaluer la performance de nos approches proposées et de les positionner par rapport aux autres approches.

Les expérimentations menées prouvent que la modélisation des tâches de désambiguïsation et d'expansion de requêtes, à base des réseaux possibilistes, raffine les performances des SRI par rapport aux approches probabilistes telles que PROX [Ben Khiroun *et al.*, 2012, Elayeb *et al.*, 2015b] et l'approche à base de dénombrement de circuit [Ben Khiroun *et al.*, 2014, Elayeb *et al.*, 2015a, Ben Khiroun *et al.*, 2018].

Perspectives

Les spécificités des langues influencent la performance des SRIs ainsi que la facilité de leurs mises en œuvre. A titre d'exemple, la langue arabe est souvent écrite sans utilisation explicite des diacritiques qui sont l'équivalent des voyelles en français [Elayeb et Bounhas, 2016]. La disponibilité des ressources dans une langue donnée, tels que les dictionnaires ou les texte alignés pour la RI translinguistique, a également son impact sur le domaine de RI. Nous envisageons ainsi à étendre le système possibiliste proposé de désambiguïsation et d'expansion de requêtes pour supporter d'autres langues et étudier l'effet de varier la langue sur les approches proposées dans le cadre de cette thèse.

Le problème de couverture reste un défis à résoudre en présence de nouveaux termes de recherche qui ne sont pas nécessairement inclus dans les ressources classique tels que les dictionnaires ou les ontologies. Ceci s'applique également sur les ressources proposées dans cette thèse et qui sont extraites à partir de l'analyse de corpus tels que les graphes de co-occurrence ou le $\mathcal{DSC}$. Une phase d'enrichissement automatique de ces ressources voire l'hybridation de différentes sources de connaissance s'avère intéressante.

L'extension du modèle possibiliste sur d'autres domaines d'application telle que la classification ou l'extraction des entités nommées semble prometteuse. Citons, à titre d'exemple, les travaux de [Ayed *et al.*, 2014] sur l'analyse morphologique des textes arabes via classifieur possibiliste ainsi que les travaux de [Lahbib *et al.*, 2015] sur l'extraction de terminologie de domaine. L'intégration d'autres approches telles que les approches à base de règles d'association et les approches de plongements de mots (« *word embeddings* ») pourrait se révéler adaptée à ces propositions .

Des extensions de notre système SPEEDSER peuvent également avoir lieu en intégrant de nouveaux modules de visualisation de données et d'analyse de texte. La variation de nouvelles ressources pour alimenter le système proposé contribuera à son amélioration dans le but de faciliter la tâche de désambiguïsation et d'expansion de requêtes. A titre d'exemple, les travaux de [Navigli et Ponzetto, 2012] ont profité de la combinaison de WordNet et des articles issus de Wikipedia pour proposer un système interactif de désambiguïsation sémantique tout en profitant de la puissance d'outil de visualisation. La migration du système SPEEDSER vers une architecture orientée service Web et la développement d'APIs d'interfaçage distants contribueront à une meilleure interaction et réutilisation du système.

Bibliographie

[Abdelali *et al.*, 2003] Abdelali, A., Cowie, J., Farwell, D. et Ogden, W. (2003). UCLIR : a Multilingual Information Retrieval Tool. *Inteligencia Artificial : revista iberoamericana de inteligencia artificial*, 8(22):103–110. *Cité page 96*

[Abdul-Jaleel *et al.*, 2004] Abdul-Jaleel, N., Allan, J., Croft, W. B., Diaz, F., Larkey, L., Li, X., Smucker, M. D. et Wade, C. (2004). UMass at TREC 2004 : Novelty and HARD. Rapport technique, Massachusetts Univ Amherst Center For Intelligent Information Retrieval, Massachusetts Univ Amherst Center For Intelligent Information Retrieval. *Cité page 64*

[Adam *et al.*, 2013] Adam, C., Fabre, C. et Tanguy, L. (2013). Etude des relations sémantiques dans les reformulations de requêtes sous la loupe de l'analyse distributionnelle. *In SemDis (enjeux actuels de la sémantique distributionnelle) dans le cadre de TALN 2013*, pages 140–153, Sables d'Olonne, France. *2 citations pages 54 et 55*

[Adriani et Rijsbergen, 1999] Adriani, M. et Rijsbergen, C. J. v. (1999). Term Similarity-Based Query Expansion for Cross-Language Information Retrieval. *In Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science, pages 311–322. Springer, Berlin, Heidelberg. *Cité page 70*

[Agirre *et al.*, 2010] Agirre, E., Arregi, X. et Otegi, A. (2010). Document Expansion Based on WordNet for Robust IR. *In Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, COLING '10, pages 9–17, Stroudsburg, PA, USA. Association for Computational Linguistics. *3 citations pages 40, 63 et 128*

[Agirre et Edmonds, 2007] Agirre, E. et Edmonds, P. (2007). *Word Sense Disambiguation : Algorithms and Applications*, volume 33. Springer Publishing Company, Incorporated, 1 édition. *Cité page 42*

[Agirre *et al.*, 2014] Agirre, E., López de Lacalle, O. et Soroa, A. (2014). Random Walks for Knowledge-based Word Sense Disambiguation. *Comput. Linguist.*, 40(1):57–84. *2 citations pages 39 et 49*

[Agirre *et al.*, 2006] Agirre, E., Martínez, D., de Lacalle, O. L. et Soroa, A. (2006). Two Graph-based Algorithms for State-of-the-art WSD. *In Proceedings of the 2006 Confe-*

rence on Empirical Methods in Natural Language Processing, EMNLP '06, pages 585–593, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cité page 49*

[Agosti *et al.*, 2012] Agosti, M., Crivellari, F. et Nunzio, G. M. D. (2012). Web log analysis : a review of a decade of studies about information acquisition, inspection and interpretation of user interaction. *Data Mining and Knowledge Discovery*, 24(3):663–696. *Cité page 67*

[Ahmed et Nürnberger, 2010] Ahmed, F. et Nürnberger, A. (2010). Multi Searcher : Can We Support People to Get Information from Text They Can'T Read or Understand ? *In Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 837–838, New York, NY, USA. ACM. *Cité page 96*

[Allan *et al.*, 2003] Allan, J., Aslam, J., Belkin, N., Buckley, C., Callan, J., Croft, B., Dumais, S., Fuhr, N., Harman, D., Harper, D. J., Hiemstra, D., Hofmann, T., Hovy, E., Kraaij, W., Lafferty, J., Lavrenko, V., Lewis, D., Liddy, L., Manmatha, R., McCallum, A., Ponte, J., Prager, J., Radev, D., Resnik, P., Robertson, S., Rosenfeld, R., Roukos, S., Sanderson, M., Schwartz, R., Singhal, A., Smeaton, A., Turtle, H., Voorhees, E., Weischedel, R., Xu, J. et Zhai, C. (2003). Challenges in Information Retrieval and Language Modeling : Report of a Workshop Held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, September 2002. *SIGIR Forum*, 37(1):31–47. *Cité page 54*

[Almasri *et al.*, 2014] Almasri, M., Chevallet, J.-P. et Berrut, C. (2014). Exploiting Wikipedia Structure for Short Query Expansion in Cultural Heritage. *In CORIA 2014 - Conférence en Recherche d'Infomations et Applications- 11th French Information Retrieval Conference. CIFED 2014 Colloque International Francophone sur l'Ecrit et le Document*, pages 287–302, Nancy, France. *Cité page 67*

[Anaya-Sánchez *et al.*, 2006] Anaya-Sánchez, H., Pons-Porrata, A. et Berlanga-Llavori, R. (2006). Word Sense Disambiguation Based on Word Sense Clustering. *In Advances in Artificial Intelligence - IBERAMIA-SBIA 2006*, pages 472–481. Springer, Berlin, Heidelberg. DOI : 10.1007/11874850_51. *Cité page 36*

[Atkins et Rundell, 2008] Atkins, B. T. S. et Rundell, M. (2008). *The Oxford Guide to Practical Lexicography.* OUP Oxford, Oxford ; New York. *2 citations pages 19 et 99*

[Audeh, 2014] Audeh, B. (2014). *Reformulation sémantique des requêtes pour la recherche d'information ad hoc sur le Web.* Thèse de doctorat, Ecole Nationale Supérieure des Mines de Saint-Etienne. *2 citations pages 68 et 69*

[Audibert, 2003] Audibert, L. (2003). *Outils d'exploration de corpus et désambiguïsation lexicale automatique.* Thèse de doctorat, University of Provence Aix-Marseille I, France. *2 citations pages 38 et 42*

[Audibert, 2004] Audibert, L. (2004). Word sense disambiguation criteria : a systematic study. *In COLING 2004, 20th International Conference on Computatio-*

nal Linguistics, Proceedings of the Conference, pages 910–916, Geneva, Switzerland. *2 citations pages 40 et 99*

[Ayed *et al.*, 2014] Ayed, R., Bounhas, I., Elayeb, B., Bellamine Ben Saoud, N. et Evrard, F. (2014). Improving Arabic Texts Morphological Disambiguation Using a Possibilistic Classifier. *In* Métais, E., Roche, M. et Teisseire, M., éditeurs : *Natural Language Processing and Information Systems*, numéro 8455 de Lecture Notes in Computer Science, pages 138–147. Springer International Publishing. DOI : 10.1007/978-3-319-07983-7_18. *2 citations pages 102 et 155*

[Azarbonyad *et al.*, 2013] Azarbonyad, H., Shakery, A. et Faili, H. (2013). Exploiting Multiple Translation Resources for English-Persian Cross Language Information Retrieval. *In Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, Lecture Notes in Computer Science, pages 93–99. Springer, Berlin, Heidelberg. *Cité page 50*

[Baccini *et al.*, 2012] Baccini, A., Déjean, S., Lafage, L. et Mothe, J. (2012). How many performance measures to evaluate information retrieval systems ? *Knowledge and Information Systems*, 30(3):693–713. *2 citations pages 23 et 150*

[Baeza-Yates et Ribeiro-Neto, 1999] Baeza-Yates, R. et Ribeiro-Neto, B. (1999). *Modern information retrieval*, volume 463. ACM press New York. *2 citations pages 14 et 95*

[Ballesteros et Croft, 1997] Ballesteros, L. et Croft, W. B. (1997). Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval. *In Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '97, pages 84–91, New York, NY, USA. ACM. *Cité page 71*

[Banerjee et Pedersen, 2002] Banerjee, S. et Pedersen, T. (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, pages 136–145, London, UK, UK. Springer-Verlag. *2 citations pages 35 et 139*

[Bannour *et al.*, 2011] Bannour, S., Audibert, L. et Nazarenko, A. (2011). Mesures de similarité distributionnelle entre termes. *In IC 2011 : 22es journées Ingénierie des Connaissances (Proceedings of the 22nd French Knowledge Engineering Conference), Chambery, France, May 16-20, 2011*, pages 523–538. *Cité page 37*

[Barathi et Valli, 2010] Barathi, M. et Valli, S. (2010). Ontology Based Query Expansion Using Word Sense Disambiguation. *arXiv :1003.1460 [cs]*, 7(2):22–27. arXiv : 1003.1460. *Cité page 45*

[Barque et Chaumartin, 2008] Barque, L. et Chaumartin, F.-R. (2008). La polysémie régulière dans WordNet. *In TALN 2008 - 15eme conférence sur le Traitement Automatique des Langues Naturelles*, pages 101–108, Avignon, France. *2 citations pages 32 et 99*

[Basile *et al.*, 2014] Basile, P., Caputo, A. et Semeraro, G. (2014). An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model. *In COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference : Technical Papers*, pages 1591–1600, Dublin, Ireland. *Cité page 30*

[Basile *et al.*, 2012] Basile, V., Bos, J., Evang, K. et Venhuizen, N. (2012). Developing a large semantically annotated corpus. *In LREC 2012, 8th International Conference on Language Resources and Evaluation.* *Cité page 33*

[Batchkarov *et al.*, 2016] Batchkarov, M., Kober, T., Reffin, J., Weeds, J. et Weir, D. (2016). A critique of word similarity as a method for evaluating distributional semantic models. *In The First Workshop on Evaluating Vector Space Representations for NLP*, Berlin. *Cité page 38*

[Beeferman et Berger, 2000] Beeferman, D. et Berger, A. (2000). Agglomerative Clustering of a Search Engine Query Log. *In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, pages 407–416, New York, NY, USA. ACM. *Cité page 67*

[Bellot *et al.*, 1998] Bellot, D. L., Loupy, C. D., Bellot, P., El-bèze, M. et Marteau, P.-f. (1998). Query Expansion and Classification of Retrieved Documents. *In Proceedings of the 7th Text Retrieval Conference (TREC-7)*, pages 382–389. *Cité page 65*

[Ben Khiroun *et al.*, 2018] Ben Khiroun, O., Elayeb, B. et Bellamine Ben Saoud, N. (2018). Towards a Query Translation Disambiguation Approach using Possibility Theory. *In Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018)*, pages 606–613, Funchal, Madeira, Portugal. *6 citations pages 3, 85, 141, 143, 145 et 154*

[Ben Khiroun *et al.*, 2011] Ben Khiroun, O., Elayeb, B., Bounhas, I., Evrard, F. et Bellamine Ben Saoud, N. (2011). A Possibilistic Approach for Semantic Query Expansion. *In The 4th international conference on Internet Technologies and Applications (ITA 2011)*, pages 308–316, Wrexham Wales (UK). *5 citations pages 2, 84, 89, 90 et 103*

[Ben Khiroun *et al.*, 2012] Ben Khiroun, O., Elayeb, B., Bounhas, I., Evrard, F. et Bellamine Ben Saoud, N. (2012). A Possibilistic Approach for Automatic Word Sense Disambiguation. *In Proceedings of the 24th Conference on Computational Linguistics and Speech Processing (ROCLING)*, pages 261–275, Taiwan. *9 citations pages 3, 84, 103, 107, 109, 110, 118, 144 et 154*

[Ben Khiroun *et al.*, 2014] Ben Khiroun, O., Elayeb, B., Bounhas, I., Evrard, F. et Bellamine Ben Saoud, N. (2014). Improving query expansion by automatic query disambiguation in intelligent information retrieval. *In The 6th International Conference on Agents and Artificial Intelligence (ICAART 2014)*, pages 153–160, Angers, Loire Valley, France. *7 citations pages 3, 85, 131, 136, 144, 151 et 154*

[Ben Romdhane *et al.*, 2017] Ben Romdhane, W., Elayeb, B. et Bellamine Ben Saoud, N. (2017). A Discriminative Possibilistic Approach for Query Translation Disambiguation. *In Natural Language Processing and Information Systems*, Lecture Notes in Computer Science, pages 366–379. Springer, Cham. *2 citations pages 84 et 150*

[Benferhat *et al.*, 1999] Benferhat, S., Dubois, D., Garcia, L. et Prade, H. (1999). Possibilistic logic bases and possibilistic graphs. *In Proceedings of the Fifteenth Conference on*

Uncertainty in Artificial Intelligence, pages 57–64, Stockholm, Sweden. Morgan Kaufmann Publishers Inc. *Cité page 79*

[Bestgen, 2006] Bestgen, Y. (2006). Improving Text Segmentation Using Latent Semantic Analysis : A Reanalysis of Choi, Wiemer-Hastings, and Moore. *Computational Linguistics*, 32(3):455. *Cité page 36*

[Bhogal *et al.*, 2007] Bhogal, J., Macfarlane, A. et Smith, P. (2007). A Review of Ontology Based Query Expansion. *Inf. Process. Manage.*, 43(4):866–886. *2 citations pages 64 et 65*

[Bian et Teng, 2005] Bian, G.-w. et Teng, S.-y. (2005). Integrating query translation and text classification in a cross-language patent access system. *In In : Proceedings of NTCIR-7 Workshop Meeting*, pages 16–19. *Cité page 96*

[Biau *et al.*, 2010] Biau, D. J., Jolles, B. M. et Porcher, R. (2010). P value and the theory of hypothesis testing : an explanation for new researchers. *Clinical Orthopaedics and Related Research*, 468(3):885–892. *2 citations pages 123 et 149*

[Billerbeck *et al.*, 2003] Billerbeck, B., Scholer, F., Williams, H. E. et Zobel, J. (2003). Query Expansion Using Associated Queries. *In Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 2–9, New York, NY, USA. ACM. *Cité page 67*

[Borgelt *et al.*, 2000] Borgelt, C., Gebhardt, J. et Kruse, R. (2000). Possibilistic graphical models. *Computational Intelligence in Data Mining*, 408:51–68. *Cité page 79*

[Borlund, 2003] Borlund, P. (2003). The concept of relevance in IR. *J. Am. Soc. Inf. Sci. Technol.*, 54(10):913–925. *Cité page 8*

[Bouchoucha *et al.*, 2014] Bouchoucha, A., Liu, X. et Nie, J.-Y. (2014). Integrating Multiple Resources for Diversified Query Expansion. *In Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings*, pages 437–442. *Cité page 54*

[Bouchoucha *et al.*, 2015] Bouchoucha, A., Liu, X. et Nie, J.-Y. (2015). Towards Query Level Resource Weighting for Diversified Query Expansion. *In* Hanbury, A., Kazai, G., Rauber, A. et Fuhr, N., éditeurs : *Advances in Information Retrieval - 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015. Proceedings*, volume 9022 de *Lecture Notes in Computer Science*, pages 1–12. *Cité page 54*

[Bounhas *et al.*, 2015] Bounhas, I., Ayed, R., Elayeb, B., Evrard, F. et Bellamine Ben Saoud, N. (2015). Experimenting a Discriminative Possibilistic Classifier with Reweighting Model for Arabic Morphological Disambiguation. *Computer Speech and Language*, 33(1):67–87. *Cité page 83*

[Braschler et Peters, 2004] Braschler, M. et Peters, C. (2004). CLEF 2003 Methodology and Metrics. *In* Peters, C., Gonzalo, J., Braschler, M. et Kluck, M., éditeurs : *Comparative Evaluation of Multilingual Information Access Systems*, numéro 3237 de Lecture Notes in Computer Science, pages 7–20. Springer Berlin Heidelberg. *3 citations pages 27, 132 et 146*

[Braschler et Schäuble, 2000] Braschler, M. et Schäuble, P. (2000). Experiments with the Eurospider Retrieval System for CLEF 2000. *In Cross-Language Information Retrieval and Evaluation*, Lecture Notes in Computer Science, pages 140–148. Springer, Berlin, Heidelberg. *Cité page 72*

[Brin et Page, 1998] Brin, S. et Page, L. (1998). The Anatomy of a Large-scale Hypertextual Web Search Engine. *In Proceedings of the Seventh International Conference on World Wide Web 7*, WWW7, pages 107–117, Amsterdam, The Netherlands, The Netherlands. Elsevier Science Publishers B. V. *2 citations pages 39 et 49*

[Brini *et al.*, 2004a] Brini, A., Boughanem, M. et Dubois, D. (2004a). Towards a Possibilistic Approach for Information Retrieval . *In EUROFUSE 2004, Data and Knowledge Engineering*, pages 92–102, Varsovie, Pologne. *2 citations pages 80 et 82*

[Brini *et al.*, 2004b] Brini, A., Boughanem, M. et Dubois, D. (2004b). Vers une approche possibiliste pour la recherche d'information. *In Veille Stratégique Scientifique & Technologique (VSST 2004)*, pages 55–65, Toulouse, France. *Cité page 80*

[Brini *et al.*, 2007] Brini, A., Boughanem, M. et Dubois, D. (2007). Un modèle de réseau possibiliste pour la recherche d'information. *Information-Interaction-Intelligence, Cépaduès Editions*, 7(1):31–54. *2 citations pages 82 et 83*

[Bruce et Wiebe, 1994] Bruce, R. et Wiebe, J. (1994). Word-sense Disambiguation Using Decomposable Models. *In Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 139–146, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cité page 33*

[Brun, 2000] Brun, C. (2000). A Client/Server Architecture for Word Sense Disambiguation. *In 18th International Conference on Computational Linguistics*, pages 132–138. Morgan Kaufmann. *Cité page 40*

[Brun *et al.*, 2001] Brun, C., Jacquemin, B. et Segond, F. (2001). Exploitation de dictionnaires électroniques pour la désambiguïsation sémantique lexicale. *Traitement Automatique des Langues*, 42(3):667–691. *Cité page 40*

[Buckley *et al.*, 1995] Buckley, C., Salton, G., Allan, J. et Singhal, A. (1995). Automatic Query Expansion Using SMART : TREC 3. *In In Proceedings of The third Text REtrieval Conference (TREC-3)*, pages 69–80. *2 citations pages 60 et 66*

[Cao *et al.*, 2005] Cao, G., Nie, J.-Y. et Bai, J. (2005). Integrating Word Relationships into Language Models. *In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 298–305, New York, NY, USA. ACM. *2 citations pages 128 et 142*

[Capstick *et al.*, 2000] Capstick, J., Diagne, A. K., Erbach, G., Uszkoreit, H., Leisenberg, A. et Leisenberg, M. (2000). A system for supporting cross-lingual information retrieval. *Information Processing & Management*, 36(2):275–289. *Cité page 96*

[Carbonell et Goldstein, 1998] Carbonell, J. et Goldstein, J. (1998). The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. *In Pro-*

ceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '98, pages 335–336, New York, NY, USA. ACM. *Cité page 54*

[Carpineto *et al.*, 2001] Carpineto, C., de Mori, R., Romano, G. et Bigi, B. (2001). An Information-theoretic Approach to Automatic Query Expansion. *ACM Transactions on Information Systems (TOIS)*, 19(1):1–27. *Cité page 64*

[Carpineto et Romano, 2012] Carpineto, C. et Romano, G. (2012). A Survey of Automatic Query Expansion in Information Retrieval. *ACM Computing Surveys (CSUR)*, 44(1):1–50. *3 citations pages 53, 62 et 95*

[Chan et Ng, 2007] Chan, Y. S. et Ng, H. T. (2007). Word sense disambiguation improves statistical machine translation. *In In 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, pages 33–40. *Cité page 127*

[Chebil *et al.*, 2016] Chebil, W., Soualmia, L. F., Omri, M. N. et Darmoni, S. J. (2016). Indexing biomedical documents with a possibilistic network. *Journal of the Association for Information Science and Technology*, 67(4):928–941. *Cité page 83*

[Chen *et al.*, 2015] Chen, T., Xu, R., He, Y. et Wang, X. (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 (Short Papers)*, volume 2, pages 15–20. Association for Computational Linguistics. *Cité page 43*

[Chen *et al.*, 2014] Chen, X., Liu, Z. et Sun, M. (2014). A Unified Model for Word Sense Representation and Disambiguation. *In EMNLP*, pages 1025–1035. Citeseer. *Cité page 36*

[Chifu et Ionescu, 2012] Chifu, A.-G. et Ionescu, R.-T. (2012). Word sense disambiguation to improve precision for ambiguous queries. *Central European Journal of Computer Science*, 2(4):398–411. *Cité page 127*

[Chifu et Mothe, 2014] Chifu, A.-G. et Mothe, J. (2014). Expansion sélective de requêtes par apprentissage. *In CORIA-CIFED*, pages 257–272. ARIA-GRCE. *Cité page 65*

[Chklovski et Mihalcea, 2002] Chklovski, T. et Mihalcea, R. (2002). Building a Sense Tagged Corpus with Open Mind Word Expert. *In Proceedings of the ACL-02 Workshop on Word Sense Disambiguation : Recent Successes and Future Directions - Volume 8*, WSD '02, pages 116–122, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cité page 33*

[Church et Hanks, 1990] Church, K. W. et Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29. *2 citations pages 36 et 142*

[Cleverdon, 1967] Cleverdon, C. (1967). The Cranfield tests on index language devices. *Aslib Proceedings*, 19(6):173–194. *Cité page 26*

[Cohen, 1968] Cohen, J. (1968). Weighted kappa : Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220. *Cité page 118*

[Cohn, 2003] Cohn, T. (2003). Performance Metrics for Word Sense Disambiguation. *In Proceedings of the Australasian Language Technology Workshop*, pages 86–93, Melbourne, Australia. *Cité page 124*

[Croft *et al.*, 2009] Croft, B., Metzler, D. et Strohman, T. (2009). *Search Engines : Information Retrieval in Practice*. Pearson, Boston. *Cité page 60*

[Cruys, 2010] Cruys, T. V. d. (2010). *Mining for Meaning : The Extraction of Lexico-semantic Knowledge from Text*. *Cité page 43*

[Cruys et Apidianaki, 2011] Cruys, T. V. d. et Apidianaki, M. (2011). Latent Semantic Word Sense Induction and Disambiguation. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies - Volume 1*, HLT '11, pages 1476–1485, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cité page 43*

[Cui *et al.*, 2003] Cui, H., Wen, J.-R., Nie, J.-Y. et Ma, W.-Y. (2003). Query Expansion by Mining User Logs. *IEEE Trans. on Knowl. and Data Eng.*, 15(4):829–839. *Cité page 67*

[Damani, 2013] Damani, O. (2013). Improving Pointwise Mutual Information (PMI) by Incorporating Significant Co-occurrence. *In Proceedings of the Seventeenth Conference on Computational Natural Language Learning, CoNLL 2013, Sofia, Bulgaria, August 8-9, 2013*, pages 20–28. *Cité page 36*

[Daoud *et al.*, 2010] Daoud, M., Tamine, L. et Chebaro, B. (2010). Proposition d'un système de RI personnalisé à base de sessions intégrant un profil utilisateur sémantique. *Document numérique*, 13(1):137–160. *Cité page 28*

[Demšar, 2006] Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7:1–30. *2 citations pages 122 et 148*

[Diaz *et al.*, 2016] Diaz, F., Mitra, B. et Craswell, N. (2016). Query Expansion with Locally-Trained Word Embeddings. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, 1:367–377. *2 citations pages 65 et 66*

[Doddington *et al.*, 2004] Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. et Weischedel, R. M. (2004). The Automatic Content Extraction (ACE) Program - Tasks, Data, and Evaluation. *In LREC*. European Language Resources Association. *Cité page 30*

[Dongen, 2000] Dongen, S. (2000). A Cluster Algorithm for Graphs. Rapport technique, CWI (Centre for Mathematics and Computer Science), Amsterdam, The Netherlands. *Cité page 44*

[Dramé *et al.*, 2014] Dramé, K., Mougin, F. et Diallo, G. (2014). Query Expansion using External Resources for Improving Information Retrieval in the Biomedical Domain. *In CLEF (Working Notes)*, volume 1180 de *CEUR Workshop Proceedings*, pages 189–194. *Cité page 64*

[Dubois et Prade, 2011] Dubois, D. et Prade, H. (2011). Possibility theory and its application : Where do we stand. *Mathware and Soft Computing*, 18(1):18–31. *2 citations pages 2 et 77*

[Dubois et Prade, 2012] Dubois, D. et Prade, H. (2012). Possibility theory. *In* Meyers, R. A., éditeur : *Computational Complexity*, pages 2240–2252. Springer New York. *2 citations pages 77 et 130*

[Duque *et al.*, 2015] Duque, A., Araujo, L. et Martinez-Romo, J. (2015). CO-graph : A new graph-based technique for cross-lingual word sense disambiguation. *Natural Language Engineering*, 21(5):743–772. *2 citations pages 45 et 49*

[Edmonds et Hirst, 2002] Edmonds, P. et Hirst, G. (2002). Near-Synonymy and Lexical Choice. *Computational Linguistics*, 28(2):105–144. *Cité page 119*

[Elayeb, 2009] Elayeb, B. (2009). *SARIPOD : Système multi-Agent de Recherche Intelligente POssibiliste des Documents Web*. Thèse de doctorat, Institut National Polytechnique de Toulouse, France & Ecole Nationale des Sciences de l'Informatique, Université de la Manouba, Tunisie. *9 citations pages 2, 9, 82, 90, 96, 135, 136, 149 et 153*

[Elayeb *et al.*, 2018] Elayeb, B., Ben Romdhane, W. et Bellamine Ben Saoud, N. (2018). Towards a new possibilistic query translation tool for cross-language information retrieval. *Multimedia Tools and Applications*, 77(2):2423–2465. *3 citations pages 84, 96 et 150*

[Elayeb et Bounhas, 2016] Elayeb, B. et Bounhas, I. (2016). Arabic Cross-Language Information Retrieval : A Review. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 15(3):1–44. *2 citations pages 151 et 154*

[Elayeb *et al.*, 2015a] Elayeb, B., Bounhas, I., Ben Khiroun, O. et Bellamine Ben Saoud, N. (2015a). Combining Semantic Query Disambiguation and Expansion to Improve Intelligent Information Retrieval. *In* Duval, B., van den Herik, J., Loiseau, S. et Filipe, J., éditeurs : *Agents and Artificial Intelligence*, numéro 8946 de Lecture Notes in Computer Science, pages 280–295. Springer International Publishing. *4 citations pages 3, 135, 151 et 154*

[Elayeb *et al.*, 2011] Elayeb, B., Bounhas, I., Ben Khiroun, O., Evrard, F. et Bellamine Ben Saoud, N. (2011). Towards a possibilistic information retrieval system using semantic query expansion. *International Journal of Intelligent Information Technologies*, 7(4):1–25. *3 citations pages 2, 90 et 130*

[Elayeb *et al.*, 2015b] Elayeb, B., Bounhas, I., Ben Khiroun, O., Evrard, F. et Bellamine Ben Saoud, N. (2015b). A Comparative Study Between Possibilistic and Probabilistic Approaches for Monolingual Word Sense Disambiguation. *Knowl. Inf. Syst.*, 44(1):91–126. *9 citations pages 3, 103, 105, 107, 112, 113, 114, 120 et 154*

[Elayeb *et al.*, 2009] Elayeb, B., Evrard, F., Zaghdoud, M. et Ben Ahmed, M. (2009). Towards an intelligent possibilistic web information retrieval using multiagent system. *The International Journal of Interactive Technology and Smart Education (ITSE), Special issue : New learning support systems*, 6(1):40–59. *4 citations pages 82, 83, 103 et 135*

[Erk et Strapparava, 2010] Erk, K. et Strapparava, C., éditeurs (2010). *Proceedings of the 5th International Workshop on Semantic Evaluation*. The Association for Computer Linguistics, Uppsala, Sweden. *Cité page 31*

[Ermakova et Mothe, 2016] Ermakova, L. et Mothe, J. (2016). Query Expansion by Local Context Analysis. *In CORIA-CIFED*, pages 235–250. ARIA-GRCE. *Cité page 65*

[Eugenio, 2000] Eugenio, B. D. (2000). On the usage of Kappa to evaluate agreement on coding tasks. *In Proceedings of the Second International Conference on Language Resources and Evaluation*, pages 441–444. *Cité page 118*

[Fabiani, 1996] Fabiani, P. (1996). *Représentation dynamique de l'incertain et stratégie de perception pour un système autonome en environnement évolutif*. Thèse de doctorat, Toulouse, ENSAE. *Cité page 78*

[Fang, 2008] Fang, H. (2008). A Re-examination of Query Expansion Using Lexical Resources. *In Proceedings of ACL-08 : HLT*, pages 139–147. *2 citations pages 63 et 128*

[Faralli et Navigli, 2012] Faralli, S. et Navigli, R. (2012). A New Minimally-Supervised Framework for Domain Word Sense Disambiguation. *In* Tsujii, J., Henderson, J. et Pasca, M., éditeurs : *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL*, pages 1411–1422. ACL. *Cité page 39*

[Fitzpatrick et Dent, 1997] Fitzpatrick, L. et Dent, M. (1997). Automatic Feedback Using Past Queries : Social Searching ? *In Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '97, pages 306–313, New York, NY, USA. ACM. *Cité page 67*

[Francis, 1965] Francis, W. N. (1965). A Standard Corpus of Edited Present-Day American English. *College English*, 26(4):267–273. *Cité page 33*

[Fu *et al.*, 2005] Fu, G., Jones, C. B. et Abdelmoty, A. I. (2005). Ontology-Based Spatial Query Expansion in Information Retrieval. *In On the Move to Meaningful Internet Systems 2005 : CoopIS, DOA, and ODBASE*, Lecture Notes in Computer Science, pages 1466–1482. Springer, Berlin, Heidelberg. *Cité page 64*

[Galeas et Freisleben, 2008] Galeas, P. et Freisleben, B. (2008). Word Distribution Analysis for Relevance Ranking and Query Expansion. *In Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, pages 500–511. Springer, Berlin, Heidelberg. *Cité page 66*

[Gan et Hong, 2015] Gan, L. et Hong, H. (2015). Improving query expansion for information retrieval using wikipedia. *International Journal of Database Theory and Application*, 8(3):27–40. *Cité page 67*

[Garrouch et Omri, 2015] Garrouch, K. et Omri, M. N. (2015). Possibilistic Network based Information Retrieval Model. *In 2015 15th International Conference on Intelligent Systems Design and Applications (ISDA)*, pages 25–30. *Cité page 83*

[Gaume, 2004] Gaume, B. (2004). Balades Aléatoires dans les Petits Mondes Lexicaux. *I3 Information Interaction Intelligence*, 4(3). *3 citations pages 89, 90 et 113*

[Gaume *et al.*, 2004] Gaume, B., Hathout, N. et Muller, P. (2004). Word Sense Disambiguation using a dictionary for sense similarity measure. *In Proceedings of the 20th international conference on Computational Linguistics*, page 1194. Association for Computational Linguistics. *4 citations pages 89, 108, 112 et 113*

[Golub et Van Loan, 1989] Golub, G. H. et Van Loan, C. F. (1989). Matrix computations. Johns Hopkins series in the mathematical sciences. *Johns Hopkins University Press, Baltimore, MD.* *Cité page 43*

[Graff *et al.*, 2007] Graff, D., Junbo, K., Ke, C. et Kazuaki, M. (2007). English Gigaword Third Edition LDC2007t07. *Cité page 33*

[Haines et Croft, 1993] Haines, D. et Croft, W. B. (1993). Relevance Feedback and Inference Networks. *In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '93, pages 2–11, New York, NY, USA. ACM. *Cité page 59*

[Harispe *et al.*, 2015] Harispe, S., Ranwez, S., Janaqi, S. et Montmain, J. (2015). Semantic Similarity from Natural Language and Ontology Analysis. *Synthesis Lectures on Human Language Technologies*, 8(1):1–254. *Cité page 34*

[Harman, 1992] Harman, D. (1992). Relevance Feedback and Other Query Modification Techniques. *In* Frakes, W. B. et Baeza-Yates, R., éditeurs : *Information Retrieval*, pages 241–263. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. *Cité page 59*

[Harris, 1970] Harris, Z. S. (1970). Distributional Structure. *In Papers in Structural and Transformational Linguistics*, Formal Linguistics Series, pages 775–794. Springer, Dordrecht. DOI : 10.1007/978-94-017-6059-1_36. *Cité page 42*

[Hedlund *et al.*, 2004] Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A. et Järvelin, K. (2004). Dictionary-Based Cross-Language Information Retrieval : Learning Experiences from CLEF 2000–2002. *Information Retrieval*, 7(1-2):99–119. *2 citations pages 71 et 96*

[Heer *et al.*, 2005] Heer, J., Card, S. K. et Landay, J. A. (2005). Prefuse : A Toolkit for Interactive Information Visualization. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, pages 421–430, New York, NY, USA. ACM. *Cité page 90*

[Herbert *et al.*, 2011] Herbert, B., Szarvas, G. et Gurevych, I. (2011). Combining Query Translation Techniques to Improve Cross-language Information Retrieval. *In Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 712–715, Berlin, Heidelberg. Springer-Verlag. *Cité page 50*

[Hersh *et al.*, 2003] Hersh, W. R., Bhupatiraju, R. T. et Price, S. (2003). Phrases, Boosting, and Query Expansion Using External Knowledge Resources for Genomic Information Retrieval. *In* Voorhees, E. M. et Buckland, L. P., éditeurs : *Proceedings of The Twelfth Text REtrieval Conference, TREC 2003, Gaithersburg, Maryland, USA, November 18-21, 2003*, volume Special Publication 500-255, pages 503–509. National Institute of Standards and Technology (NIST). *Cité page 64*

[Hosseinzadeh Vahid *et al.*, 2015] Hosseinzadeh Vahid, A., Arora, P., Liu, Q. et Jones, G. J. (2015). A Comparative Study of Online Translation Services for Cross Language Information Retrieval. *In Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 859–864, New York, NY, USA. ACM. *Cité page 150*

[Huang *et al.*, 2012] Huang, E. H., Socher, R., Manning, C. D. et Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics. *Cité page 36*

[Ide, 1971] Ide, E. (1971). New experiments in relevance feedback. *The SMART retrieval system*, pages 337–354. *Cité page 58*

[Ide et Suderman, 2006] Ide, N. et Suderman, K. (2006). Integrating linguistic resources : The american national corpus model. *In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy. *Cité page 33*

[Ide et Wilks, 2007] Ide, P. N. et Wilks, P. Y. (2007). Making Sense About Sense. *In* Agirre, A. P. E. et Edmonds, R. S. P., éditeurs : *Word Sense Disambiguation*, numéro 33 de Text, Speech and Language Technology, pages 47–73. Springer Netherlands. DOI : 10.1007/978-1-4020-4809-8_3. *2 citations pages 98 et 99*

[Islam et Inkpen, 2009] Islam, A. et Inkpen, D. (2009). Managing the Google Web 1t 5-gram data set. *In 2009 International Conference on Natural Language Processing and Knowledge Engineering*, pages 1–5. *Cité page 34*

[Jayakody, 2016] Jayakody, J. R. K. C. (2016). Natural language processing framework : WordNet based sentimental analyzer. *In Proceedings of the International Research Symposium on Pure and Applied Sciences (IRSPAS 2016)*, Faculty of Science, University of Kelaniya, Sri Lanka. *Cité page 38*

[Jiang *et al.*, 2017] Jiang, Z., Wen, J.-R., Dou, Z., Zhao, W. X., Nie, J.-Y. et Yue, M. (2017). Learning to diversify search results via subtopic attention. *In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 545–554, New York, NY, USA. ACM. *Cité page 54*

[Jimeno-Yepes *et al.*, 2011] Jimeno-Yepes, A. J., McInnes, B. T. et Aronson, A. R. (2011). Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. *BMC bioinformatics*, 12(1):223. *Cité page 45*

[Joho *et al.*, 2004] Joho, H., Sanderson, M. et Beaulieu, M. (2004). A Study of User Interaction with a Concept-Based Interactive Query Expansion Support Tool. *In Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 42–56. Springer, Berlin, Heidelberg. *Cité page 53*

[Jones *et al.*, 2006] Jones, R., Rey, B., Madani, O. et Greiner, W. (2006). Generating Query Substitutions. *In Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 387–396, New York, NY, USA. ACM. *Cité page 67*

[Kadri et Nie, 2008] Kadri, Y. et Nie, J.-Y. (2008). A Comparative Study for Query Translation using Linear Combination and Confidence Measure. *In The Third International Joint Conference on Natural Language Processing (IJCNLP)*, pages 181–188. *Cité page 96*

[Kammoun Bouzaiene, 2006] Kammoun Bouzaiene, H. (2006). *Collaboration de modèles symbolique et numérique pour une recherche d'information adaptative, évolutive et coopérative.* Thèse de doctorat, Ecole Nationale des Sciences de l'Informatique (ENSI), Université de Manouba, Tunisie. *Cité page 11*

[Keikha *et al.*, 2017] Keikha, A., Ensan, F. et Bagheri, E. (2017). Query expansion using pseudo relevance feedback on wikipedia. *Journal of Intelligent Information Systems*, pages 1–24. *2 citations pages 61 et 62*

[Khapra *et al.*, 2009] Khapra, M. M., Shah, S., Kedia, P. et Bhattacharyya, P. (2009). Projecting Parameters for Multilingual Word Sense Disambiguation. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 459–467. ACL. *Cité page 46*

[Kilgarriff, 1998] Kilgarriff, A. (1998). SENSEVAL : An Exercise in Evaluating Word Sense Disambiguation Programs. *In LREC*, pages 581–588. *Cité page 30*

[Kilgarriff, 2003] Kilgarriff, A. (2003). Thesauruses for natural language processing. *In International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003*, pages 5–13. *Cité page 32*

[Kilgarriff *et al.*, 2014] Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. et Suchomel, V. (2014). The Sketch Engine : ten years on. *Lexicography*, 1(1):7–36. *Cité page 96*

[Kim *et al.*, 2015] Kim, S., Ko, Y. et Oard, D. W. (2015). Combining lexical and statistical translation evidence for cross-language information retrieval. *Journal of the Association for Information Science and Technology*, 66(1):23–39. *2 citations pages 20 et 50*

[Koehn, 2005] Koehn, P. (2005). Europarl : A Parallel Corpus for Statistical Machine Translation. *In Proceedings of the 10th Machine Translation Summit*, volume 5, pages 79–86. *2 citations pages 33 et 142*

[Kohavi, 1995] Kohavi, R. (1995). A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. *In Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, pages 1137–1143, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. *Cité page 109*

[Koppula *et al.*, 2017] Koppula, N., Rani, B. P. et Rao, K. S. (2017). Graph Based Word Sense Disambiguation. *In Proceedings of the First International Conference on Computational Intelligence and Informatics*, Advances in Intelligent Systems and Computing, pages 665–670. Springer, Singapore. DOI : 10.1007/978-981-10-2471-9_64. *Cité page 45*

[Ksentini *et al.*, 2016] Ksentini, N., Tmar, M. et Gargouri, F. (2016). The Impact of Term Statistical Relationships on Rocchio's Model Parameters For Pseudo Relevance Feedback. *IJCISIM*, 8:135–144. *2 citations pages 58 et 68*

[Kuzi *et al.*, 2016] Kuzi, S., Shtok, A. et Kurland, O. (2016). Query Expansion Using Word Embeddings. *In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 1929–1932, New York, NY, USA. ACM. *2 citations pages 65 et 66*

[Lafourcade, 2007] Lafourcade, M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. *In SNLP'07 : 7th international symposium on natural language processing*, page 7. *Cité page 45*

[Lafourcade et Brun, 2017] Lafourcade, M. et Brun, N. L. (2017). Extracting semantic relations via the combination of inferences, schemas and cooccurrences. *In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017*, pages 417–423. *Cité page 45*

[Lahbib *et al.*, 2015] Lahbib, W., Bounhas, I. et Slimani, Y. (2015). Arabic Terminology Extraction and Enrichment Based on Domain-Specific Text Mining. *In Proceedings of The 27th International Conference on Tools with Artificial Intelligence*, pages 340–347, Vietri sul Mare, Italy. IEEE. *Cité page 155*

[Landauer et Dooley, 2002] Landauer, T. et Dooley, S. (2002). Latent Semantic Analysis : Theory, Method and Application. *In Proceedings of the Conference on Computer Support for Collaborative Learning : Foundations for a CSCL Community*, CSCL '02, pages 742–743, Boulder, Colorado. International Society of the Learning Sciences. *2 citations pages 36 et 66*

[Landis et Koch, 1977] Landis, J. R. et Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174. *2 citations pages 119 et 120*

[Lavrenko, 2008] Lavrenko, V. (2008). *A Generative Theory of Relevance*. Springer Science & Business Media. *2 citations pages 57 et 61*

[Leacock *et al.*, 1998] Leacock, C., Miller, G. A. et Chodorow, M. (1998). Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*, 24(1):147–165. *Cité page 33*

[Lee et Croft, 2014] Lee, C.-J. et Croft, W. B. (2014). Cross-Language Pseudo-Relevance Feedback Techniques for Informal Text. *In Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 260–272. Springer, Cham. *Cité page 70*

[Leech, 1993] Leech, G. (1993). 100 million words of English. *English Today*, 9(1):9–15. *Cité page 33*

[Lefever et Hoste, 2010] Lefever, E. et Hoste, V. (2010). SemEval-2010 Task 3 : Cross-Lingual Word Sense Disambiguation. *In* Erk, K. et Strapparava, C., éditeurs : *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 15–20. The Association for Computer Linguistics. *Cité page 48*

[Lefever et Hoste, 2013] Lefever, E. et Hoste, V. (2013). SemEval-2013 Task 10 : Cross-lingual Word Sense Disambiguation. *In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2 : Proceedings of the Seventh International Workshop*

on Semantic Evaluation (SemEval 2013), pages 158–166, Atlanta, Georgia, USA. Association for Computational Linguistics. *2 citations pages 48 et 123*

[Lesk, 1986] Lesk, M. (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, pages 24–26, New York, NY, USA. ACM. *3 citations pages 30, 34 et 40*

[Levow *et al.*, 2005] Levow, G.-A., Oard, D. W. et Resnik, P. (2005). Dictionary-based techniques for cross-language information retrieval. *Information Processing & Management*, 41(3):523 – 547. *Cité page 71*

[Lin, 1998] Lin, D. (1998). An information-theoretic definition of similarity. *In ICML*, volume 98, pages 296–304. Citeseer. *2 citations pages 38 et 43*

[Lin et Pantel, 2001] Lin, D. et Pantel, P. (2001). Induction of Semantic Classes from Natural Language Text. *In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, pages 317–322, New York, NY, USA. ACM. *Cité page 43*

[Liu *et al.*, 2015] Liu, D.-R., Omar, H., Liou, C.-H., Chi, H.-C. et Hsu, C.-H. (2015). Recommending blog articles based on popular event trend analysis. *Information Sciences*, 305:302–319. *Cité page 32*

[Liu *et al.*, 2004] Liu, S., Liu, F., Yu, C. et Meng, W. (2004). An Effective Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases. *In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 266–272, New York, NY, USA. ACM. *Cité page 63*

[Liu *et al.*, 2005] Liu, S., Yu, C. et Meng, W. (2005). Word sense disambiguation in queries. *In Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 525–532, New York, NY, USA. ACM. *Cité page 128*

[Liu *et al.*, 2014] Liu, X., Bouchoucha, A., Sordoni, A. et Nie, J.-Y. (2014). Compact Aspect Embedding for Diversified Query Expansions. *In Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec, Canada.*, pages 115–121. *Cité page 54*

[López-Ostenero *et al.*, 2002] López-Ostenero, F., Gonzalo, J., Peñas, A. et Verdejo, F. (2002). Interactive Cross-Language Searching : Phrases Are Better than Terms for Query Formulation and Refinement. *In Advances in Cross-Language Information Retrieval*, Lecture Notes in Computer Science, pages 416–429. Springer, Berlin, Heidelberg. *Cité page 96*

[Loupy, 2000] Loupy, C. d. (2000). *Evaluation de l'apport de connaissances linguistiques en désambiguïsation sémantique et recherche documentaire.* Thèse de doctorat, ANRT, Grenoble. *Cité page 113*

[Lovins, 1968] Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(1-2):22–31. *Cité page 14*

[Luca *et al.*, 2006] Luca, E. W. D., William, E., Luca, D., Hauke, S., Nürnberger, A. et Schlechtweg, S. (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*, pages 17–21. *Cité page 96*

[Lund et Burgess, 1996] Lund, K. et Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208. *Cité page 66*

[Lv *et al.*, 2015] Lv, C., Qiang, R., Fan, F. et Yang, J. (2015). Knowledge-Based Query Expansion in Real-Time Microblog Search. *In Information Retrieval Technology*, Lecture Notes in Computer Science, pages 43–55. Springer, Cham. *Cité page 65*

[Macdonald *et al.*, 2012] Macdonald, C., McCreadie, R., Santos, R. L. et Ounis, I. (2012). From puppy to maturity : Experiences in developing Terrier. *In Proceedings of the SIGIR 2012 Workshop in Open Source Information Retrieval*, pages 60–63. *3 citations pages 88, 132 et 147*

[Manning *et al.*, 2008] Manning, C. D., Raghavan, P. et Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA. *14 citations pages 8, 9, 14, 15, 23, 24, 25, 26, 27, 36, 53, 58, 133 et 150*

[Manning et Schütze, 1999] Manning, C. D. et Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA. *Cité page 92*

[Masterman, 1961] Masterman, M. (1961). Semantic message detection for machine translation using an interlingua. pages 438–475, UK. Her Majesty's Stationery Office. *Cité page 38*

[Mbarek *et al.*, 2017] Mbarek, R., Tmar, M., Hattab, H. et Boughanem, M. (2017). Pseudo-Relevance Feedback Method based on the Cross Product of Irrelevant Documents. *IJWA*, 9(1):8–15. *Cité page 62*

[McNamee et Mayfield, 2002] McNamee, P. et Mayfield, J. (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*, SIGIR '02, pages 159–166, New York, NY, USA. ACM. *2 citations pages 69 et 71*

[McRoy, 1992] McRoy, S. W. (1992). Using Multiple Knowledge Sources for Word Sense Discrimination. *Comput. Linguist.*, 18(1):1–30. *Cité page 45*

[Miao *et al.*, 2012] Miao, J., Huang, J. X. et Ye, Z. (2012). Proximity-based Rocchio's Model for Pseudo Relevance. *In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 535–544, New York, NY, USA. ACM. *Cité page 58*

[Mihalcea, 2005] Mihalcea, R. (2005). Unsupervised Large-vocabulary Word Sense Disambiguation with Graph-based Algorithms for Sequence Data Labeling. *In Proceedings of the*

Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05, pages 411–418, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cité page 49*

[Mihalcea, 2007] Mihalcea, R. (2007). Using Wikipedia for Automatic Word Sense Disambiguation. *In Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 196–203, Rochester, New York, USA. *Cité page 41*

[Mihalcea et Moldovan, 1998] Mihalcea, R. et Moldovan, D. (1998). Word sense disambiguation based on semantic density. *In Proceedings of COLING/ACL Workshop on Usage of WordNet in Natural Language Processing*, pages 16–22. *Cité page 45*

[Mikolov *et al.*, 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. et Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. *In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc. *Cité page 66*

[Miller et Charles, 1991] Miller, G. A. et Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28. *Cité page 38*

[Miller *et al.*, 1994] Miller, G. A., Chodorow, M., Landes, S., Leacock, C. et Thomas, R. G. (1994). Using a Semantic Concordance for Sense Identification. *In Proceedings of the Workshop on Human Language Technology*, HLT '94, pages 240–243, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cité page 33*

[Milne et Witten, 2008] Milne, D. et Witten, I. H. (2008). Learning to Link with Wikipedia. *In Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 509–518, New York, NY, USA. ACM. *Cité page 67*

[Mitra et Craswell, 2017] Mitra, B. et Craswell, N. (2017). Neural Text Embeddings for Information Retrieval. *In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pages 813–814, New York, NY, USA. ACM. *Cité page 22*

[Montgomery *et al.*, 2004] Montgomery, J., Si, L., Callan, J. et Evans, D. A. (2004). Effect of Varying Number of Documents in Blind Feedback : Analysis of the 2003 NRRC RIA Workshop "Bf_numdocs" Experiment Suite. *In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 476–477, New York, NY, USA. ACM. *Cité page 68*

[Monti *et al.*, 2013] Monti, J., Monteleone, M., Buono, M. P. d. et Marano, F. (2013). Natural Language Processing and Big Data - An Ontology-Based Approach for Cross-Lingual Information Retrieval. *In 2013 International Conference on Social Computing*, pages 725–731. *Cité page 21*

[Nasiruddin, 2013] Nasiruddin, M. (2013). État de l'art de l'induction de sens : une voie vers la désambiguïsation lexicale pour les langues peu dotées. *In* Boudin, F. et Barrault, L., éditeurs : *Rencontres des Étudiants Chercheurs en Informatique pour le Traitement*

Automatique des Langues, RÉCITAL, pages 192–205. The Association for Computer Linguistics. *Cité page 47*

[Navigli, 2009] Navigli, R. (2009). Word sense disambiguation : A survey. *ACM Computing Surveys (CSUR)*, 41(2):1–69. *6 citations pages 31, 33, 40, 44, 98 et 124*

[Navigli et Lapata, 2010] Navigli, R. et Lapata, M. (2010). An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(4):678–692. *2 citations pages 39 et 49*

[Navigli et Ponzetto, 2012] Navigli, R. et Ponzetto, S. P. (2012). BabelNet : The Automatic Construction, Evaluation and Application of a Wide-coverage Multilingual Semantic Network. *Artif. Intell.*, 193:217–250. *2 citations pages 143 et 155*

[Neale *et al.*, 2016] Neale, S., Gomes, L., Agirre, E., de Lacalle, O. L. et Branco, A. (2016). Word sense-aware machine translation : Including senses as contextual features for improved translation models. *In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2777–2783. *Cité page 31*

[Ng et Lee, 1996] Ng, H. T. et Lee, H. B. (1996). Integrating Multiple Knowledge Sources to Disambiguate Word Sense : An Exemplar-based Approach. *In Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 40–47, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cité page 33*

[Ngom, 2015] Ngom, A. N. (2015). Étude des mesures de similarité sémantique basées sur les arcs. *In CORIA 2015 - Conférence en Recherche d'Infomations et Applications - 12th French Information Retrieval Conference.*, pages 535–544, Paris, France. *Cité page 38*

[Nguyen et Ock, 2013] Nguyen, K.-H. et Ock, C.-Y. (2013). Word sense disambiguation as a traveling salesman problem. *Artificial Intelligence Review*, 40(4):405–427. *2 citations pages 98 et 101*

[Nie, 2003] Nie, J.-Y. (2003). Query expansion and query translation as logical inference. *Journal of the American Society for Information Science and Technology*, 54(4):335–346. *3 citations pages 71, 72 et 73*

[Nie, 2010] Nie, J.-Y. (2010). *Cross-language Information Retrieval*. Morgan & Claypool Publishers. *8 citations pages 3, 19, 22, 50, 69, 70, 73 et 153*

[Nilsson *et al.*, 2006] Nilsson, K., Hjelm, H. et Oxhammar, H. (2006). SUiS–cross-language ontology-driven information retrieval in a restricted domain. *In Proceedings of the 15th Nordic Conference of Computational Linguistics (NODALIDA 2005)*, pages 139–145. *Cité page 64*

[Niu *et al.*, 2004] Niu, C., Li, W., Srihari, R. K., Li, H. et Crist, L. (2004). Context clustering for word sense disambiguation based on modeling pairwise context similarities. *In SENSEVAL-3 : Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Spain*, pages 187–190. *Cité page 36*

[Oard *et al.*, 2008] Oard, D. W., He, D. et Wang, J. (2008). User-assisted Query Translation for Interactive Cross-language Information Retrieval. *Information Processing and Management*, 44(1):181–211. *2 citations pages 77 et 96*

[Ogden *et al.*, 1999] Ogden, W., Cowie, J., Davis, M., Ludovik, E., Nirenburg, S., Molina-Salgado, H. et Sharples, N. (1999). Keizai : An interactive cross-language text retrieval system. *In Proceeding of the MT SUMMIT VII workshop on machine translation for cross language information retrieval*, volume 416. *Cité page 96*

[Ogilvie *et al.*, 2009] Ogilvie, P., Voorhees, E. et Callan, J. (2009). On the number of terms used in automatic query expansion. *Information Retrieval*, 12(6):666–679. *3 citations pages 132, 133 et 137*

[Ounis *et al.*, 2006] Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C. et Lioma, C. (2006). Terrier : A high performance and scalable information retrieval platform. *In Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006).* *2 citations pages 88 et 132*

[Ounis *et al.*, 2007] Ounis, I., Lioma, C., Macdonald, C. et Plachouras, V. (2007). Research directions in terrier : a search engine for advanced retrieval on the web. *CEPIS Upgrade Journal*, 8(1). *Cité page 147*

[Pal *et al.*, 2014] Pal, D., Mitra, M. et Datta, K. (2014). Improving query expansion using WordNet. *Journal of the Association for Information Science & Technology*, 65(12):2469–2478. *Cité page 64*

[Panchenko *et al.*, 2017] Panchenko, A., Ruppert, E., Faralli, S., Ponzetto, S. P. et Biemann, C. (2017). Unsupervised Does Not Mean Uninterpretable : The Case for Word Sense Induction and Disambiguation. *In In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL'2017)*, Valencia, Spain. *Cité page 46*

[Pantel et Lin, 2002] Pantel, P. et Lin, D. (2002). Discovering Word Senses from Text. *In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 613–619, New York, NY, USA. ACM. *Cité page 43*

[Paskalis et Khodra, 2011] Paskalis, F. et Khodra, M. (2011). Word sense disambiguation in information retrieval using query expansion. *In 2011 International Conference on Electrical Engineering and Informatics (ICEEI)*, pages 1–6. *3 citations pages 135, 139 et 151*

[Paternostre *et al.*, 2002] Paternostre, M., Francq, P., Lamoral, J., Wartel, D. et Saerens, M. (2002). Carry, un algorithme de désuffixation pour le français. Rapport technique du projet Galilei. *Cité page 14*

[Paul et Baker, 1992] Paul, D. B. et Baker, J. M. (1992). The Design for the Wall Street Journal-based CSR Corpus. *In Proceedings of the Workshop on Speech and Natural Language*, HLT '91, pages 357–362, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cité page 33*

[Pennington *et al.*, 2014] Pennington, J., Socher, R. et Manning, C. (2014). Glove : Global vectors for word representation. *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543. *Cité page 66*

[Picton *et al.*, 2008] Picton, A., Fabre, C. et Bourigault, D. (2008). Méthodes linguistiques pour l'expansion de requêtes. Une expérience basée sur l'utilisation

du voisinage distributionnel. *Revue française de linguistique appliquée*, 13(1):83–95. *2 citations pages 63 et 64*

[Pilehvar *et al.*, 2013] Pilehvar, M. T., Jurgens, D. et Navigli, R. (2013). Align, Disambiguate and Walk : A Unified Approach for Measuring Semantic Similarity. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1341–1351. The Association for Computer Linguistics. *Cité page 39*

[Pincemin, 2006] Pincemin, B. (2006). Concordances et concordanciers : de l'art du bon KWAC. pages 33–42. CALS-CPST. *Cité page 92*

[Pincemin *et al.*, 2006] Pincemin, B., Issac, F., Chanove, M. et Mathieu-Colas, M. (2006). Concordanciers : Thème et variations. volume 2, pages 773–784. Presses Universitaires de Franche-Comté. *Cité page 92*

[Pinto et Pérez-sanjulián, 2008] Pinto, F. J. et Pérez-sanjulián, C. F. (2008). Automatic query expansion and word sense disambiguation with long and short queries using WordNet under vector model. *Actas de los Talleres de las Jornadas de Ingeniería del Software y Bases de Datos*, 2(2):17–23. *4 citations pages 133, 135, 138 et 139*

[Ploux et Victorri, 1998] Ploux, S. et Victorri, B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *Traitement Automatique des Langues*, (39):161–182. *Cité page 101*

[Ponzetto et Navigli, 2010] Ponzetto, S. P. et Navigli, R. (2010). Knowledge-Rich Word Sense Disambiguation Rivaling Supervised Systems. *In ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1522–1531, Uppsala, Sweden. *2 citations pages 39 et 41*

[Porter, 1997] Porter, M. F. (1997). Readings in Information Retrieval. pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. *Cité page 14*

[Reisinger et Mooney, 2010] Reisinger, J. et Mooney, R. J. (2010). Multi-prototype vector-space models of word meaning. *In Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117. Association for Computational Linguistics. *Cité page 36*

[Resnik, 2004] Resnik, P. (2004). Exploiting Hidden Meanings : Using Bilingual Text for Monolingual Annotation. *In Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, pages 283–299. Springer, Berlin, Heidelberg. *Cité page 49*

[Reymond, 2002] Reymond, D. (2002). Méthodologie pour la création d'un dictionnaire distributionnel dans une perspective d'étiquetage lexical semi-automatique. *6ème Rencontre des étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL-2002)*, pages 405–414. *Cité page 42*

[Robertson et Jones, 1976] Robertson, S. E. et Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the Association for Information Science and Technology*, 27(3): 129–146. *Cité page 59*

[Rocchio, 1971] Rocchio, J. (1971). Relevance Feedback in Information Retrieval. *In The SMART Retrieval System*, pages 313–323. Prentice-Hall, Englewood Cliffs, New Jersey, USA. *Cité page 57*

[Roche et Chauché, 2006] Roche, M. et Chauché, J. (2006). LSA : les limites d'une approche statistique. *Proceedings of atelier FDC*, 6:95–106. *Cité page 36*

[Role et Nadif, 2011] Role, F. et Nadif, M. (2011). Handling the Impact of Low Frequency Events on Co-occurrence based Measures of Word Similarity - A Case Study of Pointwise Mutual Information. *In KDIR 2011 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, Paris, France, 26-29 October, 2011*, pages 226–231. *Cité page 36*

[Romberg, 2017] Romberg, J. (2017). Comparing Relevance Feedback Techniques on German News Articles. *In Datenbanksysteme für Business, Technologie und Web (BTW 2017), 17. Fachtagung des GI-Fachbereichs „Datenbanken und Informationssysteme" (DBIS), 6.-10. März 2017, Stuttgart, Germany, Workshopband*, pages 301–310. *Cité page 68*

[Salton et Buckley, 1988] Salton, G. et Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523. *Cité page 12*

[Sandri, 1991] Sandri, S. (1991). *La combinaison de l'information incertaine et ses aspects algorithmiques*. Thèse de doctorat, Toulouse 3. *Cité page 78*

[Saneifar *et al.*, 2010] Saneifar, H., Bonniol, S., Laurent, A., Poncelet, P. et Roche, M. (2010). Passage Retrieval in Log Files : An Approach Based on Query Enrichment. *In IceTAL*, volume 6233 de *Lecture Notes in Computer Science*, pages 357–368. Springer. *Cité page 67*

[Saneifar *et al.*, 2014] Saneifar, H., Bonniol, S., Poncelet, P. et Roche, M. (2014). Enhancing passage retrieval in log files by query expansion based on explicit and pseudo relevance feedback. *Computers in Industry*, 65(6):937–951. *Cité page 67*

[Santos *et al.*, 2010] Santos, R. L., Macdonald, C. et Ounis, I. (2010). Exploiting Query Reformulations for Web Search Result Diversification. *In Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 881–890, New York, NY, USA. ACM. *Cité page 53*

[Santos *et al.*, 2015] Santos, R. L. T., Macdonald, C. et Ounis, I. (2015). Search Result Diversification. *Foundations and Trends in Information Retrieval*, 9(1):1–90. *Cité page 54*

[Savoy, 1997] Savoy, J. (1997). Ranking schemes in hybrid Boolean systems : A new approach. *Journal of the American Society for Information Science*, 48(3):235–253. *Cité page 13*

[Sawhney et Kaur, 2014] Sawhney, R. et Kaur, A. (2014). A modified technique for Word Sense Disambiguation using Lesk algorithm in Hindi language. *In 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2745–2749. *Cité page 30*

[Schamoni *et al.*, 2014] Schamoni, S., Hieber, F., Sokolov, A. et Riezler, S. (2014). Learning Translational and Knowledge-based Similarities from Relevance Rankings for Cross-Language Retrieval. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014*, volume 2, pages 488–494, Baltimore, MD, USA. *Cité page 50*

[Schütze, 1993] Schütze, H. (1993). Word space. *In Advances in neural information processing systems*, pages 895–902. *Cité page 65*

[Segond, 2000] Segond, F. (2000). Framework and results for French. *Computers and the humanities*, 34(1-2):49–60. *6 citations pages 33, 108, 109, 120, 121 et 183*

[Sharma *et al.*, 2016] Sharma, P., Tripathi, R. et Tripathi, R. C. (2016). Finding similar patents through semantic expansion. *In 2016 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–5. *Cité page 38*

[Sharma et Mittal, 2016] Sharma, V. K. et Mittal, N. (2016). Cross Lingual Information Retrieval (CLIR) : Review of Tools, Challenges and Translation Approaches. *In* Satapathy, S. C., Mandal, J. K., Udgata, S. K. et Bhateja, V., éditeurs : *Information Systems Design and Intelligent Applications*, numéro 433 de Advances in Intelligent Systems and Computing, pages 699–708. Springer India. DOI : 10.1007/978-81-322-2755-7_72. *Cité page 127*

[Silberer et Ponzetto, 2010] Silberer, C. et Ponzetto, S. P. (2010). UHD : Cross-Lingual Word Sense Disambiguation Using Multilingual Co-Occurrence Graphs. *In* Erk, K. et Strapparava, C., éditeurs : *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval@ACL 2010, Uppsala University, Uppsala, Sweden, July 15-16, 2010*, pages 134–137. The Association for Computer Linguistics. *Cité page 49*

[Singh *et al.*, 2013] Singh, V. K., Piryani, R., Uddin, A. et Waila, P. (2013). Sentiment analysis of Movie reviews and Blog posts. *In 2013 3rd IEEE International Advance Computing Conference (IACC)*, pages 893–898. *Cité page 32*

[Sinha et Mihalcea, 2007] Sinha, R. et Mihalcea, R. (2007). Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity. *In Proceedings of the International Conference on Semantic Computing*, ICSC '07, pages 363–369, Washington, DC, USA. IEEE Computer Society. *Cité page 39*

[Sokal et Michener, 1958] Sokal, R. R. et Michener, C. D. (1958). A statistical method for evaluating systematic relationship. *University of Kansas science bulletin*, 38(22):1409–1438. *Cité page 43*

[Sorg et Cimiano, 2012] Sorg, P. et Cimiano, P. (2012). Exploiting Wikipedia for Cross-lingual and Multilingual Information Retrieval. *Data Knowl. Eng.*, 74:26–45. *Cité page 21*

[Sparck Jones, 1972] Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21. *Cité page 12*

[Steinbach *et al.*, 2000] Steinbach, M., Karypis, G., Kumar, V. et others (2000). A comparison of document clustering techniques. *In KDD workshop on text mining*, volume 400, pages 525–526. Boston. *Cité page 43*

[Stevenson et Wilks, 2001] Stevenson, M. et Wilks, Y. (2001). The Interaction of Knowledge Sources in Word Sense Disambiguation. *Computational Linguistics*, 27(3):321–349. *Cité page 45*

[Stojanovic *et al.*, 2001] Stojanovic, N., Maedche, A., Staab, S., Studer, R. et Sure, Y. (2001). SEAL : A Framework for Developing SEmantic PortALs. *In Proceedings of the 1st International Conference on Knowledge Capture*, K-CAP '01, pages 155–162, New York, NY, USA. ACM. *Cité page 38*

[Sugawara *et al.*, 2015] Sugawara, H., Takamura, H., Sasano, R. et Okumura, M. (2015). Context Representation with Word Embeddings for WSD. *In Computational Linguistics*, Communications in Computer and Information Science, pages 108–119. Springer, Singapore. *Cité page 102*

[Symonds *et al.*, 2014] Symonds, M., Bruza, P., Zuccon, G., Koopman, B., Sitbon, L. et Turner, I. (2014). Automatic query expansion : A structural linguistic perspective. *JASIST*, 65(8):1577–1596. *Cité page 66*

[Symonds *et al.*, 2011] Symonds, M., Bruza, P. D., Sitbon, L. et Turner, I. (2011). Tensor query expansion : a cognitively motivated relevance model. *In 16th Australasian Document Computing Symposium*, Canberra, ACT. Australasian Document Computing Symposium. *Cité page 66*

[Tamine, 2000] Tamine, L. (2000). *Optimisation de requêtes dans un système de recherche d'information : Approche basée sur l'exploitation de techniques avancées de l'algorithme génétique*. Thèse de doctorat, Université Paul Sabatier - Toulouse III. *Cité page 10*

[Tchechmedjiev, 2012] Tchechmedjiev, A. (2012). État de l'art : mesures de similarité sémantique locales et algorithmes globaux pour la désambiguïsation lexicale à base de connaissances. *In Actes de la confrence conjointe JEP-TALN-RECITAL*, pages 295–308. ATALA. *Cité page 38*

[Tobin, 2015] Tobin, A. (2015). Is Google Translate Good Enough for Commercial Websites ? : A Machine Translation evaluation of text from English websites into four different languages. *Reitaku review : English culture research*, 21:94–116. *Cité page 150*

[Türe et Boschee, 2014] Türe, F. et Boschee, E. (2014). Learning to Translate : A Query-Specific Combination Approach for Cross-Lingual Information Retrieval. *In* Moschitti, A., Pang, B. et Daelemans, W., éditeurs : *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 589–599. ACL. *Cité page 50*

[Turney, 2001] Turney, P. D. (2001). Mining the Web for Synonyms : PMI-IR versus LSA on TOEFL. *In Machine Learning : ECML 2001*, pages 491–502. Springer, Berlin, Heidelberg. *Cité page 36*

[Turney et Pantel, 2010] Turney, P. D. et Pantel, P. (2010). From Frequency to Meaning : Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188. *4 citations pages 36, 43, 65 et 66*

[Tyar et Win, 2015] Tyar, S. M. et Win, T. (2015). Jaccard coefficient-based word sense disambiguation using hybrid knowledge resources. *In 2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pages 147–151. *Cité page 37*

[Van Rijsbergen, 2004] Van Rijsbergen, C. J. (2004). *The geometry of information retrieval*. Cambridge University Press. *Cité page 30*

[Véronis, 1998] Véronis, J. (1998). A Study of Polysemy Judgements and Inter-Annotator Agreement. *In In Programme and advanced papers of the Senseval workshop, Herstmonceux*, pages 2–4. *2 citations pages 121 et 123*

[Véronis, 2001] Véronis, J. (2001). Sense tagging : does It make sense ? *In Corpus Linguistics' 2001 Conference*. Citeseer. *Cité page 42*

[Véronis, 2003] Véronis, J. (2003). Cartographie lexicale pour la recherche d'information. *Actes de TALN 2003*, pages 265–274. *Cité page 44*

[Véronis, 2004] Véronis, J. (2004). HyperLex : lexical cartography for information retrieval. *Computer Speech & Language*, 18(3):223–252. *2 citations pages 44 et 49*

[Véronis et Ide, 1990] Véronis, J. et Ide, N. M. (1990). Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries. *In Proceedings of the 13th Conference on Computational Linguistics - Volume 2*, COLING '90, pages 389–394, Stroudsburg, PA, USA. Association for Computational Linguistics. *3 citations pages 30, 40 et 101*

[Vidhu Bhala et Abirami, 2014] Vidhu Bhala, R. V. et Abirami, S. (2014). Trends in Word Sense Disambiguation. *Artificial Intelligence Review*, 42(2):159–171. *3 citations pages 38, 98 et 122*

[Viera et Garrett, 2005] Viera, A. J. et Garrett, J. M. (2005). Understanding interobserver agreement : the kappa statistic. *Family Medicine*, 37(5):360–363. *Cité page 118*

[Voorhees, 1994] Voorhees, E. M. (1994). Query Expansion Using Lexical-semantic Relations. *In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 61–69, New York, USA. Springer-Verlag New York, Inc. *2 citations pages 63 et 128*

[Voorhees, 1999] Voorhees, E. M. (1999). The TREC-8 Question Answering Track Report. *In In Proceedings of TREC-8*, pages 77–82. *2 citations pages 25 et 67*

[Voorhees et Harman, 2005] Voorhees, E. M. et Harman, D. K. (2005). *TREC : Experiment and Evaluation in Information Retrieval*. The MIT Press. *2 citations pages 28 et 30*

[Vulic et Moens, 2014] Vulic, I. et Moens, M.-F. (2014). Probabilistic models of cross-lingual semantic similarity in context based on latent cross-lingual concepts induced from comparable data. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 349–362. ACL. *Cité page 48*

[Vulić et Moens, 2015] Vulić, I. et Moens, M.-F. (2015). Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings. *In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 363–372, New York, NY, USA. ACM. *Cité page 22*

[Wang *et al.*, 2008] Wang, X., Fang, H. et Zhai, C. (2008). A Study of Methods for Negative Relevance Feedback. *In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 219–226, New York, NY, USA. ACM. *Cité page 58*

[Wang *et al.*, 2015] Wang, X., Zhang, Q., Wang, X. et Li, J. (2015). Cross-lingual Pseudo Relevance Feedback Based on Weak Relevant Topic Alignment. *In Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 529–534. *Cité page 70*

[Weiss, 1973] Weiss, S. F. (1973). Learning to disambiguate. *Information Storage and Retrieval*, 9(1):33–41. *Cité page 42*

[Wilks *et al.*, 1990] Wilks, Y., Fass, D., Guo, C.-m., McDonald, J. E., Plate, T. et Slator, B. M. (1990). Providing machine tractable dictionary tools. *Machine Translation*, 5(2):99–154. *Cité page 30*

[Williams et Giles, 2016] Williams, K. et Giles, C. L. (2016). Improving Similar Document Retrieval Using a Recursive Pseudo Relevance Feedback Strategy. *In Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, JCDL '16, pages 275–276, New York, NY, USA. ACM. *Cité page 61*

[Wu et Palmer, 1994] Wu, Z. et Palmer, M. (1994). Verbs Semantics and Lexical Selection. *In Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA. Association for Computational Linguistics. *2 citations pages 37 et 38*

[Xiong et Zhang, 2014] Xiong, D. et Zhang, M. (2014). A Sense-Based Translation Model for Statistical Machine Translation. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1 : Long Papers*, pages 1459–1469. *Cité page 31*

[Xu et Croft, 2000] Xu, J. et Croft, W. B. (2000). Improving the Effectiveness of Information Retrieval with Local Context Analysis. *ACM Transactions on Information Systems (TOIS)*, 18(1):79–112. *Cité page 61*

[Yan *et al.*, 2003] Yan, R., Hauptmann, A. et Jin, R. (2003). Multimedia Search with Pseudo-relevance Feedback. *In Proceedings of the 2Nd International Conference on Image and Video Retrieval*, CIVR'03, pages 238–247, Berlin, Heidelberg. Springer-Verlag. *Cité page 58*

[Yarowsky, 2000] Yarowsky, D. (2000). Hierarchical Decision Lists for Word Sense Disambiguation. *Computers and the Humanities*, 34(1-2):179–186. *2 citations pages 45 et 98*

[Yin *et al.*, 2009] Yin, Z., Shokouhi, M. et Craswell, N. (2009). Query Expansion Using External Evidence. *In Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 362–374. Springer, Berlin, Heidelberg. *Cité page 67*

[Yoo et Choi, 2010] Yoo, S. et Choi, J. (2010). On the query reformulation technique for effective MEDLINE document retrieval. *Journal of Biomedical Informatics*, 43(5):686–693. *Cité page 61*

[Yuret et Yatbaz, 2010] Yuret, D. et Yatbaz, M. A. (2010). The Noisy Channel Model for Unsupervised Word Sense Disambiguation. *Computational Linguistics*, 36(1):111–127. *2 citations pages 45 et 101*

[Zadeh, 1978] Zadeh, L. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1(1):3–28. *2 citations pages 77 et 130*

[Zhang *et al.*, 2009] Zhang, J., Deng, B. et Li, X. (2009). Concept Based Query Expansion Using WordNet. *In Proceedings of the 2009 International e-Conference on Advanced Science and Technology*, AST '09, pages 52–55, Washington, DC, USA. IEEE Computer Society. *Cité page 63*

[Zhou *et al.*, 2012] Zhou, D., Truran, M., Brailsford, T., Wade, V. et Ashman, H. (2012). Translation techniques in cross-language information retrieval. *ACM Computing Surveys (CSUR)*, 45(1):1 :1–1 :44. *5 citations pages 16, 17, 18, 19 et 22*

[Zhou et Han, 2005] Zhou, X. et Han, H. (2005). Survey of Word Sense Disambiguation Approaches. *In FLAIRS Conference*, pages 307–313. *Cité page 98*

ANNEXES

Annexe

A Description du corpus de test ROMANSEVAL

A.1 Corpus documentaire

L'ensemble de documents *JOC*, utilisé dans le standard de test ROMANSEVAL, est composé de textes parallèles en neuf langues faisant partie du Journal Officiel de la Commission européenne (série C, année 1993).). Les textes (au nombre de plusieurs milliers) sont constitués de *questions* écrites des parlementaires européens sur un large éventail de sujets, et des *réponses* correspondantes de la Commission européenne [Segond, 2000].

La taille totale du corpus est d'environ 10,2 millions de mots (environ 1,1 millions de mots par langue) correspondant à l'année 1993, qui ont été recueillis et préparés au sein des projets MLCC-MULTEXT.

ROMANSEVAL intègre la version en langue française du corpus JOC sous un format qui sera présenté dans la section qui suit.

A.2 Format des documents

Les documents de ROMANSEVAL sont écrit en texte plat (et non pas dans un format XML). Les identifiants des revues originales sont marqués avec un en-tête commençant par `&&&` suivi de l'identifiant de document. Chaque question débute par le symbole `$`. [1]

Exemple [2] :

1. Site de ROMANSEVAL : http ://aune.lpl.univ-aix.fr/projects/romanseval/
2. Source : http ://aune.lpl.univ-aix.fr/projects/multext/samples/word-joc-fr.gz

```
&&&JOCWQ-006.fr&&&
$1
Objet: Organigramme de la Commission
La Commission peut-elle:
1) indiquer le nombre d'agents temporaires travaillant dans ses
   services,
2) en dresser la liste et préciser les critères selon lesquels ils
   sont sélectionnés,
3) indiquer le nombre de candidats ayant réussi à un concours qui
   figurent sur des listes de réserve
$2
Réponse donnée par  M.  Cardoso e Cunha au nom de la Commission (22
    septembre 1992) 1 et 2. La Commission transmet directement à l'
   honorable parlementaire et au Secrétariat général du Parlement
   européen des tableaux reprenant le nombre d'agents temporaires
   en service à la Commission.
```

A.3 Préparation du standard de test ROMANSEVAL

La préparation de la compétition a concerné l'extraction et le choix des mots à désambiguïser. Le corpus a été découpé en mots, étiqueté (avec, en particulier, des étiquettes catégorielles permettant de distinguer les noms N, les adjectifs A et les verbes V). Ensuite, les 600 mots les plus fréquents (200 N, 200 A, 200 V) ont été extraits, ainsi que leurs contextes d'apparition. Ces mots ont été annotés parallèlement par 6 étudiants en Linguistique, conformément aux sens du *Petit Larousse*, chaque occurrence de mot pouvant recevoir une étiquette de sens, plusieurs ou aucune.

Par exemple, dans le *Petit Larousse*, le mot *biologique* a trois sens possibles :

1. *Relatif à la biologie*
2. *Sans engrais ni pesticides chimiques. Pain biologique. Abrév. (fam.) : bio.*
3. *Arme biologique, utilisant des organismes vivants ou des toxines.*

Dans le contexte suivant, *« Frontières humaines (HFSP) est une initiative japonaise proposée en 1987 dans le cadre du Sommet économique (G7) dans le but de favoriser la coopération internationale dans le domaine de la recherche sur le cerveau et sur les mécanismes moléculaires dans les fonctions biologiques* », le mot *biologique* a reçu l'étiquette 1 par les annotateurs.

Après cette première étape, les 60 mots les plus polysémiques ont été conservés (20 N, 20 A, 20 V). Le corpus proposé aux participants pour l'expérience était donc formé de 60 mots et des 3624 contextes où ils apparaissent, chaque mot ayant environ 60 occurrences.

A.4 Mots de test

- **Noms** : barrage, chef, communication, compagnie, concentration, constitution, degré, détention, économie, formation, lancement, observation, organe, passage, pied, restauration, solution, station, suspension, vol
- **Adjectifs** : biologique, clair, correct, courant, exceptionnel, frais, haut, historique, plein, populaire, régulier, sain, secondaire, sensible, simple, strict, sûr, traditionnel, utile, vaste
- **Verbes** : arrêter, comprendre, conclure, conduire, connaître, couvrir, entrer, exercer, importer, mettre, ouvrir, parvenir, passer, porter, poursuivre, présenter, rendre, répondre, tirer, venir

A.5 Format des définitions issues de *Le Petit Larousse*

Exemple d'un extrait de définition du terme degré issu du dictionnaire *Le Petit Larousse* :

```
degré I Litt. Marche d'un escalier.
degré II0 Chacun des états intermédiaires pouvant conduire d'un
   état à un autre. - Par degrés : progressivement.
degré II1 Échelon, grade, etc., dans une hiérarchie.
degré II2 Dr. Degré d'une juridiction : chacun des tribunaux devant
    lesquels une affaire peut être successivement portée.
degré II3 Degré de parenté : distance qui sépare des parents
   consanguins ou par alliance. (Chaque génération forme un degré ;
    en ligne collatérale, les degrés se comptent en remontant d'un
   parent à l'ancêtre commun et en redescendant de celui-ci à l'
   autre parent : deux frères sont parents au deuxième degré, l'
   oncle et le neveu au troisième, etc.)
degré II45 Gramm. Degré de comparaison ou de signification : chacun
    des degrés (relatif ou absolu) de la qualité exprimée par un
   adjectif ou un adverbe (positif, comparatif ou superlatif).
```

```
degré II4a Math. Degré d'une équation entière ou d'un polynôme:
   degré du monôme composant ayant le plus haut degré.
degré II4b Math. Degré d'un monôme entier par rapport à une
   variable : exposant de la puissance à laquelle se trouve élevée
   cette variable dans le monôme.
degré II4c Math. Degré d'une courbe, d'une surface algébrique :
   degré de son équation algébrique.
```

A.6 Évaluation

Le standard de test contient la liste des mots à désambiguïser sous le format (décomposé de 7 champs) suivant : *catégorie, n° de l'occurrence, lemme, n° du paragraphe, position dans le paragraphe, longueur de l'occurrence, occurrence.*

Soit par exemple :

```
A, 1, biologique, 1608, 264, 11, biologiques
A, 2, biologique, 1645, 682, 10, biologique
```

L'objectif de la tâche de désambiguïsation est de retourner l'étiquetage de sens fourni par le système, conformément aux sens du dictionnaire *Le Petit Larousse.*