

---

# La recherche d'information multilingue : désambiguïsation et expansion de requêtes basées sur WordNet

Mustapha Baziz, Mohand Boughanem, Nawel Nassr.

Laboratoire IRIT/SIG  
Campus Universitaire Toulouse III  
118, Route de Narbonne  
F-31062 Toulouse Cedex 4  
Email : {baziz, boughane, nassr}@irit.fr  
Tel : 05-61-55-68-99 Fax : 05-61-55-62-58

---

## Résumé :

*Cet article présente une approche de Recherche d'Information Multilingue. Elle aborde le problème majeur de ce domaine, à savoir la désambiguïsation et l'expansion des requêtes traduites. Pour réaliser la traduction automatique des requêtes, des dictionnaires bilingues sont utilisés. La désambiguïsation de ces requêtes, notamment quand plusieurs traductions possibles sont proposées, s'appuie sur des concepts issus d'une base de données lexicographique externe (WordNet). Des liens sémantiques dénotant des relations telles que spécifique/générique ou partie/tout sont ensuite utilisés pour l'expansion de ces requêtes. Toutes nos expérimentations ont été effectuées sur la base CLEF et utilisent notre moteur de recherche d'information "Mercure".*

**Mots-clés :** croisement de langues, système de recherche d'information, dictionnaires, traduction de requêtes, désambiguïsation et expansion de requêtes, relations sémantiques, WordNet.

## Abstract :

*This paper presents our approach to Cross Language Information Retrieval (CLIR). It deals with the major problem of this field, namely the disambiguation and the expansion of translated queries. Bilingual dictionaries are used for query translation. The disambiguation of these queries, when several translations are possible is based on concepts issued from a large extern linguistic database like WordNet. Semantic links like generalization/specialization or part/whole are used for expanding these queries. All the experiments were done on CLEF data, using our connectionist information retrieval system, Mercure.*

**Keywords :** cross-language, information retrieval system, query translation, dictionary, disambiguation and expansion, semantic relations, WordNet.

# 1. Introduction

L'utilisation d'une langue universelle, vieux rêve philosophique, semble être encore pour longtemps une utopie. La multitude de langues actuellement présentes sur notre planète restera encore une source de problèmes pour tous ceux désirant trouver des informations. Ces problèmes apparaîtront qu'elle que soit la langue dans laquelle celles-ci s'expriment.

Le développement que connaît Internet, avec la prolifération de collections d'informations écrites dans différentes langues, conduit le réseau des réseaux vers un multilinguisme de fait. La sélection d'informations pertinentes, répondant à un besoin en information de l'utilisateur, est donc confrontée à un double problème. Le premier, spécifique à la recherche d'information (RI), réside dans la capacité du système de recherche d'informations (SRI) à séparer les informations pertinentes de celles qui ne le sont pas. Le second, lié au multilinguisme, correspond à la capacité du système d'aller au-delà de la langue de la requête.

La question principale inhérente à la recherche d'information par croisement de langues est : *comment à partir d'une requête exprimée dans une langue donnée, récupérer des documents écrits dans des langues différentes de celle de la requête ?* en d'autres termes quelles représentations ou quelles transformations faut-il faire aux documents et/ou aux requêtes pour pouvoir les comparer.

La plupart des solutions proposées aujourd'hui ont adopté la traduction des documents et/ou des requêtes comme moyen pour mettre ces documents et ces requêtes dans un même référentiel. Ceci revient :

- Soit à traduire la requête vers la langue des documents. Il s'agit de présenter au moteur de recherche les traductions de cette requête dans les différentes langues souhaitées. Le système récupérera alors les différents documents correspondants à chaque traduction,
- Soit à traduire les documents vers la langue de la requête. Les documents sont traduits dans la langue de la requête à l'aide d'outils de traduction. Le système de recherche d'information procède ensuite à une simple interrogation monolingue. Son principal inconvénient est lié à la taille de la base. Il n'est pas concevable de traduire une collection de documents dans toutes les langues souhaitées pour l'interrogation.
- Ou encore, à traduire la requête et les documents. Dans ce cas, il s'agit de représenter la requête et les documents dans un même référentiel. Ce référentiel est souvent un vocabulaire multilingue prédéfini qui peut être par exemple un thesaurus (exemple **EuroWordNet** ). Cependant l'inconvénient de ce type de vocabulaire est qu'il n'est pas toujours disponible.

Actuellement la plupart des travaux dans ce domaine se focalisent sur la traduction de la requête. Cette traduction est moins coûteuse que celle de tous les documents de la collection [Oard 96], [Davis 96], [Ballestros 96], [Sanderson 00], [Shauble 00] et [Boughanem 00].

Pour notre part, étant donné que les requêtes sont plus courtes que les documents, nous pensons aussi qu'il est plus réaliste de traduire la requête seulement. La requête est souvent une suite de termes, situation que l'on rencontre couramment dans les moteurs de recherche.

Cependant, la traduction de ses requêtes n'est pas sans engendrer des problèmes. Le problème d'expressivité de la requête traduite est posé quand les termes issus de la traduction ne sont

pas suffisants pour représenter la requête initiale. D'où nécessité d'expansion pour enrichir la requête avec des termes plus courants. Mais le problème le plus crucial à résoudre est sans doute le problème d'ambiguïté, notamment quand plusieurs traductions pour un ou plusieurs termes de la requête sont possibles.

Nous nous intéresserons dans cet article à l'étude d'une technique de désambiguïsation et d'expansion des termes de traductions de requêtes. Cette désambiguïsation s'appuie sur des concepts issus d'une base de données lexicographique externe (WordNet). Des termes issus de concepts liés sémantiquement à ceux de la requête traduite sont ensuite utilisés pour l'expansion de ces requêtes.

La section 2 de cet article présente un état de l'art sur le croisement de langues en recherche d'information. La section 3 explique pourquoi on est amené à désambiguïser et à étendre les requêtes. La section 4, décrit la technique de désambiguïsation, que nous avons proposée pour le croisement de langues. La section 5 présente les expérimentations effectuées sur la base CLEF'2001<sup>1</sup> ainsi que les résultats obtenus.

## 2. Etat des recherches dans le domaine de la RI multilingue

La notion de multilinguisme en RI peut se présenter sous différentes facettes [Oard 96]. La facette à laquelle nous nous intéressons dans ce papier est la recherche d'information par croisement de langues ou cross-language Information Retrieval (CLIR) [Grefenstette 98]. La recherche d'information par croisement de langues est la possibilité offerte à un SRI de sélectionner des documents exprimés dans une langue différente de celle de la requête. Dans le contexte de la traduction de requêtes, les systèmes tentent de résoudre deux problèmes majeurs.

Le premier problème concerne la *traduction des termes de la requête*. Dans ce cas, on essaie de substituer à chaque terme exprimé dans la langue source (L1), un ou plusieurs terme(s) sensé(s) le représenter dans la langue cible (L2). Le second problème est posé dans le cas où un terme possède plusieurs traductions et est lié au choix de la ou des meilleure(s) traduction(s), c'est le *problème de la désambiguïsation*.

Les méthodes de traduction de requêtes proposées sont basées sur l'utilisation de :

– *Dictionnaires bilingues* : l'idée principale de ces techniques Davis [Davis 96], Ballestros [Ballestros 96], Hull et al. [Hull 96] et Sanderson [Sanderson 00], est de remplacer chaque terme de la requête par le(s) terme(s) approprié(s) dans la langue cible. Les dictionnaires bilingues tels que ceux développés par les humains sont actuellement la forme la plus répandue des structures ayant une couverture suffisante pour réaliser les applications de croisement de langues. C'est aussi pour cela que les méthodes basées sur des dictionnaires sont les plus utilisées dans la recherche d'information par croisement de langues.

– *Corpus alignés (parallèles ou comparables)* : les méthodes basées sur le corpus Davis [Davis 96], Ballestros [Ballestros 98], Sheridan [Sheridan 96], Braschler et al [Schauble00], Boughanem [Boughanem 00] et Nassr [Nassr 02] utilisent directement le contenu d'un ensemble de documents, regroupés dans un corpus soit pour la traduction ou pour la désambiguïsation des requêtes. Un corpus aligné est constitué d'un ensemble de documents

---

<sup>1</sup> CLEF: Cross Language Evaluation Forum, est un programme de recherche pour l'évaluation des techniques et approches de croisement de langues en recherche d'information, (<http://galileo.iei.pi.cnr.it/DELOS/CLEF/>)

exprimés dans une langue, alignés avec des documents dans une autre langue. L'alignement entre ces documents consiste à mettre en correspondance les documents de langues différentes selon un critère donné. Il peut être parallèle ou comparable.

– L'alignement parallèle consiste à mettre en correspondance chaque document d'une langue source L1 avec le document représentant sa traduction dans la langue cible L2. Dans ce cas l'alignement peut être fait sur : le document, les paragraphes, les phrases ou les termes. Les corpus basés sur ce type d'alignement sont appelés les **corpus parallèles**.

– L'alignement comparable plus délicat à réaliser [Sheridan 96], revient à mettre en correspondance des documents en se basant sur des critères comme par exemple la présence de mêmes dates, de mêmes noms de personnes dans des documents de langues différentes [Grefenstette 98], [Oard 96]. Les corpus basés sur ce type d'alignement sont appelés les **corpus comparables**.

**Traducteurs automatiques** : les techniques basées sur les traducteurs automatiques nécessitent l'intégration d'un logiciel de traduction automatique dans le système de recherche d'information [Radwan 1994], [Pirkola 98]. Les systèmes basés sur les traducteurs automatiques sont utilisés pour obtenir un même texte dans plusieurs langues, avec ou sans l'aide d'un expert. Ces systèmes sont généralement plus complexes et loin d'être parfaits, car ils s'appuient sur des grammaires et autres méthodes linguistiques ; même s'ils donnent des résultats satisfaisants pour la traduction des documents, leur utilisation pour la traduction de requêtes n'a pas connu le même succès, du fait que ces dernières, sont souvent courtes et exprimées par des mots indépendants Yamabana [Yamabana 98], Oard [Oard 96] et Gey [Gey 99].

L'utilisation d'un dictionnaire, quand une version électronique de celui-ci existe et est facilement exploitable, est le moyen le plus simple pour réaliser la traduction de requêtes. Mais tous les travaux utilisant les dictionnaires pour la traduction de requête, ont démontré que pour améliorer les résultats de la recherche d'information par croisement de langues, il est nécessaire de combiner le dictionnaire avec une méthode de désambiguïsation stricte qui permet de réduire l'ambiguïté des termes fournis par le dictionnaire. Une part importante des travaux effectués actuellement, explorent cette direction et tentent de chercher des stratégies de désambiguïsation efficaces.

Différents travaux ont proposé une variété de stratégies pour la désambiguïsation des termes de la requête [Grefenstette 98], [Oard 98]. Les travaux recensés sont principalement basés sur les corpus alignés parallèles et comparables. La plupart des approches de désambiguïsation basées sur les corpus alignés utilisent des cooccurrences entre termes calculées à partir de ce corpus pour choisir la(es) meilleur(es) substitution(s) possibles pour un terme donné. Ainsi dans Ballestros [Ballestros 97], les valeurs de cooccurrences sont calculées entre les termes anglais et espagnols en se basant sur un corpus parallèle (espagnol-anglais). La désambiguïsation consiste à retenir pour chaque terme anglais le terme espagnol le plus cooccurent parmi les substitutions possibles obtenues par le dictionnaire COLLINS (anglais-espagnol) pour ce terme anglais. Elle a montré que la précision moyenne est améliorée de 31% par rapport aux résultats obtenus par le dictionnaire.

L'approche proposée par Davis et Odgen [Davis 97], n'utilise pas de valeurs de cooccurrence entre termes, mais effectue plusieurs recherches monolingues sur chacune des parties du corpus parallèle (anglais-espagnol) à l'aide d'un SRI de modèle vectoriel QUILT. Tout d'abord, une recherche monolingue est effectuée avec la requête sur une partie du corpus

parallèle pour trouver la liste ordonnée de documents résultats. Ensuite, une recherche monolingue sur l'autre partie du corpus parallèle est effectuée pour chacune des traductions possibles d'un terme de la requête. Le produit scalaire entre les différents vecteurs est ensuite calculé, entre les vecteurs de documents de chaque traduction et le vecteur de document du terme source. La traduction choisie est celle qui obtient une liste de documents la plus proche de la liste de la requête. Dans cette approche, il s'agit encore de faire une traduction mot à mot des termes de la requête. Ils ont montré que la désambiguïsation améliore de 37% les résultats obtenus par la traduction simple par dictionnaire. Ils ont remarqué également que la traduction choisie par le système ne favorise pas forcément les traductions les plus fréquentes dans le corpus.

Yamabana [Yamabana 98] a développé une méthode de désambiguïsation utilisant un corpus comparable. L'approche proposée consiste à calculer automatiquement à partir de ce corpus l'ensemble des valeurs de cooccurrence entre les termes de la langue source et les termes de la langue cible. Ce thesaurus est utilisé pour sélectionner la meilleure traduction en langue cible.

### **3. Pourquoi désambiguïser et étendre les requêtes ?**

Dans le domaine de la recherche d'information, la source principale de l'ambiguïté réside dans les variations linguistiques présentes dans le texte des requêtes et des documents. Ces variations linguistiques peuvent être interprétées par le fait que le langage n'est pas simplement une collection de mots, mais un moyen de communiquer au sujet de concepts. Ce qui rend l'hypothèse de récupération de mots clé insuffisante. Ceci est dû au fait que les termes utilisés par l'utilisateur dans sa requête, peuvent présenter par rapport à ceux des documents de la base, des variations morphologiques (comme dans « wolf » et « wolves »), des variations lexicales ou des mots différents sont utilisés pour représenter le même sens (« film » et « movie ») ou encore des variations sémantiques, où des mots ont plusieurs sens: Un pétrolier cherchant par exemple le mot «oil » sera confronté à «olive oil and kitchen». D'où l'idée de désambiguïsation des requêtes traduites avec des concepts issus d'une base lexicographique externe. Ces requêtes sont ensuite étendues avec d'autres concepts reliés à ceux des requêtes traduites par des liens sémantiques tels que, spécifique/générique ou partie/tout. Ceci permet, à titre d'exemple pour le mot « country », en utilisant la synonymie, de récupérer les mots « state » et « land » et la relation hyperonymie, de récupérer les mots « political unit ». ou encore de passer du sigle « EU » à « European Union, EU, European Community, EC, European Economic Community, EEC, Common Market, Europe ».

### **4. Description de l'approche suivie**

Comme pour tout système de RI, l'étape d'indexation est nécessaire. Les documents exprimés dans une langue L1 (français dans notre cas) y sont analysés. Les requêtes quant à elles, sont d'abord traduites en utilisant un dictionnaire bilingue vers la langue L2 (anglais), avant qu'elles ne soient indexées à leur tour.

A ces étapes classiques connues des systèmes de RI et de traduction automatique, est rajoutée une étape de désambiguïsation et expansion des requêtes traduites. De manière générale, la désambiguïsation consiste à sélectionner, parmi les différents sens possibles que puisse exprimer une requête, celui qui est sensé la représenter le mieux. L'expansion quant à elle, permet d'enrichir la requête en lui ajoutant les termes des concepts d'une base de données lexicographique. Ces concepts sont reliés à la requête par des liens sémantiques qui sont l'hyperonymie/hyponymie pour la généralisation/spécialisation et l'holonymie/meronymie

pour le lien de composition. Elles sont définies dans le paragraphe 4.3. Pour illustrer, une vue générale et schématique de cette approche est représentée dans Figure 1.

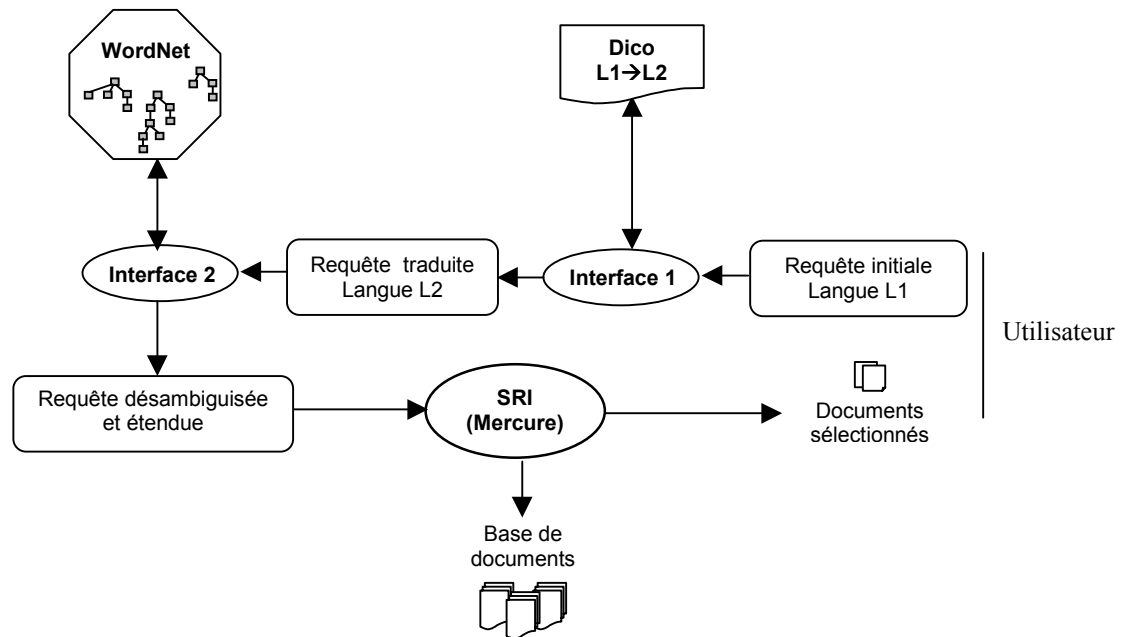


Figure 1. Schéma synoptique de la stratégie suivie.

Dans les paragraphes suivants, nous détaillons chacune de ces opérations.

## 4.1 Indexation

Cette opération automatique consiste, d'une façon générale, à extraire les termes (non vides) et leur fréquence à partir des documents, puis à les radicaliser. Concernant l'analyse de la requête, le texte libre (topics) de la requête source est analysé initialement, afin d'extraire tous les mots clés. Le résultat de cette opération est une liste de mots significatifs avec un poids associé à chaque terme.

## 4.2 Traduction de Requêtes basée sur les Dictionnaires Bilingues

Cette opération traduit chaque terme de la requête source en un ou plusieurs traductions. Ce processus est réalisé à partir des dictionnaires bilingues. Généralement, un dictionnaire est représenté par une liste de termes exprimés en langue source, qui sont alignés avec d'autres termes, en langue cible. Dans ce cas, chaque terme de la requête source est remplacé par le(s) terme(s) avec lequel (lesquels) il est (sont) associé(s).

L'idée de base de cette technique est d'exploiter la structure de réseau sémantique présente dans WordNet pour, d'une part, désambiguïser la requête issue de la traduction, et étendre cette dernière avec les meilleurs concepts de WordNet liés sémantiquement à ceux de la requête dans la langue cible, d'autres parts.

## 4.3 Désambiguïsation à l'aide des concepts de WordNet

La désambiguïsation de la requête traduite 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. Elle s'impose à partir du moment où le dictionnaire retourne plusieurs traductions possibles pour un ou plusieurs termes de la requête

initiale. Par exemple pour le terme en français «retracer», le dictionnaire retourne « to relate, account, to redraw, draw again ». La désambiguïsation consiste dans ce cas, à trancher sur le sens à prendre pour le terme « retracer ». C-à-d, le sens raconter ou dessiner.

De manière générale, la désambiguïsation de la requête peut se faire suivant deux cas :

**Cas 1 :** c'est le cas le plus favorable. Ici les termes de la requête sont liés sémantiquement et contribuent ensemble à former un même concept. Comme par exemple la requête « north american countries ».

L'exploitation dans ce cas de cette "requête-concept" est très important, du fait qu'au lieu que l'expansion de la requête traduite se fasse pour chacun des termes qui la composent, comme dans Cas2 (donc pour une requête de 3 mots par exemple, on récupérera des concepts pour  $3*5$  relations = 15 \*n concepts avec n entre 0 et quelques dizaines de concepts en général ), toute la requête est prise comme un seul concept.

Pour illustrer, considérons les sens retournés par WordNet, d'abord pour les termes pris séparément (les différents sens pour uniquement le premier terme "north" sont donnés dans l'exemple1), puis pour le concept multitermes "north american countries", et ce, pour la seule relation synonymie :

**Exemple 1:**

The noun "north" has 6 senses in WordNet.

1. North -- (the region of the United States lying north of the Mason-Dixon line)
2. Union, North -- (the United States (especially the northern states during the American Civil War); "he has visited every state in the Union"; "Lee hoped to detach Maryland from the Union"; "the North's superior resources turned the scale")
3. **north**, due north, N -- (the cardinal compass point that is at 0 or 360 degrees)
4. North, northland, septentrion -- (any region lying in or toward the north)
5. **north**, magnetic north, compass north -- (the direction in which a compass needle points)
6. North, Frederick North, Second Earl of Guilford -- (British statesman under George III whose policies led to rebellion in the American colonies (1732-1792))

The noun "north american country" has 1 sense in WordNet.

1. North American country, North American nation -- (a country on the North American continent)

Les résultats complets sont synthétisés dans Table.1.

Sans détection de concepts	<b>Cas1</b>	<b>Mot « north »</b>				
	Relation	Synonymie	Hyperonymie	Hyponymie	Meronymie	Holonymie
	Nombre de sens	6	6 (pour les 6 sens)	0	2	1 (seul sens 1 a 1 holo.)
		<b>Mot « american »</b>				
	Nbre de sens	3	3	66	0	1
		<b>Mot « countries »</b>				
	Nbre de sens	5	5	107	7	0
	<b>Total</b>	14	14	173	9	2
<b>Cas2</b>	<b>Concept « north american countries »</b>					
	Nbre de sens	1	1	4	4	1

Table.1. Deux cas de figures pour la requête " north american countries "

Il est clair (voir Table.1), que le fait de considérer le lien entre les termes utilisés dans la requête, "north american countries", réduit considérablement (dans ce cas précis au maximum : Nbre sens=1) l'ambiguïté. Par contre, si les liens entre les mots ne sont pas considérés, il suffit de sélectionner le mauvais concept pour une seule relation (ex. un sens parmi 14 ou un hyponyme parmi 173) pour entraîner la requête étendue dans le faux. Dans ce cas précis, la désambiguïsation se fait comme illustrée dans Cas2. Notons que la

reconnaissance de ses concepts dépend de la ressource sémantique utilisée (WordNet dans notre cas) et de la requête utilisateur.

**2) Cas 2 :** Dans le cas où les termes d'une requête sont indépendants, donc aucun concept multitermes n'est reconnu par WordNet, ce qui est le cas de la requête de l'exemple 2 ci-dessous, "graham bell", la désambiguïsation se fait en interrogeant WordNet, cette fois, pour chacun des termes de la requête. Le résultat est un ensemble de concepts.

Le meilleur concept qui désambiguïse la requête, est ensuite élu par une simple mesure de similitude. Pour cet exemple, les concepts retournés sont comme suit.

**Exemple 2: requête :** "graham bell"

The **noun** "graham" has 3 senses in WordNet.

1. Graham, Billy Graham, William Franklin Graham -- (United States evangelical preacher famous as a mass evangelist (born in 1918))
2. Graham, Martha Graham -- (United States dancer and choreographer whose work was noted for its austerity and technical rigor (1893-1991))
3. whole wheat flour, graham flour, **graham**, whole meal flour -- (flour made by grinding the entire wheat berry including the bran; ('whole meal flour' is British usage))

The **noun** "bell" has 9 senses in WordNet.

1. **bell** -- (a hollow device made of metal that makes a ringing sound when struck)
2. doorbell, **bell**, buzzer -- (a push button at an outer door that gives a ringing or buzzing signal when pushed)
3. **bell** -- (the sound of a bell; "saved by the bell")
4. **bell**, ship's bell -- ((nautical) each of the eight half-hour units of nautical time signaled by strokes of a ship's bell; eight bells signals 4:00, 8:00, or 12:00 o'clock, either a.m. or p.m.)
5. **bell**, bell shape, campana -- (the shape of a bell)
6. Bell, Vanessa Bell, Vanessa Stephen -- (English painter; sister of Virginia Woolf; prominent member of the Bloomsbury Group (1879-1961))
7. Bell, Alexander Bell, Alexander Graham Bell -- (American inventor of the telephone (1847-1922))
8. chime, **bell**, gong -- (a percussion instrument consisting of vertical metal tubes of different lengths that are struck with a hammer)
9. **bell** -- (the flared opening of a tubular device)

On procède ensuite par superposition et on prend le concept qui présente le plus grand nombre de mots similaires avec la requête. Dans notre cas, le meilleur synset pris correspond au sens 7 du deuxième mot de la requête avec une mesure de similitude égale à 4: 7. Bell, Alexander Bell, Alexander Graham Bell -- (American inventor of the telephone (1847-1922)),

Vient enfin l'étape de l'expansion, où l'on ajoute les concepts (sans la partie glossaire et avec restriction sur le nombre de mots) issus des 4 relations sémantiques qui sont: l'hyponymie, l'hyponymie, la meronymie et l'holonymie. Ces relations sont définies comme suit :

Relation **Hyponymie** : C'est le terme générique utilisé pour désigner une classe englobant des instances de classes plus spécifiques. Y est un *hyponyme* de X si X est un type de (kind of) Y.

Relation **Hyponymie** : C'est le terme spécifique utilisé pour désigner un membre d'une classe (relation inverse de *Hyponymie*). X est un *hyponyme* de Y si X est un type de (kind of) Y.

Relation **Holonymie** : Le nom de la classe globale dont les noms *meronymes* font partie. Y est un *holonyme* de X si X est une partie de (is a part of) Y.

Relation **Meronymie** : Le nom d'une partie constituante (part of), substance de (substance of) ou membre (member of) d'une autre classe (relation inverse de *l'holonymie*). X est un *meronyme* de Y si X est une partie de Y. exemple : {voiture} a pour *meronymes* {{porte}, {moteur}}.

On peut illustrer l'expansion, en prenant comme exemple la requête "énergie solaire" où le problème de désambiguïsation après sa traduction en anglais ne se pose pas. A titre d'exemple, pour cette requête, et pour la première relation, l'hyponymie, WordNet retourne: (ci-dessous, l'hyponyme vient après le signe =>)

Results for "Hypernyms (this is a kind of...)" search of noun "solar energy"

1 sense of solar energy

Sense 1

solar energy, solar power -- (energy from the sun that is converted into thermal or electrical energy; "the amount of energy falling on the earth is given by the solar constant, but very little use has been made of solar energy")

=> alternative energy -- (energy derived from sources that do not use up natural resources or harm the environment)

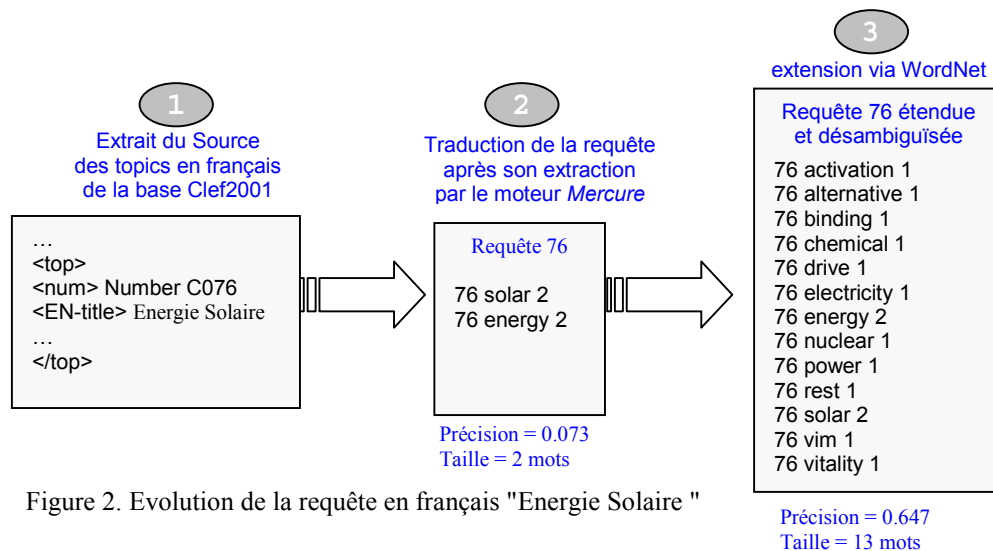


Figure 2. Evolution de la requête en français "Energie Solaire "

La requête étendue est comme dans l'étape 3 de la Figure 2 ci-dessus. Dans cet exemple, l'évolution de la requête initialement jugée pauvre (précision=0.073), apporte un gain très important à la précision (atteint 0.647). Ce qui évidemment n'est pas le cas pour toutes les requêtes, notamment celles qui sont soit mal traduites, ou très bonne initialement.

## 5. Expérimentation et évaluation

### 5.1 Environnement d'expérimentation :

#### 5.1.1 Présentation de WordNet

WordNet est une base de données lexicographique développée à l'université de Princeton par un groupe dirigé par George Miller [Miller 95]. Une définition succincte de WordNet est aussi donnée par [Habert 01]. Cette ontologie linguistique générale [Guarino & al 99] est très utilisée dans le domaine de la RI notamment pour l'expansion de requêtes [Voorhees 94] [Baziz 02], pour l'indexation [Gonzalo 98] et pour la désambiguïsation [Guarino & al 99] et [Mihalcea & al 00]. Elle couvre la grande majorité des noms, verbes, adjectifs et adverbes de la langue anglaise. Les mots dans WordNet sont organisés en ensembles appelés Synsets<sup>2</sup>. WordNet<sup>3</sup> a un vaste réseau de 144 684 mots, organisés en 109 377 synsets. La table 2 présente le nombre de noms, verbes, adjectifs et adverbes définis dans WordNet.

<sup>2</sup> Le terme synset est équivalent au terme concept dans la terminologie de WordNet. Il regroupe un ensemble de mots pour représenter une idée (sens).

<sup>3</sup> Nous avons utilisé l'interface [WordNet-QueryData 1.13](#). pour accéder aux fichiers de donnée de WordNet1.7

Part of speech	Words	Concepts	Total Word-Sense Pairs
Noun	107 930	74 488	132407
Verb	10 806	12 754	23255
Adjective	21 365	18 523	31077
Adverb	4 583	3 612	5721
Total	144 684	109 377	192460

Table2. Le nombre de mots et de concepts dans WordNet 1.7

La relation sémantique de base entre les mots codée dans WordNet est la *synonymie*. Les synsets sont liés par des relations telles que *spécifique/générique* ou *hyperonymie /hyponymie* (*is-a*), et la relation *partie-tout* ou *meronymie/holonymie* (*part-whole*).

### 5.1.2 La base de test

L'expérimentation a été effectuée sur un corpus de documents issus du projet CLEF2001. Cette collection est composée de:

- Documents en anglais CLEF-Ang. Le corpus de documents utilisé est de type "news" et provient des journaux Los Angeles Times (Etats-Unis). Le tableau 1 montre les caractéristiques de cette collection.

CLEF 2001 English Data		
Nombre de documents dans la collection	Nombre de termes dans la collection	Taille moyenne d'un document (termes)
113005	163700	282

Table3. Description de la collection de test utilisée

- 50 requêtes exprimées en français. Chacune est représentée comme une liste de termes. Elles sont numérotées de 41 à 90. Nous avons utilisé les champs titre de ces requêtes.
- Jugements de pertinence: les jugements de pertinence déterminent pour chaque requête, l'ensemble de ces documents pertinents.

L'indexation des documents est effectuée par le système Mercure développé au sein de l'équipe SIG de l'IRIT [Boughanem 00]. Le processus d'indexation produit pour chaque document une liste de termes pondérés. Le processus de recherche consiste à comparer la liste des termes de la requête avec celles des documents. Une liste ordonnée de documents est retournée à l'utilisateur en réponse à la requête. Dans nos tests, nous nous sommes restreints à un croisement sur la paire de langues français-anglais. Les expérimentations effectuées consistent à sélectionner les documents en anglais pour des requêtes exprimées en français.

## 5.2 Evaluation

Le but de nos expérimentations est de montrer la faisabilité de la technique de traduction et désambiguïsation, basée sur les concepts de WordNet. Pour ce faire, nous comparons cette technique à celle basée sur le dictionnaire bilingue<sup>4</sup>.

<sup>4</sup> Le dictionnaire utilisé est disponible gratuitement à partir de : <http://www.freedict.com>

Le dictionnaire est une liste simple de termes en français alignés avec d'autres termes en anglais.

Pour mesurer l'efficacité de notre technique de désambiguïsation, les résultats obtenus sont comparés aux résultats du test basé sur le dictionnaire. Le test du dictionnaire consiste à utiliser les requêtes en français et les traduire en anglais puis, les comparer aux documents anglais issus du corpus CLEF.

L'évaluation des performances est effectuée sur l'ensemble des documents sélectionnés pour les 50 requêtes. Elle se base sur les mesures de rappels et de précision. Soient les précisions à différents points p10, p15, p100 représentant le nombre de documents pertinents parmi les 10, 15, 100 premiers documents, et une précision moyenne (Prec.Moy.) sur l'ensemble des documents sélectionnés. Cette évaluation est effectuée selon le processus TREC<sup>5</sup> [Voorhees 94].

Deux groupes de tests ont été réalisés. Le premier groupe concerne l'utilisation des dictionnaires bilingues sans la désambiguïsation. Le deuxième groupe concerne l'utilisation des dictionnaires bilingues pour la traduction combinée avec la désambiguïsation et l'expansion, basées sur les concepts de WordNet.

1- **Dictionnaire** : la liste des termes de la requête en français est traduite via le dictionnaire bilingue en une requête en anglais. Cette requête en anglais est évaluée sur les documents exprimés en anglais.

2- **Combinaison du dictionnaire avec les concepts de WordNet (Dico + WordNet)**: la requête en français est traduite via le dictionnaire en anglais et sera désambiguïsée et étendue par les concepts de WordNet.

Id-Exec (50 requêtes)	P10	P15	P100	Prec.Moy
Monolingue	0.4085	0.3518	0.2432	0.4863
1) Dictionnaire	0.2936	0.2397	0.0987	0.3305
2) Dico + WordNet	0.3298	0.2851	0.1019	0.3589
Amélioration (2-1)	13.72%	23.96%	3.24%	8.76%

*Table3 : Impact de la désambiguïsation et de l'expansion basées sur les concepts de WordNet*

Les résultats obtenus par les deux groupes de tests (voir Table3), sont comparés au test du Monolingue. Le test du Monolingue consiste à utiliser l'ensemble des requêtes exprimées dans une langue source L1 (anglais dans notre cas). Ces requêtes sont comparées aux documents exprimés aussi dans la même langue. La liste des documents sélectionnés pour chaque requête est évaluée en terme de précision et de rappel.

D'après Table3, on remarque clairement que les résultats obtenus par cette combinaison sur toutes les précisions sont meilleurs que ceux obtenus par le dictionnaire. Plus précisément, la

---

<sup>5</sup> TREC pour Text REtrieval Conference, un programme international d'évaluation des systèmes de recherche et de filtrage d'information. site <http://trec.nist>

précision moyenne (Prec.Moy) est de 0.3305 contre 0.3589 pour le cas de la désambiguïsation + expansion (Dico + WordNet). La meilleure performance est enregistrée pour les 15 premiers documents restitués où le résultat (0.28), se rapproche de ceux obtenus avec le monolingue (0.3518), l'apport dans ce cas avoisine les 24%. Ce qui n'est pas négligeable, étant donné que l'on manipule uniquement le champ titre des topics, d'où un nombre de termes dans la requête limité. L'amélioration s'explique par le fait que pour un terme donné, la désambiguïsation a tendance à récupérer la traduction exacte du terme source et l'expansion l'enrichit avec de nouveaux termes, contrairement au dictionnaire qui propose plusieurs traductions pour un terme (problème de polysémie).

Revenons sur la comparaison effectuée entre la technique de désambiguïsation et celle basée sur le dictionnaire. Nous rappelons que nous avons utilisé un dictionnaire bilingue, récupéré gratuitement via Internet. On pouvait penser, à juste titre, que les résultats obtenus par la technique de désambiguïsation sont meilleurs que ceux du dictionnaire, car celui-ci est une version gratuite avec un vocabulaire incomplet. Ceci n'est pas le cas car ce même dictionnaire a été utilisé et testé dans d'autres travaux [Oard 98], [Adriani 00] qui montrent qu'il donne des résultats comparables aux résultats des dictionnaires commerciaux.

## 6. Conclusion

Cet article a présenté une technique de désambiguïsation et d'expansion de requêtes en recherche d'information par croisement de langues basée sur l'utilisation des concepts issus d'une base de données lexicographique externe (WordNet). Le but de cette désambiguïsation est d'améliorer la précision des requêtes traduites par le dictionnaire bilingue. La faisabilité de cette démarche a été testée en utilisant le moteur de recherche Mercure.

Nous avons montré à l'issue de ces tests, que l'utilisation des concepts d'une ressource sémantique externe, est une solution viable pour désambiguïser et étendre les requêtes traduites. La démarche proposée permet d'une part, de se focaliser sur le sens dominant de ces requêtes, et de les enrichir avec des termes reliés sémantiquement à ceux des requêtes, d'autre part. Ce qui nous permet d'envisager des perspectives à ce travail :

Il s'agit d'abord de consolider la démarche proposée, en la testant sur d'autres langues et d'autres collections, puis d'évaluer l'effet d'une variante à cette démarche sur la pertinence des réponses du SRI. Cette variante, consiste à utiliser directement un réseau sémantique multilingue telle que EuroWordNet, pour la traduction des requêtes en entier, sans passer par un dictionnaire de mots. Ce qui revient à traduire toute la requête en tant que concept et non terme à terme.

## bibliographie

[Adriani 00] : Adriani M, (2000).

Ambiguity Problem in Multilingual Information Retrieval

Workshop of the Cross language Evaluation Forum, CLEF 2000. Lecture Notes in computer Science (LNCS 2069 ), Springer Verlag, Carolperters (ED).

[Ballestros 96] : Ballesteros L., Croft W. (1996).

*Dictionary methods for cross-lingual information retrieval.*

In Proceedings of DEXA'96, pages 791-801.

**[Ballestros 98]** : Ballesteros L., Croft W. (1998).

*Resolving Ambiguity for Cross-Language Retrieval.*

In Proceedings of the 21st ACM SIGIR'98, pages, 64-71.

**[Baziz 02]** : Baziz M. (juin 2002).

« Application des Ontologies pour l'Expansion de Requêtes dans un Système de Recherche d'Informations »,

Rapport de DEA 2IL, Irit.

**[Boughanem 00]** : Boughanem M., Julien C., Mothe J., Soule-Dupuy C. (2000).

Mercure at TREC8

In Proceedings of TREC-8 (Ces proceedings sont disponibles sur le site

<http://trec.nist.gov/publications> }

**[Davis 96]** : Davis M., Dunning T. E. (1996).

*A TREC evaluation of query translation methods for multi-lingual text retrieval.*

In Proceedings of TREC-4, pages 483-497.

**[Davis 97]** : Davis M., (1997).

New Experiments in cross-language test retrieval

In Proceedings of TREC-5, pages 447-454.

**[Gey 99]** : Gey F. (1999).

Manual Queries and Machine Translation in Cross-Language Retrieval and Interactive Retrieval with Cheshire II

In Proceedings of TREC-7, page 527-540.

**[Gonzalo 98]** : Julio Gonzalo, Felisa Verdejo, Irina Chugur, and Juan Cigarran.

Indexing with wordnet synsets can improve text retrieval.

In Proceedings of the Coling/ACL '98 Workshop on Usage of WordNet for NLP, Montreal, Canada, 1998.

**[Grefenstette 98]** : Grefenstette G. (1998).

The Problem of Cross-language Information Retrieval

In Cross-Language Information Retrieval, Edited by Gregory Grefenstette, Kluwer Academic Publishers, pages 1-9.

**[Guarino & al 99]** : Nicola Guarino, Claudio Masolo, and Guido Vetere.

OntoSeek : content-based access to the web.

*IEEE Intelligent Systems, 1999.*

**[Habert 01]** : Habert B. et Monceaux L.

« WordNet, la mère (le père) de tous les réseaux de mots ? »,

Disponible sur : <http://www.limsi.fr/Individu/habert/00-01/SeminaireLMonceaux290501.ppt>, rapport LIR, Mai 2001.

**[Hull 96]** : Hull D., Grefenstette G. (1996).

Querying across languages. A dictionary-based approach to multilingual information retrieval

In Proceedings of ACM-SIGIR'96, pages 49-57.

**[Mihalcea & al 00]** : Dan I. Moldovan and Rada Mihalcea (2000).

"Improving the search on the Internet by using WordNet and lexical operators". *IEEE Internet Computing* 4(1) 34 - 43.

**[Miller 95]** : Miller G. (1995)

Wordnet: A lexical database.

Communication of the ACM, 38(11):39--41, 1995.

**[Nassr 02]** : Nassr N., Boughanem M., (2002).

Croisement de langues en recherche d'information : traduction et désambiguïsation de requêtes par phrases alignées

Inforsid 2002, XX ème Congrès inforsid IRIN, Polytech Nantes, 4-7 juin

**[Oard 96]** : Oard W.O, Dorr B. (1996).

A Survey of Multilingual Text Retrieval.

Report UMIACS-TR-96-19 CS-TR-3615 (<http://www.clis.umd.edu/dlrg/filter/papers/mlir.ps>).

**[Oard 98]** : Oard W.O, Wang j, Lin D, Soboroff I, (1998).

TREC-8 Experiment at Maryland : CLIR, QA and Routing

In proceeding of the Eighth Text Retrieval Conference (TREC-8)

**[Pirkola 98]** : Pirkola A. (1998)

The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-language Information Retrieval

In Proceedings of ACM-SIGIR'98.

**[Radwan 94]** : Radwan K.(1994).

Vers l'accès multilingue en langage naturel aux bases de données textuelles.

Thèse de l'Université Paris-sud, Centre d'Orsay.

**[Sanderson 00]** : Gollins T, Sanderson M, (2000).

Sheffield university CLEF 2000 submission Bilingual Track : German to English

Workshop of the Cross language Evaluation Forum, CLEF 2000. Lecture Notes in computer Science (LNCS 2069 ), Springer Verlag, Carol perters (ED).

**[Schauble 00]** : Schauble P, Braschler M, (2000).

Experiments with the Eurospider Retrieval System for CLEF 2000

Workshop of the Cross language Evaluation Forum, CLEF 2000. Lecture Notes in computer Science (LNCS 2069 ), Springer Verlag, Carol perters (ED).

**[Sheridan 96]** : Sheridan P., Ballerini J. P. (1996).

Experiments in multilingual information retrieval using SPIDER system

In Proceedings of ACM SIGIR'96, pages 58-65.

**[Yamabana 96]** : Yamabana K., Muraki F., Doi S., Kamei S. (1996).

A Language Conversion Front-end for Cross-Linguistic Information Retrieval In Proceedings of ACM-SIGIR'96, pages 43-39.

**[Voorhees 94]** : Voorhees E. (1994)

Query expansion using lexical-semantic relations,

Proceedings of the 17th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, pages 61-69, Dublin, Ireland.