Traduction automatique de termes biomédicaux pour la recherche d'information interlingue

Vincent Claveau

IRISA – CNRS
campus de Beaulieu
F-35042 Rennes cedex
Vincent.Claveau@irisa.fr

RÉSUMÉ. *Dans cet article, nous présentons une méthode de traduction automatique de termes biomédicaux. Cette méthode s'appuie sur une technique originale d'apprentissage supervisé de règles de réécriture et sur l'utilisation de modèles de langue. Les évaluations présentées montrent que notre technique est très performante et permet de traduire à partir et à destination de n'importe quelle langue pourvu que leurs différences soient suffisamment régulières pour être apprises. Cette méthode de traduction est appliquée et évaluée sur une tâche de RI interlingue dans le domaine biomédical avec des requêtes dans différentes langues (français, espagnol, portugais, russe, italien); les bons résultats obtenus démontrent l'intérêt de cette approche automatique pour la recherche d'information.*

ABSTRACT. *In this article, we present a new method to automatically translate biomedical terms. This method relies on an original machine-learning technique that infers rewriting rules and on the use of language models. Evaluations presented here prove that this method yields good results and allows one to translate between any two languages provided that their differences are regular enough to be learnt. This translation method is applied and evaluated on a interlingual IR task in the biomedical domain with queries in several languages (French, Spanish, Portuguese, Russian, Italian); the good results we obtain prove the interest of such an automatic approach for the information retrieval domain.*

MOTS-CLÉS : *RI interlingue, traduction artificielle, apprentissage artificiel, termes biomédicaux*

KEYWORDS: *Interlingual IR, machine translation, machine learning, biomedical terms*

CORIA 2007, pp. 303–318
Saint-Étienne, 28–30 mars 2007

1. Introduction

Dans le domaine biomédical, les problématiques de recherche et de traitement d'informations textuelles sont particulièrement importants. De nombreux documents sont en effet collectés dans des bases spécialisées et ces bases sont très consultées. Par exemple, la base PubMed regroupe actuellement 16 millions de publications scientifiques dans le domaine médical et fait face à plus de 3 millions de requêtes par jour. Dans la plupart de ces bases, chacun des documents est indexé à l'aide de terminologies de référence, notamment le thésaurus MeSH (*Medical Subject Heading*). Par ailleurs, la prédominance de l'anglais dans ces terminologies de référence rend cruciale la mise en place de stratégies multilingues pour faciliter l'accès à ces bases pour les non-anglophones. Des terminologies biomédicales multilingues existent, mais elles sont mises en défaut par l'évolution rapide des connaissances et le manque de moyens pour certaines langues.

Dans ce cadre, nous présentons et évaluons dans cet article une méthode de traduction automatique de termes biomédicaux que nous appliquons à une tâche de RI interlingue. Cette méthode doit permettre de produire des traductions de termes simples (*i.e.* composés d'un seul mot) du domaine biomédical d'une langue source dans une langue cible. Ce travail repose sur deux hypothèses majeures :

1) dans le domaine biomédical, les termes équivalents entre deux langues sont souvent morphologiquement proches ;

2) les différences entre termes de chaque langue sont régulières et peuvent être apprises automatiquement.

Ces deux hypothèses tirent parti du fait que les termes biomédicaux sont construits sur les mêmes racines grecques et latines, et leurs dérivations très régulières (*e.g.* pour le couple français-anglais *ophtalmorragie*/*ophthalmorrhagia*, *ophtalmoplastie*/*ophthalmoplasty*, *leucorragie*/*leukorrhagia*). La technique décrite et les expériences rapportées ici portent sur la traduction entre différentes langues (anglais, espagnol, français, russe…) ; l'accent sera toutefois mis sur la traduction d'une langue source quelconque vers l'anglais, qui correspond au cas le plus intéressant pour la RI interlingue dans ce domaine.

Notre approche s'appuie sur une technique d'apprentissage artificiel originale que nous avons développée. Elle nous permet d'inférer un système de traduction à partir de couples de termes langue source-langue cible traduction l'un de l'autre et morphologiquement proches. C'est ce système qui, étant donnés en entrée des termes dans la langue source – dans notre cas, une requête – doit ensuite permettre de produire les termes correspondants dans la langue source. Plus précisément, dans notre cas, le système de traduction est un ensemble de règles de réécriture (*cf.* section 3) générées à partir d'exemples de couples de termes bilingues. Il est intéressant de noter qu'à part cette phase de supervision, aucune autre connaissance, ni intervention humaine n'est requise.

Dans la section suivante, nous présentons les travaux proches et nous plaçons par rapport à ceux-ci. Nous décrivons ensuite notre technique de traduction de termes

biomédicaux que nous évaluons sur différentes paires de langues en section 4. Nous appliquons ensuite cette technique à un problème de RI interlingue dans le domaine biomédical et en présentons les résultats en section 5.

2. Travaux connexes

Peu de travaux se placent dans le cadre de la traduction directe de termes, et moins encore dans le domaine biomédical. Cette problématique a cependant déjà été abordée et une solution fonctionnelle a été proposée par Claveau et Zweigenbaum (2005). Celle-ci repose sur une technique d'apprentissage de transducteurs, que l'on peut voir comme des automates permettant la réécriture des séquences analysées. Son principal défaut est néanmoins de ne pouvoir fonctionner que sur des langues partageant le même alphabet, contrairement à l'approche nouvelle que nous présentons ici ; et nous montrons en section 4 que notre technique obtient des performances assez nettement supérieures à l'approche par transducteurs. Les travaux de Schulz, Markó, Sbrissia, Nohama et Hahn (2004) partagent également les mêmes problématiques que cet article ; ils proposent une technique de traduction de termes biomédicaux du portugais vers l'espagnol fondés sur une analyse morphologique et sur l'utilisation de règles de réécriture qu'ils testent ensuite en RI (Markó *et al.,* 2005). Cependant, à la différence de nos travaux, ces règles sont fournies manuellement et cette longue tâche doit donc être recommencée pour toute nouvelle langue.

Outre ces travaux, des problématiques proches sont parfois abordées dans le domaine de la traduction artificielle de textes. Ainsi, l'acquisition de cognats (couples de mots bilingues de formes proches) (Fluhr *et al.,* 2000, *inter alia*) s'appuie sur des opérations morphologiques simples (distance d'édition, plus longue sous-chaîne commune) parfois proches des règles de réécriture que nous inférons. D'autres études reposent quant à elles sur des recherches en corpus à l'aide de techniques statistiques de cooccurrences ou d'indices lexicaux (ponctuations, chiffres) pour trouver des alignements – et donc des relations de traduction potentielle – entre termes dans des corpus alignés (Ahrenberg *et al.,* 2000, Gale *et al.,* 1991, Tiedemann, 2004) ou comparables (Fung *et al.,* 1997b, Fung *et al.,* 1997a). Outre le problème de la rareté de corpus spécialisés alignés, ces approches diffèrent de la nôtre en cela qu'il s'agit pour ces auteurs de retrouver une traduction d'un mot dans un texte (mise en relation), alors que nous nous posons dans le cadre plus strict de la traduction (génération). De plus, bien souvent, ces techniques d'alignement automatique ont justement besoin de paires de termes traduction l'un de l'autre comme point de départ (pour un état de l'art, voir Véronis (ed.) (2000)).

D'une manière plus générale, la traduction statistique (Brown *et al.,* 1990) s'intéresse à un problème connexe ; bien sûr, la séquence à traduire dans notre cas est composée de lettres et non pas de mots. La méthode en deux temps que nous proposons (*cf.* section suivante) partage beaucoup de similarité avec l'approche classique de traduction statistique utilisant un modèle de traduction et un modèle de langue (Brown *et al.,* 1993). Cependant, la nature des données que nous manipulons induit des dif-

férences importantes. Tout d'abord, le problème de réordonnancement des mots, pris en compte dans les modèles IBM avec le paramètre de distorsion, ne se pose pas dans notre cas : l'ordre des morphes (et donc les lettres) qui composent les termes varie peu d'une langue à l'autre. L'utilisation de paramètres de fertilité ou de mots nuls dans ces mêmes modèles IBM, servant pallier la traduction mot à mot induite par le modèle traduction, n'est pas adaptée à nos données. Ces problèmes sont en effet plus naturellement pris en compte par notre technique d'inférence qui nous permet de générer des règles de réécriture traduisant non pas d'une lettre à une autre mais d'un groupe de lettres de longueur quelconque à un autre.

Les recherches sur la translittération, notamment du japonais (katakana) (Qu *et al.,* 2003, Tsuji *et al.,* 2002, Knight *et al.,* 1998, par exemple) ou de l'arabe (Al-Onaizan *et al.,* 2002a, AbdulJaleel *et al.,* 2003) et leur application à la RI interlingue partagent beaucoup de points communs avec notre approche. En effet, les techniques à l'œuvre dans ce domaine sont souvent proches de celles détaillées ici, mais ne concernent que la représentation d'imports (principalement des entités nommées) dans des langues ayant un alphabet différent de la langue source. De ce fait, ces techniques, comportant souvent une phase de représentation du terme à traduire en phonèmes, sont dites *phonetic-based* et sont à distinguer des approches *spelling-based* dans lesquelles s'inscrit la technique présentée dans cet article. Les techniques *phonetic-based* ou les techniques mixtes (Al-Onaizan *et al.,* 2002a) nécessitent donc des connaissances externes (tables de correspondances chaînes de caractères → phonèmes, correspondances phonèmes langues source → phonèmes langue cible…) qui rendent ces approches efficaces mais pas adaptables à d'autres paires de langue. Par ailleurs, dans les travaux existants sur la translittération d'entités nommés, les deux sens de traduction ne sont pas équivalents : on parle de *forward transliteration* (par exemple, translittération d'un nom arabe en alphabet latin) et de *backward transliteration* (recherche du nom source arabe à partir de sa translittération latine). Cette distinction – qui implique souvent des différences de traitement – n'a pas de sens dans notre cas ; aucune langue n'est considérée comme langue source *a priori* et l'approche que nous proposons est entièrement symétrique même si les performances de traduction peuvent varier d'un sens de traduction à l'autre.

Mentionnons enfin les travaux sur la morphologie dans lesquels des méthodes d'apprentissage artificiel ont été employées avec succès à la lemmatisation (Erjavec *et al.,* 2004), la découverte de liens morphologiques (Gaussier, 1999, Moreau *et al.,* 2006) ou encore l'analyse morphographémique (Oflazer *et al.,* 1999). La technique d'inférence de règles de réécriture présentée dans la section suivante s'inscrit dans la lignée de celles à l'œuvre dans ces études.

3. Technique de traduction artificielle de termes

La technique de traduction de termes biomédicaux que nous proposons dans cet article fonctionne en deux temps que nous exposons tour à tour ci-après. Tout d'abord, des règles de réécriture sont inférées à partir d'exemple de paires termes traduction

l'un de l'autre. Ensuite, un modèle de langue est appris à partir des termes de la langue cible de ces mêmes exemples. Une fois ces deux étapes effectuées, la traduction d'un terme inconnu consiste à appliquer les règles de réécriture inférées et à calculer à l'aide du modèle de langue la probabilité des traductions proposées afin de retenir la plus probable.

3.1. *Inférence de règles de réécriture*

La technique de traduction automatique des termes biomédicaux que nous proposons dans cet article repose sur l'apprentissage de règles de réécriture (que l'on peut aussi voir comme des règles de translittération). Ces règles, apprises à partir de listes de paires bilingues de termes du domaine (*cf.* section 4.1), sont de la forme : $\langle input\ string \rangle \rightarrow \langle output\ string \rangle$. Définissons quelques notations pour la suite de l'article : une règle de réécriture est notée r, la liste de toutes les règles inférées pendant une expérience est notée $\mathcal{R}$, $input(r)$ et $output(r)$ désignent respectivement la chaîne d'entrée et la chaîne de sortie de la règle r.

L'algorithme 1 donne un aperçu global de notre technique d'apprentissage. La pre-

Algorithm 1 Algorithme d'apprentissage

```
1: aligner les paires de termes au niveau des lettres, mettre le résultat dans L
2: for all paire de termes W dans L do
3:    for all alignement de lettre dont les 2 lettres diffèrent dans W do
4:       trouver la meilleure hypothèse de règles r dans l'espace de recherche E
5:       ajouter r à l'ensemble de règles R
6:    end for
7: end for
```

mière étape est réalisée à l'aide de DPalign, un logiciel existant mettant en œuvre l'alignement en programmation dynamique. Le principe général de DPalign est d'aligner deux séquences en cherchant à minimiser la distance d'édition entre les séquences, celle-ci étant notamment calculée à l'aide d'une matrice de substitution[1]. Il est important de noter que le logiciel nous permet d'aligner des termes ne partageant pas le même alphabet. Il utilise pour ce faire une matrice de substitution calculée directement à partir des exemples de paires fournies. Dans notre cas, une liste de paires de termes est donnée en entrée de DPalign ; à chaque terme sont ajoutés deux caractères # pour représenter le début et la fin de la chaîne de caractères. La liste de sortie $\mathcal{L}$ de DPalign contient les paires de termes alignés au niveau des lettres, telles que par exemple ('_' signifie *aucun caractère*) : $\begin{matrix} \#opht\_almologie\# \\ \#ophthalmology\_\# \end{matrix}$, pour la paire de langue français-anglais, ou $\begin{matrix} \#adenosinetriphosphatase\# \\ \#аденозин\_трифосфатаза\# \end{matrix}$ pour anglais-

1. Voir l'URL : `http://search.cpan.org/~BIRNEY/bioperl-1.4/Bio/Tools/dpAlign.pm` pour plus de détails.

russe. Par la suite, le terme d'entrée (respectivement de sortie) d'une telle paire alignée p est noté $input(p)$ (resp. $output(p)$) ; de plus, $align(x, y)$ indique que la sous-chaîne x est alignée avec la sous-chaîne y dans la paire de terme considérée.

Dans notre processus d'apprentissage, ces paires de mots alignés sont considérées comme des exemples à partir desquels les deux boucles imbriquées infèrent des règles de réécriture. Comme pour beaucoup de problèmes d'apprentissage artificiel symbolique, cette phase d'inférence (ligne 4) peut être considérée comme un problème de parcours d'espace. À chaque élément de cet espace est assigné un score ; on cherche à trouver l'élément de l'espace maximisant ce score.

Dans notre cas, l'espace de recherche est composé de toutes les règles de réécriture possibles compatibles avec l'exemple choisi. Par exemple, considérons que la paire de mots W choisie à la ligne 2 est *#opht_almologie#/#ophthalmology_#*, et supposons que c'est l'alignement *i/y* qui est choisi à la ligne 3. Quelques règles de réécriture compatibles dans ce contexte sont *i* → *y*, *gi* → *gy*, *ie* → *y* (NB : on ne note pas le caractère _), *ologie#* → *ology#*...

Le score d'une règle est calculé à partir de la liste $\mathcal{L}$; il est défini comme étant le ratio entre le nombre de fois où la règle s'applique aux termes alignés de la liste d'exemples et le nombre de fois où la prémisse de la règle apparaît dans les termes source de la liste d'exemples. Formellement, le score d'une règle r est donc défini par :

$$score(r) = \frac{|\{p \in \mathcal{L} \mid input(r) \subseteq input(p) \wedge output(r) \subseteq align(input(r), p)\}|}{|\{s \in \mathcal{L}_{input} \mid input(r) \subseteq s\}|}$$

où $\subseteq$ représente l'inclusion de chaîne de caractères.

Du fait du très grand nombre de règles possibles, chercher la règle maximisant la fonction de score pour chacun des exemples peut être une tâche très lourde en temps de calcul. Heureusement, l'espace de recherche peut être organisé hiérarchiquement pour rendre l'exploration plus efficace. En effet, les règles compatibles pour un exemple peuvent être organisées de la plus générale à la plus spécifique avec la notion de subsomption suivante :

$$r_1 \succeq r_2 \Leftrightarrow (input(r_1) \subseteq input(r_2) \wedge output(r_1) \subseteq output(r_2)).$$

Cette relation de subsomption est réflexive, antisymmétrique et transitive ; l'espace résultant est un treillis. La figure 1 présente l'espace de recherche organisé par cette subsomption construit à partir de l'exemple *i/y* dans *#opht_almologie#/#ophthalmology_#*. Dans notre cas, la recherche est effectuée du plus général au plus spécifique (*top-down*) ; cela, et les propriétés d'héritage que cette structure implique, nous permet de rechercher efficacement la meilleure règle (calcul du score d'une règle en n'examinant que les paires de termes que son père couvre, élagage de l'espace basé sur le meilleur score courant...).

Finalement, cet algorithme va potentiellement générer une règle de réécriture par différence pour chacune des paires de termes utilisées en exemple. Cela peut conduire

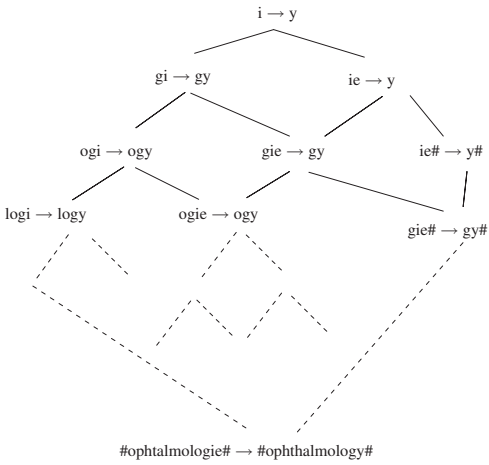

Figure 1. *Treillis de recherche de l'exemple i/y dans #opht_almologie#/#ophthalmology_#*

à obtenir un grand nombre de règles. Pour un nouveau terme à traduire, ces règles vont mener à générer différentes traductions parmi lesquelles il faut choisir la plus probable. C'est l'objet de la sous-section suivante.

3.2. *Évaluation des traductions proposées*

Pour traduire un terme inconnu dans la langue d'entrée, toutes les règles applicables à ce terme (*i.e.* les règles dont la prémisse correspond à une sous-chaîne du terme) sont effectivement appliquée. Dans le cas de règles concurrentes, toutes les possibilités sont générées. À ce niveau, un terme peut donc être traduit de différentes façons. Pour choisir parmi ces possibilités la traduction la plus probable, nous avons utilisé une approche simple basée sur les modèles de langue comme cela est fait dans la plupart des travaux de translittération cités en section 2.

Les modèles de langue (ML) sont largement utilisés pour la traduction artificielle, la transcription de l'oral ou encore la recherche d'information (Charniak, 1993). Cependant, dans ces cadres, les ML sont utilisés pour assigner une probabilité à une séquence de mots, alors que dans notre cas, la probabilité va au contraire être assignée à un mot considéré comme une séquence de lettres. Plus formellement, avec les notations standard, pour un mot w composé des lettres $l_1, \ldots, l_m$, on a :

$$P(w) = \prod_{i=1}^{m} P(l_i | l_1, ..., l_{i-1})$$

En pratique, les probabilités $P(l_i|...)$ sont estimées à partir des termes de la langue cible, décomposés en n-grammes de lettres, issus de la liste d'exemples. Pour prévenir le problème des séquences de lettres non vues, les probabilités sont en réalité calculées à partir d'un historique réduit aux $n-1$ lettres précédentes (*i.e.* $P(l_i|l_{i-n+1}, ..., l_{i-1})$) et un lissage simple est appliqué. Dans les expériences présentées ci après, n est fixé à 7 lettres.

Intuitivement, le ML va favoriser les traductions qui *ressemblent* à des mots bien formés dans la langue cible. Parmi toutes les traductions proposées pour un terme de la langue source, on conserve donc finalement celle qui obtient la probabilité la plus forte selon le ML appris. Par ailleurs, il est intéressant de noter qu'en plus du fait de choisir la traduction la plus probable, cette technique nous donne un facteur de confiance sur la traduction qui est proposée. En pratique, nous avons constaté que l'utilisation de modèles de langue donnait effectivement de très bons résultats, nous permettant de sélectionner dans la quasi totalité des cas le candidat correct parmi les traductions générées par les règles de réécriture.

4. Évaluation de la traduction de terme

4.1. *Description des données*

Deux jeux de données multilingues sont utilisées pour nos expériences de traduction. Le premier est une collection de termes français-anglais issue du dictionnaire médical Masson (`http://www.atmedica.com`). C'est la même collection que celle utilisée dans Claveau et Zweigenbaum (2005), ce qui nous permettra de comparer nos résultats. Seules les paires composées de termes simples dans la langue source et dans la langue cible, hors acronymes, sont conservées. La liste bilingue ainsi constituée contient environ 12 000 paires de termes.

Le second jeu de termes multilingues est le Métathésaurus de l'UMLS (Tuttle *et al.,* 1990, Bodenreider, 2004). Cette collection de thésaurus rassemble des termes biomédicaux dans un grand nombre de langue (allemand, anglais, danois, espagnol, finnois, français, hollandais, hongrois, italien, japonais, norvégien, portugais, russe, suédois, tchèque…). Le métathésaurus de l'UMLS associe un identifiant indépendant des langues (le Concept Unique identifier, CUI) à chacun des termes de chacun des thésaurus ; ces CUI nous permettent donc de constituer des ensembles de paires de termes bilingues. Là encore, seules les paires constituées de termes simples (non acronymes) sont conservées.

4.2. *Méthode d'évaluation*

Avant l'utilisation de notre technique de traduction pour la recherche d'information interlingue, nous évaluons tout d'abord sa précision intrinsèque. Nous suivons pour cela une approche standard : la liste de paires de termes est découpée en deux en-

sembles, le premier sert pour l'apprentissage (inférence de règle et modèle de langue), et le second, composé de 1000 paires, sert de jeu de test. Une fois les règles et le modèle de langues appris, nous l'appliquons à chaque terme d'entrée du jeu de test. Nous comparons alors la traduction proposée avec celle attendue. Si les deux chaînes de caractères sont identiques, la traduction est considérée correcte ; dans tous les autres cas, elle est considérée incorrecte.

Les résultats sont évalués en terme de précision, c'est-à-dire le pourcentage de traductions correctes générées. Cependant, puisque le modèle de langue nous fournit un indice de confiance, on peut décider de ne conserver que les traductions dont l'indice est supérieur à un certain seuil. Si ce seuil est fixé assez haut, la précision sera certainement élevée mais au détriment du nombre de traductions proposées, et vice-versa. À la manière d'une courbe rappel-précision, nous représentons donc ci-dessous, pour tous les seuils possibles d'indice de confiance, la précision suivant le pourcentage de mots traduits.

4.3. *Résultats*

4.3.1. *Traduction entre le français et l'anglais*

Pour cette première expérience, nous nous intéressons à la traduction français-anglais et anglais-français. Comme précisé précédemment, nous utilisons le jeu de données Masson qui nous permet de comparer directement nos résultats à ceux de Claveau et Zweigenbaum (2005). Les figures 2 et 3 présentent les graphes de précision des traductions générées pour les ensembles de test pour les sens français-anglais et anglais-français. Dans des langues proches comme le sont le français et l'anglais, beaucoup de termes spécialisés sont identiques. Comme simple *baseline*, nous calculons donc la précision qu'obtiendrait un système proposant systématiquement un terme comme sa propre traduction ; cette précision minimale donne ainsi une idée de la difficulté de la tâche de traduction. Sur ces courbes, nous indiquons également les résultats obtenus par la technique proposée par Claveau et Zweigenbaum (2005).

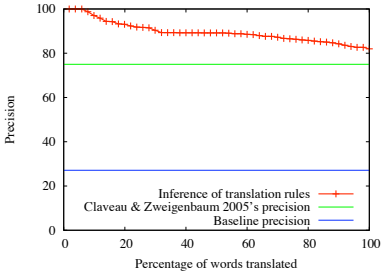

Figure 2. *Performances de traduction du français vers l'anglais*

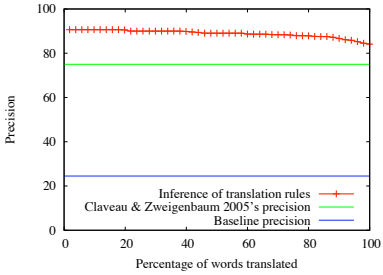

Figure 3. *Performances de traduction de l'anglais vers français*

Quel que soit le sens de traduction, les deux graphes montrent que notre technique d'apprentissage a de meilleures performances que l'approche par transducteurs de Claveau et Zweigenbaum (2005). Pour la traduction du français vers l'anglais, notre approche a une précision de 82.6% pour 100% des mots traduits (soit 8% de plus que l'approche par transducteur). Pour le sens inverse, la précision est de 84.8% à 100% des mots traduits (soit une amélioration de 10% par rapport aux transducteurs). Notre technique est très largement au-dessus de la *baseline*, mais il convient de noter que celle-ci montre que 25% des termes biomédicaux sont identiques en français et en anglais, ce qui semble montrer que les deux langues sont suffisamment proches pour rendre la tâche d'apprentissage relativement aisée.

4.3.2. *Traduction à partir d'autres langues*

Nous répétons les mêmes expériences avec d'autres paires de langues disponibles dans l'UMLS. Parmi les différentes combinaisons de langues possibles, nous n'en présentons ci-dessous que quelques unes en nous fixant comme langue cible l'anglais. C'est en effet pour cette langue que notre tâche de RI interlingue a le plus de sens. La figure 4 présente les performances de la traduction de l'espagnol vers l'anglais. Les résultats sont plutôt bons : 73.4% des termes sont correctement traduits quand toutes les traductions sont gardées (*i.e.* aucun seuil de ML n'est fixé). Des résultats très similaires sont obtenus pour la paire portugais-anglais (figure 5) avec des précisions de 71.5% pour les pires cas.

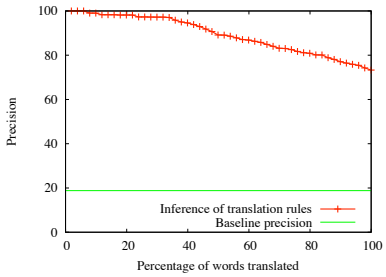

Figure 4. *Performances de traduction de l'espagnol vers l'anglais*

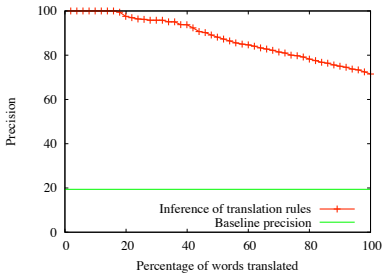

Figure 5. *Performances de traduction du portugais vers l'anglais*

La traduction de l'italien vers l'anglais (figure 6) donne également des résultats assez bons quoique que légèrement inférieurs aux précédents, notamment l'espagnol et le portugais bien que leur *baseline* soient quasiment égales. Au pire cas, ce sont tout de même 64.5% des termes qui sont correctement traduits.

Comme nous l'avons dit précédemment, notre technique de traduction permet de traiter des langues avec des alphabets différents, pourvu qu'elles montrent des régularités pouvant être apprises automatiquement. Dans cette dernière expérience, nous

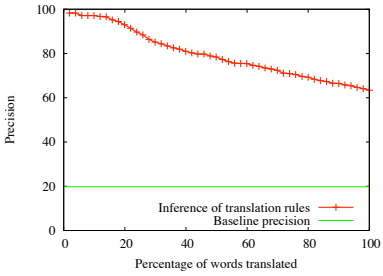

Figure 6. *Performances de traduction de l'italien vers l'anglais*

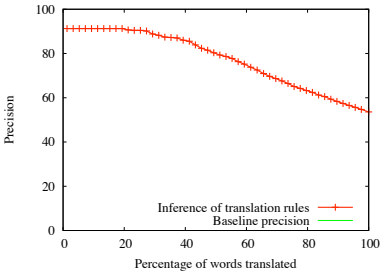

Figure 7. *Performances de traduction du russe vers l'anglais*

nous intéressons à la paire russe-anglais. La figure 7 présente les résultats obtenus. Dans le cas présent, à cause de la différence d'alphabet, la *baseline* est à 0. La précision minimale obtenue ici est de 57.5%. Bien qu'inférieurs aux autres paires de langue, ces résultats sont relativement bons étant donnée la difficulté apparente de la tâche. Cela indique qu'une grande partie des termes biomédicaux russes sont construits par translittération en cyrillique des mêmes morphèmes grecs et latins que l'anglais, le français, l'espagnol...

4.3.3. *Analyse des cas d'erreurs*

À l'examen des résultats, les différentes erreurs de traduction faites par notre technique peuvent être classées en différentes catégories. Ces cas d'erreurs sont identiques à ceux détaillés dans (Claveau *et al.,* 2005) ; nous n'en rappelons ici que les grandes lignes. La majorité des erreurs est simplement causée par le fait que certains termes traductions l'un de l'autre ne sont pas liés morphologiquement. C'est évidemment souvent le cas pour la paire russe-anglais, mais aussi pour des paires de langues pourtant proches (*e.g. asimiento/grip* pour espagnol-portugais ou *embrochage/pinning* pour français-anglais).

D'autres erreurs sont causées par des exceptions de traduction. En effet, certaines familles de termes, régulières d'un point de vue de la traduction comportent quelques cas particuliers. Par exemple, les termes français en *-rragie* sont généralement traduits en anglais par un terme en *-rrhagia*, (*e.g.*, *stomatorragie/stomatorrhagia*, *pneumorragie/pneumorrhagia*...). Il y a malheureusement des exceptions comme *hémorragie/hemorrhage* ou *pleurorragie/pleurorrhage* qui font échouer notre approche par apprentissage.

Enfin, quelques erreurs sont dues à l'apparente proximité de mots appartenant pourtant à des catégories grammaticales ou des classes sémantiques différentes, trompant ainsi les techniques d'apprentissage. Par exemple, les adjectifs français en -

ique sont généralement traduits en anglais par des adjectifs en *-ic* (*e.g.*, *spasmolytique/spasmolytic*), alors que les noms communs avec le même suffixe sont traduits en anglais par des termes en *-ics* (*thermodynamique/thermodynamics*).

5. Recherche d'information interlingue

Après l'évaluation des performances intrinsèques de traduction, nous appliquons dans cette section notre technique à la traduction de requêtes spécialisées dans le domaine biomédical. Nous décrivons tout d'abord le cadre expérimental de de cette tâche de RI interlingue, puis les résultats obtenus pour différentes langues.

5.1. *Cadre expérimental*

Dans le cadre de ces expériences, nous utilisons la collection "Filtering Task" de TREC-9, elle-même issue de la collection OHSUMED. Cette collection en anglais est composée de 350 000 résumés de MEDLINE (articles du domaine biomédical), de plus de 4000 requêtes et des jugements de pertinence associés. Les requêtes sont composées d'un sujet – un terme du thésaurus MeSH, développé pour indexer les articles biomédicaux – et d'une définition de ce terme.

Initialement construite pour le filtrage, nous utilisons ces données comme une collection de recherche standard pour nos expériences. Nous n'utilisons pour requête que le champ sujet. Pour évaluer l'intérêt de notre technique de traduction dans un contexte de RI interlingue, nous traduisons manuellement ces requêtes dans une langue donnée, à l'aide de l'UMLS. Ce dernier ne contient malheureusement pas les équivalents de tous les termes anglais ; seules les requêtes pour lesquelles une traduction existe sont conservées. Leur nombre dépend de la langue choisie mais en pratique, ce nombre reste très important (*e.g.* 2300 requêtes pour le français).

On dispose ainsi d'une collection de documents en anglais et de requêtes dans une autre langue. Notre technique de traduction est donc utilisée pour retraduire les requêtes, automatiquement cette fois-ci, de langue choisie vers l'anglais et transmet la requête à un système de recherche ; dans nos expériences, nous utilisons Lemur[2] paramétré de façon à imiter le fonctionnement du célèbre système Okapi. Bien entendu, pour ne pas biaiser les résultats, aucun des termes des requêtes et de leurs traductions ne sert d'exemple lors de l'apprentissage des règles de réécriture et du modèle de langue utilisé pour traduire les requêtes vers l'anglais.

5.2. *Résultats*

Pour chacune des expériences, nous utilisons comme point de comparaison les résultats obtenus par Lemur en utilisant les mêmes requêtes, mais dans leur ver-

2. Lemur est disponible à l'URL www.lemurproject.org.

sion anglaise originale. Nous indiquons également les résultats obtenus en traduisant les requêtes avec un outil généraliste, Systran BabelFish (`http://babelfish.altavista.com`). Les résultats sont présentés sous la forme classique de courbes rappel-précision.

Les figures 8, 9, 10, 11 et 12 présentent respectivement les courbes rappel-précision pour les traductions en anglais des requêtes en français, italien, espagnol, portugais et russe. Pour chacune d'elles, on constate tout d'abord sans surprise que les résultats avec les requêtes traduites automatiquement par notre technique sont nettement inférieurs à ceux obtenus avec les requêtes originales. La traduction automatique des requêtes donne néanmoins de bons résultats, dans tous les cas meilleurs que ceux obtenus par le biais de la traduction avec BabelFish. Seule exception, la traduction des requêtes en russe obtient environ les mêmes résultats en utilisant notre approche et BabelFish, ce qui s'explique par la relative faiblesse des performances de traduction vue dans la section précédente pour le russe.

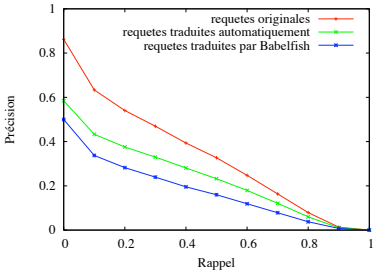

Figure 8. *Rappel-précision pour la traduction des requêtes du français vers l'anglais*

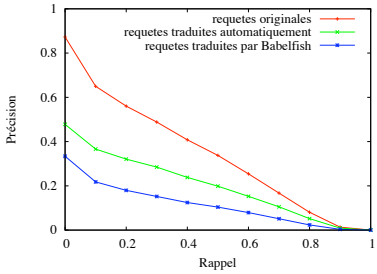

Figure 9. *Rappel-précision pour la traduction des requêtes de l'italien vers l'anglais*

Ces différents résultats de recherche d'information interlingue valident ainsi l'intérêt de notre approche de traduction artificielle des termes composant les requêtes. On remarque également sans surprise que les résultats obtenus à cette tâche de RI interlingue sont assez fidèlement corrélés avec les performances de traduction exposées en section 4.

6. Conclusion et perspectives

La recherche d'information interlingue est un problème particulièrement important dans le domaine biomédical. La méthode de traduction automatique de termes que nous avons exposée dans cet article se veut un élément de réponse à ce problème. Cette méthode est basée sur une technique d'apprentissage supervisé originale combinant règles de réécriture et modèles de langue. Elle nous permet d'apprendre à traduire des termes entre deux langues quelconques.

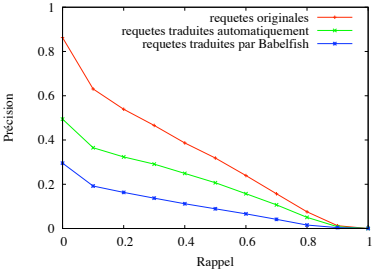

Figure 10. *Rappel-précision pour la traduction des requêtes de l'espagnol vers l'anglais*

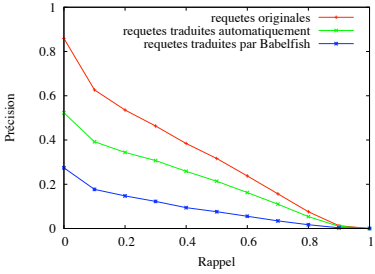

Figure 11. *Rappel-précision pour la traduction des requêtes du portugais vers l'anglais*

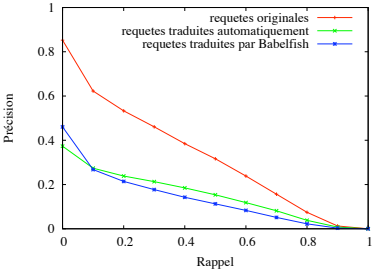

Figure 12. *Rappel-précision pour la traduction des requêtes du russe vers l'anglais*

L'évaluation que nous avons menée a montré que notre approche offre des performances de traduction variables selon les langues mais d'assez bon niveau pour être utilisée dans un contexte de RI interlingue. À partir de la collection OHSUMED telle qu'utilisée dans TREC-9, nous avons évalué les résultats qu'obtiendrait un système de RI utilisant notre technique pour traduire des requêtes en français, italien, portugais, espagnol ou russe vers l'anglais. Les résultats montrent clairement l'intérêt de notre approche bien qu'une large marge de progression existe encore.

Plusieurs perspectives sont ouvertes par ce travail. Concernant la technique de traduction, les résultats pourraient être améliorés, en tenant notamment compte de ca-

ractéristiques connues sur les termes à traduire, comme par exemple leurs parties-du-discours (nom, adjectif, verbe…). Comme nous l'avons souligné en section 4.3.3, cela permettrait d'éviter quelques erreurs de traduction, mais implique de savoir intégrer cette connaissance lors de l'apprentissage des règles de réécriture et/ou à l'utilisation du modèle de langue. À plus long terme, il serait également souhaitable d'étendre cette approche de traduction aux termes complexes (composées de plusieurs mots, comme *col du fémur*). La traduction de ces termes ne se fait pas forcément mot à mot et nécessite souvent une analyse syntaxique préalable.

Enfin, l'application à la tâche de RI interlingue de notre technique ouvre elle-aussi des perspectives. Dans les expériences présentées ici, seule la traduction obtenant le meilleur score selon le modèle de langue était gardée. Or dans ce cadre applicatif précis, on pourrait également utiliser à la place du modèle de langue ou en conjonction de celui-ci, l'index de la collection pour sélectionner le candidat le plus pertinent parmi les différentes traductions de la requête générées par les règles de réécriture comme cela est fait dans de nombreux travaux sur la translittération (Qu *et al.,* 2003, Al-Onaizan *et al.,* 2002b, par exemple).

7. Bibliographie

AbdulJaleel N., Larkey L. S., « Statistical transliteration for English-Arabic Cross Language Information Retrieval », *Proceedings of the 12th International Conference on Information and Knowledge Management, CKIM'03*, New Orleans, États-Unis, p. 139-146, 2003.

Ahrenberg L., Andersson M., Merkel M., *A knowledge-lite approach to word alignment*, in Véronis (ed.) (2000), chapter 5, p. 97-138, 2000.

Al-Onaizan Y., Knight K., « Machine Transliteration of Names in Arabic Text », *Proceedings of ACL Workshop on Computational Approaches to Semitic Languages*, Philadelphie, États-Unis, 2002a.

Al-Onaizan Y., Knight K., « Translating Named Entities Using Monolingual and Bilingual Resources », *Proceedings of the Conference of the Association for Computational Linguistics, ACL'02*, Philadelphie, États-Unis, p. 400-408, 2002b.

Bodenreider O., « The Unified Medical Language System (UMLS) : integrating biomedical terminology », *Nucleic Acids Research*, vol. 32, p. D267-D270, 2004.

Brown P. F., Cocke J., Stephen A. Della Pietra V. J. D. P., Jelinek F., Lafferty J. D., Mercer R. L., Roossin P. S., « A Statistical Approach to Machine Translation », *Computational Linguistics*, 1990.

Brown P. F., Pietra V. J. D., Pietra S. A. D., Mercer R. L., « The Mathematics of Statistical Machine Translation : Parameter Estimation », *Computational Linguistics*, 1993.

Charniak E., *Statistical Language Learning*, MIT Press, Cambridge, Massachusetts, 1993.

Claveau V., Zweigenbaum P., « Automatic Translation of Biomedical Terms by Supervised Transducer Inference », *Proceedings of the 10th Conference on Artificial Intelligence in Medicine, AIME 05*, Aberdeen, Écosse, p. 236-240, 2005.

Erjavec T., Džeroski S., « Machine learning of morphosyntactic structure : Lemmatizing unknown Slovene words », *Applied Artificial Intelligence*, vol. 18, n° 1, p. 17-41, 2004.

Fluhr C., Bisson F., Elkateb F., *Parallel text alignment using crosslingual information retrieval techniques*, in Véronis (ed.) (2000), chapter 9, 2000.

Fung P., McKeown K., « Finding terminology translations from non-parallel corpora », *Proceedings of the 5th Annual Workshop on Very Large Corpora*, Hong Kong, p. 192-202, 1997a.

Fung P., McKeown K., « A Technical Word and Term Translation Aid using Noisy Parallel Corpora Across Language Groups », *Machine Translation*, vol. 12, n° 1/2, p. 53-87, 1997b.

Gale W., Church K., « Identifying word correspondences in parallel texts », *Proceedings of the 4th Darpa Workshop on Speech and Natural Language*, Pacific Grove, États-Unis, p. 152-157, 1991.

Gaussier E., « Unsupervised Learning of Derivational Morphology from Inflectional Corpora », *Proceedings of Workshop on Unsupervised Methods in Natural Language Learning, 37th Annual Meeting of the Association for Computational Linguistics, ACL 99*, Maryland, États-Unis, p. 24-30, 1999.

Knight K., Graehl J., « Machine Transliteration », *Computational Linguistics*, vol. 24, n° 4, p. 599-612, 1998.

Markó K., Stefan Schulz O. M., Hahn U., « Bootstrapping Dictionaries for Cross-Language Information Retrieval », *Proceedings of the 28th International Conference on Research and Development in Information Retrieval, SIGIR 05*, Salvador, Brésil, p. 528-535, 2005.

Moreau F., Claveau V., « Extension de requêtes par relations morphologiques acquises automatiquement », *Revue I3 (Information - Interaction - Intelligence)*, vol. 6, n° 2, p. 31-50, 2006.

Oflazer K., Nirenburg S., « Practical Bootstrapping of Morphological Analyzers », *Proceedings of EACL Workshop on Computational Natural Language Learning, CONLL 99*, Bergen Norvège, 1999.

Qu Y., Grefenstette G., Evans D. A., « Automatic Transliteration for Japanese-to-English Text Retrieval », *Proceedings of the 26th International Conference on Research and Development in informaion Retrieval, SIGIR 03*, Toronto, Canada, 2003.

Schulz S., Markó K., Sbrissia E., Nohama P., Hahn U., « Cognate Mapping - A Heuristic Strategy for the Semi-Supervised Acquisition of a Spanish Lexicon from a Portuguese Seed Lexicon », *Proceedings of the 20th International Conference on Computational Linguistics, COLING'04*, Geneva, Switzerland, p. 813-819, 2004.

Tiedemann J., « Word to word alignment strategies », *Proceedings of the 20th International Conference on Computational Linguistics, COLING 04*, Genève, Suisse, p. 212-218, 2004.

Tsuji K., Daille B., Kageura K., « Extracting French-Japanese Word Pairs from Bilingual Corpora based on Transliteration Rules », *Proceedings of the 3rd International Conference on Language Resources and Evaluation LREC'02*, Las Palmas de Gran Canaria, Spain, p. 499-502, 2002.

Tuttle M., Sherertz D., Olson N., Erlbaum M., Sperzel D., Fuller L., Neslon S., « Using Meta-1 – the 1st Version of the UMLS Metathesaurus », *Proceedings of the 14th annual Symposium on Computer Applications in Medical Care (SCAMC)*, Washington, États-Unis, p. 131-135, 1990.

Véronis (ed.) J., *Parallel Text Processing*, Kluwer Academic Publishers, Dordrecht, 2000.