Inférence de règles de réécriture pour la traduction de termes biomédicaux

Vincent CLAVEAU
IRISA - CNRS
Campus de Beaulieu
35042 Rennes cedex, France
Vincent.Claveau@irisa.fr

Résumé. Dans le domaine biomédical, le caractère multilingue de l'accès à l'information est un problème d'importance. Dans cet article nous présentons une technique originale permettant de traduire des termes simples du domaine biomédical de et vers de nombreuses langues. Cette technique entièrement automatique repose sur l'apprentissage de règles de réécriture à partir d'exemples et l'utilisation de modèles de langues. Les évaluations présentées sont menées sur différentes paires de langues (français-anglais, espagnol-portugais, tchèque-anglais, russe-anglais...). Elles montrent que cette approche est très efficace et offre des performances variables selon les langues mais très bonnes dans l'ensemble et nettement supérieures à celles disponibles dans l'état de l'art. Les taux de précision de traductions s'étagent ainsi de 57.5 % pour la paire russe-anglais jusqu'à 85 % pour la paire espagnol-portugais et la paire français-anglais.

Abstract. In the biomedical domain, offering a multilingual access to specialized information is a major issue. In this paper, we present an original approach to translate simple biomedical terms between several languages. This fully automatic approach is based on a machine learning technique inferring rewriting rules and on language models. The experiments that are presented are done onn several language pairs (French-English, Spanish-Portuguese, Czech-English, Russian-English...). They demonstrate the efficiency of our approach by yielding translation performances that vary according to the languages but are always very good and better than those of state-of-art techniques. Indeed, the translation precision rates go from 57.5 % for translation from Russian to English up to 85 % for Spanish-Portuguese and French-English language pairs.

Mots-clés : traduction artificielle, terminologie biomédicale, apprentissage artificiel, modèles de langue.

Keywords: machine translation, biomedical terminology, machine learning, language models.

1 Introduction

Dans le domaine biomédical, les problématiques d'accès à l'information sont particulièrement importants. De nombreux documents sont en effet quotidiennement collectés dans des bases

spécialisées très consultées. La base PubMed regroupe par exemple 16 millions de publications médicales et fait face à plus de 3 millions de requêtes par jour. Dans la plupart de ces bases, les documents sont indexés à l'aide de terminologies de référence en anglais ; la mise en place de stratégies multilingues pour faciliter l'accès à ces bases pour les non-anglophones est donc cruciale. Quelques terminologies biomédicales multilingues existent, mais elles sont mises en défaut par l'évolution rapide des connaissances et le manque de moyens pour certaines langues.

En réponse à ces besoins, nous présentons et évaluons dans cet article une méthode de traduction automatique de termes biomédicaux fonctionnant sur différentes langues (anglais, espagnol, français, russe...). Cette méthode permet 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 s'appuient sur le fait que les termes biomédicaux sont construits sur les mêmes racines grecques et latines (Namer, 2005), et leurs dérivations très régulières (*e.g.* pour le couple français-anglais *ophtalmorragie*/*ophthalmorrhagia*, *ophtalmoplastie*/*ophthalmoplasty*, *leucorragie*/*leukorrhagia*).

Notre approche s'appuie sur une technique d'apprentissage artificiel simple que nous avons développée. Elle nous permet d'inférer un ensemble de règles de réécriture à partir de couples de termes langue source-langue cible traductions l'un de l'autre. Ces règles, une fois apprises, peuvent alors être appliquées à des termes de la langue source pour produire les termes équivalents dans la langue cible. Il est intéressant de noter qu'aucune connaissance, ni intervention humaine n'est requise à l'exception de la phase de supervision (*i.e.* la constitution de l'ensemble de couples de termes langue source-langue cible), celle-ci pouvant se faire simplement en exploitant les terminologies multilingues existantes.

Dans la section suivante, nous présentons les travaux connexes à notre problématique. Nous décrivons ensuite en section 3 notre technique de traduction de termes biomédicaux que nous évaluons sur différentes paires de langues en section 4.

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 & Zweigenbaum (2005b; 2005a). Celle-ci repose sur une technique d'apprentissage de transducteurs mais ne peut s'appliquer qu'aux langues partageant le même alphabet, contrairement à l'approche présentée ici. Outre sa plus grande souplesse, nous montrons en section 4 que notre technique obtient, de plus, des performances supérieures à cette approche par transducteurs. Schulz *et al.* (2004) ont également proposé une technique de traduction de termes biomédicaux, du portugais vers l'espagnol, s'appuyant sur des règles de réécriture. Cependant, ces règles sont fournies manuellement ; une telle approche n'est donc pas envisageable à grande échelle pour le traitement de plusieurs paires de langues.

Hors du domaine biomédical, des problématiques proches sont parfois abordées dans le domaine de la traduction automatique de textes. Ainsi, la détection de cognats (couples de mots bilingues de formes proches) (Fluhr *et al.*, 2000, *inter alia*) s'appuie sur des opérations morphologiques simples parfois proches des règles de réécriture que nous inférons. D'autres travaux

reposent quant à eux sur des recherches en corpus à l'aide de techniques statistiques de cooccurrences pour trouver des alignements – et donc des relations de traduction potentielle – entre termes dans des corpus alignés (Ahrenberg *et al*., 2000; Gale & Church, 1991) ou comparables (Fung & McKeown, 1997). 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). Mentionnons enfin les travaux sur la translittération, notamment du katakana ou de l'arabe (Tsuji *et al*., 2002; Knight & Graehl, 1998, par exemple). Les techniques utilisées dans ceux-ci sont parfois proches de celle proposée ici, mais ne concernent que la représentation d'imports dans des langues ayant un alphabet différent de la langue source.

3 Technique de traduction artificielle de termes

La technique de traduction de termes biomédicaux que nous proposons fonctionne en deux temps. Tout d'abord, des règles de réécriture sont inférées à partir d'exemples de paires de termes traductions l'un de l'autre (sous-section 3.1). Un modèle de langue est ensuite appris et utilisé pour choisir la traduction la plus probable parmi les possibilités générées par l'application des règles de réécriture inférées (sous-section 3.2). Nous terminons la section par quelques commentaires sur cette technique de traduction de termes.

3.1 Inférence de règles de réécriture

La technique de traduction proposée 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 : $\langle input\ string \rangle \rightarrow \langle output\ string \rangle$. Dans la suite de l'article, nous notons r une règle de réécriture, $\mathcal{R}$ est la liste de toutes les règles inférées pendant une expérience, $input(r)$ et $output(r)$ désignent respectivement la chaîne d'entrée et la chaîne de sortie de la règle r.

Algorithme 1 Apprentissage des règles de réécriture

```
1: aligner les paires de termes au niveau des lettres, mettre le résultat dans L
2: for all paire de termes W1 dans L do
3:     for all alignement de lettres dont les 2 lettres diffèrent dans W1 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
```

L'algorithme 1 donne un aperçu global de notre technique d'apprentissage. La première étape est réalisée à l'aide de DPalign (`http://www.cnts.ua.ac.be/~decadt/?section=dpalign`). Ce logiciel aligne deux séquences en minimisant leur distance d'édition par programmation dynamique selon l'algorithme de Wagner & Fischer (1974) ; les coûts de substitution des caractères sont calculés sur l'ensemble des paires à aligner. Ce logiciel ne repose donc pas sur une similarité formelle des caractères pour aligner les séquences ; il nous est ainsi possible d'aligner des termes ne partageant pas le même alphabet.

Une liste de paires de termes est donnée en entrée ; à chaque terme sont ajoutés deux caractères *#* pour représenter le début et la fin de la chaîne de caractères. La liste de sortie $\mathcal{L}$ va alors contenir les paires de termes alignés au niveau des lettres ; le tableau 1 en présente quelques exemples ('_' signifie *aucun caractère*). 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 termes considérée.

$\mathcal{L}$ portugais-anglais	$\mathcal{L}$ anglais-russe
#cetosteróides# *#ketosteroid_s#*	*#adenosinetriphosphatase#* #аденозин_трифосфатаза#
#electroporação_# *#electroporation#*	*#hydrox_ypregnenolone#* #гидроксипрегненолон_#
#encef_alograf_ia# *#encephalography_#*	*#keratoplasty__#* #кератопластика#

TAB. 1 – Exemples d'alignements produits pour deux paires de langues

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 associé 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 W1 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 $i \rightarrow y$, $gi \rightarrow gy$, $ie \rightarrow y$ (on ne note pas le caractère _ dans les règles), *ologie#* $\rightarrow$ *ology#*...

Le score d'une règle est calculé à partir de la liste $\mathcal{L}$; c'est 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 ($\subseteq$ représente l'inclusion de chaîne de caractères) :

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

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, antisymétrique 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...).

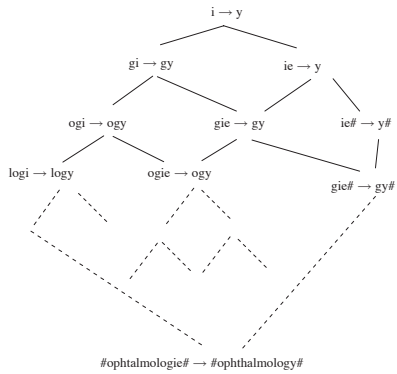

FIG. 1 – Treillis de recherche de l'exemple *i/y* dans *#opht_almologie#/#ophthalmology_#*

3.2 Évaluation des traductions proposées

L'algorithme présenté ci-dessus va potentiellement générer une règle de réécriture par différence pour chacune des paires de termes utilisées en exemple, conduisant à obtenir un grand nombre de règles. 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 appliquées. Dans le cas de règles concurrentes, toutes les possibilités sont générées. Pour choisir parmi ces possibilités la traduction la plus probable, nous avons utilisé une approche simple basée sur les modèles de langue (ML). Les 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 associer une probabilité à une séquence de mots, alors que dans notre cas, la probabilité va au contraire être associée à un mot considéré comme une séquence de lettres. Plus formellement, avec les notations standard on a :

$$P(w) = \prod_{i=1}^{m} P(l_i|l_1, ..., l_{i-1}) \quad \text{pour un terme } w \text{ composé des lettres } l_1, l_2, ..., l_m$$

En pratique, les probabilités $P(l_i|...)$ sont estimées à partir des termes biomédicaux 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 termes biomédicaux 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 de choisir la traduction la plus probable, cette technique nous donne un facteur de confiance sur la traduction retenue.

3.3 Commentaires sur la technique de traduction

Deux points concernant la technique de traduction proposée méritent d'être mentionnés. Il est tout d'abord intéressant de constater que l'approche que nous avons adoptée peut être rapprochée du cadre usuel de la traduction artificielle statistique dans lequel le but est de traduire une séquence de mots f dans une langue source en une séquence e dans la langue cible en cherchant e maximisant $P(e) \cdot P(f|e)$ (Brown *et al.*, 1993). Le terme $P(f|e)$ représente la probabilité que la séquence f soit la traduction de e. Ces probabilités, qui sont estimées à partir d'un corpus aligné, peuvent être rapprochées de nos règles de réécriture et leur score. Le terme $P(e)$ sert à vérifier que la séquence proposée soit bien formée ; son fonctionnement est en tout point similaire au modèle de langue que nous utilisons, si ce n'est que nos séquences sont composées de lettres et non de mots. Il y a cependant quelques différences importantes avec notre approche, principalement induites par la nature de nos données. Ainsi, dans le cadre de la traduction artificielle de textes, les différences d'ordonnancement des mots entre la langue source et la langue cible sont des problèmes difficiles à résoudre et mènent à construire des modèles de traduction compliqués. Dans notre cas, ce problème est quasiment inexistant : l'ordre des morphes, et donc des lettres composant les termes, varie peu d'une langue à l'autre. Le fait de manipuler des lettres et non des mots nous permet aussi d'utiliser un ML avec un historique de taille importante sans craindre de tomber trop souvent sur des séquences non observées ; les combinaisons possibles de lettres sont en effet bien moins nombreuses que les combinaisons de mots.

Le deuxième point à noter concerne une des limites évoquées par Claveau & Zweigenbaum (2005b) à propos de leur technique de traduction de termes. Ces derniers indiquent en effet que leur approche par transducteur ne peut pas prendre en compte des informations sur les termes comme les parties-du-discours (PoS). Cela a pour effet de générer des erreurs et de complexifier l'apprentissage dans le cas de mots polyfonctionnels comme *linguistique* qui se traduira différemment selon qu'il soit nom ou adjectif. Dans notre cas, cette limite est levée puisque l'ajout de ces informations se fait très naturellement avec le modèle de langues. Les probabilités peuvent en effet être calculées en incluant la partie-du-discours de la séquence en cours de traitement (on calcule alors les scores en fonction des $P(l_i|l_{i-n+1}, ..., l_{i-1}, PoS)$).

4 Évaluation de la traduction de terme

4.1 Description des données

Deux jeux de données sont utilisés 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 & Zweigenbaum (2005b), ce qui nous permettra de comparer les 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 (Bodenreider, 2004). Cette collection de thésaurus rassemble des termes biomédicaux dans 17 langues et associe à chacun des termes un identifiant de concept indépendant des langues, le Concept Unique identifier, CUI. Ces CUI nous permettent donc de constituer des ensembles de paires de termes bilingues. Là encore, seuls les termes simples non acronymes sont conservés.

4.2 Méthode d'évaluation

Pour l'évaluation de notre technique, nous suivons une approche standard : la liste initiale de paires de termes est découpée en deux ensembles, le premier sert pour l'apprentissage (inférence de règles et modèle de langue), et le second, composé de 1 000 paires, sert de jeu de test. Une fois les règles et le modèle de langue appris sur le jeu d'entraînement, 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 (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. Un seuil élevé doit favoriser la précision au détriment du nombre de traductions proposées, et vice-versa. À la manière de courbes rappel-précision, nous représentons donc ci-après la précision suivant le pourcentage de mots traduits pour tous les seuils possibles d'indice de confiance.

4.3 Résultats

4.3.1 Traduction entre le français et l'anglais

Pour cette première expérience, nous nous intéressons à la traduction entre le français et l'anglais. Comme précisé précédemment, nous utilisons le jeu de données Masson qui nous permet de comparer directement nos résultats à ceux de Claveau & Zweigenbaum (2005b; 2005a) dont nous reportons les résutats ci-après. Nous utilisons aussi les informations de parties-du-discours dans le modèle de langue. Les figures 2 et 3 présentent les graphes de précision des traductions générées sur les ensembles de test pour les deux sens de traduction. 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.

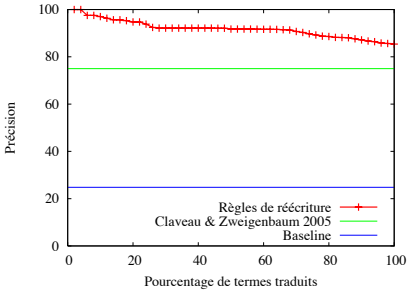

FIG. 2 – Performances de traduction du français vers l'anglais

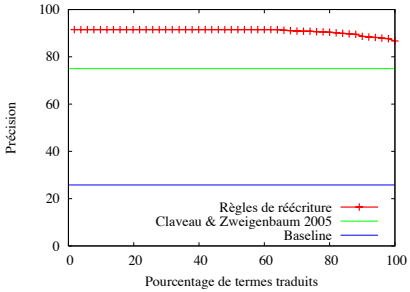

FIG. 3 – Performances de traduction de l'anglais vers français

Notre approche obtient de très bons résultats : 85.4 % de précision pour 100 % des mots traduits

pour le sens français-anglais et 84.8 % pour le sens inverse. Dans les deux cas, l'amélioration par rapport à l'approche par transducteur (Claveau & Zweigenbaum, 2005b) est de 10 %. Cela s'explique principalement par la souplesse de notre approche par règles qui lève la contrainte de déterminisme de l'approche par transducteur imposant qu'une séquence ne puisse se traduire que d'une façon, limitant et complexifiant ainsi l'apprentissage en présence de données bruitées, d'exceptions ou de mots polyfonctionnels. Concernant l'utilisation des modèles de langue, on remarque qu'utiliser les ML sans prendre en compte des parties-du-discours fournit des précisions légèrement plus faibles (82.6 % pour le sens français-anglais et 84.8 % pour l'autre sens). Et choisir le candidat au hasard parmi les différentes traductions génerées plutôt qu'utiliser les scores de ML mène à une précision d'environ 50 %. Ces deux résulats montrent bien l'intérêt des ML pour choisir la meilleure traduction et le bien-fondé de l'inclusion des parties-du-discours dans ces ML. Enfin, 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 indiquer que les deux langues sont suffisamment proches pour rendre la tâche d'apprentissage relativement aisée.

4.3.2 Autres paires de langues

Nous répétons les 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. La figure 4 présente les résultats obtenus avec deux langues réputées proches : l'espagnol et le portugais. Les résultats sont très bons : 87.9 % des termes portugais sont correctement traduits en espagnol quand on traduit tous les termes (*i.e.* quand aucun seuil pour le ML n'est fixé) ; dans le sens inverse, ce sont 85 % des termes qui sont correctement traduits. Ces bons résultats ne sont pas surprenants : l'espagnol et le portugais sont très proches, et comme le soulignent les très hautes *baselines*, beaucoup de mots sont en fait identiques.

Nous présentons maintenant les résultats obtenus par la traduction de diverses langues vers l'anglais — cas le plus à même d'être utilisé en pratique. La figure 5 présente les performances de la traduction de l'espagnol et du portugais vers l'anglais. Les résultats sont plutôt bons : 71.7 % des termes espagnols et 71.5 % des termes portugais sont correctement traduits quand toutes les traductions sont gardées (*i.e.* aucun seuil de ML n'est fixé). Ces résultats sont conformes avec la proximité de l'espagnol et du portugais illustrée dans l'expérience précédente. La traduction de l'italien et du tchèque vers l'anglais (figure 6) donne également des résultats comparables : au pire cas, 70 % des termes italiens et 75.5 % des termes tchèques sont correctement traduits.

Comme nous l'avons dit précédemment, notre technique de traduction peut traiter des langues avec des alphabets différents, pourvu qu'elles montrent des régularités pouvant être apprises automatiquement. Pour illustrer cela, nous nous intéressons à la paire russe-anglais. La figure 7 présente les résultats obtenus ; la précision minimale obtenue ici est de 57.5 % (du fait de la différence d'alphabet, la *baseline* est ici à 0). Bien qu'inférieurs aux autres paires de langues, ces résultats sont relativement bons étant donnée la difficulté apparente de la tâche. Cela met en exergue l'emploi en russe des mêmes racines gréco-latines que pour les autres langues étudiées.

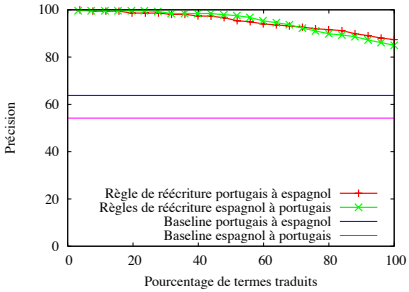

FIG. 4 – Performances de traduction espagnol - portugais

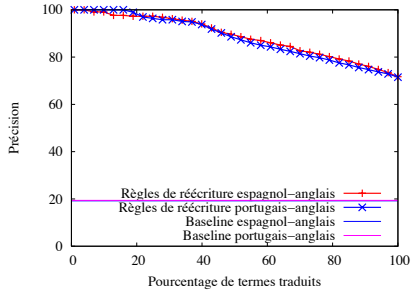

FIG. 5 – Performances de traduction portugais/espagnol vers anglais

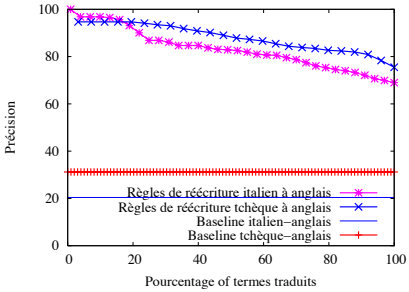

FIG. 6 – Performances de traduction de l'italien/tchèque vers l'anglais

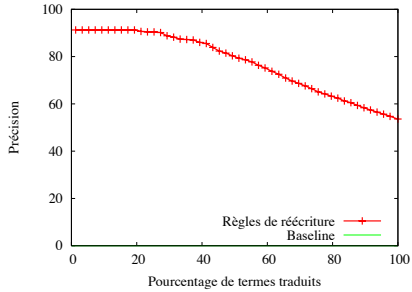

FIG. 7 – Performances de traduction du russe vers l'anglais

5 Conclusion

Notre technique de traduction à base de règles de réécriture capture automatiquement les régularités existant dans le domaine biomédical entre les termes des langues sources et cibles. De ce fait et sans surprise, la principale cause des traductions erronées observées est l'absence de lien morphologique entre le terme source et le terme cible. 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). Les évaluations menées montrent cependant que ces cas sont assez rares pour que notre technique offre des taux de précision variables selon les langues mais très bons dans l'ensemble, et supérieurs à ceux disponibles dans l'état de l'art. De plus, l'utilisation de modèles de langues pour donner un score à chaque traduction permet de contrôler la précision souhaitée en fixant ou non un seuil à dépasser.

Parmi les perspectives ouvertes par ce travail, la traduction des termes complexes (composés de plusieurs mots, comme *col du fémur*) est l'une des plus importantes pour assurer une bonne couverture des terminologies à traduire (ils représentent par exemple environ 50 % de la terminologie MeSH). Enfin, dans un cadre plus applicatif, l'utilisation de cette technique dans un cadre de recherche d'information translingue (traduction de requêtes de la base PubMed) est à l'étude et donne des premiers résultats encourageants (Claveau, 2007).

Références

AHRENBERG L., ANDERSSON M. & MERKEL M. (2000). *A knowledge-lite approach to word alignment*, chapter 5, p. 97–138. In (Véronis, 2000).

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

BROWN P. F., PIETRA V. J. D., PIETRA S. A. D. & MERCER R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, **19**(2).

CHARNIAK E. (1993). *Statistical Language Learning*. Cambridge, Massachusetts: MIT Press.

CLAVEAU V. (2007). Traduction automatique de termes biomédicaux pour la recherche d'information interlingue. In *Actes de la Conférence en Recherche d'Information et Applications, CORIA'07*, St-Étienne, France.

CLAVEAU V. & ZWEIGENBAUM P. (2005a). Automatic translation of biomedical terms by supervised transducer inference. In *Proceedings of the 10th Conference on Artificial Intelligence in Medicine, AIME 05*, Lecture Notes of Computer Science, Aberdeen, Écosse: Springer.

CLAVEAU V. & ZWEIGENBAUM P. (2005b). Traduction de termes biomédicaux par inférence de transducteurs. In *Actes de la conférence Traitement automatique des langues naturelles, TALN'05*, Dourdan, France.

FLUHR C., BISSON F. & ELKATEB F. (2000). *Parallel text alignment using crosslingual information retrieval techniques*, chapter 9. In (Véronis, 2000).

FUNG P. & MCKEOWN K. (1997). A technical word and term translation aid using noisy parallel corpora across language groups. *Machine Translation*, **12**(1/2), 53–87.

GALE W. & CHURCH K. (1991). Identifying word correspondences in parallel texts. In *Proceedings of the 4th Darpa Workshop on Speech and Natural Language*, p. 152–157, Pacific Grove, CA, États-Unis.

KNIGHT K. & GRAEHL J. (1998). Machine transliteration. *Computational Linguistics*, **24**(4), 599–612.

NAMER F. (2005). Morphosémantique pour l'appariement de termes dans le vocabulaire médical : approche multilingue. In *Actes de la conférence Traitement Automatique des langues naturelles, TALN'05*, Dourdan, France.

SCHULZ S., MARKÓ K., SBRISSIA E., NOHAMA P. & HAHN U. (2004). Cognate Mapping - A Heuristic Strategy for the Semi-Supervised Acquisition of a Spanish Lexicon from a Portuguese Seed Lexicon. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING'04*, p. 813–819, Genève, Suisse.

TSUJI K., DAILLE B. & KAGEURA K. (2002). Extracting french-japanese word pairs from bilingual corpora based on transliteration rules. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC'02*, p. 499–502, Las Palmas de Gran Canaria, Espagne.

J. VÉRONIS, Ed. (2000). *Parallel Text Processing*. Dordrecht: Kluwer Academic Publishers.

WAGNER R. A. & FISCHER M. J. (1974). The string-to-string correction problem. *Journal of the Association for Computing Machinery*, **21**(1), 168–173.