Traduction de termes biomédicaux par inférence de transducteurs

Vincent Claveau (1), Pierre Zweigenbaum (2, 3 & 4)

(1) OLST - Université de Montréal
CP 6128 succ. Centre-Ville
Montréal, QC, H3C 3J7, Canada
vincent.claveau@umontreal.ca

(2) AP-HP, STIM/DSI, Hôpital Broussais,
96, rue Didot, 75674 Paris cedex 14
(3) INSERM, U729, 15, rue de l'École de Médecine, 75006 Paris
(4) INaLCO, CRIM, 2, rue de Lille, 75343 Paris cedex 07
pz@biomath.jussieu.fr

Mots-clefs : Traduction automatique de termes, terminologie biomédicale, apprentissage artificiel, inférence de transducteurs

Keywords: Automatic translation of terms, biomedical terminology, machine learning, transducer induction

Résumé Cet article propose et évalue une méthode de traduction automatique de termes biomédicaux simples du français vers l'anglais et de l'anglais vers le français. Elle repose sur une technique d'apprentissage artificiel supervisée permettant d'inférer des transducteurs à partir d'exemples de couples de termes bilingues ; aucune autre ressource ou connaissance n'est requise. Ces transducteurs, capturant les grandes régularités de traduction existant dans le domaine biomédical, sont ensuite utilisés pour traduire de nouveaux termes français en anglais et vice versa. Les évaluations menées montrent que le taux de bonnes traductions de notre technique se situe entre 52 et 67%. À travers un examen des erreurs les plus courantes, nous identifions quelques limites inhérentes à notre approche et proposons quelques pistes pour les dépasser. Nous envisageons enfin plusieurs extensions à ce travail.

Abstract This paper presents and evaluates a method to automatically translate simple terms from French into English and English into French in the biomedical domain. It relies on a machine-learning technique that infers transducers from examples of bilingual pairs of terms; no additional resources or knowledge is needed. Then, these transducers, making the most of high translation regularities in the biomedical domain, can be used to translate new French terms into English or vice versa. Evaluations reported show that our technique achieves good successful translation rates (between 52 and 67%). When examining at the most frequent errors made, some inherent limits of our approach are identified, and several avenues are proposed in order to bypass them. Finally, some perspectives are put forward to extend this work.

1 Introduction

Dans le domaine biomédical, l'évolution rapide des connaissances et la prédominance de l'anglais comme langue de communication rendent cruciales les problématiques de production et de gestion de ressources terminologiques multilingues. Dans ce cadre, la traduction de terminologies existantes (par exemple le thésaurus MeSH), dont fait l'objet cet article, revêt une grande importance. Par ailleurs, outre leur utilité pour les professionnels du domaine, les ressources terminologiques multilingues sont aussi essentielles à beaucoup d'applications du TAL et plus particulièrement pour la traduction automatique. En effet, l'un des problèmes majeurs de cette dernière, lorsqu'elle est appliquée à des textes spécialisés, est l'absence de ressources de traduction (terminologies ou corpus alignés) portant sur le domaine. Ainsi, les expériences menées par P. Langlais et M. Carl (2004) montrent que, dans certains textes, 35% des phrases contiennent au moins un mot inconnu de leur système de traduction généraliste. Les auteurs montrent que ces mots sont en fait des termes du domaine d'étude et soulignent donc l'importance de disposer de ressources terminologiques multilingues pour mener à bien ces tâches de traduction.

Dans cet article, nous présentons et évaluons une méthode automatique tentant de répondre à ces besoins dans un cadre toutefois restreint. 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. Les expériences rapportées ici portent sur la traduction du français vers l'anglais et de l'anglais vers le français. Ce travail repose sur deux hypothèses majeures :

1. dans le domaine biomédical, les termes simples en anglais et français sont en majorité 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 français et anglais sont construits sur les mêmes racines grecques et latines, et leurs dérivations très régulières (*e.g. ophtalmorragie/ophthalmorrhagia*, *ophtalmoplastie/ophthalmoplasty*, *leucorragie/leukorrhagia*).

Notre approche s'appuie sur une technique d'apprentissage artificiel qui nous permet d'inférer un classifieur à partir de couples de termes français-anglais traduction l'un de l'autre et morphologiquement proches. C'est ce classifieur qui, étant donnés en entrée des termes français, doit ensuite permettre de produire les termes anglais correspondants ou inversement. Plus précisément, dans notre cas, le classifieur est un transducteur (*cf.* section suivante) et nous utilisons donc une technique existante d'inférence de transducteurs pour le générer 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.

Peu de travaux se placent dans le cadre de la traduction directe de termes. On peut néanmoins citer les travaux de S. Schulz *et al.* (2004) de traduction de termes biomédicaux du portugais vers l'espagnol fondés sur une analyse morphologique et l'utilisation de règles de réécriture fournies manuellement. Cependant, des problématiques proches sont souvent abordées dans le domaine de la traduction automatique 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) pour aligner des mots dans un corpus bilingue. Les transducteurs sont également parfois utilisés pour la traduction non pas de termes, mais de textes sous forme de chaînes de mots (Knight & Al-Onaizan, 1998) ou d'arbres syntaxiques (Knight & Graehl, 2005) ; les techniques de construction des transducteurs proposées dans ce cadre n'assurent cependant pas la même capacité à traiter des

séquences inconnues que celle que nous présentons ci-après. D'autres travaux reposent quant à eux sur des techniques statistiques de cooccurrences pour trouver des alignements entre mots ou 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.

La section suivante présente la technique que nous utilisons pour inférer des transducteurs. Nous décrivons ensuite en section 3 la méthodologie employée pour nos expérimentations et les données utilisées. La section 4 détaille d'un point de vue quantitatif et qualitatif les résultats obtenus et nous concluons en donnant quelques perspectives ouvertes par ce travail.

2 Inférence de transducteurs

D'un point de vue général, un transducteur est un outil qui permet d'accepter en entrée des séquences d'un certain langage (langage étant pris ici au sens le plus large) et de produire en sortie les séquences associées dans un autre langage. L'emploi de ce type d'outils à des tâches de traduction en langage naturel semble donc évident. Cependant, en pratique, la complexité des relations entre langue d'entrée et langue de sortie impose deux importantes limites à leur utilisation :

1. l'expressivité des transducteurs n'est pas assez importante pour représenter certains phénomènes complexes de traduction ;
2. la construction des transducteurs pour ce type de tâche est trop complexe pour être menée manuellement.

Ces deux raisons expliquent pourquoi ce type de techniques est en pratique peu employée en dehors de tâches restreintes.

Dans le travail que nous présentons ici, ces deux problèmes cruciaux sont atténués par le fait que notre tâche de traduction est limitée : nous tentons de traduire des termes simples, dans un domaine restreint, entre des langues proches. Par ailleurs, pour ce qui est de la première limite citée ci-dessus, les types de transducteurs que nous utilisons sont relativement expressifs (*cf.* sous-section suivante) et les résultats présentés en section 4 semblent indiquer qu'ils sont adaptés à notre tâche. De plus, ces transducteurs ne sont pas construits manuellement mais inférés automatiquement à l'aide d'un algorithme d'apprentissage artificiel (voir Cornuéjols & Miclet (2002) pour une introduction) développé par J. Oncina (1991) et nommé OSTIA (*cf.* section 2.2). C'est ce dernier point qui fait l'originalité de ce travail et permet de contourner la seconde limite énoncée ci-dessus.

2.1 Transducteurs sous-séquentiels

Les transducteurs inférés par OSTIA — utilisés pour traduire nos termes biomédicaux — sont une extension des transducteurs classiques appelés transducteurs sous-séquentiels. Des définitions formelles de ces objets peuvent être trouvées dans (Oncina *et al.*, 1993) ; nous n'en

donnons ci-dessous que des descriptions générales.

Les transducteurs sont des machines à états finis que l'on peut voir comme des graphes dans lesquels un symbole d'entrée et une séquence de sortie sont associés à chaque arc. Un transducteur a un état initial et un ou plusieurs états finals. Intuitivement, un transducteur est donc un automate auquel on associe des séquences de sortie aux symboles d'entrée sur les arcs. Une séquence d'entrée E est *reconnue* ou *acceptée* s'il existe une suite d'arcs partant de l'état initial et arrivant à un état final telle que les symboles d'entrée, concaténés dans l'ordre de parcours de ces arcs, forment exactement E. La *traduction* d'une séquence d'entrée E correspond à la concaténation, dans l'ordre, de toutes les séquences de sortie des arcs traversés pour la reconnaissance de E.

Un transducteur séquentiel est un transducteur dans lequel tous les états sont finals, et où il est impossible d'avoir deux arcs sortant d'un même état ayant le même symbole d'entrée. Cette dernière propriété est celle qui assure le déterminisme des traductions issues de ces transducteurs. Enfin, un transducteur sous-séquentiel est un transducteur séquentiel dans lequel à chaque état est associée une séquence de sortie. Celle-ci est produite lorsque la séquence d'entrée se termine sur l'état ; c'est ce qui rend les transducteurs sous-séquentiels relativement expressifs.

La figure 1 présente un transducteur sous-séquentiel simple avec les notations habituelles des automates. Il représente la fonction de traduction qui fait correspondre un mot d'entrée vide à la séquence de sortie D, $a(bc)^n$ à $A(BC)^nE$ et $a(bc)^nb$ à $A(BC)^nBF$. En revanche, un mot comme $abca$ n'est pas reconnu, et donc non traduit par ce transducteur.

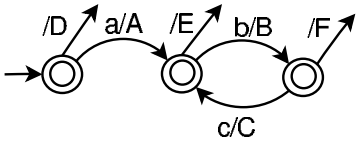

FIG. 1 – Exemple de transducteur sous-séquentiel

Dans notre cas, les séquences d'entrée sont soit les termes biomédicaux en français, vus comme des suites de lettres, et les séquences de sortie sont les termes correspondants en anglais, soit l'inverse.

2.2 Algorithme OSTIA

L'inférence de transducteurs est une technique d'apprentissage artificiel symbolique supervisée. Elle permet d'inférer, c'est-à-dire de produire automatiquement, un classifieur à partir d'exemples de couples séquence d'entrée/séquence de sortie ; dans notre cas, ce sont des couples de termes biomédicaux français-anglais. Ce classifieur est un transducteur qui reconnaît toutes les séquences d'entrée exemples et les traduit correctement en leur séquence de sortie correspondante, mais doit aussi idéalement être capable de produire la sortie correcte d'une chaîne d'entrée inconnue appartenant au même langage que les séquences d'entrée exemples. Les mécanismes de traduction des séquences d'entrée en séquences de sortie doivent donc être suffisamment réguliers pour permettre à l'algorithme d'apprentissage de *généraliser* les exemples ; on parle alors de saut inductif. Dans les travaux que nous présentons ici, cette phase d'inférence est mise en œuvre par l'algorithme OSTIA.

Cet algorithme d'inférence est formellement présenté par J. Oncina (1991), nous n'en décrivons ici que le principe général de fonctionnement. Nous l'illustrons à l'aide d'un exemple :

nous cherchons à apprendre le transducteur précédent (figure 1) avec les six exemples suivants : $\{\epsilon/D, a/AE, ab/ABF, abc/ABCE, abcb/ABCBF, abcbc/ABCBCE\}$ où ϵ représente la chaîne vide. L'algorithme OSTIA se déroule en trois étapes (Oncina, 1998) :

1. un arbre des préfixes de toutes les séquences d'entrée du jeu d'entraînement est construit. Des chaînes vides sont assignées à tous les nœuds et arcs internes de cet arbre, et à chaque nœud feuille est associée la séquence de sortie correspondant à la séquence d'entrée reconnue par cette branche (*cf.* figure 2).
2. tous les préfixes communs des séquences de sortie sont ensuite remontés des feuilles vers la racine de l'arbre (figure 3).

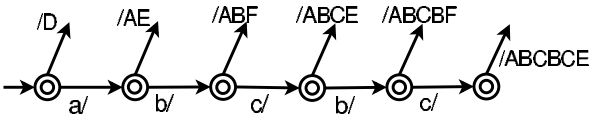

FIG. 2 – Transducteur après l'étape 1

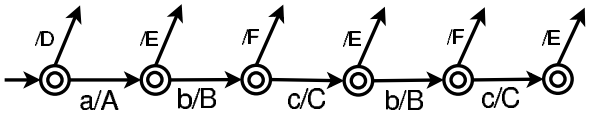

FIG. 3 – Transducteur après l'étape 2

3. enfin, en partant de la racine, tous les nœuds sont considérés deux à deux et fusionnés si le transducteur résultant n'entre pas en contradiction avec les données du jeu d'entraînement (figures 4 et 5). L'ordre de ces tentatives de fusion est généralement indiqué par une fonction heuristique. Il est possible de repousser des séquences de sortie vers les feuilles pour permettre des fusions. Quand plus aucune fusion n'est possible, l'algorithme termine.

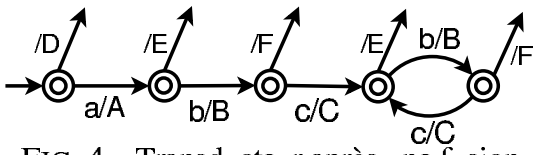

FIG. 4 – Transducteur après une fusion

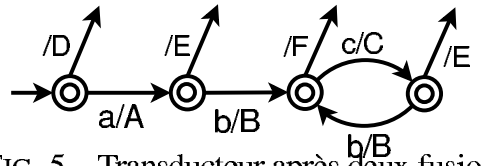

FIG. 5 – Transducteur après deux fusions

C'est bien sûr cette dernière étape qui assure le pouvoir inductif du transducteur, puisqu'elle assure de reconnaître et traduire correctement tous les couples exemples, mais permet également de reconnaître et traduire de nouvelles séquences d'entrée. Il a été montré formellement que cet algorithme converge et produit un transducteur décrivant le concept de traduction représenté par l'échantillon de données pairées constituant le jeu d'entraînement (Oncina, 1991). OSTIA a déjà été appliqué à de nombreuses tâches avec succès, dont la traduction de phrases en langage contrôlé (structures et vocabulaires restreints) (Oncina, 1998).

3 Expérimentations

Nous présentons tout d'abord les données utilisées pour nos expérimentations et la façon dont elles ont été préparées. Nous décrivons ensuite le cadre méthodologique des expériences d'inférence que nous avons menées à partir de ces données.

3.1 Constitution des données d'apprentissage et d'évaluation

Les données que nous utilisons pour évaluer notre technique sont issues d'un dictionnaire médical français en ligne (Dictionnaire Médical Masson, http://www.atmedica.com) contenant pour certaines de ses entrées les termes anglais équivalents. Parmi toutes les entrées, nous ne retenons que celles qui sont des mots simples à la fois en français et en anglais (absence d'espace et

de tiret), ne contenant pas de majuscules (pour éviter les noms propres) et d'une longueur minimale de 8 lettres (pour éviter les acronymes et se focaliser sur des termes techniques, contenant plusieurs morphes). Ce sont ainsi environ 12 000 paires de termes français-anglais qui sont obtenues des 35 000 entrées du dictionnaire.

Pour se focaliser sur les termes qui sont morphologiquement proches, la similarité formelle de chaque paire a été évaluée à l'aide d'une distance d'édition normalisée par la longueur des mots. Les paires sont classées dans une liste selon ce score en ordre décroissant de similarité :

```
1.00|zirconium|zirconium
...
0.93|ophtalmotoxine|ophthalmotoxin
0.93|ophtalmologiste|ophthalmologist
...
0.71|oschéite|oscheitis
0.71|organisé|organized
...
0.12|acouphène|tinnitus
0.11|engelures|chilblain
```

3.2 Méthodologie

Notre technique d'inférence repose entièrement sur les exemples, il est donc nécessaire de ne lui fournir pour l'entraînement que des paires effectivement morphologiquement proches, c'est-à-dire issues de la partie supérieure de la liste triée. Les données de test permettant d'évaluer notre approche peuvent quant à elles être tirées à n'importe quel niveau de la liste, même s'il semble évident qu'on ne peut pas attendre d'un quelconque système des traductions correctes des termes du bas de la liste. Aucun seuil n'apparaissant dans la distribution des mesures de similarité, on propose deux types d'expérience pour tenir compte de cela :

exp. 1. les paires d'entraînement et de test sont issues de la moitié supérieure de la liste ;

exp. 2. les paires d'entraînement sont issues de la moitié supérieure de la liste et celles de test de toute la liste.

Pour chacune des expériences, nous testons les sens de traductions français vers anglais et l'inverse, avec différentes tailles de jeux d'entraînement. Les jeux de test comportent 2000 paires (bien entendu différentes des paires d'entraînement) et les processus d'inférence et de validation sont répétés dix fois et les résultats moyennés. L'expérience 2 est une évaluation dans le pire des cas puisque les transducteurs inférés sont testés sur des paires qui peuvent n'être pas morphologiquement apparentées.

L'unique mesure utilisée pour rendre compte de la performance des transducteurs est la précision des traductions proposées. Cette précision est le ratio de termes de la langue source correctement traduits dans la langue cible (*i.e.* identiques aux termes attendus). Si le terme d'entrée n'est pas reconnu par le transducteur, il est considéré comme mal traduit.

4 Résultats

Cette section détaille les performances obtenues par les expériences présentées ci-avant. Nous en présentons d'abord les résultats d'un point vue quantitatif puis, en sous-section 4.2, nous

proposons un examen plus qualitatif des traductions issues des transducteurs. Enfin, à la lumière de ces résultats, nous mettons en évidence en sous-section 4.3 certaines limites inhérentes à notre approche.

4.1 Précision et complexité des transducteurs inférés

La précision des transducteurs est mesurée pour nos deux expériences selon le nombre d'exemples de couples d'entraînement utilisés par OSTIA. Les figures 6 et 7 présentent les courbes obtenues respectivement pour les expériences 1 et 2, dans les deux sens de traduction. Comme base de comparaison (*baseline*), nous calculons la précision qu'obtiendrait un système de génération de traduction simpliste proposant pour terme cible la même chaîne de caractères que le terme source (traduction à l'identique).

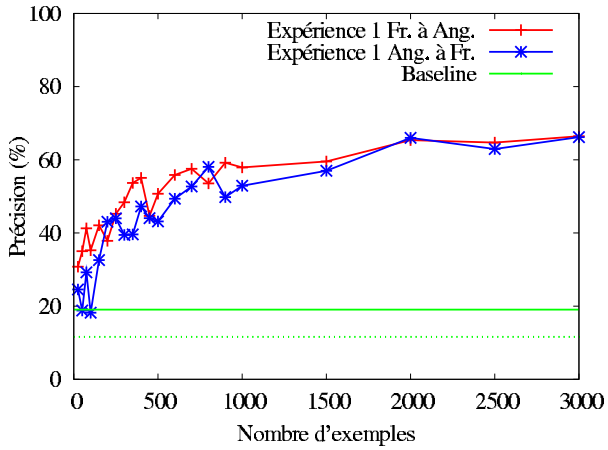

FIG. 6 – Précision selon le nombre d'exemples pour l'exp. 1

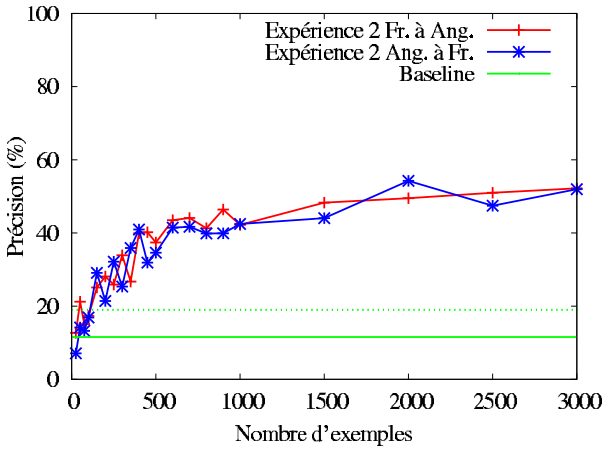

FIG. 7 – Précision selon le nombre d'exemples pour l'exp. 2

On remarque tout d'abord que le sens de traduction influe peu sur la précision. Par ailleurs, les résultats qui ressortent de ces figures sont plutôt bons. Pour 3000 exemples, la précision est d'environ 67% pour l'expérience 1 et 52% pour l'expérience 2, que ce soit du français vers l'anglais ou l'inverse. Comme l'indique la *baseline*, beaucoup de termes sont identiques en français et en anglais, mais les résultats des transducteurs dépassent largement cette base. Le taux de réussite élevé pour l'expérience 2 est particulièrement intéressant puisqu'il représente les résultats de notre technique sans aucune restriction, les tests étant effectués sur des couples parfois morphologiquement non apparentés. Ces résultats confirment donc le bien-fondé de notre approche, et notamment les deux hypothèses présentées en introduction.

Nous mesurons également la variation de la complexité des transducteurs inférés selon le nombre d'exemples. Cette complexité est mesurée en fonction du nombre de nœuds et d'arcs des transducteurs (figure 8) et en temps de calcul de la phase d'inférence (figure 9). Les informations reportées concernent l'expérience 1, dans le sens français vers l'anglais, celles des autres expériences étant similaires. On constate que la complexité arcs/nœuds est quasi linéaire en nombre d'exemples, mais très élevée. La taille de ces transducteurs est telle qu'il n'est d'ailleurs pas possible de les visualiser en totalité. La complexité en temps de calcul est plus problématique car elle augmente de manière exponentielle avec le nombre d'exemples, ce qui peut freiner l'utilisation de cette approche sur de plus larges jeux de données. Néanmoins, la précision se stabilisant assez rapidement, l'emploi de tels jeux de données semble inutile.

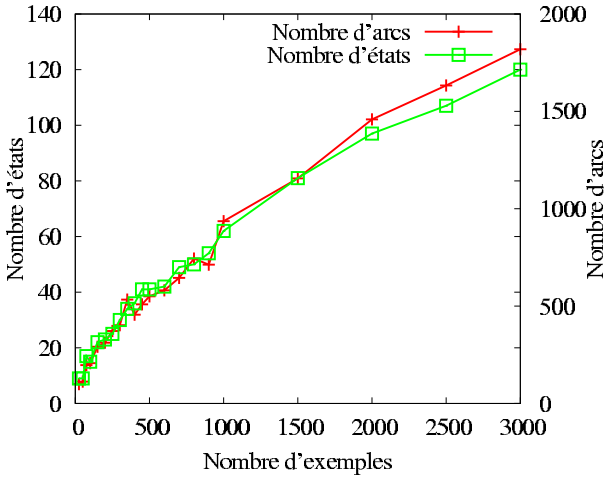

FIG. 8 – Complexité en nœuds et arcs selon le nombre d'exemples (exp. 1)

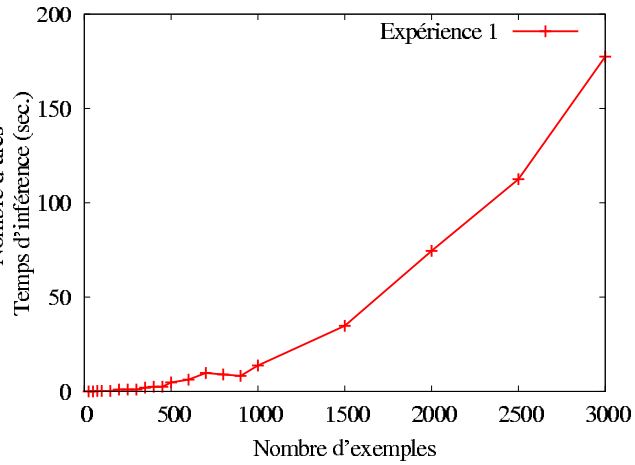

FIG. 9 – Complexité en temps selon le nombre d'exemples (exp. 1)

4.2 Examen des résultats

Les erreurs commises par les transducteurs inférés relèvent de différentes causes. Tout d'abord, certaines sont simplement dues à des termes dont la traduction n'est pas morphologiquement proche du terme d'origine. Comme attendu, ce type d'erreur est plus particulièrement présent dans l'expérience 2 (environ 10% des erreurs).

D'autres erreurs sont dues quant à elles à des exceptions aux régularités de traduction. On constate en effet que certaines traductions de familles de termes, bien que régulières dans leur majorité, présentent quelques cas particuliers. Ceux-ci font que les traductions proposées par les transducteurs inférés sont incorrectes. Par exemple, les termes français en *-rragie* se traduisent généralement en *-rrhagia* (*e.g.*, *stomatorragie/stomatorrhagia*, *pneumorragie/pneumorrhagia*...). Cependant, les couples *hémorragie/hemorrhage* et *pleurorragie/pleurorrhage* font exception à cette règle. Dans les expériences reportées précédemment, lorsqu'ils étaient rencontrés dans le jeu de test, *hémorragie* était (incorrectement) traduit en *hemorrhagia* et *pleurorragie* en *pleurorrhagia*. Il est intéressant de noter que ces termes irréguliers sont généralement d'emploi courant, appartenant presque à la langue générale, de laquelle ils ont certainement acquis cette dérivation particulière.

Enfin, certaines erreurs sont dues à l'apparente proximité de mots relevant de parties du discours ou de classes sémantiques différentes. Par exemple, les adjectifs français en *-ique* se traduisent généralement par un terme anglais en *-ic* (*e.g.*, *spasmolytique/spasmolytic*), alors que les noms avec le même suffixe se traduisent en *-ics* (*thermodynamique/thermodynamics*) ; ou encore, les noms de discipline en *-logie* se traduisent en *-logy* (*cardiologie/cardiology*) et les troubles du langage avec le même suffixe en *-logia* (*dyslogie/dyslogia*).

4.3 Limites de notre approche

Les erreurs de traduction présentées ci-avant mettent en relief certaines limitations de notre approche, ou plus précisément de la technique d'apprentissage utilisée. Tout d'abord, on ne peut apprendre que les régularités de traduction, il est donc normal que les exceptions, imprévisibles par nature, ne puissent pas être apprises. Cependant, ces exceptions, lorsqu'elles se trouvent

dans le jeu d'entraînement, risquent de provoquer de mauvaises inférences et complexifient les transducteurs. Malheureusement, OSTIA ne sait pas repérer ces paires irrégulières et ne permet pas d'apprendre des classifieurs distinguant règles générales et exceptions. Le même problème se pose pour gérer le bruit, c'est-à-dire des paires incorrectement encodées (faute d'orthographe dans l'un des termes ou erreur de traduction). Même si ce cas s'est peu présenté dans nos données, ce critère est à considérer si l'on souhaite employer cette même technique sur des paires de mots obtenues d'une source moins fiable.

Une autre limite de cette technique d'apprentissage est son incapacité à inclure des informations externes lors de sa phase d'inférence. En effet, OSTIA ne s'attache qu'à la suite de lettres composant les mots pour produire un transducteur, alors que, nous l'avons constaté, des informations catégorielles ou sémantiques permettraient de lever des ambiguïtés et d'améliorer les résultats. D'autres techniques d'apprentissage permettent d'adjoindre aisément ce type d'informations supplémentaires, comme la programmation logique inductive (Cornuéjols & Miclet, 2002, chapitre II.2), mais ne sont pas aussi performantes qu'OSTIA pour manipuler des séquences de lettres.

5 Conclusion et perspectives

Cet article présente une technique de traduction de termes simples du domaine biomédical du français vers l'anglais et inversement, s'appuyant sur une technique d'apprentissage artificiel, l'inférence de transducteurs. Les transducteurs sont inférés par l'algorithme OSTIA à partir d'exemples de paires de termes bilingues. Ils permettent ensuite, étant donné un terme dans une langue, de générer le terme correspondant dans une autre langue. Aucune connaissance ou ressource autre que les exemples n'est requise, laissant augurer une bonne portabilité de cette technique à d'autres paires de langues. Les évaluations que nous présentons montrent que cette technique obtient de bons résultats en produisant entre 50% et 66% de traductions correctes selon les expériences. On note de plus que certaines des erreurs de traduction sont dues à des mots largement utilisés, relevant presque de la langue courante. Ces mots ont par conséquent une grande chance d'apparaître dans des dictionnaires ou autres ressources de traduction, et donc ne nécessiteront pas l'emploi de transducteurs pour les traduire.

Beaucoup de perspectives sont ouvertes par ce travail. D'un point de vue technique, on peut notamment envisager d'appliquer cette même technique en considérant les termes non plus comme des séquences de lettres mais des séquences de morphes (*e.g. broncho⊕pleuro⊕pneumo⊕nie*). Des systèmes d'analyse morphologique dérivationnelle et compositionnelle existent déjà pour le domaine biomédical en français (Namer & Zweigenbaum, 2004) et pourraient ainsi servir de première étape à OSTIA. Une autre extension possible porte sur la recherche de traductions de termes complexes (à plusieurs mots). En effet, si l'on dispose de la traduction individuelle de chaque composant d'un terme complexe, on peut construire ou rechercher celle du terme pris dans sa globalité. Mais il faut pour cela tenir compte des variations possibles de ces termes (*virus de la variole/virus variolique*, *variola virus/variolic virus*) (Jacquemin, 2001; Daille, 2003).

D'un point de vue applicatif, nous envisageons d'utiliser et d'évaluer cette approche sur d'autres paires de langues (comprenant notamment l'espagnol, le portugais, l'allemand). Enfin, notre technique pourrait être utilisée pour l'alignement de corpus, les systèmes existants fonctionnant d'autant mieux que des couples de mots en relation de traduction sont connus (Véronis, 2000). Il sera alors intéressant de mesurer le gain de cette approche par rapport à une simple distance d'édition, communément utilisée dans ce contexte. Elle peut même être utilisée pour aligner

directement des ressources terminologiques. Dans ces deux cas, le problème abordé est quelque peu différent puisque, comme évoqué en introduction, il s'agit alors de mettre en relation des termes et non plus de produire des traductions.

Remerciements

Nous tenons à remercier José Oncina pour nous avoir donné accès au code d'OSTIA, et François Coste pour nous avoir fait partager son expérience sur l'inférence de langages réguliers.

Références

AHRENBERG L., ANDERSSON M. & MERKEL M. (2000). *A knowledge-lite approach to word alignment*, chapitre 5. In (Véronis, 2000).

CORNUÉJOLS A. & MICLET L. (2002). *Apprentissage artificiel*. Paris : Eyrolles.

DAILLE B. (2003). Conceptual structuring through term variation. In *Proceedings of the ACL'03 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, p. 9–16, Sapporo, Japon.

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

JACQUEMIN C. (2001). *Spotting and Discovering Terms through NLP*. Cambridge : MIT Press.

KNIGHT K. & AL-ONAIZAN Y. (1998). Translation with Finite-State Devices. In *Third Conference of the Association for Machine Translation in the Americas, AMTA'98*, p. 421–437, Langhorne, États-Unis.

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

KNIGHT K. & GRAEHL J. (2005). An Overview of Probabilistic Tree Transducers for Natural Language Processing. In *Proceedings of the 6th International Conference, CICLing 2005,*, Mexico, Mexique.

LANGLAIS P. & CARL M. (2004). General-purpose statistical translation engine and domain specific texts: Would it work ? *Terminology*, 10(1), 131–152.

NAMER F. & ZWEIGENBAUM P. (2004). Acquiring meaning for french medical terminology: contribution of morpho-semantics. In *Proceedings of the Conference MEDINFO 2004*, San-Francisco, États-Unis.

ONCINA J. (1991). *Aprendizaje de lenguajes regulares y transducciones subsecuenciales*. Thèse de doctorat, Universidad Politécnica de Valencia, Valence, Espagne.

ONCINA J. (1998). The data driven approach applied to the OSTIA algorithm. In *Proceedings of the Fourth International Colloquium on Grammatical Inference, ICGI'98*, p. 50–56, Ames, États-Unis.

ONCINA J., GARCÍA P. & VIDAL E. (1993). Learning subsequential transducers for pattern regognition interpretation tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(5), 448–458.

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 Third 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.