Constitution et exploitation d'un corpus parallèle issu du web pour l'extension d'une terminologie multilingue

Louise Deléger[1] et Pierre Zweigenbaum[1, 2, 3]
[1]Inserm, U729; [2]AP-HP, STIM; [3]Inalco, CRIM

1. Introduction

Nous présentons ici un exemple de constitution et d'exploitation d'un corpus parallèle pour enrichir des terminologies existantes. Il s'agit d'aligner ce corpus au niveau des mots afin de trouver de nouvelles traductions françaises de termes médicaux anglais. En effet, les terminologies d'un domaine de spécialité ne sont jamais exhaustives et on rencontre un besoin constant de les mettre à jour et de les enrichir. C'est particulièrement le cas en médecine où les terminologies contenues dans l'UMLS[1] sont très majoritairement en anglais. Nous cherchons à y remédier pour la partie française. En particulier, nous souhaitons enrichir le thésaurus MeSH[2] dans le cadre du projet VUMeF[3] [1].

Nous avons donc entrepris de constituer et d'utiliser un corpus parallèle bilingue français-anglais issu du web. Nous présentons d'abord notre méthode d'acquisition de corpus et les problèmes rencontrés. Nous détaillons ensuite les caractéristiques de ce corpus et les conséquences pour notre tâche. Enfin, nous évoquons le travail d'alignement effectué et montrons l'apport du corpus en exposant les résultats obtenus.

2. Acquisition du corpus

Notre corpus est issu du site web canadien bilingue français-anglais "Santé Canada / Health Canada"[4]. Plusieurs méthodes existent pour constituer un corpus parallèle à partir du web [2]. Nous avons pour notre part utilisé divers types d'informations : liens HTML explicites, méta-informations, critères de cohérence globale.

2.1. Aspiration du site

Le site a été téléchargé quasi-intégralement. Même si un tel téléchargement peut sembler lourd (il représente plusieurs gigaoctets avant filtrage des documents non textuels), nous avons considéré que cette aspiration préalable serait plus pratique qu'une aspiration sélective de pages possédant les propriétés souhaitées. En effet, le fait de disposer des fichiers « sous la main » permet d'obtenir de façon immédiate des informations (propriétés des noms de fichiers, balises employées dans les fichiers HTML, nombres respectifs des différents types de fichiers) qui orientent les traitements effectués. Il rend aussi beaucoup plus rapide la mise au point des programmes, qui n'ont plus à télécharger les fichiers utiles. Enfin, ce téléchargement initial n'empêche pas de compléter ultérieurement le corpus en suivant des liens qui n'avaient pas été exhaustivement explorés.

2.2. Génération de paires de documents parallèles : critères d'inclusion

L'appariement des documents parallèles a été effectué en se basant sur le marquage de la structure hypertextuelle des pages du site : les liens explicites vers une page de traduction. En effet, après examen de la structure des pages web du site, nous avons constaté que chaque page contenait un lien vers sa traduction, que

1 L'UMLS (Unified Medical Language System) comprend un métathésaurus qui rassemble différentes ressources terminologiques biomédicales et fait le lien entre ces terminologies.

2 Le MeSH (Medical Subjects Headings), thésaurus utilisé pour indexer la littérature scientifique biomédicale, est inclus dans l'UMLS. Sa version française est traduite par le DISC (département de l'information scientifique et de la communication) de l'Inserm.

3 Le projet VUMeF est financé dans le cadre du Réseau National des Technologies pour la Santé (Ministère de la recherche) et vise à étendre la part du français dans l'UMLS. Notamment, un des sous-projets est d'étendre la traduction française du MeSH. Parmi les partenaires, outre l'Inserm U729, on compte le DISC (Inserm) et l'équipe CISMeF.

4 http://www.hc-sc.gc.ca/

l'on suit en cliquant sur une image (image « english » sur une page en français et image « français » sur une page en anglais) ou sur un lien textuel « français » ou « anglais ». Le critère opératoire est alors un lien HTML (<a href=...>) contenant une image nommée « english » ou « français » (attribut "alt" de la balise <img> ou texte explicite). Une liste des paires de textes est ainsi générée.

2.3. Problèmes rencontrés lors de l'appariement : critères d'exclusion

Nous avons cependant constaté la présence de fichiers anglais parmi la liste des fichiers français, et vice-versa. Une étude de ces documents a montré que cela était dû à une erreur des auteurs de ces pages dans la valeur de l'attribut "alt". Des documents anglais possédaient la valeur "english" et des documents français la valeur "français". Nous avons exploité un autre aspect de la structure des pages web afin d'affiner le critère d'appariement des documents : nous nous sommes appuyés sur des méta-informations. En effet, les pages contenant aussi une balise <meta> indiquant la langue du document, nous avons ajouté un test supplémentaire : il faut que la langue du document et la langue indiquée par l'attribut "alt" soient différentes.

Il reste encore un petit nombre de fichiers mélangés, dus à une erreur et dans la balise <meta> et dans l'attribut « alt ». Nous éliminons les plus évidents d'après leur nom: un document anglais ne devrait pas comporter le mot « français » dans son nom et vice-versa. Dans ce site, nous observons que les noms de fichiers sont en anglais pour les documents anglais et en français pour les documents français. De plus une certaine partie des documents sont organisés dans des dossiers « english » et « français ». Il est donc utile de s'appuyer sur les noms de fichier. Cependant, il aurait été difficile de se baser entièrement dessus : le type de correspondance de noms n'est pas uniforme dans le site, et la mise en correspondance aurait nécessité dans certains cas d'être en mesure de traduire les mots utilisés dans les noms de fichiers.

Nous avons ensuite eu recours à des critères de cohérence globale pour éliminer les documents non parallèles. Par exemple, certains documents ont obtenu plus d'un correspondant. Or un document ne doit correspondre qu'à un seul autre. Nous avons décidé de supprimer ces documents vu leur faible proportion. De plus, il n'est pas nécessaire d'être exhaustif pour la tâche visée. Une grande quantité de documents, mais pas nécessairement la totalité du site, suffit pour trouver des termes médicaux. Nous avons aussi éliminé les quelques documents présents à la fois dans les documents français et dans les documents anglais.

Nous nous sommes également appuyés sur la taille des documents pour relever d'éventuelles erreurs d'appariement. Deux documents parallèles doivent avoir des tailles proches (en général le document français est légèrement plus long que le document anglais). Nous nous attendons donc à trouver un certain rapport de taille entre les documents et nous éliminons ceux qui s'écartent trop de ce rapport.

Il serait encore possible, après cela, d'utiliser le contenu des documents pour continuer à affiner la constitution du corpus parallèle : passer un catégoriseur de langue pour déterminer la langue de chaque document et éliminer ainsi des erreurs restantes.

Enfin, un dernier critère spécifique à notre tâche d'alignement a consisté à évaluer la qualité de l'alignement en phrases de deux documents. Ce point est détaillé dans la section 4.1.

2.4. Synthèse

Nous pouvons synthétiser notre stratégie de constitution de corpus parallèle de la sorte : nous employons à la fois des tests d'inclusion (pour construire le corpus) et des tests d'exclusion (pour réduire le corpus, éliminer le bruit).

Nous récupérons d'abord un corpus parallèle un peu plus large que nécessaire grâce au test de base sur les liens HTML dans les documents. Ceci est notre test d'inclusion auquel nous ajoutons plusieurs tests d'exclusion. On peut en distinguer deux sortes : ceux qui peuvent s'appliquer en même temps que les tests d'inclusion afin de les affiner (c'est le cas de notre test sur la balise meta que nous combinons au test principal sur les liens HTML) et ceux qui nécessitent la collecte de tous les documents du corpus (c'est le cas des tests

de cohérence globale sur le nombre de correspondants d'un document et le classement des documents dans une seule catégorie, française ou anglaise). Certains tests peuvent aussi s'appliquer dans l'un ou l'autre cas : ce sont les tests sur les noms de fichiers, la taille des documents, et la qualité de l'alignement des phrases. En ce qui nous concerne, nous avons effectué ces tests une fois la collection de documents obtenue, ce qui nous a donné des valeurs moyennes pertinentes pour le rapport de tailles ou l'alignement de phrases.

3. Caractéristiques du corpus

3.1.Quantité et qualité

Comme beaucoup de ressources issues du web, notre corpus est très large. Il compte 11041 paires de documents, soit environ 27,7 millions de mots. Ceci est un avantage pour la tâche d'alignement que nous souhaitons accomplir. En effet, celle-ci inclut des traitements statistiques (calcul de co-occurrences) qui sont plus performants sur de grandes quantités de données.

Comme pour toutes les ressources issues du web, la question de la qualité des données se pose. Celles-ci sont souvent bruitées et il s'agit de savoir dans quelle mesure ce bruit va affecter le traitement à effectuer. Dans notre cas, on constate effectivement une certaine proportion de bruit, plus particulièrement des fautes d'orthographe, des espaces oubliées ou insérées, etc. Ces mots seront traités comme inconnus, ce qui rend l'alignement plus difficile.

3.1.Encodage

Les pages Web du site sont encodées avec le code page Windows 1252 (cp 1252). Cela est sujet à problème car ce jeu de caractères utilise des positions correspondant à des caractères de contrôle dans d'autres jeux (Iso-latin 1, Unicode) qui seront donc mal affichés. Il s'agit des caractères de position 128 à 160. Dans les pages Web, ceux-ci sont soit codés tels quels, soit sous forme d'entités numériques HTML. Dans les deux cas, ils ne sont affichés correctement qu'avec le jeu de caractère cp 1252. Par souci de standardisation et de portabilité, nous désirons des textes encodés en Iso-latin 1 ou en UTF-8. Il faudra donc effectuer un recodage des documents. Notre choix s'est porté sur un encodage en UTF-8, format plus portable qui peut servir à encoder beaucoup plus de langues.

3.2.Langue

Notre corpus est bilingue français-anglais. Pour être plus précis, le site web est canadien et il s'agit donc de français du Québec. Nous relevons donc des différences par rapport au français de France et certaines traductions peuvent donc paraître étranges voire erronées. En particulier, un certain nombre de termes ou d'expressions n'existent pas en français standard ou sont peu usitées (« poser des gestes », « recherchiste »).

3.3.Public

Le site web s'adresse aux spécialistes ainsi qu'au grand public. Il est donc probable qu'il contient moins de vocabulaire spécialisé médical qu'un site entièrement consacré aux professionnels. Il contient à la fois du vocabulaire très spécialisé et du vocabulaire plus général à destination des patients (dit « vocabulaire patient »). Il serait donc intéresser de procéder à une caractérisation du contenu des documents, afin de déterminer quelle portion vise le grand public, et quelle autre vise plus particulièrement les spécialistes. Cela permettrait de caractériser et classer le vocabulaire récupéré.

3.4.Domaines

Le site « santé canada » se divise en une multitude de « sous-sites » traitant de sujets divers. À la distinction

selon le public s'ajoute donc celle sur les thèmes abordés. Nous avons déjà abordé cette problématique dans une situation proche, celle de l'exploitation du corpus médical EQUER composé de documents issus de sites web différents. Nous avons alors combiné un étiquetage sémantique et une analyse factorielle des correspondances [3] pour faire émerger les grands pôles thématiques (que nous avons baptisés organisation, bioprocess, matière) de ces différents sites. Nous envisageons d'appliquer la même méthode sur ce corpus pour mettre en lumière des sous-corpus plus homogènes.

4. Alignement et Résultats

4.1. Alignement

Nous rappelons ici brièvement la tâche à accomplir. Il s'agit d'aligner les mots du corpus parallèle. Nous avons procédé en deux grandes étapes:

- Tout d'abord, un alignement au niveau des phrases du corpus, qui est nécessaire pour l'alignement des mots. Nous avons utilisé l'outil de Dan Melamed, GMA (Geometric Mapping and Alignment) [4] qui donne de bons résultats.
- Puis nous alignons les mots du corpus grâce à une suite d'outils développée à l'université de Linköping (Suède), les I*Tools [5]. Ces outils permettent d'aligner à la fois des mots simples et des mots complexes, ce qui convient à notre but terminologique. Parmi l'ensemble des alignements, nous sélectionnons les termes médicaux qui nous intéressent.

Lors de la tâche d'alignement, il est encore possible de raffiner l'appariement des documents parallèles. En effet, l'alignement des phrases de deux documents parallèles est un bon indicateur de correspondance entre documents. Si l'alignement donne des résultats très mauvais, il y a de fortes chances que les documents ne soient pas parallèles. Nous développons une méthode de détection de ces documents en évaluant la qualité de l'alignement des phrases. Nous considérons que la majorité des correspondances entre phrases doivent être des correspondances 1:1 et accordons des pénalités graduelles pour chaque autre type d'alignement (1:2, 2:1, 0:1, 1:0 etc.). Un score d'alignement est ainsi calculé et les documents s'éloignant trop de la moyenne sont considérés non parallèles et éliminés.

4.2. Résultats

L'exploitation de ce corpus a donné des résultats intéressants.

Le corpus a été entièrement converti en texte, segmenté en phrases, étiqueté (par TreeTagger) et analysé syntaxiquement (par Syntex). Sa taille étant importante, nous n'en avons pour le moment traité jusqu'au bout qu'une partie (540 paires de documents). Sur ces 540 paires, nous avons pu récupérer 9860 termes médicaux anglais (rappelons que nous cherchons des traductions françaises de termes MeSH anglais) avec leurs traductions françaises. Nous avons étudié plus particulièrement un échantillon des résultats. Il s'agit de 145 traductions de termes MeSH. Parmi ces traductions, 66 sont déjà connues et enregistrées dans la version française du MeSH, et 79 sont de nouvelles traductions. Le fait de retrouver des traductions attestées est un bon signe d'efficacité de la méthode. Les nouvelles traductions ont été soumises à un « expert » et 64 de ces traductions ont été reconnues valides.

Nous avons examiné les traductions récupérées du corpus parallèle et les avons réparties en deux groupes (voir les exemples du tableau 1) :

- des variantes morphologiques : c'est-à-dire les pluriels et les féminins de termes déjà présents; ce sont les traductions les moins intéressantes car elles peuvent être générées sans avoir recours à un corpus parallèle ;
- des synonymes : ce sont les termes qui nous intéressent. Ce sont des termes médicaux utilisés en pratique mais auxquels un traducteur n'aurait pas forcément pensé.

1. Tableau 1 Exemples de traduction de termes MeSH

Terme anglais	Terme français	Terme(s) déjà présent(s) dans le MeSH	Type
Adipose tissue	Tissus adipeux	Tissu adipeux	Variante morphologique
Bone cancer	Cancer des os	Tumeurs osseuses / Tumeur des os	Synonyme

5. Conclusion

En utilisant le web, nous avons pu trouver un corpus bilingue correspondant à notre domaine de spécialité et d'une taille suffisante. Nous avons mis en place une stratégie d'acquisition de corpus parallèle basée sur des techniques d'inclusion et d'exclusion. L'exploitation de ce corpus nous a permis d'atteindre notre objectif : trouver des termes médicaux français qui ne sont pas recensés dans les terminologies de l'UMLS. Ces termes sont utilisés en pratique, mais un traducteur français n'y aurait pas forcément pensé. De plus, en ayant recours à un gros corpus, on récupère souvent plusieurs traductions pour un même terme.

Références

[1] S.J. Darmoni, É. Jarrousse, P. Zweigenbaum, P. Le Beux, R. Baud, M. Joubert, H. Vallée, R.A. Côté, A. Buemi, D. Bourigault, G. Recourcé, S. Jeanneau, and J.M. Rodrigues. Extending the French part of the UMLS, in M. Musen, editor, Actes AMIA Annual Fall Symposium 2003, Washington, DC, 2003.

[2] Resnik P, Smith N A. The web as a parallel corpus. Computational Linguistics, 29, 349-380. Special Issue on the Web as a Corpus. 2003.

[3] Delbecque T, Zweigenbaum P. Indexation UMLS en français: une expérience. In Régis Beuscart and Jean-Marc Brunetaud, editors, Actes Journées francophones d'informatique médicale, Lille, 2005.

[4] Melamed I. D. (2000). Bitext maps and alignments via pattern recognition. In J. Véronis Rédacteur, Parallel Text Processing : Alignment and use of translation corpora. Dordrecht : Kluwer Academic Publishers.

[5] Merkel M., Petterstedt M. & Ahrenberg L. (2003). Interactive word alignment for corpus linguistics. In Proceedings from Corpus Linguistics 2003, Lancaster, UK.