Utilisation du Web comme ressource bilingue pour la traduction de termes complexes français/anglais

Stéphanie LEON
Equipe DELIC – Université de Provence
29, Av. Robert Schuman – 13621 Aix-en-Provence Cedex 1
fanny.leon@orange.fr

Le Web, qui génère des besoins considérables de traduction, offre en même temps un réservoir gigantesque de données bilingues, qui peuvent être exploitées par des moyens automatiques, en particulier grâce à des moteurs de recherche tels que *Google* ou *Yahoo*, afin d'extraire ou de valider des traductions candidates. Bien que l'utilisation du Web en tant que grande base de données lexicale soit un phénomène récent et encore peu maîtrisé, de plus en plus de travaux en TALN s'y appuient, et la traduction automatique n'échappe pas au phénomène. Nous proposons une réflexion sur l'utilisation du Web en tant que base de données lexicale pour la traduction. Nous montrons d'abord que, bien que la littérature soit encore jeune et les tests embryonnaires, de plus en plus d'expériences aboutissent à des résultats intéressants. Nous présentons ensuite nos travaux relatifs à la traduction de termes complexes par comparaison de « mondes lexicaux » à partir du Web et nous analysons enfin les limites du Web auxquelles nous avons été confrontées.

1- Le Web, une méga-base de données lexicales bilingues

Jusqu'à présent, une majorité des travaux en traduction ont proposé des méthodes basées sur l'exploitation de corpus parallèles (pour un état de l'art, voir Véronis, 2000) ou comparables (citons, entre autres, (Rapp, 1999), (Fung, McKeown, 1997) (Fung, Yee, 1998) et (Morin et al., 2004)). Toutefois, les corpus parallèles constituent des ressources rares. De plus, ces derniers, tout comme les corpus comparables, concernent des domaines restreints. Il est difficile de savoir combien de mots sont indexés par les moteurs de recherche sur le Web dans chaque langue, étant donné le caractère commercialement sensible de cette information, mais des tests indirects (voir Kilgarriff & Grefenstette, 2003) permettent d'estimer à environ 100 milliards le nombre de mots indexés par *Google* pour la seule langue anglaise. Cette quantité est considérable : le British National Corpus, qui est de loin le plus grand corpus linguistique au monde, et a servi de base à de nombreuses études (Burnard, 1995), ne comporte que 100 millions de mots, c'est-à-dire une taille environ 1000 fois inférieure. Même si les données du Web sont moins contrôlées, et donc plus « bruitées », elles permettent d'envisager un changement radical d'échelle pour la linguistique empirique. Nous présentons deux grands courants principaux quant aux travaux relatifs à l'acquisition de traductions à partir du Web : ceux qui utilisent le Web en tant que base de données quantitative et ceux qui exploitent le Web tel un corpus « parallèle » ou « partiellement parallèle ».

1-1 Une source d'informations quantitatives pour la traduction

En traduction automatique, le Web est particulièrement utile afin de tester si un mot, une co-occurrence ou une expression est utilisée. Nous avons présenté dans (Léon, Millon, 2005) une méthode de traduction semi-automatique de co-occurrences lexicales, basée sur la fréquence sur le Web des traductions candidates. Prenons pour exemple la co-occurrence lexicale *commettre-vol* (voir Figure 1), *Google* nous permet de valider les traductions correctes, grâce à leur nombre d'occurrences. Par exemple, la requête "*commit a flight*" OR "*commit the flight*" retourne seulement 13 résultats. La requête "*commit a theft*" OR "*commit the theft*" retourne quant à elle 5110 résultats. Parmi ces deux traductions candidates, les résultats sélectionnent de façon écrasante la co-occurrence lexicale satisfaisante (*to commit-theft*) (Léon, Millon, 2005).

	Effectifs absolus		***Effectifs par million***	
	flight	**theft**	**flight**	**theft**
commit	13	5510	0	306

Figure 1 : Exemples de résultats sur *Google* (janvier 2005)

Les travaux de (Léon, Millon, 2005) se situent dans la lignée de ceux de (Grefenstette, 1999) qui utilisent le nombre de fréquences d'une co-occurrence sur le Web afin de sélectionner les meilleures traductions de co-occurrences lexicales, ainsi que ceux de (Cao, Li, 2002). Bien que les fréquences sur le Web ne soient que des approximations et qu'il faille manipuler celles-ci avec une grande attention, l'utilisation de plus en plus fréquente du Web afin d'extraire de façon quantitative des phénomènes lexicaux permet d'envisager le critère de la fréquence sous un nouveau jour pour la linguistique de corpus. De plus en plus de travaux combinent des méthodes statistiques à des résultats obtenus sur le Web. Cette méthode a été appliquée pour la première fois par (Turney, 2001), qui a proposé la méthode Web-based Mutual Information (WMI), c'est-à-dire l'application de calculs d'information mutuelle (Church & Hanks, 1989) à des résultats de fréquences de co-occurrences à partir de requêtes Internet. La difficulté en traduction automatique est de ne pas pouvoir tester les relations d'équivalences entre la co-occurrence lexicale source et sa traduction : certains cas de polysémie ne peuvent être résolus.

1-2 Le Web, un « corpus parallèle »

Au-delà de l'aspect quantitatif des données lexicales, le Web est riche d'un grand nombre de documents bilingues. En effet, de nombreux documents sont des textes parallèles (manuels, catalogues, etc.). Dans d'autres cas, il s'agit de traductions ponctuelles données dans le texte d'un document monolingue, comme dans l'exemple :

"Further support was guaranteed by large loans from the World Bank, the Saudi Fund, France's Central Fund for Economic Cooperation (Caisse Centrale de Coopération Economique--CCCE)"

Certains travaux ont exploité le Web en tant que « corpus parallèle » ou « partiellement parallèle ». Citons, entre autres, les travaux de (Ma et Liberman, 1999) et le système BITS (Bilingual Internet Text Search) qui permet d'acquérir des textes parallèles multilingues, à partir du Web ; mais aussi les travaux de (Nagata, 2001) pour le japonais et l'anglais et ceux de (Resnik, 1999 ; Resnik et Smith, 2002) pour l'anglais et l'allemand. Toutefois, la difficulté de telles méthodes est que le Web est susceptible de contenir des traductions erronées, qu'il s'agisse d'emplois de locuteurs non natifs, ou de textes générés par des systèmes de traduction automatique.

2- Une expérience d'acquisition de traductions de termes complexes par comparaison de « mondes lexicaux »

Nous présentons une méthode de traduction automatique français/anglais de termes complexes, basée sur une comparaison de « mondes lexicaux » (ensemble de co-occurrents), à partir du Web. A partir de termes complexes récupérés via le moteur de recherche *Exalead*, nous générons toutes leurs traductions potentielles par combinaison des traductions des unités simples au sein d'un dictionnaire. L'intérêt de notre méthode est qu'elle permet de désambiguïser de façon efficace les cas de polysémie, en exploitant l'entourage linguistique sur le Web des termes complexes et de leurs traductions candidates.

2-1- Génération des « mondes lexicaux » à partir du Web

Le moteur de recherche Yahoo est interrogé automatiquement afin de récupérer les 1000 premiers titres et résumés renvoyés pour chaque requête des termes complexes. Nous générons la liste des cinquante mots les plus fréquents pour les résultats de chaque terme complexe, liste qui constitue leur monde lexical. Les mêmes étapes sont répétées pour les traductions potentielles des termes complexes.

Une comparaison entre les mondes lexicaux français et anglais nous permet enfin de désambiguïser les cas de polysémie, à l'aide du coefficient de Jacquard (degré de similitude entre deux ensembles). La figure 2 montre un exemple de mondes lexicaux très proches pour le français et l'anglais pour le terme complexe *appareil numérique* et sa traduction *digital camera*.

Appareil numérique	canon, photographie, nikon,informatique, produits,accessoires, digital, mémoire, kodak, pc, olympus, flash, zoom, cartes, argentique, prises, gamme, reflex, prise, matériel
Digital camera	photography, film, computer, kodak, technology, olympus,canon, right, zoom, sony, memory, resolution, lens, imaging, fuji, ratings, pc, brands, product, battery

Figure 2 : Extrait des mondes lexicaux d'*appareil numérique* et de *digital camera*

2-2- Génération des requêtes des couples de traduction

Nous utilisons enfin un filtre lors duquel nous utilisons le Web tel un corpus partiellement « parallèle ». Notre hypothèse est que si une traduction candidate est correcte, elle doit apparaître au moins une fois sur le Web dans un même document que le terme complexe source. Nous ne conservons que les couples français/anglais ayant une fréquence sur le Web supérieure ou égale à un, comme dans l'exemple :

"quartier latin" "Latin quarter" (Fréquence du couple : 6)

2-3- Premiers résultats

A l'heure actuelle, notre méthode a été évaluée sur un échantillon de 132 termes complexes. Notre expérience a retourné 57 traductions correctes sur les 132 termes complexes proposés, soit un rappel d´un peu moins de la moitié (43,2 %). La précision, elle, est excellente, puisqu´aucune des traductions n'était erronée, ce qui permet d'ajouter les expressions à notre dictionnaire sans autre intervention. Notre but étant la production automatique de ressources bilingues, nous privilégions volontairement la précision au détriment du rappel, de façon à obtenir une ressource de bonne qualité avec un effort manuel minimal. D'autres tests en cours visent à évaluer la méthode sur un plus grand nombre de données.

3- Limites et perspectives de l'utilisation du Web pour la traduction

Nous présentons les limites auxquelles nous avons été confrontée lors de l'utilisation du Web en tant que ressource lexicale pour la traduction.

3-1 Qualité des données

Les ressources disponibles sur Internet sont plus « bruitées » que celles des corpus « traditionnels », car les textes sont publiés librement. En effet, le Web renferme des biais (ou bizarreries) langagiers émis par exemple par des locuteurs non natifs ou des locuteurs non-spécialistes. Ce problème est d'autant plus présent pour la langue anglaise que de nombreux locuteurs non natifs sont conduits à l'utiliser. Ces combinatoires lexicales erronées "bruitent" le Web, et pourraient provoquer des résultats erronés en traduction. Grâce à la fréquence de leurs occurrences, des méthodes statistiques permettent d'éliminer (ou du moins de réduire) automatiquement celles-ci. De plus, certains documents peuvent être dupliqués, notamment à cause de la présence de « spams ». Depuis ces derniers mois, le spam ne touche pas uniquement les textes monolingues, il est également présent d'un point de vue multilingue, car les sites traduits de façon automatique sont de plus en plus nombreux. Enfin, contrairement aux corpus « traditionnels », les données du Web contiennent un nombre important de mots liés au Web qui ne présentent pas d'utilité dans l'étude linguistique d'un terme donné, comme par exemple en français, *lien*, *blog*, *site*, etc. Il est important d'avoir recours à des « anti-dictionnaires » permettant de ne pas prendre en compte ces mots.

3-2 Analyse syntaxique

Les données textuelles du Web sont "brutes", c'est-à-dire qu'aucune information morpho-syntaxique n'est adjointe. Or, ce manque d'accès à des informations morpho-syntaxiques lors des requêtes ne permet pas de réduire les ambiguïtés catégorielles, ainsi que de catégoriser les lexèmes en cadres syntaxiques de type "nom adjectif", "nom-verbe", adverbe-adjectif", etc. Prenons l'exemple suivant :

The Library of Congress ***set the changeover*** *date.*

Dans ce cas, *changeover* est régi par le nom *date* et non par la forme verbale *set*. Toutefois, lors de requêtes de la forme *set the changeover* (en tant que verbe/déterminant/nom), de tels cas feront partie des résultats bruités. De plus, Google et Yahoo ne prennent pas en compte des phénomènes tels que la ponctuation ou encore les majuscules. Dans l'exemple suivant, il n'est pas possible de discriminer *reserve, a theft* de *reserve a theft* :

*A man will face court next month charged with stealing three date palms from a **Swansea reserve, a theft** which sparked three months of community outrage.*

Dans notre cas, cette difficulté syntaxique s'applique également à l'extraction des mondes lexicaux. Par exemple, la traduction du terme *aspirateur* est *vacuum cleaner*. A l'heure actuelle, les deux mots *vacuum* et *cleaner* sont recensés de façon séparée au sein de nos listes, mais il serait important de les regrouper, par le biais d'une analyse des dépendances syntaxiques des résumés anglais. Une perspective d'évolution à ces problèmes est d'appliquer une analyse morpho-syntaxique aux résumés ou aux pages Web extraites, à l'aide d'analyseurs syntaxiques.

Références

BURNARD, L. (1995). *The BNC Reference Manual*. Oxford :Oxford University Computing Service

CAO Y., LI H. (2002). Base noun phrase translation using web data and the EM algorithm. Actes de *International Conference on Computational Linguistics (COLING)*, 127-133.

CHUQUET H., PAILLARD M. (1987). *Approche linguistique des problèmes de traduction: anglais-français*. Gap : Ophrys.

CHURCH, K. W.; HANKS, P. (1989). Word association norms, mutual information, and lexicography. In:Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics. Vancouver, British Columbia, 76-83.

FUNG P., MCKEOWN K. (1997). Finding terminology translations from non-parallel corpora. Actes de *Annual Workshop on Very Large Corpora*, 192-202.

FUNG P., YEE L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. Actes de *International Conference on Computational Linguistics (COLING)*, 414-420.

GREFENSTETTE G. (1999). The WWW as a resource for example-based MT tasks. Actes de *Translating and the Computer Conference*.

KILGARRIFF A., GREFENSTETTE G. (2003). Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics, 29*(3): 333-348.

LEON S., MILLON C. (2005). Acquisition semi-automatique de relations lexicales bilingues (français-anglais) à partir du Web. Actes de *Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)*, 595-604.

MA, XIAOYI AND MARK LIBERMAN. (1999). Bits: A method for bilingual text search over the Web. In Machine Translation Summit VII.

MORIN E., DUFOUR-KOWALSKI S., DAILLE B. (2004). Extraction de terminologies bilingues à partir de corpus. Actes de *Traitement Automatique des Langues Naturelles (TALN)*.

NAGATA, M., SAITO, T., & SUZUKI, K. (2001). *Using the Web as a bilingual dictionary*. Association of Computational Linguistics (ACL) Workshop on human language technology and knowledge management, Toulouse, France.

RAPP R. (1999). Automatic Identification of Word Translations from Unrelated English and German Corpora. Actes de *Association for Computational Linguistics*, 519-525.

RESNIK, P. (1999). *Mining the web for bilingual text*. Actes de *Association of Computational Linguistics*, 527-534.

RESNIK, P., & SMITH, N. (2002). *The Web as a parallel corpus*. Report technique UMIACS-TR-2002, Maryland, Etats-Unis, Université de Maryland.

TURNEY, P.D. (2001). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. Actes de *Twelfth European Conference on Machine Learning*, 491-502.

VÉRONIS J. (2000). Parallel Text Processing: Alignment and use of translation corpora, Dordrecht : Kluwer Academic Publishers.