TALN 2008, Avignon, 9–13 juin 2008

Traduction multilingue : le projet MulTra

Eric Wehrli, Luka Nerima
LATL-Département de linguistique
Université de Genève
Eric.Wehrli@lettres.unige.ch, Luka.Nerima@lettres.unige.ch

Résumé. L'augmentation rapide des échanges et des communications pluriculturels, en particulier sur internet, intensifie les besoins d'outils multilingues y compris de traduction. Cet article décrit un projet en cours au LATL pour le développement d'un système de traduction multilingue basé sur un modèle linguistique abstrait et largement générique, ainsi que sur un modèle logiciel basé sur la notion d'objet. Les langues envisagées dans la première phase de ce projet sont l'allemand, le français, l'italien, l'espagnol et l'anglais.

Abstract. The increase of cross-cultural communication triggered notably by the Internet intensifies the needs for multilingual linguistic tools, in particular translation systems for several languages. The LATL has developed an efficient multilingual parsing technology based on an abstract and generic linguistic model and on object-oriented software design. The proposed project intends to apply a similar approach to the problem of multilingual translation (German, French, Italian and English).

Mots-clés : Traduction automatique multilingue, approche par objets, génération de lexiques bilingues.

Keywords: Multilingual machine translation, object design, bilingual dictionary derivation.

1 Introduction

Malgré l'avancée spectaculaire des systèmes de traduction stochastiques (cf. Ney, 2005), de nombreux chercheurs dans le domaine de la traduction automatique considèrent qu'une variante du modèle de transfert basé sur des règles linguistiques constitue encore pour l'heure l'approche la plus satisfaisante au problème de la traduction. Il y a par contre un large désaccord sur la question du choix du modèle linguistique, sur la nature des représentations qui serviront à exprimer les correspondances d'une langue à l'autre, ainsi que sur le niveau d'abstraction auquel le transfert devrait s'effectuer (voir Eberle, 2001).

Un des problèmes que nous souhaitons aborder dans ce projet est celui de la traduction multilingue, c'est-à-dire du développement d'un système de traduction capable de traiter non pas quelques paires de langues, mais potentiellement un large nombre de langues. Rappelons, pour illustrer ce point, que dans le contexte européen pas moins d'une quinzaine de langues sont couramment utilisées dans la vie quotidienne mais aussi pour des activités commerciales ou politiques. Et comme le signale Boitet (2001), "Despite considerable investment over the past 50

years, only a small number of language pairs is covered by MT (...), and even fewer are capable of quality translation or speech translation."

Le projet MulTra repose sur un niveau de représentation linguistique abstrait, inspiré des travaux récents en linguistique générative (cf. Chomsky, 1995, Culicover et Jackendoff, 2005, Bresnan 2001), suffisamment riche pour exprimer toutes la diversité des langues prises en considération, mais suffisamment abstrait pour permettre de rendre compte de généralisations profondes qui se cachent parfois derrière des différences de surface. Une telle approche a été adoptée avec succès pour l'analyseur Fips (cf. Wehrli, 2007). Au plan informatique, c'est une modélisation "objets" qui a été retenue, très semblable à celle définie pour l'analyseur multilingue Fips.

La rapide prolifération des modules de transfert a souvent été reprochée au modèle de traduction basé sur le transfert syntaxique comme modèle pour la traduction multilingue (voir par exemple Arnold et al. 1995 :81). L'argument repose sur le fait que le nombre de modules de transfert croît de manière quadratique en fonction du nombre n de langues. Cet argument est cependant considérablement affaibli si l'on peut montrer que les modules de transfert sont relativement peu importants par rapport aux modules d'analyse et de génération, qui eux croissent de manière linéaire au nombre de langues prises en considération. C'est bien l'objectif du projet MulTra, qui s'appuyant sur un algorithme de transfert générique, restreint les modules de transfert spécifiques aux aspects non-isomorphes des structures sources et cibles, comme nous le verrons plus bas. L'autre élément de complexité non linéraire dans le développement d'un système de traduction multilingue est celui des dictionnaires bilingues. Nous verrons dans la section 3, une méthode de dérivation partielle des dictionnaires bilingues par transitivité qui semble en mesure de réduire drastiquement l'ampleur du travail à accomplir.

2 Architecture du système

L'algorithme de traduction obéit au schéma traditionnel du transfert. Le texte source est tout d'abord analysé par l'analyseur multilingue Fips (Wehrli, 2007), qui produit une représentation abstraite de la phrase précisant d'une part le découpage syntagmatique (structure arborescente) et d'autre part une représentation des relations prédicat-arguments (un peu sur le modèle des structures d'arguments utilisées en LFG). Etant donné le caractère abstrait de ce niveau de représentation, la mise en correspondance d'une langue à une autre est un processus relativement simple, qui peut être esquissé comme suit : parcours récursif de l'arborescence source dans l'ordre tête, sous-arbre gauche, sous-arbre droit. Le transfert lexical intervient au moment du transfert de la tête (non vide) d'un constituant et livre un objet lexical cible, souvent mais pas obligatoirement, de la même catégorie que l'objet lexical source. Le mécanisme de projection syntaxique – qui permet la création d'une structure syntaxique sur la base d'un item lexical – est ensuite invoqué pour déterminer la structure syntaxique cible dont l'item lexical constitue la tête. Dans le cas de constituants qui ont été interprétés comme des arguments, le processus de transfert est un peu différent dans la mesure où certaines propriétés de la langue cible peuvent être déterminées par les informations lexicales associées au prédicat qui gouverne cet argument. Pour prendre un exemple simple, l'objet direct du verbe français *regarder* dans l'example (1a) sera traduit en anglais comme un syntagme prépositionnel avec la préposition *at* comme tête, comme illustré en (2a). Cette information vient de la base de données lexicale. Plus spécifiquement, notre dictionnaire bilingue spécifie non seulement les correspondances entre items lexicaux, mais également dans le cas d'items de nature prédicative, la correspondance entre les arguments. Autrement dit, le lexique français-anglais établit une relation entre le lexème

français [NP *regarder* NP] et le lexème anglais [NP *look* [PP at NP]].

(1)a. Paul regardait la voiture.

b. [$_{\texttt{TP}}$ [$_{\texttt{DP}}$ Paul] regardait$_i$ [$_{\texttt{VP}}$ e$_i$ [$_{\texttt{DP}}$ la [$_{\texttt{NP}}$ voiture]]]]

(2)a. Paul was looking at the car.

b. [$_{\texttt{TP}}$ [$_{\texttt{DP}}$ Paul] was [$_{\texttt{VP}}$ looking [$_{\texttt{PP}}$ at [$_{\texttt{DP}}$ the [$_{\texttt{NP}}$ car]]]]]

La génération se fait sur la base du mécanisme de projection, ainsi que sur la base de la structure source, puisque par défaut l'algorithme de transfert fait l'hypothèse de structures source et cible isomorphes. Les règles de transfert spécifiques à une paire de langues donnée sont régies par le mécanisme de redéfinition de méthode. C'est le cas, par exemple, pour le transfert entre le français et l'anglais des adjectifs postnominaux, dont les équivalents anglais se placent en position prénominales, ou pour prendre un exemple plus drastique, la gestion de l'ordre des mots entre l'allemand (langue à verbe en position finale de phrase) et le français. La génération de la phrase cible s'achève par la génération morphologique, c'est-à-dire la sélection de la forme morphologique d'un mot appropriée aux contraintes locales telles que nombre, cas, genre, personne, temps ou mode.

2.1 Ajout d'une nouvelle langue

L'ajout d'une langue additionnelle au système exige les composantes suivantes : (i) l'analyseur Fips pour cette langue, (ii) un module de génération (principalement morphologique), (iii) un dictionnaire bilingue pour chacune des autres langues du système, et (iv) un module (éventuellement vide) de transfert pour chacune des paires de langues additionnelles[1].

Traduction `L`-à-`L` Comme nous l'avons vu ci-dessus, l'ajout d'une nouvelle langue à notre système exige, en plus des lexiques monolingues et bilingues, un analyseur pour cette langue, un générateur, c'est-à-dire deux composantes qui sont complétement indépendantes des applications de traductions particulières. Autrement dit, un texte en langue `L` est analysé de la même manière, quelle que soit la langue cible dans laquelle il sera traduit. De même, la génération d'un texte cible `L` se fait complètement indépendamment de la langue source du texte. Cela revient à dire que les modules les plus difficiles à réaliser sont indépendants, respectivement des langues cibles et des langues sources avec lesquels ils seront amenés à interagir.

Un outil de développement original de notre projet a été d'ajouter à la combinatoire traditionnelle $n \cdot (n-1)$ pour n langues distinctes, la traduction d'une langue vers elle-même, soit $n \cdot n$. Etant donné notre modèle général de traduction, la traduction d'un texte disons du français vers le français exige (i) un analyseur du français, (ii) un générateur du français et (iii) un dictionnaire bilingue français-français. Comme dans ce cas les langues source et cible sont isomorphes, le module de transfert français-français sera vide, le transfert étant intégralement pris en charge par le module de transfert générique. L'avantage méthodologique que procure cette traduction `L`-à-`L` est de permettre une validation de l'analyseur et du générateur de la langue `L` de manière extrêmement simple puisqu'il suffira de comparer l'entrée et la sortie de ce "traducteur" pour déceler des problèmes dans l'un ou l'autre de ces deux modules. Autre avantage pratique

[1] Ce module sera vide dans le cas (hypothétique) de deux langues dont les structures syntaxiques seraient isomorphes.

de cette approche, le travail sur la langue L ne nécessite évidemment aucune connaissance bilingue particulière. Le développement et la validation systématique deux modules d'analyse et de génération de la nouvelle langue `L` simplifie grandement le développement des modules bilingues `L-L'`.

L'exemple suivant illustre le fonctionnement de ce traducteur L-L pour l'italien. La phrase (3a) est la phrase source et (3b) la phrase cible. Un petit outil de développement permet de mettre en gras (ou en couleur) toutes les différences entre une phrase source et la phrase cible correspondante.

(3)a. Raffaello **ci** ha lasciato magnifici quadri.
b. Raffaello ha lasciato **ci** magnifici quadri.

Dans cet exemple, volontairement simplifié, c'est la génération du pronom clitique *ci* qui fait problème, puisqu'il apparaît en position postverbale plutôt que préverbale. De telles erreurs sont plus faciles à identifier et à corriger dans un contexte L-L que dans une traduction à partir d'une autre langue.

3 Dictionnaires bilingues

La base de données lexicale utilisée dans le projet MulTra est composée pour chaque langue (i) d'un lexique de mots, contenant toutes les formes fléchies des mots de la langue, (ii) d'un lexique de lexèmes, contenant les informations syntaxiques des mots (un lexème correspondant à peu près à une entrée d'un dictionnaire classique) et (iii) un lexique de collocations (en réalité des expressions à mots multiples). Nous appelons items lexicaux les lexèmes et les collocations d'une langue.

La base de données lexicale multilingue contient les informations lexicales pour le transfert d'une langue à une autre. Pour les fins de stockage, nous utilisons un système de gestion de base de données relationnelle. Pour chaque paire de langues, le dictionnaire bilingue est implémenté sous forme de table relationnelle contenant les associations entre les items lexicaux des deux langues. Le dictionnaire bilingue est bi-directionnel, c'est-à-dire que l'association est valable dans les deux sens. En plus de ces liens, la table contient des informations sur le contexte de traduction, les préférences de traduction dans les cas de traductions multiples, les descripteurs sémantiques pour les mots ambigus, les correspondances entres arguments dans les deux langues (c'est le cas notamment des verbes). La structure des tables est identique pour toutes les paires de langues.

Bien que les lexiques bilingues soient bidirectiennels, ils ne sont pas symétriques. Si un mot *m* de la langue A a une seule traduction *t* dans la langue B, cela n'implique pas que *t* n'a qu'une seule traduction *m*. Par exemple, le mot anglais *tongue* est traduit en français par *langue*, alors que *langue* est traduit vers l'anglais par *tongue* et par *language*. Dans ce cas, l'attribut "descripteur" mentionnera respectivement "partie du corps" ou "language". Un autre élément d'asymétrie est l'attribut "préférence" destiné à ordonner la traduction par des synonymes. Par exemple, le lexicographe pourra, pour traduire *lovely*, donner un ordre de préférence pour *charmant* par rapport à *agréable*. Les préférences de traduction dans l'autre sens devront bien entendu être considérées indépendamment.

Un des problèmes de la traduction automatique multilingue réside dans le fait que le nombre de dictionnaires bilingues requis est une fonction quadratique du nombre de langues considérées.

En d'autres termes, il est nécessaire d'avoir autant de tables bilingues qu'il y a de paires de langues considérées, c'est-à-dire $n \cdot (n-1)/2$ tables. Par exemple, pour 4 langues, il faut 6 dictionnaires bilingues, pour 5 langues 10 dictionnaires bilingues, pour 6 langues 15 dictionnaires bilingues, etc. Dans le cadre de ce projet, nous prenons en compte 5 langues (français, anglais, allemand, italien, espagnol) ce qui requiert idéalement, si nous visons la traduction de ces langues vers toutes les autres, 10 dictionnaires bilingues. Nous verrons dans la section suivante que nous contournons en partie ce problème en générant automatiquement une partie de ces dictionnaires bilingues.

Nous considérons qu'une couverture bilingue satisfaisante (pour les fins de traduction générale) exige au moins 40 à 50'000 correspondances par paire de langues. Pour le moment, nous disposons de 4 dictionnaires bilingues sur les 10 nécessaires avec les nombres d'entrées listés ci-dessous :

paire de langues	nombre d'entrées
anglais-français	75'000
allemand-français	45'000
français-italien	37'000
espagnol-français	20'000

FIG. 1 – Nombre de correspondances dans les dictionnaires bilingues

Il est important de mentionner que ces 4 dictionnaires bilingues ont été créés manuellement par des lexicographes et que la qualité des entrées peut être considérée comme bonne.

3.1 Génération automatique

Pour parvenir à la constitution de la base de données bilingue pendant la durée du projet, nous envisageons d'utiliser une génération semi-automatique pour une partie des dictionnaires bilingues. À cette fin, nous allons dériver par transitivité un lexique bilingue en utilisant deux lexiques existants. Si, par exemple, nous disposons d'un lexique bilingue pour la paire de langues anglais, français et un autre pour la paire français, allemand, nous pouvons dériver un lexique bilingue pour la paire anglais, allemand. Les correspondances générées sont ensuite validées par corpus comme nous allons le voir ci-dessous. Les correspondances qui ne peuvent pas être validées de cette façon nécessitent un contrôle manuel.

L'idée d'utiliser une langue pivot pour dériver un nouveau lexique bilingue par transitivité n'est pas nouvelle (voir par exemple, Paik & al. 2004, Ahn & Frampton 2006). Le recours à une langue pivot a aussi été utilisé dans d'autre domaine comme la recherche d'information interlangue (*Cross Language Information Retrieval*) (voir par exemple Gollins & Sanderson 2001, Lu & al. 2004).

Le schéma général de la dérivation est d'utiliser deux lexiques bilingues qui partagent la même langue et de faire une jointure sur les tables relationnelles correspondantes en utilisant l'item lexical de la langue en partagée comme pivot. Par exemple, la validation des correspondances ainsi obtenues est essentielle et seulement celles qui sont de bonne qualité doivent être conservées dans le lexique générés.

Plus précisément, le processus se déroule comme suit :

1. Choisir deux tables bilingues pour les paires de langues (A, B) et (B, C) et effectuer une jointure naturelle.
2. Valider toutes les correspondances non ambiguë. Est considérée comme correspondance non ambiguë toute correspondance obtenue par transitivité à partir de correspondances dont le champ "descripteur" est vide.
3. Valider toutes les correspondances obtenues par un pivot de type collocation. Nous considérons en effet comme très improbable qu'une collocation soit ambiguë.
4. Toutes les autres correspondances sont soumises à deux filtres successifs : a) seules les correspondances avec une préférence supérieure ou égale à la moyenne sont prises en considération[2], b) les correspondances sont ensuite recherchées dans un corpus parallèle, seules les correspondances effectivement utilisées comme traduction dans le corpus sont validées.

La dernière étape mérite d'être un peu détaillée. Tout d'abord, le corpus parallèle est étiqueté par FipsTag, un étiqueteur basé sur l'analyseur syntaxique Fips. Cette manière de procéder a le grand avantage de lemmatiser les mots du corpus, ce qui permet d'effectuer le traitement au niveau des items lexicaux (lexèmes et collocations) qui sont les éléments de base des correspondances bilingues plutôt qu'au niveau des mots. Cette approche est particulièrement précieuse pour les verbes à particules, comme on en trouve par exemple en anglais et en allemand. Afin de vérifier la validité des correspondances, nous avons développé un outil qui recherche et qui compte dans le corpus parallèle le nombre d'occurrences de chaque correspondance générée ainsi que celles de ses deux correspondants pris séparément. A la fin du processus, nous effectuons un calcul de rapport de vraisemblance (*Likelihood ratio*) afin de décider de la validation ou du rejet des correspondances rarement utilisées dans le corpus. Un autre avantage (inattendu) de travailler sur un corpus étiqueté est de pouvoir utiliser le numéro d'index unique associé par FipsTag à chaque lexème (ou collocation) dans l'algorithme de comptage et de vérification. Nous avons principalement exploité le fait que les numéros d'index peuvent être utilisés comme indices de tableaux pour accélérer de façon significative le fonctionnement de l'algorithme.

3.2 Résultats de la génération

À ce stade du projet, nous avons généré automatiquement deux lexiques bilingues : (1) le lexique anglais-allemand sur la base de l'anglais-français et allemand-français et (2) le lexique anglais-italien sur la base de l'anglais-français et français-italien. Pour la validation des correspondances, nous avons utilisé le corpus parallèle des débats du Parlement européen pour la période 1996 à 2001, EuroParl Version 1 (Koehn, 2005). Le tableau ci-dessous résume les résultats :

	anglais-allemand	anglais-italien
correspondances candidates	56'640	49'686
correspondances non ambigües	30'259	28'322
obtenues avec une collocation comme pivot	2'220	1'778
correspondances validées par corpus	6'282	7'051
total des correspondances obtenues	**38'741**	**37'151**

La qualité des correspondances dérivés est très bonne, mais le nombre de correspondances que nous avons réussi à vérifier dans le corpus est en deçà de nos attentes : nous avons constaté

[2]La préférence est un entier de 0 à 6 (6 le maximum) qui permet d'ordonner les correspondances multiples pour un item donné.

queseulement 26% des correspondances générées ont pu être validées dans le corpus EuroParl pour l'anglais - en allemand et 36% pour l'anglais – italien. Les corpus EuroParl utilisés varient entre 17M et 19.6M mots selon la langue. Le temps nécessaire pour étiqueter le corpus avec FipsTag, le temps pour effectuer la requête SQL qui génère les correspondances par transitivité et le temps nécessaire pour vérifier les correspondances dans le corpus sont indiqués dans le tableau ci-dessous. Nous avons effectué ces mesures avec un processeur double cœur de 2,67 GHz. :

anglais	43.8h
allemand	37.7h
italien	26.3h

FIG. 2 – Temps requis pour l'étiquetage

	requête SQL	validation dans le corpus
anglais-allemand	0.5min	3.2min
anglais-italien	0.4min	3.0min

FIG. 3 – Temps requis pour la validation par corpus

Nous pouvons observer que l'étiquetage nécessite beaucoup de temps. Ce n'est toutefois pas trop pénalisant car l'étiquetage des corpus n'est effectué qu'une seule fois. Ce qui est intéressant, par contre, c'est que la requête dans la base de données et la validation des correspondances dans le corpus sont rapides, permettant ainsi grand nombre d'essais afin d'ajuster les paramètres de filtrage de la requête.

4 Conclusion

Le projet MulTra vise au développement d'un système de traduction multilingue obéissant au modèle classique analyse-transfert-génération sans succomber aux problèmes de complexité due à la prolifération des modules de transferts et à celle des dictionnaires bilingues par le choix d'une part d'un modèle linguistique abstrait favorisant les représentations génériques et d'autre part une architecture logicielle basée sur la notion d'objet, permettant elle aussi l'expression de processus génériques susceptibles d'être redéfinis (ou étendus) pour satisfaire les besoins spécifiques des langues particulières en ce qui concerne l'analyseur et le générateur, ou de paires de langues dans le cas du transfert. De même, un effort particulier a été entrepris pour réduire le travail considérable que représente l'élaboration des dictionnaires bilingues. La dérivation de correspondances bilingues par transitivité, validées grâce à un corpus bilingue, semble en mesure de réduire de près de 40% (et plus de 90% pour les expressions à mots multiples) la tâche des lexicographes. Le projet n'est pas suffisamment avancé pour que l'on puisse véritablement comparer les résultats de traduction avec des systèmes commerciaux, mais les premiers pas sont prometteurs. A l'heure actuelle, les paires suivantes ont été développées ou sont en développement (anglais-français, allemand-français, français-anglais, allemand-anglais, français-italien, italien-français, français-espagnol, espagnol-français), avec des dictionnaires bilingues de 20'000 (espagnol-français) à 75'000 entrées (anglais-français).

Remerciements

Cette recherche a été financée en partie par le Fonds national suisse de la recherche scientifique, subside no 100012-113864.

5 Bibliographie

Ahn, K. et M. Frampton, 2006. "Automatic Generation of Translation Dictionaries Using Intermediary Languages" in *Cross-Language knowledge Induction Workshop of the EACL-06*, 41-44.

Arnold, D., L. Balkyn, S. Meijer, R. Humphreys et L. Sadler, 1995. *Machine Translation : An Introductory Guide*, http ://www.essex.ac.uk/linguistics/clmt/MTbook/HTML/book.html

Boitet, C. 2001. "Four technical and organizational keys to handle more languages and improve quality (on demand) in MT" in *Proceedings of MT-Summit VIII*, 18-22.

Bresnan, J. 2001. *Lexical Functional Syntax*, Oxford, Blackwell.

Culicover, P. et R. Jackendoff, 2005. *Simpler Syntax*, Oxford, Oxford University Press.

Eberle, K. 2001. "FUDR-based MT, head switching and the lexicon" in *Proceedings of MT-Summit VIII*, 93-98.

Gollins, T. et M. Sanderson, 2001. "Improving cross language retrieval with triangulated translation", *Proceedings of the 24th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, 90-95.

Koehn, Philipp, 2005. "Europarl : A Parallel Corpus for Statistical Machine Translation", *MT Summit 2005*, 79-86.

Lu, W.H., L.F. Chien et H.J. Lee, 2004. "Anchor Text Mining for Translation of Web Queries : A Transitive Translation Approach", *ACM Transactions on Information Systems (TOIS)*, vol. 22.2, 242-269.

Ney, H. 2005. "One Decade of Statistical Machine Translation", *Proceedings of MT-Summit X*, 12-17.

Paik, K., S. Shirai et H. Nakaiwa, 2004. "Automatic Construction of a Transfer Dictionary Considering Directionality", *COLING 2004 Multilingual Linguistic Resources Workshop*, 25-32.

Seretan, V. et E. Wehrli, 2007. "Collocation translation based on sentence alignment and parsing", *Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2007)*, 401-410.

Wehrli, E. 1998. "Translating Idioms", *Proceedings of COLING-98*, Montreal, 1388-1392.

Wehrli, E. 2007. "Fips, a "Deep" Linguistic Multilingual Parser", *Deep Linguistic Processing Workshop*, ACL-2007, Prague, 120-127.