Pour une heuristique de la traduction automatique basée sur la traduction pédagogique. Session 6 - Didactique, enseignement, apprentissage

Charles Barone, Valeria Franzelli

TRALOGY

Pour une heuristique de la traduction automatique basée sur la traduction pédagogique

Charles Barone
Université de Pise (Italie)
barone@rom.unipi.it

Valeria Franzelli
Université de Pise (Italie)
v.franzelli@alice.it

TRALOGY II - Session 6
Date d'intervention : 18/01/2013

Nous présentons ici une recherche centrée sur la traduction du français vers l'italien dont l'objectif est double : il s'agit, d'une part, de tester l'hypothèse que l'apprentissage d'un système de traduction automatique peut se faire de manière progressive, selon des modalités comparables à celles de l'apprentissage humain d'une langue étrangère, d'autre part de comparer les erreurs produites par des étudiants au cours d'activités de traduction pédagogique avec les erreurs relevables en traduction automatique, afin de juger si l'identification des problèmes de la traduction humaine peuvent contribuer à une amélioration des résultats de la traduction automatique.

lien video :http://webcast.in2p3.fr/videos-pour_une_heuristique_de_la_traduction_automatique

Introduction

La recherche que nous présentons est centrée sur la traduction du français vers l'italien et s'est fixée deux objectifs : il s'agissait en premier lieu de tester l'hypothèse que l'apprentissage d'un système de traduction automatique peut se faire de manière progressive, selon des modalités comparables à celles de l'apprentissage humain d'une langue étrangère, en adaptant les contenus à ceux du Cadre Européen Commun de Référence pour les Langues ; en second lieu, notre tâche a consisté à dresser un inventaire des erreurs produites par des étudiants au cours d'activités de traduction pédagogique, afin d'orienter la traduction de la machine sur les difficultés les plus significatives à un stade donné de l'apprentissage, puis de comparer les résultats de notre programme de TA avec ceux des traductions fournies par nos étudiants, mais aussi avec ceux d'autres systèmes de TA. Nous focaliserons ici notre attention sur le premier niveau du Cadre Européen, le niveau A1.

1. Des contenus d'apprentissage progressifs

Les systèmes de traduction automatique statistique reposent sur l'utilisation de textes parallèles bilingues et c'est de l'accumulation des données dans ces deux corpus que l'on fait dépendre les performances de ces systèmes. Or, faute d'une collaboration suivie entre linguistes, traducteurs et experts de traitement automatique des langues, la quantité des textes prévaut sur leur qualité, d'où les résultats hasardeux auxquels aboutit souvent la traduction automatique. Dans la recherche que nous menons depuis près de deux ans à l'université de Pise, nous avons voulu expérimenter les performances de la machine avec une progression raisonnée des contenus, comme il advient depuis longtemps dans l'apprentissage d'une langue étrangère.

Cette progression en didactique des langues étrangères est aujourd'hui évaluée selon les niveaux du CECRL, un cadre qui devrait permettre à l'enseignant de disposer d'une ossature apparemment solide. Ainsi peut-on lire en introduction aux ouvrages de la collection *Activités pour le Cadre Européen Commun de Référence* (Parizet M.-L. et al. 2005 et s.) :

> Organisé en six niveaux qui vont de la découverte à la maîtrise de la langue-culture, le CECRL établit une progression plus réaliste et plus précise que l'habituelle distinction entre élémentaire, intermédiaire et avancé.

Ce système ambitieux n'est cependant pas sans poser des problèmes dès qu'on tente de s'y référer scrupuleusement, comme nous avons dû le faire pour répartir les éléments de morphologie et de morphosyntaxe en fonction de ces niveaux. Certes, les manuels pédagogiques nous mettent explicitement en garde :

> Attention [...] à bien comprendre l'utilisation qui doit être faite de ce référentiel : *les savoir-faire, actes de parole et contenus linguistiques préconisés pour chaque niveau correspondent aux éléments que l'étudiant doit acquérir à un moment donné de son apprentissage* [c'est l'auteur qui souligne]. Pour un niveau donné, il est donc bien évidemment possible de proposer un document contenant des éléments plus complexes que ce qui est décrit dans le référentiel. Un document doit permettre à l'apprenant d'acquérir des savoirs et savoir-faire correspondant à son niveau de compétence, ce qui ne signifie en aucun cas que le document en question doive se réduire aux contenus préconisés [...]. (Chauvet 2008)

Toujours est-il que, à bien y regarder, des questions essentielles demeurent :

> Le CECR indique clairement la nécessité d'une progression [...] mais les rédacteurs du CECR n'entrent jamais dans le détail, ne font aucune proposition concrète relative aux progressions [...]. (Robert 2008 : 177)

Pour nous en tenir à un exemple au niveau A1, l'emploi de l'imparfait est ainsi présenté chez Beacco (2007 : 97) :

> Au niveau A1, l'apprenant/utilisateur est capable [...] d'utiliser des verbes au présent de l'indicatif [...] et, éventuellement à l'imparfait, exclusivement pour un nombre très limité de verbes (*il/c'était, il (y) avait, il faisait... : c'était bien, il faisait beau*), qui permettent de construire un discours narratif.

C'est donc selon les niveaux du Cadre Européen qu'ont été soumis à nos étudiants les phrases et brefs récits en français qu'ils devaient traduire en italien et c'est selon ces mêmes niveaux qu'a été structuré le corpus destiné à l'apprentissage de notre programme de traduction automatique. À devoir choisir entre un système à base de règles ou hybride et un système statistique, nous avons opté pour ce dernier, qui connaît aujourd'hui un certain succès en traduction automatique, dès lors que la paire de langues considérées appartient à la même famille de langues, voire quand ce sont deux langues proches (Koehn 2005). Tel est le cas pour le français et l'italien. Le système que nous avons adopté est un produit dérivé de la plateforme logicielle libre Moses (http ://www.statmt.org/moses) : *DoMy Pro*[1], qui, par une séquence de routines (« graphes ») sous système d'exploitation Ubuntu, permet à l'enseignant comme au traducteur de bénéficier d'une approche plus conviviale de Moses, à tout le moins de ne pas recourir à l'informaticien expert en traitement automatique des langues.

Notre modèle de traduction est ainsi organisé en six répertoires, correspondant chacun à un niveau du Cadre, du A1 au C2, avec ses contenus spécifiques, qui englobent à mesure ceux du répertoire/niveau précédent. Dans sa version actuelle, le programme doit être modifié dans ses paramètres pour son « entraînement », selon le niveau/répertoire à tester ; nous prévoyons à terme de créer un réseau local composé de six postes de travail, correspondant chacun à un niveau du Cadre Européen, afin d'éviter la réécriture des paramètres au sein de chaque « graphe ».

Le corpus parallèle est constitué de phrases en français alignées manuellement avec leur traduction italienne ; retranscrites à partir de divers manuels de français langue étrangère et traduites, en tant qu'exercices de traduction pédagogique, par nos étudiants de français à l'université (des apprenants débutants de Licence aux étudiants de Master 2 en traduction), elles ont fait l'objet d'un contrôle systématique avant leur insertion dans le corpus : car si la qualité de l'alignement est essentielle pour obtenir de bons résultats en TA (Och & Ney 2000 ; Lambert et alii 2006), l'intervention du linguiste l'est elle aussi :

> Theoretically, when using SMT, no linguistic knowledge is required. In practice, once the system is built, linguistic knowledge becomes necessary to achieve perfect translations at all grammatical levels […] (Farrús et al. 2011)

L'entraînement du modèle de traduction est également assuré par les contenus de divers romans et pièces de théâtre, alignés eux aussi au niveau de la phrase, en langue originale et dans leur traduction italienne publiée ; pour chaque répertoire, le choix des œuvres reflète le degré de difficulté du niveau correspondant du Cadre Européen.

Le modèle de langue de la machine, sa langue cible, correspond à la langue maternelle de la grande majorité de nos étudiants, l'italien ; le corpus est ici constitué des ressources textuelles les plus diverses, soigneusement sélectionnées dans la littérature contemporaine et la presse italiennes.

Pour le modèle de traduction, après de nombreux essais, le système a pris en compte les *n*-grammes jusqu'à $n = 6$, le modèle de langue ceux jusqu'à $n = 5$; dans nos derniers essais sur le niveau A1, les *n*-grammes à 5, pour les deux modèles, semblent fournir de meilleurs résultats.

(1) http://www.precisiontranslationtools.com : en novembre 2012, le produit a évolué dans la version DoMT Desktop.

2. Le relevé des erreurs dans la traduction pédagogique

Nous avons tout d'abord exploité la distinction entre « faute » et « erreur » (Corder 1980), qui a permis de diviser en deux groupes les phénomènes repérés dans les travaux des étudiants : d'un côté les « erreurs », ou phénomènes systématiques, dictées par un manque de compétences morphosyntaxiques, stylistiques et lexicales ; de l'autre les « fautes », ou erreurs non systématiques, qui sont donc involontaires, telles que les coquilles et souvent les omissions, et que l'auteur pourrait corriger de manière autonome. Les conditions de travail ont été les facteurs principaux considérés pour cette première distinction : les travaux rédigés devant l'ordinateur et à la maison présentent en effet des coquilles (de l'inversion des lettres à l'absence d'espace de ponctuation) que les épreuves d'examen, écrites à la main et en classe, ne présentent pas ; ces dernières comportent, en revanche, des omissions (de mots ou même de phrases) qui sont en général absentes dans les premiers, sans doute en raison d'un manque de temps pour la relecture finale.

Nous avons ensuite focalisé notre attention sur les « erreurs », en essayant de les classer non seulement en fonction de l'unité de traduction qui était à leur origine (unité grammaticale, stylistique ou lexicale), mais aussi en fonction de la langue. Nous avons pu observer, par exemple, que, comme les étudiants maîtrisent la langue d'arrivée, ils peuvent aisément reformuler les énoncés, et que pour cela il devient difficile d'associer certaines erreurs grammaticales ou lexicales à une unité précise du texte de départ, mais plutôt au sens général reproduit dans le texte d'arrivée. La catégorie « erreur lexicale » a donc été réajustée en « erreur lexico-sémantique française » : l'étudiant n'a pas compris le sens général de l'énoncé, soit à cause de sa formulation en langue de départ soit à cause d'un lexème en particulier contenu dans le texte de départ. Il a été en outre parfois difficile de comprendre si certaines erreurs relèvent d'un manque de compétence de l'étudiant en italien ou bien en français, par exemple dans la traduction du langage soutenu : l'apprenant ne sait-il pas reproduire la variation en italien ou bien n'a-t-il pas reconnu la variation en français ?

Grâce à la comparaison des erreurs repérées dans un premier corpus constitué de 150 travaux, nous avons pu néanmoins structurer une grille de classement définitive présentant les étiquettes suivantes :

- grammaire française, catégorie qui regroupe les erreurs dictées par un manque de compétences grammaticales en langue de départ (temps verbaux, prépositions, adverbes, accents grammaticaux, syntaxe, etc.) ;
- grammaire italienne, pour les erreurs dues à des lacunes grammaticales en langue d'arrivée (temps verbaux, prépositions, adverbes, orthographe, syntaxe, etc.), mais aussi pour des mots ou des expressions n'ayant aucun sens en italien ;
- lexico-sémantique française qui concerne toutes les erreurs relevant d'une interprétation erronée du sens d'un mot, d'une locution ou d'une phrase en langue de départ ;
- stylistique italienne, où l'on inclut toutes les erreurs de connotation, registre et ponctuation en langue d'arrivée ;
- fautes, pour toute omission ou coquille non intentionnelles, que l'auteur pourrait corriger de manière autonome par une relecture attentive de son travail.

3. Application au niveau A1

Quant aux phénomènes les plus représentatifs de notre « petit musée des erreurs » au niveau débutant, la comparaison a été effectuée avec notre programme de TA, ainsi qu'avec d'autres systèmes : pour représenter ceux à base de règles, nous avons recouru à la version 7 d'*Idiomax*

(http ://www.idiomax.com), pour les hybrides à *Systran* (http ://www.systranet.com/systranet-services/translate) et pour les statistiques au traducteur de *Google* (http ://translate.google.fr), ces deux derniers dans leur version gratuite en ligne.

La traduction italienne de *assez* par *assai* chez les apprenants italophones est bien connue dans la littérature relative à l'enseignement du FLE en Italie, son occurrence est de 47 % dans le corpus analysé, à savoir 21 devoirs effectués en classe par des étudiants débutants[2] :

Texte source	Ce soir elle va rentrer à la maison **assez** tard.
Texte cible	Stasera tornerà a casa **abbastanza** tardi.
Erreur humaine	Stasera tornerà a casa **molto/assai** tardi.
TA règles	Questa sera lei va riporre alla casa **abbastanza** tardi.
TA hybride	Questa sera rientrerà alla casa **abbastanza** tardi.
TA stat.	Stasera andrà a casa **piuttosto** tardi.
Moses (DoMy)	Questa sera tornerà a casa **abbastanza** tardi.

Nous observerons que la traduction de Google avec *piuttosto* est des plus acceptables dans ce contexte.

Un problème qui se pose pour l'apprenant italien, qu'il s'agisse d'une erreur véritable ou d'une faute de distraction, mais aussi pour la machine, est celui de la traduction du *vous* de politesse : dans 62 % des cas, les étudiants opèrent un calque du français en conjuguant le verbe à la deuxième personne du pluriel au lieu de la troisième personne du singulier (avec implicitation du *Lei*) :

Texte source	Vous avez choisi, Madame ?
Texte cible	Ha scelto, Signora ?
Erreur humaine	Avete scelto, signora ?
TA règles	Voi avete scelto, signora ?
TA hybride	Avete scelto, signora ?
TA stat.	Hai selezionato, signora ?
Moses (DoMy)	Ha scelto, Signora ?

Même si le phénomène est répandu dans l'italien parlé contemporain, c'est également la faute de distraction qui justifie l'emploi d'un pronom personnel complément inapproprié dans la traduction du *lui* français quand celui-ci représente un nom féminin ; on relève cette faute dans 20 % des copies d'étudiants :

Texte source	Elle parle à David, mais lui, il ne **lui** parle pas.
Texte cible	Parla a David, ma lui non **le** parla.
Erreur humaine	Parla a David, ma lui non **gli** parla.
TA règles	Lei parla a David, ma gli, egli non **gli** parla.
TA hybride	Parla a David, ma lui, non **gli** parla.
TA stat.	Parla con Davide, ma lui, lui non parla.
Moses (DoMy)	Parla con David, ma lui non **le** parla.

(2) Précisons que les étudiants débutants en français sont peu nombreux dans nos formations en langues étrangères à l'université de Pise.

En dernier exemple, relatif au domaine du lexique, citons celui de *légumes*, bien connu lui aussi des enseignants de français en Italie, que les apprenants associent, dans 47 % des cas, à *legumi* :

Texte source	La cantine propose du poisson avec du riz ou **des légumes**.
Texte cible	La mensa propone pesce con riso o **verdura**.
Erreur humaine	La mensa propone pesce con riso o **legumi**.
TA règles	La mensa fa proponimenti del pesce con del riso o delle **verdure**.
TA hybride	La mensa propone pesce con riso o **verdure**.
TA stat.	La mensa offre pesce con riso o **verdure**.
Moses (DoMy)	La mensa con il pesce propose il riso o delle **verdure**.

Conclusion

Ce ne sont certes pas des résultats obtenus au niveau A1, et avec un corpus aussi restreint, qui nous permettront d'aboutir ici à quelque affirmation que ce soit. Ces quelques exemples suffisent néanmoins à illustrer comment, par le biais d'une activité telle que la traduction pédagogique vers l'italien, qui évalue, en termes de didactique, la compétence passive en français de l'apprenant, sa L2, il est possible de concevoir une heuristique de la TA orientée, quant à elle, sur une compétence active en italien, langue cible de la machine.

Bibliographie

Beacco, Jean-Claude et Porquier, Rémy, (2007), *Niveau A1 pour le français. Un référentiel + CD audio*, Éd. Didier

Chauvet, Aude (2008), *Référentiel pour le Cadre Européen Commun*. *A1-A2-B1-B2-C1-C2*, Alliance Française/CLE International

Corder, S. Pit (1980), « Que signifient les erreurs des apprenants ? », in *Langages*, 57, pp. 9-15

Farrús, Mireia, Costa-Jussà, Marta R., Mariño, José B., Poch, Marc, Hernández, Adolfo, Henríquez, Carlos., Fonollosa, José A. R. (2011), "Overcoming statistical machine translation limitations: error analysis and proposed solutions for the Catalan-Spanish language pair", in *Language Resources and Evaluation*, 45, 2, pp. 181-208

Koehn, Philipp (2005), "Europarl: A parallel Corpus for Statistical Machine Translation", in *Proceedings of MT Summit X*, pp. 79-86

Lambert, Patrick, De Gispert, Adrià, Banchs, Rafael E., Mariño, José B. (2005), "Guidelines for Word Alignment Evaluation and Manual Alignment", in *Language Resources and Evaluation*, 39, pp. 267-285

Och, Franz-Josef, Ney, Hermann (2000), "A Comparison of Alignment Models for Statistical Machine Translation", in *Proceedings of the 18th Int. Conf. on Computational Linguistics*, Saarbrücken (Germany), pp. 1086-1090

Parizet, Marie-Louise, Grandet, Éliane, Corsain, Martine (2005), *Activités pour le Cadre Européen Commun De Référence Niveau A1*, Cle International, « Activités Pour Le Cadre Commun »

Parizet, Marie-Louise (2005), *Activités pour le Cadre Européen Commun De Référence Niveau A2*, Cle International, « Activités Pour Le Cadre Commun »

Parizet, Marie-Louise (2006), *Activités pour le Cadre Européen Commun De Référence Niveau B1*, Cle International, « Activités Pour Le Cadre Commun »

Robert, Jean-Pierre (2008) [2002[1]], *Dictionnaire pratique de didactique du FLE*, Paris, Ophrys, « L'Essentiel français »