Mesures de confiance en traduction automatique

Proposition de thèse en collaboration avec Lingua Custodia
Draft - Mai 2013

François Yvon et Guillaume Wisniewski
{yvon,wisniews}@limsi.fr

7 juin 2013

1 Éléments de Contexte

Les récents progrès des systèmes de traduction automatique (TA), imputables, pour partie, au développement des méthodes d'exploitation statistique de grands corpus parallèles [9], les rendent aujourd'hui à même d'apporter une aide réelle, y compris pour des services de traduction professionnels. C'est particulièrement le cas lorsque les documents à traduire appartiennent à un domaine précis (documents financiers, textes sur le sport, ...). Le scénario le plus répandu dans ce contexte consiste à commencer par traduire automatiquement les textes. Ces traductions sont ensuite « post-éditées », c'est-à-dire révisées par des traducteurs ou des correcteurs professionnels.

La post-édition permet de réduire le temps nécessaire à la traduction [10] et ce champ se développe de manière très rapide, aussi bien comme problématique de recherche (des ateliers sur la question ont été organisés dans le cadre des principales conférences de TA, AMTA et MT Summit), que comme activité industrielle. C'est, par exemple, l'approche retenue par Lingua Custodia, qui est partenaire de ce projet et entreprise d'accueil pour la thèse, pour garantir la qualité des traductions. Lingua Custodia est une startup fondée en septembre 2011 qui développe des moteurs de traductions spécialisés dans le langage de la finance, destinés en particulier à traduire des documentations de fonds d'investissement. Pour ces documents, si les traductions sont répétives, elles se doivent impérativement être révisée par des éditeurs experts ; le domaine se prête bien au cadre général de post-édition.

1.1 Mesurer la confiance

Une fonctionnalité importante dans le contexte de la post-édition est la capacité d'un système de traduction à s'auto-évaluer, c'est-à-dire à délivrer, en plus des traductions automatiques, des éléments d'appréciation quantitatives qui rendraient compte de la confiance qu'a le système dans sa sortie. De telles informations pourraient être utilisées par des post-éditeurs pour organiser et

faciliter leur travail de révision (par exemple en identifiant les parties de la traduction posant problème) ou pour concevoir des scénarios de traduction interactive [8]. Au fur et à mesure de l'amélioration des traductions automatiques, la problématique de l'*estimation de confiance* a ainsi progressivement émergé comme une question de première importance pour les applications opérationnelles.

La littérature scientifique sur la question de l'estimation de confiance montre que ces travaux se caractérisent par un certain nombre de partis-pris :

- la confiance est le plus souvent mesurée au niveau des phrases [7, 14] : il s'agit alors de fournir un score pour chacune des hypothèses de traduction caractérisant la qualité de celle-ci. On note pourtant quelques travaux visant à l'évaluer au niveau des mots [3, 2, 11] ou des segments [3, 5], tâche manifestement bien plus délicate. Encore plus rares sont les travaux s'intéressant à l'estimation de confiance au niveau d'un document entier [12].
- le système de traduction est le plus souvent considéré comme une boîte noire, ce qui permet de découpler presqu'entièrement la génération des hypothèses et l'évaluation de leur qualité. Cette manière de procéder facilite la comparaison de diverses manières de prédire la confiance [4], mais prive l'estimation de confiance d'informations importantes.
- la qualité qu'il s'agit de prédire est rarement bien caractérisée. Selon les études, il peut s'agir de prédire des mesures automatique de qualité telles que BLEU [], ou bien des jugements de qualités émis par des humains, voire des mesures plus directes de l'*utilisabilité* des traductions comme par exemple la difficulté apparente de la post-édition [4], voire le temps nécessaire à corriger la phrase [13].
- ainsi formulée, l'estimation de confiance se modélise très directement sous la forme d'un problème d'apprentissage supervisé : partant d'une base de phrases en langue source $(s_i)_{i=1...N}$, dont les traductions $(t_i)_{i=1...N}$ ont été évaluées sous la forme de grandeurs numériques discrètes ou continues $(e_i)_{i=1...N}$, il s'agit de construire une représentation $F(s,t)$ du couple source-cible, puis une fonction de prédiction $h(F(s,t)$ qui reproduise au mieux (au sens de l'erreur moyenne) les exemples d'apprentissage disponibles. Selon la nature des scores e, ce programme correspond à un programme de classification supervisée ou de régression [15].
- la question qui a alors été le plus étudiée [14, 1, 6] dans ce contexte est celle du choix d'une représentation idoine $F(s,t)$. Cette représentation inclut naturellement des descripteurs numériques qui sont utilisés par le système de traduction (par exemple le score du modèle de traduction, ou encore celui du modèle de langue). Pour être efficace elle doit également inclure des descripteurs qui ne sont pas accessibles lors de la traduction (par exemple la bonne formation syntaxique ou la cohérence sémantique globales de la phrase cible). Par comparaison, le choix d'une fonction de prédiction s'avère relativement secondaire, tant que l'on s'en tient à des outils d'apprentissage ayant fait leurs preuves (séparateurs à large marge, arbres de classification ou de régression, etc.)

2 Objectifs scientifiques de la thèse

Ce travail de thèse porte sur la question de l'estimation de confiance. Comme il se déroule dans un cadre opérationnel de développement de systèmes de traduction automatique dans le domaine financier, il semble approprié de s'intéresser à certaines questions qui sont aujourd'hui peu étudiées :

- l'utilisation de descripteurs endogènes, c'est-à-dire dérivés de caractéristiques observables du système de traduction ;
- l'utilisation de descripteurs externes de « haut-niveau », c'est-à-dire reposant sur des analyses profondes des traductions et mobilisant des ressources riches liées à la connaissance fine du domaine (en particulier terminologiques ou ontologiques) ;
- la prédiction de la confiance au niveau de tout le document.

Plus fondamentalement, une autre question importante qui sera étudiée est celle de la définition de la qualité. On pourra pour ce faire envisager d'expérimenter avec plusieurs manières de concevoir et de noter la qualité des traductions pour étudier lesquelles sont les plus stables (c'est-à-dire correspondent à des jugements consensuels), les plus utiles (pour l'application de révision), ou les plus faciles à prédire.

2.1 Utilisation de descripteurs endogènes

L'estimation de confiance se fonde sur une représentation du couple (phrase source, hypothèse de traduction) qui inclut des descripteurs multiples. Certains sont disponibles durant la traduction et font partie des scores utilisés pour construire la meilleure hypothèse de traduction ; d'autres ne peuvent être calculés qu'une fois la traduction intégralement construite. En revanche, de nombreuses informations relatives au système ne sont pas utiles pour la traduction, et sont le plus souvent ignorées ou inconnues lorsqu'il s'agit de calculer la confiance. Or ces informations peuvent s'avérer cruciales, et renforcer (ou au contraire affaiblir) la validité des choix opérés par le système de traduction.

Ainsi, connaître le nombre d'exemples attestant une association bilingue peut s'avérer très utile : si cette traduction est la seule dont peut se servir le système, elle sera choisie avec probabilité 1 ; pourtant il ne semble pas indifférent pour comparer les traductions de phrases différentes de savoir si cette estimation de probabilité repose sur une seule, ou bien sur des centaines d'observations. La grandeur statistique qui permet de formaliser cette intuition est celle de la variance de l'estimation des scores qui valuent les associations bilingues qu'exploite le modèle de traduction. Une autre information importante, est la similarité entre les exemples d'apprentissage qui servent à extraire les associations bilingues et la traduction à produire.

Dans cet axe, on s'efforcera donc de construire et d'étudier des indicateurs statistiques qui relient la qualité de l'estimation et la qualité des traductions qu'ils permettent de calculer.

2.2 Utilisation de descripteurs riches

Une démarche très commune consiste à essayer de mettre en relation les résultats d'une analyse linguistique poussée de l'hypothèse en langue cible avec sa fluidité, sous l'hypothèse qu'une phrase grammaticale serait plus fluide d'une phrase qui ne le serait pas. Les analyses linguistiques envisageables sont nombreuses : analyse syntaxique, identification des entités nommées, des relations entre entités, ou bien encore des implications logiques entre clauses.

Les expériences décrites dans la littérature ont pour l'instant produit des résultats plutôt décevants en la matière [1, 6] et ces descripteurs n'améliorent pas les performances des systèmes n'utilisant que des descripteurs de surface. À cet échec relatif, il est possible d'avancer deux causes :

- les analyseurs linguistiques utilisés sont le plus souvent des analyseurs robustes, développés pour la « langue générale » (exemplairement l'analyse de dépêches de presse), et peinent à fournir des analyses correctes y compris pour des phrases grammaticales d'un domaine particulier (par exemple par manque de couverture lexicale, ou par méconnaissance de la terminologie ou de la phraséologie) ;
- les analyseurs utilisés sont développés (appris) dans le but de traiter des énoncés humains, c'est-à-dire dans l'ensemble plutôt corrects ; leur utilisation sur les sorties de systèmes de traduction, qui peuvent comprendre de nombreuses déviations par rapport à la norme, donne des résultats souvent imprévisibles. Exprimé sous un terme statistique, la difficulté est que les modèles statistique de grammaticalité (ou de sémanticité) sont appris sur des corpus qui sont très différents des sorties réelles des systèmes de traduction automatique ; les modèles qu'ils estiment sont donc parfois peu pertinentes pour jauger de la qualité plus ou moins médiocre des hypothèses de traduction.

Deux axes d'études sont envisagés qui correspondent d'une part à l'étude de l'intérêt d'analyses profondes utilisant des analyseurs linguistiques spécialisés et construits à partir de ressources linguistiques « riches » ; d'autre part au réexamen de l'aporie décrite ci-dessus. Notons pour finir que ces problèmes sont d'une certaine manière reliés et proviennent au fond d'un écart entre les données d'apprentissage et les données de test. Cette question difficile a fait l'objet de nombreuses études en traitement des langues, en particulier dans le cadre de l'adaptation au domaine ou au registre. La question est ici un peu différente et se rattache d'une certaine manière à celle de la robustesse des analyseurs.

2.3 Confiance au niveau document

Comme expliqué ci-dessus, dans les méthodes de l'art, la confiance est le plus souvent calculée au niveau des mots ou des phrases, à l'exception notoire de [12]. Les auteurs de ce dernier travail se posent en effet la question d'ordonner par qualité des documents automatiquement traduits. Comme pour la prédiction de la confiance au niveau des phrases, ce problème est résolu en associant des outils de calculs d'une représentation (ici très surfacique) du texte et de sa traduction

avec des méthodes d'apprentissage artificiel.

Plusieurs perspectives seront étudiées pour prolonger ce travail :

- prendre en compte des indicateurs linguistiques plus riches portant sur le niveau du document, en particulier en ajoutant des traits relatifs au suivi des références ou à l'organisation du discours. D'un point de vue pratique, cette orientation ne requiert en principe que la disponibilité des moyens idoines de calculer ces représentations.
- modéliser le caractère non-additif des mesures de confiance dans une optique de post-édition : corriger un segment qui est mal traduit « de manière systématique » peut en fait être réalisé de manière très simple par des fonctions de rechercher/remplacer. À l'inverse des erreurs plus aléatoires pourront entraîner un effort plus grand de post-édition.

Pour ces différents travaux, on notera que le projet pourra bénéficier de plusieurs facteurs facilitateurs : l'intégration du doctorant au sein de l'équipe de développement des systèmes de traduction ; la disposition de traductions post-éditées qui seront accumulées tout au long du projet en interaction avec des traducteurs ou des éditeurs professionnels ; des ressources dictionnairiques ou terminologiques pourront également être utilisées.

3 Déroulement de la thèse

Les travaux se dérouleront à parité dans les locaux du LIMSI et dans ceux de Lingua Custodia. Le plan de travail prévu est le suivant.

Année 1 :

- prise en main des outils de traduction automatique de la société ;
- rédaction d'un état de l'art sur l'estimation de confiance en traduction automatique ;
- spécification, développement et documentation d'une chaine de traitement linguistique pour les textes à base de composants open-source ;
- développement d'un système de base pour l'estimation de confiance.

Année 2

- formalisation du problème de l'estimation de confiance ;
- modélisation et étude de descripteurs endogènes ;
- amélioration de la chaine de traitement, extraction de descripteurs plus riches, au niveau phrase et au niveau document ;
- étude de l'impact de ces nouveaux descripteurs.

Année 3

- mesures de confiance au niveau du document ;
- rédaction et défense du mémoire de thèse.

Références

[1] Eleftherios Avramidis. Quality estimation for machine translation output using linguistic analysis and decoding features. In *Proceedings of the Se-*

venth Workshop on Statistical Machine Translation, WMT '12, pages 84–90, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[2] Nguyen Bach, Fei Huang, and Yaser Al-Onaizan. Goodness : A method for measuring machine translation confidence. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, pages 211–219, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[3] John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. Confidence estimation for machine translation. In *Proceedings of Coling 2004*, pages 315–321, Geneva, Switzerland, 2004.

[4] Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada, June 2012. Association for Computational Linguistics.

[5] Adrià de Gispert, Graeme Blackwood, Gonzalo Iglesias, and William Byrne. N-gram posterior probability confidence measures for statistical machine translation : an empirical study. *Machine Translation*, pages 1–30, 2012.

[6] Mariano Felice and Lucia Specia. Linguistic features for quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 96–103, Montréal, Canada, June 2012. Association for Computational Linguistics.

[7] Michael Gamon, Anthony Aue, and Martine Smets. Sentence-level mt evaluation without reference translations : Beyond language modeling. In *Proceedings of EAMT*, pages 103–111, 2005.

[8] Jesús González Rubio, Daniel Ortiz Martínez, and Francisco Casacuberta. Balancing user effort and translation error in interactive machine translation via confidence measures. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 173–177, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[9] Philipp Koehn. *Statistical Machine Translation*. 2010.

[10] François Masselot Mirko Plitt. A productivity test of statistical machine translation post-editing in a typical localisation contex. *The Prague Bulletin of Mathematical Linguistics*, (93) :7–16, 2010.

[11] Sylvain Raybaud, David Langlois, and Kamel Smaïli. "this sentence is wrong." detecting errors in machine-translated sentences. *Machine Translation*, 25(1) :1–34, 2011.

[12] Radu Soricut and Abdessamad Echihabi. Trustrank : Inducing trust in automatic translations via ranking. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 612–621, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[13] Lucia Specia. Exploiting objective annotations for measuring translation post-editing effort. In *Proceedings of the 15th conference of EAMT*, pages 73–80, Leuven, Belgium, 2011.

[14] Lucia Specia, Craig Saunders, Marco Turchi, Zhuoran Wang, and John Shawe-Taylor. Improving the confidence of machine translation quality estimates. In *MT Summit XII : proceedings of the twelfth Machine Translation Summi*, pages 136–143, Ottawa, Ontario, Canada, 2009.

[15] Yong Zhuang, Guillaume Wisniewski, and François Yvon. Non-linear models for confidence estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 157–162, Montréal, Canada, June 2012. Association for Computational Linguistics.