

Développement à la demande des systèmes de traduction automatique statistiques

Li Gong

La traduction automatique statistique produit des résultats qui en font un choix privilégié dans la plupart des scénarios de traduction assistée par ordinateur. Cependant, le développement de tels systèmes à haute performance implique des traitements très coûteux sur des données à grande échelle. De nouvelles données sont continuellement disponibles, alors que les systèmes construits de manière standard sont typiquement statiques, ce qui rend l'utilisation de nouvelles données coûteuse car les systèmes doivent être reconstruits en intégralité. En outre, le processus d'adaptation des systèmes de traduction est généralement fondé sur un corpus de développement et est effectué une fois pour toutes. Dans cette thèse, nous proposons un cadre informatique pour répondre à ces trois problèmes conjointement. Ce cadre permet de développer des systèmes de traduction à la demande avec des mises à jour incrémentales et permet d'adapter les systèmes construits à chaque nouveau texte à traduire.

La première contribution importante de cette thèse concerne une nouvelle méthode d'alignement sous-phrastique qui peut aligner des paires de phrases en isolation. Cette propriété permet aux systèmes de traduction de calculer des informations à la demande afin d'intégrer de façon transparente de nouvelles données disponibles sans re-entraînement complet des systèmes. La deuxième contribution importante consiste en l'intégration de stratégies d'échantillonnage contextuel pour sélectionner des exemples de traduction à partir de corpus à grande échelle sur la base de leur similarité avec le texte à traduire afin d'obtenir des tables de traduction adaptées.

Chapitre 1

Dans le première chapitre, nous donnons un aperçu de la traduction automatique statistique, qui se concentre sur la traduction automatique statistique fondée sur les segments.

Dans la traduction automatique statistique, la génération de traductions est abordée comme un problème de recherche. Étant donné une phrase *source* à traduire $\mathbf{s} = s_1^I = s_1 \dots s_I$, le but est de trouver l'hypothèse de traduction la plus probable $\mathbf{t} = t_1^J = t_1 \dots t_J$, où I et J représentent les longueurs de la phrase source et de la phrase cible. La meilleure traduction est obtenue en maximisant $P(\mathbf{t}|\mathbf{s})$ et en appliquant le modèle du canal bruité [Brown et al., 1990], défini dans l'Équation (1) :

$$\begin{aligned} \mathbf{t}_{best} &= \operatorname{argmax}_{J, t_1^J} p(t_1^J | s_1^I) \\ &= \operatorname{argmax}_{J, t_1^J} p(s_1^I | t_1^J) p(t_1^J) \end{aligned} \tag{1}$$

où $p(s_1^I | t_1^J)$ correspond au modèle de traduction et $p(t_1^J)$ au modèle de langage.

Le modèle de traduction $p(s_1^I | t_1^J)$ est souvent décomposé en un ensemble de modèles, comprenant les probabilités de traduction, des *pondérations lexicales* et des probabilités de *distorsion*. Ces modèles sont estimés sur la base des alignements mot-à-mot qui sont calculés à l'aide des modèles IBM [Brown et al., 1993] avec l'algorithme espérance-maximisation (EM) [Dempster et al., 1977]. Au lieu d'utiliser les mots comme unités de traduction, on utilise désormais les *segments* (groupes de mots continus) pour estimer les modèles de traduction, ce qui procure les avantages suivants :

- en utilisant des segments, un système peut prendre en compte des situations dans lesquelles les mots ne correspondent pas à de bonnes unités de traduction ;
- l'utilisation de segments peut réduire l'ambiguïté lexicale lors de la traduction ;
- les segments peuvent contenir des réorganisations dans l'ordre des mots de la traduction.

Un modèle de langage $p(t_1^J)$ est utilisé pour mesurer la probabilité des hypothèses en langue *cible* produites par le système de traduction automatique statistique. Celui-ci fournit une estimation de la probabilité pour toutes les

séquences de mots possibles, et est défini comme :

$$p(t_1^J) = p(t_1)p(t_2|t_1) \dots p(t_J|t_1, t_2 \dots t_{J-1}) \quad (2)$$

La probabilité donnée par le modèle de langage $p(t_1^J)$ est obtenue en calculant le produit des probabilités d'un mot étant donné l'historique des mots précédents.

Un autre aspect important concerne l'estimation de la qualité des traductions produites. En pratique, on utilise souvent des métriques d'évaluation automatique, BLEU (BiLingual Evaluation Understudy [Papineni et al., 2002]) étant l'une des plus utilisées. Celle-ci compare une hypothèse de traduction avec une ou plusieurs traductions de *référence* en comptant les n -grammes en commun. Formellement, la métrique est définie comme :

$$\text{BLEU} = \text{brevity-penalty} \times \exp \sum_{n=1}^N \lambda_n \log \text{precision}_n \quad (3)$$
$$\text{brevity-penalty} = \min\left(1, \frac{\text{longueur de candidat}}{\text{longueur de référence}}\right)$$

où precision_n est la précision n -gramme pour une valeur de n donnée; la *pénalité de concision* (brevity-penalty) est utilisée pour pénaliser des situations où l'hypothèse de traduction est plus courte que la traduction de référence.

D'autres métriques sont également régulièrement utilisées pour évaluer la qualité des traductions, telles que TER (Translation Error Rate) [Snover et al., 2006], NIST [Dodington, 2002], ou METEOR (Metric for Evaluation of Translation with Explicit Ordering) [Banerjee and Lavie, 2005, Lavie and Agarwal, 2007].

Chapitre 2

Dans le deuxième chapitre, nous présentons les trois problèmes principaux qui concernent les approches à l'état de l'art de développement de systèmes de traduction automatique statistique.

Premièrement, le développement de systèmes de traduction s'avère très coûteux en calculs. Pour construire des systèmes performants, il est souvent nécessaire d'utiliser de très grandes quantités de données, ce qui requiert

un temps de pré-traitement important. De plus, les modèles obtenus occupent une place de stockage importante, ce qui nécessite fréquemment un pré-filtrage pour les utiliser.

Deuxièmement, une fois un système construit, celui-ci est statique. Or, de nouvelles données deviennent continuellement disponibles. Les intégrer au système est alors difficile et requiert typiquement une reconstruction intégrale du système avec l'ensemble des données.

Troisièmement, dans un système de traduction à l'état de l'art, aucune distinction n'est faite entre parties des données d'apprentissage. Ces parties correspondent toutefois à des origines différentes, et une certaine distinction entre données utiles et non appropriées s'avère nécessaire.

Il existe déjà des méthodes et des pistes de recherche pour améliorer la performance du système de traduction relativement à ces trois problèmes.

Concernant le premier problème, les systèmes peuvent extraire des tables de traduction spécialisées pour chaque texte à traduire, permettant d'utiliser de grands corpus ainsi que des longs segments. Une sélection par échantillonnage des données permet de réduire davantage les calculs nécessaires par la construction des tables de traduction.

Concernant le second problème, des méthodes d'alignement incrémental et de combinaison de modèles permettent d'intégrer progressivement de nouvelles données disponibles dans des systèmes existants. Cependant, les approches d'alignement incrémental existantes nécessitent de pré-entraîner des modèles d'alignement puis d'attendre des quantités de données supplémentaires suffisantes. L'entraînement d'un bon modèle d'alignement, généralement entraîné sur une grande quantité de données, est en soi déjà très coûteux.

Concernant enfin le troisième problème, différentes approches ont été décrites pour l'utilisation sélective de données d'entraînement. Par exemple, la sélection de données et la pondération d'exemples permettent d'adapter des modèles de traduction. Mais cette adaptation doit dépendre de chaque tâche de traduction particulière, et requiert la disponibilité d'un corpus de développement approprié. La qualité de l'adaptation obtenue dépend alors de la similarité entre le corpus de développement et les futurs textes à traduire.

Chapitre 3

Dans le chapitre précédent, nous avons présenté les trois problèmes principaux qui touchent les approches à l'état de l'art pour le développement

des systèmes de traduction automatique statistique. Dans cette thèse, nous allons répondre à ces trois problèmes de façon conjointe. Nous introduisons dans le Chapitre 3 un nouveau cadre informatique pour le développement de systèmes de traduction qui permet le développement de systèmes à partir de très grands corpus de manière très efficace. L'ensemble des informations sont calculées *au besoin* ; l'intégration de nouvelles données est très simple ; et il devient possible d'adapter facilement les modèles de traduction aux textes à traduire.

Notre approche de développement de systèmes à la demande repose sur deux éléments principaux : l'alignement de mots à la demande, et l'estimation à la volée des modèles (hors modèle de langage).

La première originalité de notre système est la possibilité de calculer des alignements de mots à la demande. Dans les approches à l'état de l'art, cet alignement est indispensable pour estimer les modèles de traduction. Notre système le calcule à la demande à partir de données sélectionnées par échantillonnage. Deux étapes sont nécessaires : pour chaque paire de phrases à aligner, un traitement de repérage de traductions candidates (*transpottting*) est réalisé afin de collecter des statistiques entre mots dans la phrases source et cible, qui sont ensuite utilisées pour calculer les scores d'association entres mots. Ces scores sont à leur tour utilisés dans une seconde étape par un algorithme de segmentation binaire récursive qui produit des alignement mot-à-mot pour chaque paire de phrases.

La deuxième différence majeure entre notre système et un système classique est la capacité du premier à estimer les modèles de traduction en fonction du besoin. Les corpus d'entraînement sont stockés dans un *tableau de suffixes* qui permet de localiser rapidement les occurrences d'un segment dans l'ensemble du corpus. Les modèles de traduction d'un segment \bar{s} sont estimés sur un sous-ensemble de ses occurrences dans le corpus.

En combinant ces deux techniques, la procédure suivie par notre système est décrite dans l'Algorithme 1.

Une séries d'expériences est ensuite utilisée pour valider notre système. Les résultats montrent que :

- l'estimation à la volée des modèles de traduction fondée sur l'échantillonnage nous permet de construire très rapidement des systèmes et d'obtenir des résultats aussi bons que ceux obtenus par les systèmes à l'état de l'art ;
- notre approche d'alignement à la demande s'avère aussi performante que la technique d'alignement état de l'art (Giza++) sur de petites

Algorithm 1 Procédure de développement à la demande d'un système de traduction

Corpus d'entraînement
Un document à traduire
Extraire tous les segments dans le document
for all segment \bar{s} extrait **do**
 sélectionner les exemples de traduction pour \bar{s}
 calculer les alignements pour les exemples sélectionnés
 estimer les modèles de traduction en utilisant les alignements calculés pour \bar{s}
end for

tâches ;

- système entièrement développé à la demande s'avère légèrement inférieur selon les métriques de qualité des traductions que le système état de l'art (Moses) pour des expériences standard, tout en étant nettement plus rapide.

Dans ces premières expériences, nous n'avons cependant pas mis en œuvre les avantages potentiels de notre approche. Dans le chapitre suivant, ces avantages seront mis en lumière pour produire de façon très rapide des hypothèses de traduction meilleurs que celle du système état de l'art.

Chapitre 4

Dans le chapitre précédent, nous avons présenté un nouveau cadre pour le développement de systèmes de traduction. Bien que les systèmes construits produisent des traductions légèrement moins bonnes que le système état de l'art, ses caractéristiques ont de nombreux avantages potentiels que nous avons explorés *via* des scénarios plus pragmatiques.

Tout d'abord, le système état de l'art pré-calcule toutes les informations qu'il utilise, ce qui le rend statique selon deux aspects :

- premièrement, les modèles de traduction sont statiques : ils sont estimés à partir d'un corpus connu à l'avance puis utilisés pour traduire tous les textes à traduire ;
- deuxièmement, un corpus de développement est nécessaire, lequel est supposé provenir de la même distribution de la traduction que l'ensemble des textes à traduire, ce qui correspond à une hypothèse généralement

trop simpliste.

Les deux raisons ci-dessus nous amènent à chercher des moyens pour développer progressivement des systèmes à la demande qui seront mis à jour de manière légère dès que de nouvelles données seront disponibles. Dans ce chapitre, nous choisissons d'introduire deux cas d'utilisation rendus possible par notre cadre (*traduction pour les communautés* et *traduction de textes quelconques Any-text*) pour illustrer ses capacités et sa flexibilité.

Dans le scénario de traduction pour les communautés, il existe une séquence de documents à traduire qui sont d'une même origine inconnue. Cette dernière contrainte exclut la possibilité de préparer un corpus de développement à l'avance. Dans nos expériences, les documents sont traduits les uns après les autres. Après la traduction de chaque document, la sortie du système est supposée être post-éditée par un traducteur humain. Le document et sa traduction sont ensuite intégrés dans le système en vue d'améliorer les traductions pour les documents suivants.

Nous avons évalué la performance de notre système sous plusieurs conditions ; les résultats confirment que notre système est très rapide, et montrent qu'il s'améliore rapidement au fur et à mesure des traductions. Grâce à l'estimation des modèles de traduction à la volée, notre système devient très rapide ; s'il n'y a par exemple qu'une seule phrase à traduire, moins d'une minute est nécessaire pour obtenir un système. Cette propriété permet d'intégrer de façon très efficace toutes nouvelles données disponibles, ce qui permet une amélioration rapide de la qualité des traductions produites.

Le deuxième scénario de traduction étudié, traduction de textes quelconques (*Any-text*), est plus difficile : on ne traduit que des documents d'origine inconnue et sans relation les uns avec les autres. Cela reprend essentiellement la situation d'un service web de traduction (comme Google Translate¹). Dans nos expériences, nous appliquons la même stratégie que précédemment mais phrase par phrase, ce qui signifie que pour chaque document, après la traduction de chaque phrase, sa traduction automatique est post-éditée par un traducteur humain puis intégrée dans le système. Neuf documents de différentes origines ont été utilisés, et notre système s'est montré plus performant pour 5 d'entre eux que le système état-de-l'art.

Ce chapitre nous a permis d'illustrer les caractéristiques marquantes de notre approche : l'estimation à la demande nous permet d'estimer des modèles de traduction très rapidement, et l'alignement incrémental autorise l'intégration

1. <https://translate.google.com/>

très rapide de nouvelles données pour améliorer la qualité des traductions produites. Notre cadre implémente une approche innovante qui pourrait également être adaptée pour permettre une traduction interactive. Par ailleurs, notre cadre permet le développement de systèmes même lorsqu’aucun corpus de développement n’est disponible, ce qui constitue une propriété très appréciable pour de nombreux cas réels.

Chapitre 5

Dans les chapitres précédents, les exemples d’entraînement sont collectés par échantillonnage aléatoire, c’est-à-dire sans adaptation particulière au texte à traduire. Dans ce chapitre, nous proposons l’introduction d’une méthode d’*échantillonnage contextuel* qui permet de sélectionner les exemples de traduction les plus cohérents avec chaque texte à traduire afin d’adapter utilement les modèles de traduction. Nous avons essayé deux types de similarités pour évaluer notre approche de prise en compte du contexte : la précision n -gramme et le TF-IDF.

La précision n -gramme entre deux phrases est définie comme le rapport du nombre de n -grammes en commun par rapport au nombre total de n -gramme dans les deux phrases. Ici nous utilisons une version modifiée de cette mesure, $Sim(\mathbf{s}, \mathbf{s}^d)$:

$$Sim(\mathbf{s}, \mathbf{s}^d) = \sum_{n=1}^N \frac{1}{N} p_n(\mathbf{s}, \mathbf{s}^d) \quad (4)$$

$$p_n(\mathbf{s}, \mathbf{s}^d) = \frac{|\mathbf{s}_{n\text{-grammes}} \cap \mathbf{s}^d_{n\text{-grammes}}|}{(1 - Z) * |\mathbf{s}_{n\text{-grammes}}| + Z|\mathbf{s}^d_{n\text{-grammes}}|} \quad (5)$$

où \mathbf{s} et \mathbf{s}^d représentent les deux phrases ; $\mathbf{s}_{n\text{-grammes}}$ et $\mathbf{s}^d_{n\text{-grammes}}$ représentent l’ensemble de n -grammes dans \mathbf{s} et \mathbf{s}^d , respectivement ; et Z est un facteur de pénalité sur la longueur de \mathbf{s} . La précision n -gramme mesure une certaine similarité entre les chaînes de mots de deux phrases ; les phrases qui partagent plus de n -grammes obtiendront des scores plus forts. Toutes les tailles de n -grammes contribuent ici de la même manière au calcul du score.

TF-IDF est une mesure de similarité très utilisée en Recherche d’Information. Chaque document ou phrase est représenté par un vecteur $\mathbf{v}_\mathbf{s} = (w_1 \dots w_j \dots w_n)$, où n est la taille de vocabulaire. Chaque entrée w_j dans ce

vecteur est calculée par :

$$w_j = tf_j \times \log(idf_j) \quad (6)$$

où, tf_j est le nombre d'occurrences dans le document du j ème mot dans le vocabulaire, et idf_j est la fréquence documentaire inverse du j ème mot dans le vocabulaire. Nous avons utilisé la définition suivante :

$$idf_j = \frac{\# \text{ documents in corpus}}{\# \text{ documents containing the } j\text{-th vocabulary word}} \quad (7)$$

La similarité entre deux documents peut alors être définie comme la distance entre les deux vecteurs correspondants.

Ces deux types de similarité contextuelle ont été intégrées pour permettre l'adaptation des systèmes de traduction développés. Nous avons en outre proposé un modèle de confiance en l'adaptation réalisée.

Nous avons évalué notre stratégie d'échantillonnage contextuel ainsi que notre modèle de confiance. Nos résultats montrent notamment que l'utilisation de TF-IDF permet de bien adapter nos systèmes au domaine du texte à traduire. L'analyse des sorties du système révèle que les mots et segments qui ont un score de confiance plus important sont plus souvent bien traduits que ceux dont le score de confiance est plus faible. Toutefois, l'utilisation de ce modèle de confiance ne nous a pas permis d'améliorer la qualité des traductions produites. La précision n -grammes donne des résultats inférieurs qui semblent dépendre davantage des tâches de traduction, ce qui est peut-être dû au paramètre Z de cette mesure qui peut ne pas être bien adapté à chaque document individuel.

Conclusion

Le très grand coût de calcul associé au développement de systèmes de traduction de haute performance est reconnu depuis longtemps, et il est problématique pour plusieurs raisons. Tout d'abord, ce coût limite sérieusement le nombre d'expériences qui peuvent effectivement être lancées sur de très grands ensembles de données et la fréquence de mise à jour des systèmes lorsque de nouvelles données deviennent disponibles. De plus, il constitue une barrière à l'entrée pour les nouveaux arrivants dans ce domaine de recherche, qui doivent investir un montant plus élevé de ressources pour construire des systèmes à l'état de l'art.

L'estimation de modèles à la volée avait préalablement été étudié dans [Callison-Burch et al., 2005, Lopez, 2008], mais nous avons proposé une approche rapide et compétitive pour extraire des informations de traduction. Dans nos travaux, les modèles de traduction n'ont plus besoin d'être pré-calculés, et ne sont calculés que pour le texte à traduire, ce qui réduit considérablement les coûts de développement. Ces coûts peuvent être davantage réduits par l'utilisation d'une stratégie d'échantillonnage, qui n'utilise qu'un nombre limité d'exemples pour estimer les modèles. Les études précédentes reposaient sur la disponibilité *a priori* des alignements mot-à-mot nécessaire, ce qui est toutefois obtenu par l'étape la plus coûteuse du développement de tels systèmes.

La première contribution importante de notre thèse a donc été consacré à la proposition d'une nouvelle méthode d'alignement mot-à-mot qui peut aligner chaque paire de phrases isolément. Cette propriété était nécessaire pour permettre le développement de systèmes de traduction à la demande, avec des gains supplémentaires importants en temps de traitement, en particulier lorsque juste quelques textes doivent être traduits. Dans notre cadre, toutes les informations nécessaires pour les estimations des modèles de traduction sont calculées à la demande, et le temps de traitement nécessaire dépend principalement de la taille du texte à traduire plutôt que de la taille des données d'entraînement. Comme illustration concrète, notre système peut créer des tables de traduction pour des phrases individuelles en moins d'une minute lorsque le corpus parallèle utilisé contient de l'ordre de 400 millions de tokens dans la langue source.

Une autre propriété essentielle est l'intégration transparente des données nouvellement disponibles. Cette propriété permet de faire usage de toutes nouvelles données, mais également d'exploiter immédiatement des textes post-édités par un traducteur humain. Les systèmes développés peuvent donc éviter des erreurs répétitives et s'adapter au domaine et au style spécifique dans un scénario de traduction interactif.

Une autre question importante que nous avons identifiée dans les travaux précédents de Callison-Burch et al. [2005] et Lopez [2008] est que les exemples de traduction sont choisis aléatoirement dans le corpus d'entraînement. Bien que les résultats expérimentaux ont montré que cette approche peut atteindre des performances de traduction compétitifs, celle-ci n'est clairement pas optimale, en particulier lorsque la majorité des données d'entraînement sont hors-domaine relativement au domaine du texte à traduire.

La deuxième contribution principale de notre thèse a donc été d'intégrer

des stratégies d'échantillonnage contextuels pour sélectionner des exemples de traduction dans un corpus à grande échelle qui sont semblables au texte à traduire, et donc créer des tables de traduction adaptées. De même que pour les approches traditionnelles de sélection de données, notre approche d'échantillonnage contextuel sélectionne les sous-parties les plus pertinentes du corpus d'entraînement. Cependant, contrairement à la plupart des approches de sélection de données, notre approche s'applique au niveau des segments, ce qui permet à la fois d'exploiter l'ensemble du corpus d'entraînement, ainsi que d'assurer une meilleure couverture lexicale.

Références

- Satanjeev Banerjee and Alon Lavie. METEOR : An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, USA, June 2005. URL <http://www.aclweb.org/anthology/W/W05/W05-0909>.
- Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2) :79–85, June 1990. ISSN 0891-2017. URL <http://dl.acm.org/citation.cfm?id=92858.92860>.
- Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation : Parameter estimation. *Computational Linguistics*, 19(2) :263–311, June 1993. ISSN 0891-2017. URL <http://dl.acm.org/citation.cfm?id=972470.972474>.
- Chris Callison-Burch, Colin Bannard, and Josh Schroeder. Scaling phrase-based statistical machine translation to larger corpora and longer phrases. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 255–262, Ann Arbor, USA, June 2005. doi : 10.3115/1219840.1219872. URL <http://www.aclweb.org/anthology/P05-1032>.

- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1) :pp. 1–38, 1977. ISSN 00359246. URL <http://www.jstor.org/stable/2984875>.
- George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research (HLT)*, pages 138–145, San Francisco, USA, 2002. Morgan Kaufmann Publishers Inc. URL <http://dl.acm.org/citation.cfm?id=1289189.1289273>.
- Alon Lavie and Abhaya Agarwal. METEOR : An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation (WMT)*, pages 228–231, Prague, Czech Republic, June 2007. URL <http://www.aclweb.org/anthology/W/W07/W07-0734>.
- Adam Lopez. Tera-scale translation models via pattern matching. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, pages 505–512, Manchester, UK, August 2008. URL <http://www.aclweb.org/anthology/C08-1064>.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU : a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, USA, July 2002. doi : 10.3115/1073083.1073135. URL <http://www.aclweb.org/anthology/P02-1040>.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231, Boston, USA, August 2006.