Système de traduction automatique statistique combinant différentes ressources

Sadat, F.; Foster, George; Kuhn, Roland

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

Système de traduction automatique statistique combinant différentes ressources *

Sadat, F., Foster, G., and Kuhn, R.
Février 2006

* publié à la conférence du traitement automatique des langues naturelles (TALN 2006). Leuven, Belgique. Du 10 au 13 avril 2006. NRC 48758.

Canada

TALN 2006, Leuven, 10-13 avril 2006

Système de traduction automatique statistique combinant différentes ressources

Fatiha Sadat, George Foster and Roland Kuhn

Institut de Technologie de l'Information
101 rue st-Jean Bosco, Gatineau, QC K1A 0R6, Québec, Canada
prénom.nom@cnrc-nrc.gc.ca

Résumé Cet article décrit une approche combinant différents modèles statistiques pour la traduction automatique basée sur les segments. Pour ce faire, différentes ressources sont utilisées, dont deux corpus parallèles aux caractéristiques différentes et un dictionnaire de terminologie bilingue et ce, afin d'améliorer la performance quantitative et qualitative du système de traduction. Nous évaluons notre approche sur la paire de langue français-anglais et montrons comment la combinaison des ressources proposée améliore de façon significative les résultats.

Abstract This paper describes an approach combining different statistical models for phrase-based machine translation. Different knowledge resources are used, such as two parallel corpora with different characteristics and a bilingual dictionary of terminology, in order to improve the qualitative and quantitative performance of the translation system. We evaluate our approach on the French-English language pair and show how combining the proposed resources significantly, improves the results.

Mots-clés : Traduction automatique statistique basée sur les segments, corpus parallèle, dictionnaire de terminologie bilingue.

Keywords: Statistical phrase-based machine translation, parallel corpora, bilingual dictionary of terminology.

1 Introduction

Le nombre d'approches en traduction automatique s'est multiplié dans les dernières années. Il existe entre autres la traduction par les règles, la traduction statistique et la traduction guidée par l'exemple. Dans cet article, nous décrivons l'approche statistique adoptée pour le projet *Portage* au sein du Groupe de Technologies Langagières Interactives (GTLI)[1] au Conseil national de recherches, Canada[2].

Un système de traduction automatique a pour fonction de traduire un texte *S* dans une langue source en un texte *T* dans une langue cible. Dans cette étude, nous adoptons une approche statistique utilisant différentes ressources, dont deux corpus parallèles des plus connus pour la paire de langue français-anglais, où l'un des textes est la traduction de l'autre, souvent appelé *bi-textes*, ainsi qu'un dictionnaire de terminologie bilingue pour la même paire de langues.

Nous avons participé à la compétition d'évaluation des systèmes de traduction (*shared task of the ACL 2005 Workshop on Building and Using Parallel Texts*[3]) (Koehn and Monz, 2005) avec les paires de langues fournies aux participants (Sadat et al., 2005). Notre système de traduction, nommé *Portage,* a montré une des meilleurs performances lors de la compétition et a été classé troisième parmi les 11 équipes participantes. Dans cet article, nous décrivons la suite de l'évaluation en utilisant une approche combinant des corpus parallèles et un dictionnaire de terminologie bilingue et montrons des résultats qui auraient pu nous classer premier dans cette compétition.

Le contenu de cet article se résume comme suit. La section 2 décrit la traduction statistique en général. La section 3 présente les différentes étapes et ressources utilisées dans *Portage*. La section 4 décrit l'intégration de ressources terminologiques au système de traduction. Nous montrons dans la section 5 les analyses et les performances obtenues en implémentant notre approche et discutons en section 6 d'autres travaux auxquels la présente étude est liée. La section 7 conclue cet article et suggère des extensions et travaux futurs.

2 La Traduction statistique (SMT)

La *traduction statistique* (SMT) se base sur la théorie mathématique de distribution et d'estimation probabiliste développée par Frederick Jelinek au *IBM* T.J. Watson Research Center et—en particulier — sur un article de (Brown et al., 1990), (Carl, 2003). Les systèmes statistiques apprennent un modèle probabiliste de traduction *P(t|s)* à partir d'un texte bilingue et un modèle probabiliste de la langue cible *P(t)* à partir d'un texte monolingue. En général, la qualité des traductions générées par un tel système est proportionnelle à la quantité des données sur lesquelles les paramètres du système sont estimés. Par opposition à l'approche traditionnelle de « système expert », l'approche statistique de la traduction automatique est capable de s'améliorer automatiquement au fur et à mesure que de nouvelles données d'entraînement deviennent disponibles.

En temps d'exécution, la meilleure traduction pour une phrase nouvelle est recherchée grâce à la maximisation de ces deux modèles probabilistes.

---

[1] http://iit-iti.nrc-cnrc.gc.ca/projects-projets/portage-tech_f.html

[2] http://www.nrc-cnrc.gc.ca/

[3] http://www.statmt.org/wpt05/mt-shared-task/

$$arg\ maxP(t\ |\ s) = arg\ max\left\{P(\ t\ ) \times P(\ s\ |\ t\ )\right\}$$

Typiquement, la traduction statistique génère la phrase cible à partir des traductions de mots simples et isolés. La « meilleure » traduction est déterminée en SMT par les probabilités *P(s|t)* et *P(t)* qui sont générées indépendamment l'une de l'autre et représentent le modèle de traduction et le modèle de langue. En pratique, les deux modèles, de langue et de traduction, sont représentés par des ensembles de tables contenant les valeurs de probabilité de certains paramètres.

Parmi les caractéristiques de la traduction automatique, notons la nécessité de disposer de grandes quantités de textes bilingues alignés nécessaires pour l'entraînement, le décodage et le ré-ordonnancement des hypothèses de traduction, tel qu'expliqué dans la section 3.

Dans notre cas, nous avons opté pour l'approche basée sur les segments qui utilise un modèle log-linéaire, décrit par la formule suivante :

$$arg\ maxP(t \mid s) = arg\ max\left[\ exp\left(\sum_i \alpha_i \times f_i(s,t)\right)\right]$$

Les traits f_i représentent le modèle de langue P(t|), le modèle de traduction P(s|t), le modèle de distorsion et la pénalité sur les mots, parmi d'autres.

Pour évaluer les performances de la traduction, on utilise les métriques traditionnelles dans le domaine de la traduction automatique. Parmi ces métriques, on retrouve le *score BLEU* (Papineni et al., 2002), une mesure pour laquelle la traduction candidate et celle de référence sont comparées non seulement au niveau des mots mais également au niveau des bigrammes, trigrammes etc. Pour calculer le score BLEU entre une traduction candidate *c* et une traduction de référence *r*, nous avons la formule suivante (Patry et Langlais, 2005) :

$$Bleu(\ c,r\ ) = BP \times e^{\sum_{n=1}^{N} \frac{n-grammes_c \cap n-grammes_r}{N \times |n-grammes_c|}}$$

où, N est la taille maximale des n-grammes considérés (par exemple 4) et $n\text{-}grammes_c$ et $n\text{-}grammes_r$ sont respectivement les ensembles de n-grammes des phrases *c* et *r*.

BP (la *brevity penalty*) est définie comme suit :

$$BP\ = min\left(1, e^{\frac{|c|}{|r|}}\right)$$

Le coefficient BP est là pour éviter que le score BLEU ne favorise les traductions candidates courtes pour lesquelles |n-grammes$_c$| est petit, ce qui augmente artificiellement le quotient dans l'exponentielle de la formule de BLEU. Par contre, le score BLEU favorise les traductions candidates contenant les n-grammes les plus courants alors que ce sont rarement ceux qui sont les plus porteurs de sens.

Le score BLEU est normalisé entre 0 et 1 et exprimé souvent en pourcentage.

3 Portage

Portage est développé par le Groupe de Technologies Langagières Interactives (GTLI) du Conseil national de recherche Canada (CNRC). Ce système de traduction comprend quatre phases principales : *prétraitement* des données bruitées en tokens avec traduction de quelques mots ou segments générés à partir de règles; *entraînement et décodage* afin de produire les modèles de

langue en langue cible et de traduction basé sur les segments et traduire le texte source en langue cible utilisant une hypothèse de traduction; *ré-ordonnancement* afin de produire une ou plusieurs hypothèses de traduction et ré-ordonnancement de ces hypothèses afin de maximiser la performance du système de traduction tel que mesuré avec les métriques (ex. NIST/BLEU/Wer); et finalement *post-traitement* des données de sortie, consistant à redonner un format adéquat du texte obtenu après traduction, suivant la langue cible.

3.1 Prétraitement

Le prétraitement est une phase nécessaire et indispensable pour la conversion des données brutes dans les deux langues, source et cible en un format adéquat à l'entraînement des modèles et au décodage (Foster et al., 2003). Nous avons utilisé deux corpus avec des caractéristiques différentes : d'une part le *Hansard*, un corpus français/anglais préalablement aligné et rassemblant les débats de la Chambre des Communes du Parlement canadien; d'autre part l'*Europarl,* qui est extrait des registres du Parlement européen et comporte plusieurs paires de langues européennes dont le français-anglais.

Dans cette étude, nous distinguons trois ensembles de textes : ceux utilisés dans l'entraînement, ceux utilisés dans le développement et finalement le test pour les deux langues source et cible. Le test consiste en un texte source et une ou plusieurs référence(s) (c'est-à-dire traduction(s) produite(s) par des humains).

La première opération de prétraitement est la *tokenisation*, dont l'objectif est justement de transformer la chaîne de caractères en tokens. Cette opération s'applique aux textes source et cible et va prendre en compte les espaces pour séparer les mots, les nombres et la ponctuation. L'opération suivante va tout mettre en minuscule pour simplifier le travail lors de l'entraînement.

La *tokenisation* est une opération assez simple pour certaines langues (ex. français, anglais) mais assez compliquée pour d'autres, surtout quand la segmentation des mots est nécessaire, comme c'est le cas pour le chinois et l'allemand.

En plus de ces opérations simples, nous avons développé un module basé sur des règles pour la détection des nombres et des dates dans le texte source et l'identification de leurs traductions dans le texte cible, qu'on appellera module *nombre et date*. Ce module a été appliqué sur les textes d'entraînement (Hansard et Europarl), de développement dans les langues source et cible et de test (en langue source seulement).

L'algorithme de Moore (Moore, 2002) est utilisé pour aligner les textes en langues source et cible afin de synchroniser les lignes, i.e., la ième ligne dans le texte cible est la traduction de la *ième* ligne dans le texte source. Deux textes alignés auront le même nombre de lignes.

3.2 Entraînement et décodage

Le décodage en SMT comprend la recherche des hypothèses *t* ayant les plus grandes probabilités pour être considérées comme étant les traductions de la phrase source en question, suivant le modèle de traduction *P(t|s)*. Le modèle *P(t|s)* est une combinaison log-linéaire de quatre principaux composants : un ou plusieurs modèles trigramme de la langue cible, un ou plusieurs modèles de traduction basés sur les segments, un modèle de distorsion qui rend compte des différences dans l'ordre des mots en langues source et cible, et un modèle de longueur qui rend compte des différences de longueur entre les deux langues.

Le modèle trigramme de langue a été implanté en utilisant le programme SRILM (Stolcke, 2002). Le modèle de traduction basé sur les segments utilise le modèle symétrique IBM2 pour l'induction

des paires de segments comme décrit par Koehn (Koehn, 2004). Les modèles de distorsion est aussi ceux de longueur sont similaires à ceux de Koehn.

Pour établir les poids des composants dans le modèle log-linéaire, nous avons implémenté l'algorithme de Och (Och, 2003). Ceci implique essentiellement la génération des *N meilleurs* hypothèses de traduction dans un processus itératif, représentant l'espace entier de recherche pour un ensemble donné de phrases sources du corpus de développement. Une variante de l'algorithme de Powell est utilisée pour trouver les poids qui optimisent le score BLEU sur ces hypothèses, comparés aux traductions de référence du corpus de développement. Ces opérations de décodage sont accomplies par le décodeur *Canoe*, qui implémente un algorithme de recherche en faisceau en programmation dynamique (dynamic-programming beam search algorithm), comme l'algorithme décrit par (Koehn, 2004) avec quelques extensions comme la capacité de décoder vers l'avant et/ou vers l'arrière.

3.3 Ré-ordonnancement

Afin d'améliorer les sorties du décodeur *Canoe*, nous avons utilisé une technique de ré-ordonnancement (*rescoring*), qui consiste à générer une liste des *N* hypothèses de traduction les plus probables puis de réordonner celles-ci en utilisant un modèle entraîné avec la méthode Och pour optimiser le score BLEU (Och, 2002). Ceci est similaire à la dernière phase de l'algorithme décrit dans la section précédente, excepté pour l'utilisation d'un modèle log-linéaire plus riche que celui utilisé dans le décodeur.

En plus des quatre composants du modèle initial de décodage, la phase de ré-ordonnancement inclut d'autres traits, tel que les probabilités du modèle IBM2 dans les deux directions (c.-à-d., $P(s|t)$ et $P(t|s)$); et des traits basés sur le modèle IBM1 conçu pour détecter les imperfections de la traductions, c.-à-d., si un mot dans la langue source a une traduction imparfaite dans la langue cible. Ce trait de mots inconnus (*missing-word feature*) est aussi appliqué dans les deux directions.

3.4 Post-traitement

La dernière phase dans *Portage* consiste en un post-traitement et implique deux opérations : la *restauration des majuscules* (Agbago et al., 2005) aux endroits appropriés et la *détokenisation* afin de redonner au texte traduit son format normal suivant la langue cible.

4 Intégration de ressources terminologiques

Plusieurs stratégies sont envisageables afin d'améliorer la performance du système de traduction. Il est possible d'augmenter la taille des corpus parallèles en rajoutant des bitextes et/ou des textes comparables d'un domaine spécialisé collectés sur la toile. Cependant, ces stratégies s'avèrent très coûteuses en temps et en calcul et les corpus comparables se sont distingués comme étant bruités. Il semble plus intéressant d'ouvrir un système de traduction à des ressources terminologiques existantes, telles que la base de données terminologique et linguistique du gouvernement du Canada *Termium*[4] (Langlais, TALN 2002). Notons que nous n'avons pas encore le droit d'utiliser *Termium* dans nos recherches sur la traduction.

L'Office québécois de la langue française nous a permis d'utiliser le *Grand Dictionnaire Terminologique* (GDT) pour des fins de recherche. Ce lexique terminologique anglais-français contient des termes, des synonymes, des acronymes, des définitions, des unités phraséologiques, des exemples d'utilisation et des observations dans des domaines très variés; il donne accès à près

[4] www.termium.com/

de 3 millions de termes français et anglais du vocabulaire industriel, scientifique et commercial, dans 200 domaines d'activité.

Dans notre implémentation, nous avons considéré un modèle de traduction probabiliste construit à partir des paires de termes en langue source et des alternatives de traduction proposées dans le lexique terminologique. Les scores de probabilité sont considérés équiprobables pour chaque terme ou phrase source. Exemple, si le lexique contient trois traductions du mot français « port » en anglais comme suit : haven, harbor et port - le modèle *P(t|s)* aura les probabilités suivantes : *P(haven|port)*=0.33, *P(harbor|port*)=0.33, *P(port|port)* = 0.33. Ce modèle de traduction, comme les autres modèles de traduction, reçoit un poids log-linéaire qui lui est donné par l'algorithme d'Och, comme expliqué dans la section 3.2.

On peut aussi envisager de rajuster ces scores de probabilités par une méthode statistique de désambiguïsation qui tient compte du contexte des termes ou des segments dans les corpus d'entraînement.

5 Analyse et résultats

Nous présentons dans cette section les résultats de l'évaluation du processus de traduction automatique statistique en utilisant les ressources présentées précédemment. Les tailles des données d'entraînement sont comme suit :

- 688 031 phrases extraites du corpus Europarl dans les deux langues français et anglais, fournis lors de la compétition d'évaluation de systèmes de traduction WPT-ACL 2005 (Koehn and Monz, 2005).
- 6 056 014 phrases extraites du corpus Hansard dans les deux langues français et anglais.

Le corpus de développement consiste en un ensemble de 2000 phrases dans les deux langues, fournis lors de la compétition d`évaluation WPT-ACL 2005. Un corpus contenant 1000 phrases sert à optimiser les poids des modèles de décodage et de ré-ordonnancement. Le nombre des meilleures hypothèses de traduction a été fixé à *N*=1000. Les 1000 phrases restantes étaient réservées à l'évaluation de la performance des modèles de traduction.

Les résultats illustrés dans le tableau 1 sont basés sur la traduction d'un corpus de test contenant 2000 phrases en langue source et obtenu suivant les règles de la compétition WPT-ACL 2005. La partie référence du corpus de test en langue cible nous a été fournie après que les résultats soient établis par les organisateurs de la compétition WPT-ACL 2005.

Le tableau 1 montre les résultats des méthodes de traduction suivantes :

1. La méthode *E* utilise le corpus Europarl dans l'entraînement pour apprendre un modèle de langue et un modèle de traduction.
2. La méthode *E_p* utilise le corpus Europarl dans l'entraînement et la traduction des nombres et des dates basée sur les règles dans la phase prétraitement. Au total, un modèle de langue et un modèle de traduction sont générés.
3. La méthode *E-H* utilise les deux corpus, Europarl et Hansard, dans l'entraînement pour apprendre deux modèles de langue et deux modèles de traduction liés à chaque corpus.
4. La méthode *E-H_p* utilise les deux corpus, Europarl et Hansard, dans l'entraînement et la traduction des nombres et des dates basée sur les règles dans la phase prétraitement. Au total, deux modèles de langue et deux modèles de traduction sont générés.

5. En plus des ressources utilisées par la méthode *E-H*, la méthode *E-H-GDT* utilise le GDT pour générer un troisième modèle de traduction.
6. La méthode *E-H-GDT_p* est similaire à la méthode *E-H-GDT*, excepté pour la traduction des nombres et des dates basée sur les règles dans la phase prétraitement.

Finalement, l'analyse et correction des erreurs de traduction, dans la phase de post-traitement, dues principalement à la couverture du vocabulaire, est représentée par la méthode *E-H-GDT+analyse_correction*.

Pour évaluer les différentes approches citées, nous avons utilisé la métrique de précision des *n-gramme,* par rapport à un ensemble de traduction de références, *BLEU* (Papineni et al., 2002) que l'on voudra maximiser.

Le tableau 1 montre une amélioration en terme de score BLEU lors de l'utilisation des deux modèles de langue et des deux modèles de traduction (méthode *E-H*). Aussi, on remarque une dégradation d'environ un point BLEU en utilisant la traduction des nombres et des dates basée sur les règles (méthode *E_P* et *E-H_P* et *E-H-GDT_P*). Cette dégradation est principalement causée par l'optimisation des règles de traduction suivant les conventions du corpus Hansard et appliquées au corpus Europarl; par exemple, une date dans le corpus Hansard est généralement écrite « *mois jour année* » alors que dans le corpus Europarl, elle peut aussi être écrite « *jour mois année* ». Cette différence de conventions est due à l'aspect culturel des deux régions géographiques (Amérique du Nord vs. Europe) d'où les deux corpus sont issus.

Nos performances sont très bonnes si on les compare aux résultats des autres participants à WPT-ACL 2005. La méthode E-H a été classée troisième lors de la compétition avec une différence de 0,74 en terme de score BLEU par rapport au premier participant (30,27) et une différence de 0,67 en terme de score BLEU par rapport au deuxième participant (30,20).

Les performances que nous avons relevées avec notre décodeur Canoe sur les corpus Hansard et Europarl sont cependant comparables à celles relevées en utilisant le décodeur Pharaoh (Koehn, 2004). Les scores BLEU obtenus en utilisant le décodeur Pharaoh sur une hypothèse de traduction (sans ré-ordonnancement) sont listés comme suit : 26,98 (méthode *E*), 20,12 (méthode *E_p*), 22,21(méthode *E-H*), 24,84 (méthode *E-H_p*).

Méthode	Décodage	Décodage+Ré-ordonnancement
E (1 LM, 1 TM)	27,71	29,22
E_p (1 LM, 1 TM)	26,45	28,21
E-H (2 LMs, 2 TMs)	28,71	29,53
E-H_p (2 LMs, 2 TMs)	28,29	28,56
E-H-GDT (2 LMs, 3 TMs)	29,12	**30,23**
E-H-GDT_p (2 LMs, 3 TMs)	27,90	29,35
E-H-GDT+analyse_correction (2 LMs, 3 TMs)	-	**30,28**

Tableau 1 : Résultats en terme de score BLEU sur le corpus de test

Finalement, l'approche intégrant le GDT (méthode *E-H-GDT*) montre une meilleure performance - également légèrement supérieure au résultat du deuxième participant (0,03) et légèrement inférieure au résultat du premier participant (0,04). L'analyse des mots non traduits par l'approche *E-H-GDT* nous a permis d'adopter l'idée d'extraire les lemmes des mots en langue source et de les traduire en utilisant le GDT. Nous avons utilisé l'outil *Treetagger*[5] pour l'extraction des lemmes des mots non traduits et laissés en langue source par el décodeur. La traduction des lemmes des quelques mots détectés a permis d'améliorer la performance du système légèrement - suffisamment pour surpasser le résultat du participant classé premier lors de la compétition WPT-ACL2005.

Un exemple d'une phrase en français traduite à l'aide du système *Portage* en anglais, utilisant les deux corpus Hansard et Europarl et comparé à la référence (traduction manuelle faite par un expert), est comme suit :

Source (FR)	*Sortie de Portage (EN)*	*Référence (EN)*
qui pourrait en effet dénier l' entrée , dans la maison européenne , de la pologne de copernic et de jean-paul ii , de la hongrie martyrisée en 1956 à budapest ou de la capitale symbole , prague , lorsque les démocraties populaires défenestraient le chef d' état et écrasaient le peuple de jan palach , immolé pour la liberté .	that could in fact , denied entry into the european house , poland 's with copernicus pope john paul ii , and in hungary 's martyrisée in budapest in 1956 or the symbol of capital , prague , when the democracies grass-roots défenestraient the head of state and the people of écrasaient jan palach , declining for freedom .	for indeed , who could deny entry into the european house to states such as the poland of copernicus and john paul ii , the hungary bullied in budapest in 1956 or the capital that has become a symbol , prague , when popular democracies threw the head of state out of the window and crushed the nation of jan palach , sacrificed in the name of freedom .

6 Intégration de l'information morphologique

Il est envisageable d'intégrer de l'information morphologique au système de traduction statistique (Ueffing and Ney, 2003.). D'après le gain obtenu dans la performance de notre système de traduction après analyse et correction des mots non traduits, nous avons envisagé la lemmatisation. Des évaluations préliminaires sur le corpus Europarl ont montré qu'une approche combinant le modèle standard au modèle impliquant la lemmatisation est plus performante que l'approche du modèle standard seul. Par modèle standard, nous faisons référence à l'approche utilisant un modèle de langue (LM) et un modèle de traduction (TM), telle que décrite dans la section 3. Par modèle impliquant la lemmatisation standard, nous faisons référence à une approche qui utilise un modèle de langue (LM_lemm) et un modèle de traduction (TM_lemm) générés à partir d'un corpus lemmatisé, c-a-d un corpus constitué des lemmes de tous les noms, verbes, adjectifs et adverbes. La combinaison des deux approches utilisera deux modèles de langues représentés par LM:LM_lem et deux modèles de traduction représentés par TM:TM_lemm. Nous avons obtenu un résultat prometteur et envisageons d'explorer cette piste dans nos travaux futurs afin d'intégrer la lemmatisation à la phase de décodage.

7 Conclusion

Nous avons participé à la compétition WPT ACL 2005 dans le but d'évaluer notre système de traduction, *Portage*, nouvellement implémenté, de cerner les problèmes lié à sa conception et de

[5] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html/

comparer sa performance aux autres systèmes de traduction existants sur la scène internationale. En adoptant une approche combinant les deux corpus Europarl et Hansard, nous avons été classé troisième avec une légère différence dans la performance avec le participant classé premier.

L'intégration du *Grand Dictionnaire Terminologique* (GDT) améliore les résultats de la traduction en terme de score BLEU et la qualité de traduction. Ceci suggère qu'une liste d'équivalents bilingues extraite du GDT et utilisée également dans l'entraînement d'un modèle de traduction statistique, en plus des modèles de traduction et de langue générés à partir des corpus parallèles, peut améliorer la performance du système de traduction.

Dans nos travaux futurs, en plus de l'intégration d'une analyse morphologique au système de traduction, nous envisageons d'exploiter les corpus comparables et de raffiner les règles de traduction des nombres et des dates. La translittération des entités nommées est parmi les extensions que nous souhaitons explorer dans le futur afin d'améliorer la performance de notre système de traduction.

Le système de traduction *Portage* est opérationnel pour les deux langues officielles du Canada (français, anglais) dans les deux directions. Le système de traduction a été testé dans des compétitions internationales pour la traduction vers l'anglais de textes en chinois (NIST[6] 2005), en espagnol, en allemand et en finlandais (WPT-ACL 2005).

Remerciements

Les auteurs tiennent à remercier l'Office Québécois de la Langue Française pour l'utilisation du Grand Dictionnaire Terminologique dans leurs recherches, l'équipe Portage du GTLI au Conseil National de Recherches Canada pour sa participation dans le développement du système de traduction automatique, Dr. Michel Simard et Eric Joanis pour leurs discussions et commentaires fructueux.

Références

AGBAGO A., KUHN R., ET FOSTER G. (2005). Truecasing for the PORTAGE System. Actes de *International Conference on Recent Advances in Natural Language Processing RANLP 2005, Borovets, Bulgaria,* 21-23 Septembre 2005.

BROWN P.F., COCKE J., DELLA PIETRA S.A., DELLA PIETRA V.J., JELINEK F., LAFFERTY J.D., MERCER R.. L. ET ROOSSIN P.S. (1990). A Statistical Approach to Machine Translation. Computational Linguistics, 16(2):79-85, 1990.

FOSTER G., GANDRABUR S., LANGLAIS P., PLAMONDON P., RUSSELL G. ET SIMARD M. (2003). Statistical Machine Translation: Rapid Development with Limited Resources. Actes de *MT Summit IX 2003, New Orleans*, Septembre 2003.

KNIGHT K., CHANDER I., HAINES M., HATZIVASSILOGLOU V., HOVY E., IIDA M., LUK S.K., WHITNEY R., ET YAMADA K. (1995). Filling Knowledge Gaps in a Broad-Coverage MT System. Actes de *International Joint Conference on Artificial Intelligence (IJCAI)*, 1995.

KOEHN P. (2004). Pharaoh: a Beam Search Decoder for Phrase-based Statistical Machine Translation Models. Actes de *Association for Machine Translation in the Americas AMTA,* 2004.

[6] http://www.nist.gov/speech/tests/mt/mt05eval_official_results_release_20050801_v3.html

KOEHN P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. Actes de *MT Summit,* 2005.

KOEHN P., ET MONZ C. (2005) Shared Task: Statistical Machine Translation between European Languages. Actes de *Association for Computational Linguistics ACL Workshop on Building and Using Parallel Texts*, *Ann Arbor,* June 2005. 119–124.

LANGLAIS P. (2002). Ressources Terminologiques et Traduction Probabiliste : Premiers pas Positif vers un Systeme Adaptatif. Actes de *TALN 2002, Nancy,* 24-27 Juin 2002.

MOORE R. (2002). Fast and Accurate Sentence Alignment of Bilingual Corpora. In Machine Translation: From Research to Real Users. Actes de *5th Conference of the Association for Machine Translation in the Americas, Tiburon, California, Springer-Verlag, Heidelberg*, *Germany*. 135-244.

OCH F.J., ET NEY. H. (2000). Improved Statistical Alignment Models. Actes de *38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, China,* Octobre 2000. 440-447.

OCH, F. J. ET NEY, H. (2002). Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. Actes de *40th Annual Meeting of the Association for Computational Linguistics, Philadelphia*. 295–302.

OCH F.J. (2003). Minimum Error Rate Training for Statistical Machine Translation. Actes de *41st Annual Meeting of the Association for Computational Linguistics, Sapporo*, Juillet 2003.

OCH F.J., GILDEA D., KHUDANPUR S., SARKAR A., YAMADA K., FRASER A., KUMAR S., SHEN L., SMITH D., ENG K., JAIN V., JIN Z., ET RADEV D. (2004). A Smorgasbord of Features for Statistical Machine Translation. Actes de *HLT/NAACL 2004, Boston, MA*. May 2004.

PAPINENI K., ROUKOS S., WARD T., ET ZHU W.J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. Actes de *40th Annual Meeting of the Association for Computational Linguistics ACL, Philadelphia,* Juillet 2002, 311-318.

PATRY A. ET LANGLAIS P. (2005). Paradocs : un système d'identification automatique de documents parallèles. Actes de *TALN 2005 Dourdan, 6-10 Juin* 2005.

SADAT F., JOHNSON H., AGBAGO A., FOSTER G., KUHN R., MARTIN J. ET TIKUISIS A.. (2005). Portage: A Phrase-based Machine Translation System. Actes de *ACL 2005 Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond. Ann Arbor, Michigan, USA*. Juin 29-30 2005.

STOLCKE A. (2002). SRILM - an Extensible Language Modeling Toolkit. Acets de *ICSLP-2002.* 901-904.

UEFFING N., ET NEY H. (2003). Using POS Information for Statistical Machine Translation into Morphologically Rich Languages. Actes de *10th Conference of the European Chapter of the Association for Computational Linguistics, Budapest, Hungary*. April 2003. 347-354.