De la Chambre des communes à la chambre d'isolement : adaptabilité d'un système de traduction basé sur les segments de phrases

Philippe Langlais, Fabrizio Gotti, Alexandre Patry

Université de Montréal, RALI/DIRO
{felipe ; gottif ; patryale}@iro.umontreal.ca

Résumé

Nous présentons notre participation à la deuxième campagne d'évaluation de CESTA, un projet EVALDA de l'action Technolangue. Le but de cette campagne consistait à tester l'aptitude des systèmes de traduction à s'adapter rapidement à une tâche spécifique. Nous analysons la fragilité d'un système de traduction probabiliste entraîné sur un corpus hors-domaine et dressons la liste des expériences que nous avons réalisées pour adapter notre système au domaine médical.

Mots-clés : traduction probabiliste, adaptabilité, aspiration de bitextes, mémoire de traduction.

Abstract

We present our participation in the second evaluation campaign of CESTA, an EVALDA project within the framework of Technolangue. The goal of this task consisted in testing the adaptability of translation systems. We analyze the inadequacy of a statistical phrase-based system trained on legislative texts to translate medical corpora. We describe the experiments we conducted in order to adapt our engine to the new task.

Keywords: statistical translation, adaptation, Web crawling, memory-based translation.

1. Introduction

Depuis plusieurs années, de nombreuses équipes œuvrant en traduction comparent leurs systèmes dans le cadre de campagnes d'évaluation dédiées. Malgré le nombre grandissant de ces campagnes, leur thème est presque toujours le même, soit tester l'aptitude des systèmes à traduire des dépêches d'information ou des textes législatifs.

En marge de ces campagnes démarrait en 2004 la première campagne d'évaluation du projet CESTA, partie de la plate-forme d'évaluation EVALDA de l'action Technolangue[1]. Cette campagne avait deux buts principaux : reproduire pour la langue française des évaluations comparables à celles menées par NIST pour la langue anglaise et évaluer des solutions de rechange aux métriques dites objectives couramment calculées pour évaluer la qualité d'une traduction. Les résultats de cette première campagne sont présentés dans (Surcin *et al.*, 2005).

La seconde campagne à laquelle nous avons participé récemment visait à tester l'adaptabilité

[1] http://www.technolangue.net

des systèmes participants et était organisée comme suit. Un bitexte (un texte et sa traduction alignés au niveau des phrases) représentatif du domaine de test était distribué aux participants qui étaient alors libres d'adapter leur système en conséquence, notamment à la terminologie du domaine. Dix jours plus tard, les participants recevaient un corpus à traduire (FINAL dans la suite) d'environ 200 000 mots dans lequel se cachent 20 000 mots sur lesquels portera l'évaluation.

Nous rapportons les expériences que nous avons réalisées dans le but d'adapter un système de traduction basé sur les segments au domaine médical qui était celui du test. Nous décrivons premièrement les grandes lignes de notre système (section 2). Nous résumons les efforts que nous avons menés pour obtenir différentes ressources du domaine médical (section 3). Nous montrons ensuite que la traduction d'un texte hors-domaine par notre système de traduction probabiliste pose de sérieux problèmes (section 4) et décrivons les différentes expériences que nous avons réalisées pour adapter notre système de manière aveugle (section 5) et informée (section 6), c'est-à-dire en l'absence ou en présence du texte à traduire. Nous montrons en particulier que, dans le cadre de cette campagne, notre course aux corpus s'est avérée, certes de manière fortuite, particulièrement fructueuse, et validons la pertinence de combiner mémoire de traduction phrastique et moteur de traduction statistique. Nous concluons en section 7.

2. Moteur de traduction

Dans cette étude, nous avons utilisé un système de traduction statistique basé sur les séquences de mots (*Phrase-based SMT*) appelé RAMSES (Patry *et al.*, 2006). Ce système[2] implémente la maximisation (argmax) formulée dans l'équation 1 où $p(e|f)$ est un lexique bilingue (probabilisé) de séquences de mots (transducteur), $p(f)$ est un modèle de langue trigramme et $\mathcal{F}$ représente l'ensemble des phrases françaises.

$$\hat{f} = \underset{f \in \mathcal{F}}{\mathrm{argmax}}\, p(e|f) \times p(f) \tag{1}$$

Les modèles de langue utilisés ici ont été entraînés avec l'implémentation *Kneser-Ney-discounting* de la boîte à outils SRILM (Stolcke, 2002). L'obtention du transducteur nécessite quant à lui plus d'efforts : un bitexte d'entraînement doit tout d'abord être aligné au niveau des mots. Nous utilisons à cet effet les alignements de Viterbi produits par les modèles IBM 2 de la boîte à outils GIZA++ (Och et Ney, 2000). Les paramètres du transducteur sont alors extraits des alignements de mots à l'aide d'un outil développé à l'interne[3].

Deux raisons expliquent le choix d'un tel système pour nos expériences. Primo, l'unique ressource nécessaire pour dériver un tel moteur est un bitexte. Secundo, les systèmes *phrase-based* sont les systèmes statistiques les plus performants actuellement (Koehn *et al.*, 2003).

3. La course aux corpus

Les organisateurs de la seconde campagne CESTA ont distribué 10 jours avant les tests un corpus du même domaine de spécialité que le jeu de test officiel. Les expériences décrites ici ont été menées sur 200 paires de phrases sélectionnées aléatoirement dans ce corpus (corpus

[2] RAMSES fait partie du package MOOD (http ://www-etud.iro.umontreal.ca/ patryale/mood/)

[3] FE-PHRASE est disponible sur simple demande ; THOT (Ortiz-Martínez *et al.*, 2005), un programme du domaine public offre des fonctionnalités comparables.

TEST dans la suite). Le reste, soit 722 paires de phrases, a été utilisé pour l'entraînement (corpus CESTA). Chaque concepteur de système pouvait alors à loisir (automatiquement ou non) adapter son système à ce domaine. Notre équipe a donc entrepris une course aux corpus, d'autant plus intéressante que nos ressources pour le domaine médical étaient alors fragmentaires.

Santé Canada Par simple inspection, nous avons identifié que CESTA provenait au moins en partie du site internet de Santé Canada. Nous avons donc aspiré ce site[4] à l'aide du programme `wget` et avons récupéré 20 000 pages web. En visitant ce site, nous avons également remarqué que de nombreux liens étaient faits vers le site de l'Agence de Santé publique au Canada. Nous avons donc également aspiré ce site[5] (14 000 pages web).

L'organisation bilingue de ces deux sites est exceptionnelle : presque toutes les pages sont disponibles en anglais et en français. De plus, une politique quasi-systématique de dénomination des pages permet d'identifier, dans 80 % des cas, la traduction d'une page donnée. En l'absence d'une telle organisation, nous aurions utilisé le système PARADOCS (Patry et Langlais, 2005).

Notons que nous avons cependant utilisé notre identificateur de langue SILC[6] afin d'éliminer certaines paires de documents pathologiques (textes non traduits, traduction partielle, etc.). Le texte des paires parallèles a ensuite été aligné au niveau des phrases par notre aligneur JAPA[7].

Finalement, en quelques heures, nous avons réuni un bitexte de plus de 800 000 paires de phrases, dont la qualité (mesurée par des contrôles manuels sporadiques) nous semble adéquate. Les caractéristiques principales de ce bitexte (SANTÉ) sont indiquées en table 1.

MeSH Nous avons également fait la demande du MeSH bilingue[8] , une version bilingue du thésaurus biomédical de la *National Library of Medicine*. Nous avons été surpris de la simplicité avec laquelle nous avons pu bénéficier de cette ressource : une simple demande par courriel nous a permis de l'obtenir en moins de deux jours. Ses caractéristiques sont rapportées en table 1.

Le bouche à oreille Nous avons également fait la demande à nos collègues du département de linguistique et de traductologie de corpus spécialisés dans le domaine médical. Nous avons de cette façon obtenu le corpus monolingue (français) PATRICK (voir table 1).

Sans surprise, SANTÉ semble le corpus le plus adapté à la tâche avec seulement 3.2 % de formes inconnues et moins d'un cinquième (17.5 %) des phrases de TEST en contenant. L'union de ces corpus donnent évidemment les statistiques les plus favorables : moins de 2 % ($\%\text{unk}_f^c$) des formes de TEST sont inconnues et moins de 10 % des phrases ($\%\text{unk}_p^c$) en contiennent.

4. Traduction hors-domaine

La traduction automatique d'un texte de spécialité pose différents problèmes qui ont été étudiés dans le cadre d'un moteur probabiliste basé sur les mots par (Langlais et Carl, 2004). Les auteurs identifiaient comme un problème majeur la présence de nombreux mots inconnus des modèles embarqués. Ces mots inconnus sont souvent constitutifs des termes du domaine.

Nous avons reproduit cette étude en utilisant cette fois notre moteur de traduction basé sur les segments. Deux systèmes entraînés sur des bitextes parlementaires ont été testés, l'un sur les

[4] http://www.hc-sc.gc.ca
[5] http://www.phac-aspc.sc.ca
[6] http://rali.iro.umontreal.ca/Silc/index.jsp
[7] http://rali.iro.umontreal.ca/Japa
[8] http://disc.vjf.inserm.fr

corpus	$\|S\|$	$\|e\|$	$\|f\|$	%unk$_p$	%unk$_f$	%unk$_o$	%unk^c_p	%unk^c_f
HANSARD	1 753 443	85 810	106 987	47.5	13.3	6.2	47.5	13.3
EUROPARL	901 676	92 874	106 530	51.0	13.4	5.9	39.5	10.7
SANTÉ	812 309	142 791	149 393	17.5	3.2	1.2	14.5	2.9
MESH	401 332	55 008	42 189	93.5	38.5	24.8	12.5	2.1
CESTA	722	3 192	3 631	67.0	29.2	10.5	9.0	1.5
PATRICK	45 730	—	30 565					
TEST	200	1 555	1 696					

Tableau 1. Caractéristiques principales des corpus que nous avons utilisés dans cette étude. $|S|$ désigne le nombre de phrases (resp. de paires de phrases) du corpus (resp. bitexte), $|e|$ et $|f|$ désignent respectivement le nombre de formes anglaises et françaises (décompte fait sur les textes en minuscules). %unk$_p$, %unk$_f$ et %unk$_o$ désignent respectivement, pour TEST*, le pourcentage de phrases contenant au moins un mot inconnu, de formes inconnues et d'occurrences inconnues. %unk^c_p et %unk^c_f indiquent les pourcentages de phrases et formes mesurés sur l'union des corpus : si l'on accole les 5 premiers corpus listés, seulement 9 % des phrases de* TEST *possèdent au moins un mot inconnu, et 1.5 % des formes de* TEST *sont inconnues de ce corpus.*

débats de la Chambre des communes canadienne (HANSARD), l'autre sur les débats de la commission européenne (EUROPARL). Les performances de ces deux systèmes pour traduire des phrases du domaine médical sont présentées en table 2 en termes des métriques automatiques BLEU, NIST, WER et SER.

Le lecteur est invité à lire au sujet de ces métriques (Surcin *et al.*, 2005), qui traite de la possible inadéquation de certaines d'entre elles à rendre compte de la qualité d'un système. Pour notre part, nous croyons que le calcul des ces métriques offre une solution viable à l'évaluation de systèmes imparfaits. Rappelons que SER (*Sentence Error Rate*) et WER (*Word Error Rate*) sont des taux d'erreur que l'on souhaite donc minimiser, tandis que BLEU et NIST sont deux mesures de précision que l'on cherche à maximiser.

corpus	WER	SER	NIST	BLEU	%unk$_p$	%unk$_f$	%unk$_o$
EUROPARL	67.4	100.0	4.72	13.94	63.5	16.6	7.8
HANSARD	68.2	100.0	4.89	14.96	57	16.6	7.8

Tableau 2. Performance de deux moteurs de traduction mesurée sur le corpus TEST*. %unk$_p$, %unk$_f$ et %unk$_o$ désignent respectivement les pourcentages de phrases contenant au moins un mot inconnu, de formes inconnues et d'occurrences inconnues.*

114 des 200 phrases du corpus TEST (57 %[9]) contiennent au moins un mot inconnu des modèles entraînés sur HANSARD pour un total de 258 mots inconnus (376 occurrences). Les types inconnus présents plus d'une fois dans TEST sont principalement des mots du domaine médical (les 10 types les plus fréquents sont *vivo, metabolic, whooping, renal, ocular, ars, substrates,*

[9] La différence entre les chiffres comparables des tables 1 et 2 s'explique par le bruit dans la chaîne de traitement aboutissant à un modèle (bruit souvent principalement lié à un problème d'alignement phrastique).

postoperative, *plasma* et *hepatic*). Environ 20 % des mots inconnus ne font pas partie (du moins de manière évidente) du domaine médical (principalement des mots du vocabulaire général et des données chiffrées). Le nombre de formes inconnues des modèles entraînés sur EUROPARL est de manière fortuite identique à celui observé sur HANSARD, et ce, même si les listes ne sont pas identiques (des noms propres comme *ontario* sont par exemple absents d'EUROPARL).

À titre indicatif, nous présentons à la figure 1 des exemples tirés aléatoirement de la session de traduction réalisée avec le système RAMSES entraîné sur différents corpus dont HANSARD (le meilleur des deux systèmes selon les métriques BLEU et NIST).

5. Adaptation aveugle au domaine

Nous décrivons ici nos tentatives visant à adapter automatiquement notre système au domaine, sans regard au texte que nous avions à traduire ; nous appelons cet exercice une adaptation aveugle à un domaine. Dans la table 3 sont résumées les performances de RAMSES lorsqu'entraîné sur les différents bitextes présentés en section 3. De manière prévisible, nous observons de meilleures performances pour les modèles entraînés sur des corpus d'un domaine proche de celui du jeu de test. Il est intéressant de noter que le bitexte CESTA qui ne contient que 722 paires de phrases permet d'obtenir des traductions de « meilleure qualité » que celles produites par le système entraîné sur le corpus HANSARD qui contient 1,7 million de paires de phrases.

En revanche, il est surprenant que l'ajout de MESH dans le corpus d'entraînement amène une baisse presque systématique des performances. Le vocabulaire très spécialisé (voire savant) de cette ressource est peut-être inadapté à la traduction de textes vulgarisés. La petite taille de TEST peut également expliquer cela.

L'entraînement des modèles sur l'union de tous les corpus donnerait certainement quelques améliorations, au prix de temps de calculs accrus. À l'inverse, nous pourrions chercher à sélectionner un sous-corpus d'entraînement qui maximise la taille des vocabulaires source et cible, comme cela a été proposé par (Eck *et al.*, 2005). Nous verrons en section 6.3 que nous avions mieux à faire.

corpus	WER	SER	NIST	BLEU	corpus	WER	SER	NIST	BLEU
EUROPARL	67.4	100.0	4.72	13.94	CESTA	65.2	97.0	4.85	16.19
+ MESH	68.5	100.0	4.66	13.87	+ MESH	62.5	97.0	5.03	16.89
+ PATRICK	66.9	99.5	4.83	14.67	+ PATRICK	65.7	96.0	4.73	15.76
HANSARD	68.2	100.0	4.89	14.96	SANTÉ	48.5	93.5	6.78	34.42
+ MESH	67.1	100.0	4.83	14.88	+ MESH	50.3	95.5	6.53	32.59
+ PATRICK	67.7	100.0	4.88	15.32	+ PATRICK	48.3	94.0	6.77	34.85

Tableau 3. Performances de RAMSES *lorsqu'entraîné sur les différents bitextes*

6. Adaptation informée

La section précédente a montré l'importance d'obtenir un bitexte d'entraînement proche du domaine des textes à traduire. Nous cherchons ici à savoir si l'on peut affiner la sélection d'un corpus d'entraînement en prenant cette fois en considération le texte à traduire. Seul le modèle de traduction a fait l'objet de ré-entraînements ; le modèle de langue le plus performant, partagé par tous les systèmes testés, est obtenu en combinant le corpus SANTÉ et le corpus PATRICK.

Sélectionner un corpus proche d'un texte à traduire à partir d'un *corpus de base* peut être abordé comme un problème de recherche d'information (Hildebrand *et al.*, 2005). Les phrases du texte à traduire sont des requêtes dans une base constituée des paires de phrases du corpus de base. Les paires dont la phrase source est proche d'une phrase à traduire sont simplement versées dans le corpus d'entraînement du moteur de traduction.

Le corpus de base utilisé ici est constitué de la concaténation des bitextes SANTÉ, EUROPARL et HANSARD, soit un total de 3,4 millions de paires de phrases. Deux techniques éprouvées de recherche d'information ont été étudiées ici : les modèles de langue et le modèle vectoriel.

6.1. Recherche des phrases proches par modèle de langue

Depuis les travaux de (Ponte et Croft, 1998), nous savons que l'utilisation de modèles de langue pour identifier des documents proches d'une requête dans une collection donnent des résultats état de l'art. Si quelques études montrent qu'il est possible d'entraîner des modèles bigrammes (adjacents ou non) sur des documents de petite taille (Alvarez *et al.*, 2004), nombreux sont les chercheurs à s'intéresser aux modèles unigrammes, ce que nous faisons ici.

Chacune des $N = 3,4$ millions de phrases anglaises de notre corpus de base est représentée par un modèle unigramme entraîné par fréquence relative. Afin de contourner le problème aigu de sous-représentation de données (une phrase anglaise contient en moyenne 19,12 mots dans notre corpus), nous lissons chaque modèle à l'aide d'un modèle unigramme entraîné sur l'ensemble du corpus de base. Formellement :

$$p_i(w) = \lambda \frac{{}_i|w|}{N_i} + (1-\lambda)\frac{\sum_{j=1}^{N} {}_j|w|}{\sum_{j=1}^{N} N_j} \quad (2)$$

où i désigne une des N phrases du corpus de base, N_i sa longueur (comptée en mots) et ${}_i|w|$ désigne la fréquence de w dans la phrase i. Dans cette étude, nous avons fixé λ à 0.8.

Pour chaque phrase s^j à traduire, nous sélectionnons les n phrases les plus proches $v^j_{1..n}$, c'est-à-dire celles dont le modèle unigramme associé note le mieux la phrase s^j selon l'équation 3. Nous réunissons toutes ces phrases en un corpus d'entraînement. Puisque nous avons 200 phrases à traduire dans TEST, nous obtenons $200 \times n$ phrases d'entraînement.

$$v^j_{1..n} = \underset{i\in[1,N]}{\text{argmax}_{1..n}}\, p_i(s^j) = \underset{i\in[1,N]}{\text{argmax}_{1..n}} \prod_{k=1}^{|s^j|} p_i(s^j_k) \quad (3)$$

Les performances de cette approche à la sélection d'un corpus d'entraînement sont résumées dans la colonne de gauche de la table 4, pour la traduction du corpus TEST. On observe que, pour $n = 1$ (un corpus d'entraînement de seulement 200 phrases), nous obtenons des performances comparables à celles mesurées sur le corpus CESTA de 722 paires de phrases du domaine. Les performances obtenues en retenant de plus en plus de phrases pour chaque phrase à traduire semblent plafonner autour de celles mesurées pour le système entraîné sur le corpus SANTÉ au complet. Le meilleur système que nous avons entraîné semble celui obtenu en retenant les 5 000 phrases les plus proches de chaque phrase à traduire, ce qui constitue un bitexte de 544 388 paires de phrases (des mêmes phrases sont sélectionnées pour plusieurs phrases à traduire). Si les performances ne sont pas significativement meilleures que celles obtenues en entraînant le système sur les 812 309 paires de phrases du corpus SANTÉ, la réduction de la taille du

	modèle unigramme				modèle vectoriel			
n	WER	SER	NIST	BLEU	WER	SER	NIST	BLEU
1	68.8	95.5	4.48	16.55	64.9	96.0	4.84	17.71
5	60.7	96.0	5.74	23.13	60.3	96.5	5.81	24.92
10	55.9	95.5	6.05	26.48	53.8	96.5	6.17	27.59
100	50.5	96.5	6.68	32.61	51.1	96.5	6.73	32.24
1000	48.8	95.5	6.74	34.41	50.4	95.5	6.78	34.21
5000	49.3	96.5	6.81	34.91	49.6	94.0	6.81	34.43

Tableau 4. Performance de RAMSES *en fonction du nombre des n meilleures paires retenues par phrase source à traduire, soit par un modèle de langue unigramme, soit par un modèle vectoriel*

modèle de traduction sous-jacent présente un intérêt pratique, notamment dans les systèmes de traduction embarqués. Les performances semblent de plus plafonner dès la sélection de 1 000 paires de phrases par phrase à traduire.

6.2. Recherche des phrases proches par modèle vectoriel

Afin de mesurer l'impact de la technique permettant d'identifier les phrases les plus proches dans le corpus de base des phrases à traduire nous avons également testé l'approche populaire du modèle vectoriel. Dans notre contexte, chaque phrase s_i du corpus de base est représentée par un vecteur v_d dont la dimension est égale au vocabulaire (source) de l'application. Chaque coefficient de ce vecteur correspond à un mot particulier de ce vocabulaire et mesure grossièrement le pouvoir discriminant de ce mot. Nous utilisons ici le critère $tf \cdot idf$ qui stipule qu'un mot est d'autant plus caractéristique d'un document (ici une phrase) qu'il est fréquent dans ce document et peu fréquent en général.

Chaque phrase à traduire est représentée dans cet espace vectoriel et les n vecteurs les plus proches au sens d'une mesure de cosinus désignent les phrases sources qui seront retenues avec leur traduction pour l'entraînement du système. Nous avons utilisé l'indexeur et le moteur de recherche de LUCENE[10] pour réaliser cette expérience. Les résultats sont consignés dans la colonne de droite de la table 4. Globalement, cette seconde approche semble offrir de meilleures performances que l'approche unigramme, plus particulièrement lorsque n est faible. Ceci peut s'expliquer par le faible pouvoir discriminant de la méthode de lissage que nous avons utilisée.

6.3. Utilisation d'une mémoire de traduction phrastique

Comme nous l'avons mentionné à la section 3, nous nous trouvons dans une situation où des phrases à traduire du corpus CESTA et celles du corpus FINAL se trouvent dans notre corpus SANTÉ. Nous avons donc constitué une mémoire de traduction phrastique à partir du corpus SANTÉ pour traduire FINAL. Cette mémoire est interrogée pour chaque phrase à traduire, et si elle y est retrouvée, son pendant anglais est récupéré.

En ignorant les espaces entre les mots, nous faisons abstraction des différences de segmentation des mots ; 66,3 % des phrases de FINAL sont ainsi repérables *verbatim* dans la mémoire. Les 33,7 % restants n'ont pas été identifiés à cause de différences de segmentation du texte original en phrases entre nos outils et ceux des organisateurs. Cette disparité est particulièrement impor-

[10] http://lucene.apache.org

tante pour les énumérations de plusieurs points, pour lesquelles nos outils considèrent chaque point comme une phrase alors que FINAL les concatène en une seule.

Après harmonisation partielle de ces segmentations, nous obtenons une couverture de 88,2 % de FINAL par notre mémoire. Sa création et le développement des programmes qui l'interrogent a requis 20 heures-personnes. Un examen manuel suggère que d'importants efforts supplémentaires pour réconcilier les techniques de segmentation des phrases auraient pu permettre une couverture de l'ordre de 95 %, ce qui dénote l'importance des outils de segmentation lors de la création de ce genre de mémoires.

Ultimement, nous avons donc traduit FINAL à l'aide d'un hybride entre une mémoire de traduction rudimentaire et les outils de traduction statistique décrits dans cette étude. Ce dernier produit une traduction chaque fois que la mémoire est muette (11,8 % des phrases).

7. Discussion

Nous avons analysé l'inadéquation d'un système statistique de traduction basé sur les segments lorsqu'il est employé à traduire des textes hors-domaine. Nous avons montré qu'il était assez simple d'adapter notre système au domaine médical, et ce, grâce au fait que ce domaine est bien représenté pour la paire de langues anglais-français sur la toile.

Nous avons étudié le couplage de la recherche d'information et de la traduction et avons observé qu'il était pertinent dans le cas de petits corpus et autorisait une réduction importante des corpus d'entraînement à qualité égale, voire supérieure. Nos conclusions à ce sujet sont compatibles avec celles faites par (Hildebrand *et al.*, 2005), et ce, même si notre quête de corpus s'est avérée anormalement fructueuse (plus de 80 % de couverture phrastique).

Nous avons également observé que la combinaison d'une mémoire de traduction phrastique et de notre moteur de traduction était des plus attrayantes. Cela corrobore les résultats obtenus dans le cadre des traductions des bulletins météorologiques d'environnement Canada (Langlais *et al.*, 2005) ; une autre situation où la mémoire était particulièrement adaptée à la tâche.

Le fait que nous ayons réussi à obtenir un corpus d'entraînement particulièrement adapté à la tâche peut introduire un biais dans les résultats présentés. Nous l'admettons. Cependant, cette situation n'est pas éloignée de celles des campagnes d'évaluations où le corpus de test est habituellement issu de textes apparentés à ceux utilisés pour l'entraînement.

Références

ALVAREZ C., LANGLAIS P. et NIE J.-Y. (2004). « Word Pairs in Language Modeling for Information Retrieval ». In *7th Conference on RIAO*. Avignon, France, p. 686-705.

ECK M., VOGEL S. et WAIBEL A. (2005). « Low Cost Portability for Statistical Machine Translation based on N-gram Coverage ». In *Tenth Machine Translation Summit*. Phuket, Thailand, p. 227–324.

HILDEBRAND A. S., ECK M., VOGEL S. et WAIBEL A. (2005). « Adaptation of the Translation Model for Statistical Machine Translation based on Information Retrieval ». In *Proceedings of the EAMT Conference*. Budapest, Hugary, p. 133–142.

KOEHN P., OCH F. et MARCU D. (2003). « Statistical Phrase-Based Translation ». In *Proceedings of the HLT-NAACL Conference*. Edmonton, Canada, p. 127–133.

LANGLAIS P. et CARL M. (2004). « General-purpose Statistical Translation Engine and Do-

main Specific Texts : Would it Work ? ». In *Terminology*, 10(1), 131–153.

LANGLAIS P., GANDRABUR S., LEPLUS T. et LAPALME G. (2005). « The Long-Term Forecast for Weather Bulletin Translation ». In *Machine Translation*. À paraître.

OCH F. et NEY H. (2000). « Improved Statistical Alignment Models ». In *Proceedings of the ACL Conference*. Hongkong, p. 440–447.

ORTIZ-MARTÍNEZ D., GARCÍA-VAREA I. et CASACUBERTA F. (2005). « Thot : a Toolkit To Train Phrase-based Statistical Translation Models ». In *Tenth Machine Translation Summit*. Phuket, Thailand, p. 141–148.

PATRY A., GOTTI F. et LANGLAIS P. (2006). « MOOD : A Modular Object-Oriented Decoder for Statistical Machine Translation ». In *LREC*. Gênes, Italie. À paraître.

PATRY A. et LANGLAIS P. (2005). « Paradocs : un système d'identification automatique de documents parallèles ». In *12e Conference sur le Traitement Automatique des Langues Naturelles (TALN)*. Dourdan, France, p. 223-232.

PONTE J. M. et CROFT W. B. (1998). « A Language Modeling Approach to Information Retrieval ». In *21st annual international ACM SIGIR conference on Research and Development in Information Retrieval*. Melbourne, Australia, p. 275-281.

STOLCKE A. (2002). « SRILM - An Extensible Language Modeling Toolkit ». In *Proceedings of the ICSLP Conference*. Denver, USA.

SURCIN S., HAMON O., HARTLEY A., RAJMAN M., POPESCU-BELIS A., HADI W. M. E., TIMIMI I., DABADDIE M. et CHOUKRI K. (2005). « Evaluation of Machine Translation with Predictive Metrics beyond BLEU/NIST : CESTA Evaluation Campaign #1 ». In *Tenth Machine Translation Summit*. Phuket, Thailand, p. 117–124.

Src : *tips : put your baby on his or her back to sleep .*	
Ref : *conseils : couchez votre bébé sur le dos .*	
HANSARD	pratiques : mettez votre baby sur son retour à dormir
CESTA	tips : put votre bébé sur his ou s' back à sleep .
SANTÉ	tips : couchez votre bébé sur le dos pour dormir .
Src : *attention : cardiologists*	
Ref : *à l' attention des : cardiologues*	
HANSARD	attention cardiologists :
CESTA	l' attention cardiologists :
SANTÉ	attention : cardiologues
Src : *autopsy revealed signs of intrauterine growth retardation and hypoxia but no malformations .*	
Ref : *l' autopsie a révélé des signes de retard de croissance intra-utérine et d' hypoxie , mais aucune malformation .*	
HANSARD	autopsie révèle des signes de croissance intrauterine arriération et hypoxia , mais ce n' malformations .
CESTA	autopsy revealed symptômes de intra-utérine growth retardation et hypoxia mais aucun malformations .
SANTÉ	l' autopsie a révélé des signes de croissance intra-utérine et hypoxie (mais non malformations .
Src : *the vaccine for whooping cough is safe and effective*	
Ref : *le vaccin contre la coqueluche est sûr et efficace*	
HANSARD	le vaccin pour toux aspirante est sûrs et efficaces
CESTA	le vaccin contre la coqueluche est sûr et efficace
SANTÉ	le vaccin contre la coqueluche est sûrs et efficaces
Src : *it is estimated that every day in canada , at least one child is born with fetal alcohol syndrome (fas) which can lead to a variety of life long disabilities .*	
Ref : *on estime que chaque jour , au canada , au moins un enfant naît avec le syndrome d' alcoolisme foetal (saf) qui peut mener à une variété d' incapacités permanentes .*	
HANSARD	on estime que tous les jours au canada , au moins un enfant est né le syndrome d' alcoolisme dd) , qui peut conduire à une variété de la vie des personnes handicapées . longtemps avec fetal cpsa
CESTA	il est coût que chaque journée au canada , à least un enfant est born avec syndrome d' alcoolisme foetal (saf) peut mènera à un variety de life long lequel disabilities .
SANTÉ	on estime que chaque jour au canada , au moins un enfant naît avec le syndrome d' alcoolisme foetal (saf) , ce qui peut mener à une variété de la vie à long apprentissage .

Figure 1. Traductions produites par RAMSES *entraîné sur différents corpus.*