Etude de l'impact d'un lexique bilingue spécialisé sur la performance d'un moteur de traduction à base d'exemples

Nasredine Semmar Othman Zennaki Meriama Laib
CEA, LIST, Laboratoire Vision et Ingénierie de Contenus, F-91191, Gif-sur-Yvette, France
nasredine.semmar@cea.fr, othman.zennaki@cea.fr, meriama.laib@cea.fr

RÉSUMÉ

La traduction automatique statistique bien que performante est aujourd'hui limitée parce qu'elle nécessite de gros volumes de corpus parallèles qui n'existent pas pour tous les couples de langues et toutes les spécialités et que leur production est lente et coûteuse. Nous présentons, dans cet article, un prototype d'un moteur de traduction à base d'exemples utilisant la recherche d'information interlingue et ne nécessitant qu'un corpus de textes en langue cible. Plus particulièrement, nous proposons d'étudier l'impact d'un lexique bilingue de spécialité sur la performance de ce prototype. Nous évaluons ce prototype de traduction et comparons ses résultats à ceux du système de traduction statistique Moses en utilisant les corpus parallèles anglais-français Europarl (European Parliament Proceedings) et Emea (European Medicines Agency Documents). Les résultats obtenus montrent que le score BLEU du prototype du moteur de traduction à base d'exemples est proche de celui du système Moses sur des documents issus du corpus Europarl et meilleur sur des documents extraits du corpus Emea.

ABSTRACT

Studying the impact of a specialized bilingual lexicon on the performance of an example-based machine translation engine

Non-availability of parallel corpora for several languages and for specific domains is a major challenge for domain adaptation in statistical machine translation. We present, in this paper, an Example-Based Machine Translation engine based on cross-language information retrieval and which needs only a monolingual corpus in the target language. In particular, we investigate the impact of using a domain-specific bilingual lexicon on the performance of this translation engine. We evaluate and compare this translation engine to the statistical machine translation system Moses using English-French parallel corpora Europarl (European Parliament Proceedings) and Emea (European Medicines Agency Documents). The obtained results show that the BLEU score of the Example-Based Machine Translation engine is close to the ones of Moses for the Europarl corpus and better for the Emea corpus.

MOTS-CLÉS : Traduction automatique, recherche d'information interlingue, lexique bilingue, modèle de traduction, modèle de langue, automate d'états finis, champs conditionnels aléatoires.

KEYWORDS: Machine translation, cross-language information retrieval, bilingual lexicon, translation model, language model, finite-state machine, conditional random fields.

1 Introduction

La performance de la Traduction Automatique Statistique (TAS) dépend considérablement de la qualité et de la quantité des corpus parallèles d'apprentissage disponibles. Ces corpus n'existent pas pour tous les couples de langues et toutes les spécialités et leur production est lente et coûteuse. Les sources des corpus d'apprentissage actuellement disponibles sont souvent issues d'organisations internationales (Organisation des Nations Unies, Parlement Européen, etc.) et concernent des domaines d'application généralistes. Cela a un impact très important sur la qualité de la traduction de documents de spécialité. Plusieurs travaux ont montré que les systèmes construits à partir de données d'apprentissage généralistes ne sont pas appropriés pour traduire des documents dans des domaines spécifiques. Nous présentons, dans cet article, un prototype d'un moteur de traduction à base d'exemples utilisant la recherche d'information interlingue et ne nécessitant qu'un corpus de textes en langue cible. Nous montrons, en particulier, que l'ajout d'un lexique bilingue spécialisé au dictionnaire général de ce prototype améliore significativement la qualité de traduction des textes du domaine de spécialité sans perte de qualité de traduction pour les textes du domaine général.

La suite de l'article est organisée comme suit : dans la section 2, nous présentons un bref état de l'art sur les approches de traduction automatique et d'adaptation multi-domaines en traduction statistique. Puis nous décrivons dans la section 3, les principaux modules et étapes composant le processus de traduction utilisé par notre prototype de traduction à base d'exemples. La section 4 est consacrée aux expérimentations effectuées ainsi que la présentation des résultats obtenus et la section 5 conclut notre étude et présente nos travaux futurs.

2 Etat de l'art

Il existe principalement deux approches pour la traduction automatique: celle à base de règles et celle s'appuyant sur des corpus (Hutchins, 2003). La combinaison de ces approches a permis le développement de systèmes hybrides dont l'objectif est de parvenir à tirer profit des avantages de chaque approche (Simard et al., 2007). La traduction automatique à base de règles privilégie le traitement linguistique. Cette approche repose généralement sur un traitement à trois phases. Une phase d'analyse qui calcule une représentation syntaxique de la phrase source. Une phase de transfert qui transforme cette représentation syntaxique en une représentation correspondante en langue cible et une phase de synthèse qui, à partir de la structure syntaxique résultat du transfert, produit la phrase cible. Une variante interlingue de cette approche consiste à calculer une représentation syntaxico-sémantique suffisamment abstraite pour être indépendante de toute langue. Le principal avantage de l'approche à base de règles est qu'elle fournit des résultats présentant un minimum de qualité lexicale et grammaticale due à l'utilisation de ressources linguistiques monolingues et bilingues (dictionnaires, règles de grammaire, etc.) mais ces ressources sont coûteuses car généralement construites à la main. La traduction à base de corpus est issue de l'idée de la possibilité d'utiliser des traductions existantes pour traduire de nouveaux textes. Deux approches sont nées de cette idée : La traduction statistique ou probabiliste et la traduction à base d'exemples. La traduction statistique (Brown et al., 1990) utilise un corpus de textes bilingues pour apprendre le modèle de traduction et un corpus monolingue pour apprendre le modèle de langue. Les deux modèles appris servent ensuite à calculer la probabilité qu'une phrase donnée en langue cible soit une traduction de la phrase source. Plusieurs méthodes sont utilisées en modélisation : les méthodes basées sur les mots, sur les séquences de mots et enfin sur la syntaxe de la phrase. La traduction à base d'exemples (Somers, 1999) consiste à parcourir une base de données de phrases et génère une traduction composée de fragments communs (des groupes de mots) avec la phrase à traduire. Ces

approches à base de corpus sont efficaces lorsque, d'une part, les langues source et cible ont une morphologie proche, et d'autre part, les corpus utilisés en apprentissage ont une taille suffisante.

Plusieurs idées ont été explorées pour l'adaptation de la traduction automatique statistique à un domaine cible particulier. (Langlais, 2002) a intégré des lexiques spécialisés dans le modèle de traduction d'un moteur de traduction statistique. Cet ajout a permis une amélioration de la qualité de traduction des documents du domaine de spécialité. (Lewis et al., 2010) ont développé un système de traduction statistique spécifique à un domaine en mettant en commun toutes les données d'apprentissage dans un grand groupe de données, en incluant de plus en plus de données monolingues du domaine. Ils ont entrainé des modèles très spécifiques à la langue sur les données monolingues du domaine afin de réduire l'effet d'amortissement des données hétérogènes sur la qualité de traduction. (Hildebrand et al., 2005) ont utilisé une approche qui consiste essentiellement à choisir des échantillons du corpus d'apprentissage qui ressemblent le plus aux données de test pour améliorer la qualité de traduction lors du changement du domaine. (Bertoldi, Federico, 2009) ont utilisé des corpus monolingues et (Snover et al., 2008) ont utilisé des corpus comparables pour adapter des systèmes de traduction statistique appris sur les corpus Europarl pour traduire des dépêches. Les résultats obtenus ont montré un gain significatif en performance. (Banerjee et al., 2010) ont combiné deux modèles de domaines distincts. Chaque modèle est appris à partir de petites quantités de données spécifiques à un domaine. Ces données sont collectées à partir d'un seul site Web. Les auteurs ont utilisé des techniques de filtrage de documents et de classification pour réaliser la détection automatique de domaines. (Daumé, Jagarlamudi, 2011) ont utilisé des techniques de fouille de données pour trouver des traductions pour les mots inconnus à partir de corpus comparables et ils ont intégré ces traductions dans un système de traduction statistique. L'ajout de ces traductions a permis d'améliorer la qualité de la traduction (entre 0,5 et 1,5 points BLEU) sur quatre domaines et deux paires de langues. (Pecina et al., 2011) ont exploité des corpus spécifiques acquis en crawlant le web dans le but d'adapter un système de traduction statistique à de nouveaux domaines. Ils ont observé que l'impact de petites quantités de corpus parallèles dans le domaine est plus important pour la qualité de la traduction que de grandes quantités de corpus monolingues. (Wang et al., 2012) ont utilisé un modèle de traduction unique et ont généralisé un décodeur conçu pour un seul domaine pour traiter différents domaines. Ils ont utilisé cette méthode pour adapter un système de traduction statistique généraliste pour traduire des brevets dans 20 paires de langues. Les auteurs ont rapporté un gain de 0,35 points BLEU pour la traduction des brevets et une perte de seulement 0,18 points BLEU pour la traduction de documents dans le domaine général.

Notre approche pour l'adaptation au domaine est proche de celle décrite dans (Langlais, 2002), mais propose l'intégration du lexique bilingue spécialisé dans les deux principaux composants du système de traduction à base d'exemples: le moteur de recherche interlingue et le reformulateur bilingue.

3 Traduction automatique basée sur la recherche d'information interlingue

L'approche que nous avons développée pour implémenter notre prototype de traduction à base d'exemples (nous utilisons par la suite l'acronyme EBMT -Example-Based Machine Translation- pour désigner ce prototype) consiste, d'une part, à construire une base de données de phrases en langue cible et considérer chaque phrase à traduire comme une requête en langue source à cette base, et d'autre part, à combiner les réponses fournies par le moteur de recherche interlingue avec le résultat d'un reformulateur bilingue en vue de générer la traduction de la requête (Figure 1). Pour illustrer le fonctionnement de notre prototype de traduction à base d'exemples (Semmar et al.,

2015), nous avons indexé un corpus composé de 1127 phrases en langue cible (français) issue du corpus anglais-français du projet Arcade II (Véronis et al., 2008) et nous avons considéré la phrase « Social security funds in Greece encourage investment in innovation. » comme texte à traduire.

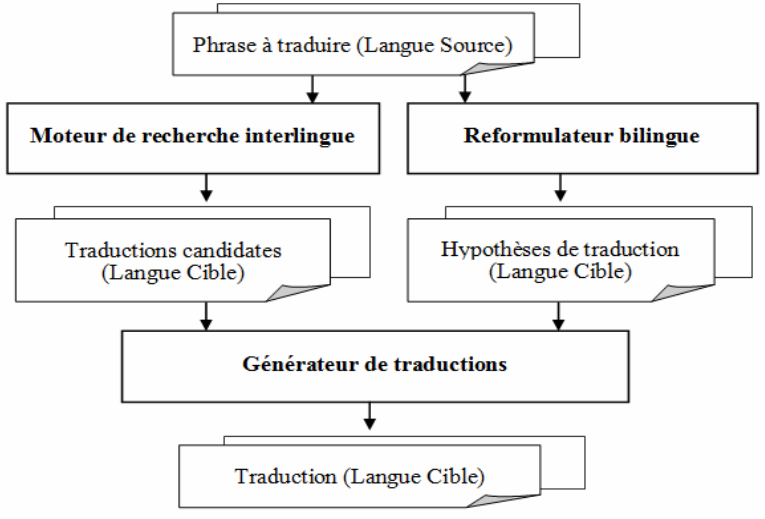

FIGURE 1: Principaux composants du prototype de traduction à base d'exemples

3.1 Moteur de recherche interlingue

La recherche d'information interlingue consiste, à partir d'une requête en une seule langue, à fournir des réponses trouvées dans des documents qui sont dans d'autres langues (Grefenstette, 1998). Dans notre utilisation de la recherche d'information interlingue en traduction automatique, un document correspond à une phrase. Le rôle du moteur de recherche interlingue de notre prototype de traduction est d'extraire pour chaque phrase à traduire (la requête de l'utilisateur) des phrases ou des sous-phrases depuis un corpus monolingue indexé dans la langue cible. Ces phrases ou sous-phrases correspondent à une traduction totale ou partielle de la phrase à traduire. Ce moteur de recherche interlingue est basé sur une analyse linguistique profonde de la requête et du corpus monolingue à indexer et utilise un modèle vectoriel pondéré (Salton, McGill, 1986). Il est composé :

1. d'un analyseur linguistique basé sur la plate-forme libre LIMA[1] (Besançon et al., 2010) composé d'un analyseur morphologique, d'un désambiguïseur morpho-syntaxique et d'un analyseur syntaxique. Cet analyseur linguistique permet d'identifier les mots présents dans la phrase requête et les phrases à indexer (mots simples ou composés), de fournir leurs lemmes avec leur information linguistique (étiquette morpho-syntaxique, genre, nombre, etc.) ainsi que leurs relations de dépendance syntaxique. Ces relations sont détectées par l'analyseur syntaxique en utilisant des automates d'états finis définis à l'aide de règles à base d'expressions régulières exprimant les successions possibles d'étiquettes morpho-syntaxiques.
2. d'un analyseur statistique qui consiste à attribuer un poids aux mots simples et aux mots composés de la requête et de l'ensemble des phrases à indexer. Le poids d'un mot représente à la fois son importance et le fait qu'il est discriminant ou non. Pour calculer les poids des

[1] https://github.com/aymara/lima/wiki

mots, nous avons utilisé une combinaison entre la fréquence du mot (TF) et sa fréquence documentaire inverse (IDF). Le poids w_{ij} du mot *j* dans la phrase *i* est défini avec la formule $w_{ij}=tf_{ij}logN/n_j$, où tf_{ij} est la fréquence du mot *j* dans la phrases *i*, *N* est le nombre total des phrases à indexer, et n_j est le nombre de phrases dans lesquelles le mot *j* apparaît.

3. d'un indexeur utilisant la librairie Lucene[2] pour construire la base de données textuelles composée des phrases à indexer en langue cible en se basant sur le résultat de l'analyseur linguistique.
4. d'un comparateur qui consiste à calculer la similarité entre le vecteur de la phrase requête et les vecteurs des phrases indexées en mesurant le cosinus de l'angle formé par ces vecteurs. Les phrases en langue cible pertinentes pour la requête sont triées par ordre décroissant de similarité. Les correspondances de vecteurs permettent des correspondances partielles entre les phrases de la base de données textuelles et la requête, ce qui permet de retrouver des phrases qui ne satisfont la phrase requête qu'approximativement.
5. d'un reformulateur qui permet l'expansion de la phrase requête pendant la recherche. La reformulation consiste à formuler la requête autrement en remplaçant les mots qui la composent par des variantes afin de récupérer des phrases pertinentes dans lesquelles les mots saisis ne sont pas toujours présents. Cette reformulation se fait à l'aide de dictionnaires monolingues qui permettent la reformulation dans une même langue (synonymes, antonymes, etc.) et les dictionnaires bilingues qui permettent la reformulation dans des langues différentes.

Par exemple, pour la requête (phrase à traduire) « Social security funds in Greece encourage investment in innovation. », deux chaînes nominales ont été identifiées par l'analyseur syntaxique : « Social security funds in Greece » et « investment in innovation ». A partir de la première chaîne nominale, l'analyseur syntaxique a reconnu trois mots composés: Social security funds in Greece (Greece_fund_security_social), Social security funds (fund_security_social) et Social security (security_social). Le tableau ci-dessous (Table 1) montre les deux premières traductions candidates fournies par le moteur de recherche interlingue pour cette requête.

Classe	Termes de la requête	Traduction candidate
1	fund_security_social, Greece, investment	Les caisses de sécurité sociale de Grèce revendiquent l'indépendance en matière d'investissements.
2	fund_security_social	Objet: Caisses de sécurité sociale grecques.

TABLE 1 : Traductions candidates retournées par le moteur de recherche interlingue pour la requête (phrase à traduire) « Social security funds in Greece encourage investment in innovation. »

Les traductions candidates ont en commun avec la requête les termes « fund_security_social, Greece, investment » pour la première classe et les termes « fund_security_social » pour la deuxième classe. De plus, le moteur de recherche interlingue fournit les informations linguistiques (étiquette morpho-syntaxique, genre, nombre, etc.) pour chaque mot des traductions candidates. Ces traductions candidates sont représentées sous forme de graphes de mots et encodées à l'aide d'automates d'états finis. Chaque transition de l'automate correspondant au lemme du mot et ses informations linguistiques (Figure 2).

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

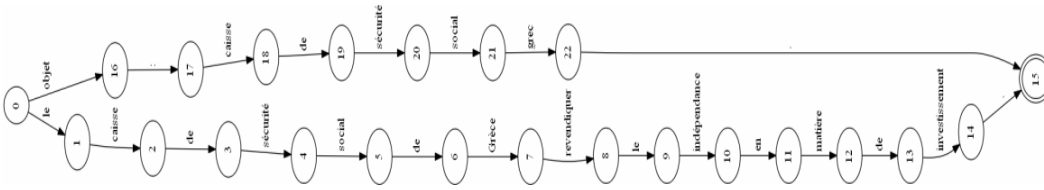

FIGURE 2: Automates d'états finis représentant les deux premières traductions candidates

3.2 Reformulateur bilingue

Le rôle du reformulateur bilingue est de produire pour chaque phrase à traduire un ensemble d'hypothèses de traduction. Le processus de reformulation consiste, d'une part, à transformer dans la langue cible la structure syntaxique de la phrase à traduire, et, d'autre part, à traduire ses mots. Le reformulateur est basé sur des automates d'état finis, un ensemble de règles linguistiques pour transformer les structures syntaxiques de la langue source vers la langue cible (Transfert syntaxique) et un dictionnaire bilingue pour traduire les mots de la phrase à traduire (Transfert lexical). Les règles de transfert syntaxique sont construites à la main et sont fondées sur des patrons morpho-syntaxiques (Table 2).

N° de la règle	Succession d'étiquettes du patron morpho-syntaxique (Langue source)	Succession d'étiquettes du patron morpho-syntaxique (Langue cible)
1	AN	NA
2	ANN	NNA
3	NN	NN
4	AAN	NAA
5	NAN	NNA
6	NPN	NPN
7	NNN	NNN
8	ANPN	NAPN
9	NPAN	NPNA
10	TN	TN

TABLE 2: Patrons morpho-syntaxiques fréquents utilisés pour transformer les structures syntaxiques de la phrase à traduire de l'anglais vers le français. Dans ces patrons, A indique un Adjectif, P pour Préposition, T pour Participe Passé et N pour Nom

Les expressions qui correspondent à chaque patron sont identifiées par l'analyseur syntaxique lors de la phase de reconnaissance des chaines nominales et verbales. Ces expressions déterminent les phrases acceptées par l'automate d'états finis dont les sorties sont des instances de ces phrases dans la langue cible. Par exemple, l'analyseur syntaxique a identifié à partir de la phase à traduire « Social security funds in Greece encourage investment in innovation. », deux chaines nominales : « Social security funds in Greece » et « investment in innovation ». Ces chaines nominales sont reliées par le verbe « encourage ». Une transformation possible du mot composé « social security funds » est l'expression « the funds of the security social » en utilisant la deuxième règle (Table 2). Il est important de mentionner ici que lorsque deux noms communs se suivent (security funds), nous

pouvons ajouter à la règle sélectionnée une préposition, une préposition accompagnée d'un déterminant ou un déterminant tout en gardant la possibilité d'ignorer ces ajouts. Ainsi, les mots de liaison « the » (déterminant) et « of » (préposition) ont été ajoutés à la deuxième règle afin de compléter la transformation. L'automate d'états finis de l'étape du transfert syntaxique produit un treillis de mots (graphe de mots acyclique orienté) en langue source. Chaque mot est représenté avec son lemme dans le treillis et est associé avec ses informations linguistiques (étiquette morpho-syntaxique, genre, nombre, etc.). L'étape du transfert lexical consiste à traduire en langue cible les lemmes des mots des structures syntaxiques en utilisant le dictionnaire bilingue du moteur de recherche interlingue. Le dictionnaire anglais-français est composé de 243539 entrées[3]. Ces entrées sont représentées dans leurs formes normalisées (lemmes). Pour obtenir les lemmes des mots des structures syntaxiques de la première étape de la reformulation, nous avions utilisé le lemmatiseur fourni par la plate-forme d'analyse linguistique LIMA. Cette étape produit un nombre important d'hypothèses de traduction dû à la combinaison des règles de transfert syntaxique et la polysémie dans le dictionnaire bilingue (Figure 3). Le résultat du reformulateur bilingue est un ensemble de treillis dans lesquels les mots sont en langue cible. Pour récupérer les meilleures hypothèses de traduction, ce treillis de mots a été évalué en utilisant les champs markoviens conditionnels ou CRF (Lafferty et al., 2001). Cette évaluation est effectuée en utilisant un modèle de langue appris sur le corpus monolingue de la langue cible annoté en lemmes et en étiquettes morpho-syntaxiques. Notons que cette évaluation a été possible après que chaque chemin du treillis de mots a été transformé au « format CRF » puisque les CRF ne peuvent pas opérer directement sur des graphes.

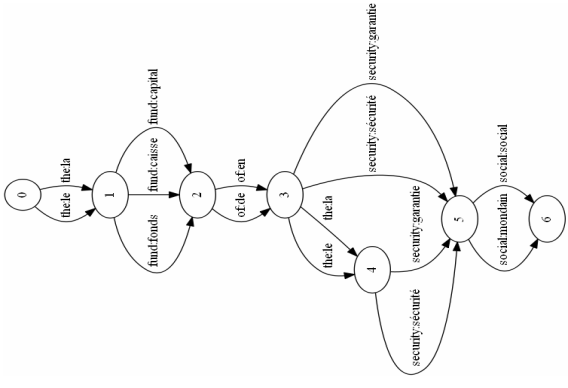

FIGURE 3: Une partie du treillis de mots correspondant à la transformation syntaxique et lexical entre l'anglais et le français du mot composé "Social security funds"

3.3 Générateur de traductions

Le rôle du générateur de traductions est de produire les n-meilleures traductions en utilisant les traductions candidates fournies par le moteur de recherche interlingue, les hypothèses de traduction produites par le reformulateur bilingue et le modèle de langue appris à partir du corpus en langue cible. Le processus de génération de traductions se déroule en deux étapes :

1. La première étape consiste à composer les automates d'états finis correspondants aux traductions candidates retournées par le moteur de recherche interlingue avec les automates d'états finis correspondants aux hypothèses de traduction produites par le reformulateur

[3] http://catalog.elra.info/product_info.php?products_id=666

bilingue. Les états où la composition doit être réalisée sont déterminés par les mots qui relient les chaînes nominales des traductions candidates aux chaînes nominales identiques ou équivalentes dans les hypothèses de traduction. Dans notre exemple, le mot « encourage » qui relie les deux patrons morpho-syntaxiques impliqués dans la transformation syntaxique de la phrase à traduire, et le mot « revendiquer » qui relie les deux chaînes nominales de la première traduction candidate (Table 1) déterminent ces états. Toutes les opérations sur les automates d'états finis des traductions candidates et hypothèses de traduction ont été réalisées en utilisant la librairie AT&T FSM Library[4] (Mohri et al., 2002).

2. La deuxième étape permet de sélectionner les n-meilleures traductions obtenues à partir d'un modèle de langue appris sur un corpus en langue cible. Pour trouver les lemmes des n-meilleures traductions à partir du résultat de composition d'automates de la première étape, nous avons construit un modèle de langue à l'aide du corpus monolingue lemmatisé en langue cible. Nous avons utilisé la boite à outils CRF++[5] en considérant les lemmes et les étiquettes morpho-syntaxiques des mots comme des traits. La lemmatisation de ce corpus a été réalisée à l'aide de la plate-forme linguistique LIMA. Par conséquent, les mots des n-meilleures traductions sont dans leurs formes normalisées (lemmes). Pour générer les n-meilleures traductions avec des mots dans leurs formes de surface (fléchies), nous avons appliqué un générateur morphologique (fléchisseur) qui utilise l'information linguistique (étiquette morpho-syntaxique, genre, nombre, etc.) des mots. Comme un lemme peut avoir plusieurs formes fléchies, nous avons évalué le nouveau treillis de mots (des n-meilleures traductions) en utilisant un modèle de langue appris sur un corpus monolingue en langue cible annoté en formes fléchies. Cette évaluation permet de fournir les n-meilleures traductions de la phrase à traduire.

Notons que dans les deux étapes précédentes, le treillis des lemmes et le treillis des formes fléchies ont été transformés au « format CRF » pour être évalués par les deux modèles de langue appris respectivement sur des lemmes et des formes fléchies. Le tableau ci-dessous (Table 3) présente les deux meilleures traductions générées par notre prototype de traduction.

Rang	Traduction
1	les caisses de la sécurité sociale en Grèce encouragent l'investissement dans l'innovation.
2	les fonds de la sécurité sociale en Grèce encouragent l'investissement en l'innovation.

TABLE 3 : Les deux meilleures traductions générées par notre prototype pour la phrase exemple

4 Résultats expérimentaux et évaluation

4.1 Corpus et protocole expérimental

Afin d'étudier l'impact de l'utilisation d'un lexique bilingue (spécifique au domaine) sur les performances de notre prototype de traduction (désigné par EBMT), nous avons réalisé des

[4] Librairie disponible à partir du site AT&T pour une utilisation non commerciale
[5] http://wing.comp.nus.edu.sg/~forecite/services/parscit-100401/crfpp/CRF++-0.51/doc/

expérimentations sur deux corpus parallèles anglais-français (Table 4): Europarl (European Parliament Proceedings) et Emea (European Medicines Agency Documents). Ces deux corpus ont été extraits de la base libre de corpus parallèles OPUS (Tiedemann, 2012).

N° du run	Apprentissage (nombre de phrases)	Tuning (nombre de phrases)
1	150K (Europarl)	3,75K (Europarl)
2	150K+10K (Europarl+Emea)	1,5K (Europarl)
3	150K+20K (Europarl+Emea)	1,5K (Europarl)
4	150K+30K (Europarl+Emea)	1,5K (Europarl)
5	500K (Europarl)	2,5K (Europarl)
6	500K+10K (Europarl+Emea)	2K+0,5K (Europarl+Emea)
7	500K+20K (Europarl+Emea)	2K+0,5K (Europarl+Emea)
8	500K+30K (Europarl+Emea)	2K+0,5K (Europarl+Emea)

TABLE 4 : La taille (en nombre de phrases) de l'ensemble des corpus bilingues utilisés pour l'apprentissage et le développement du système de traduction Moses

L'évaluation consiste à comparer les résultats de la traduction produite par le système de traduction libre Moses[6] avec ceux produits par notre prototype EBMT à la fois sur des textes issus du domaine et sur d'autres hors-domaine. Le corpus d'apprentissage anglais-français est utilisé pour construire les modèles de langue et de traduction de Moses. Les phrases en français de ce même corpus d'apprentissage sont utilisées pour créer la base de données textuelles du moteur de recherche interlingue intégré dans le prototype EBMT. Nous avons effectué huit *runs* avec, pour chacun d'entre eux, un test avec des données du domaine et un autre hors-domaine. Pour cela, nous avons extrait de manière aléatoire un ensemble de 500 paires de phrases à partir du corpus Europarl comme un corpus correspondant au domaine et 500 autres paires de phrases à partir du corpus Emea comme un corpus hors-domaine. Ces tests ont pour but de montrer l'impact que peut avoir le lexique du domaine sur les résultats de traduction. Dans le cas de Moses, le lexique du domaine est représenté par le corpus parallèle spécialisé (Emea) qui est injecté dans les données d'apprentissage (Europarl). Dans le cas du prototype EBMT, le lexique du domaine est extrait du corpus parallèle spécialisé (Emea) en utilisant notre outil d'alignement de mots (Semmar et al., 2010; Bouamor et al., 2012). Cet outil utilise, d'une part, un modèle linguistique pour l'appariement de mots simples en utilisant un dictionnaire bilingue amorce, les cognats et les catégories grammaticales, et d'autre part, un modèle utilisant les relations de dépendance syntaxique pour l'appariement de mots composés se traduisant mot à mot et un modèle statistique pour la mise en correspondance d'expressions multi-mots. Le lexique spécialisé ainsi constitué est injecté dans le lexique anglais-français utilisé à la fois par le moteur de recherche interlingue et par la reformulateur bilingue. Pour évaluer la performance des deux systèmes EBMT et Moses, nous avons utilisé le score BLEU (Papineni et al., 2002).

4.2 Résultats et discussion

Nous avons évalué les performances des deux systèmes Moses et EBMT en utilisant le score BLEU sur les deux ensembles de test pour les huit *runs* décrits dans la section précédente. Sachant que nous considérons une référence par phrase, les résultats obtenus sont présentés dans le tableau ci-dessous (Table 5). Ces résultats montrent que pour les tests réalisés sur le corpus du domaine, les

[6] http://www.statmt.org/moses/

deux systèmes réalisent un score BLEU relativement élevé avec un meilleur score pour Moses dans tous les tests. Pour les tests sur le corpus hors domaine, les performances du prototype EBMT sont nettement meilleures que celles de Moses, notamment pour les premier et cinquième *runs* où ce dernier obtient un très faible score (13,62 et 14,74). Par ailleurs, nous pouvons noter un impact significatif du lexique anglais-français utilisé par le moteur de recherche interlingue et par le reformulateur bilingue sur les résultats du prototype EBMT qui en a bénéficié pour améliorer régulièrement son score BLEU dans tous les *runs*. De la même manière, nous pouvons noter que pour améliorer les résultats de la traduction du système Moses, il est plus important d'avoir un corpus parallèle qui soit du domaine même s'il est réduit, plutôt que d'avoir un large corpus qui soit hors domaine. Par exemple, l'ajout d'un corpus parallèle spécialisé composé de 30 000 phrases aux 500 000 phrases du corpus Europarl a entrainé un gain de 14,52 points de score BLEU. Cependant, pour le corpus de test du domaine, le score BLEU de Moses pour les *runs* 7 et 8 (en ajoutant respectivement 20 000 et 30 000 phrases pour les 500 000 phrases d'Europarl) est inférieur à son score BLEU pour le *run* 6 (en ajoutant seulement 10 000 phrases pour les 500 000 phrases d'Europarl).

N° du run	**Domaine (Europarl)**		**Hors-Domaine (Emea)**	
	Moses	**EBMT**	**Moses**	**EBMT**
1	34,79	30,57	13,62	24,27
2	32,62	30,10	22,96	27,80
3	33,81	29,60	23,30	28,70
4	34,25	28,70	24,55	29,50
5	37,25	33,12	14,74	26,94
6	37,62	32,10	22,68	29,02
7	37,40	31,03	26,50	33,26
8	37,43	29,92	29,26	36,84

TABLE 5 : Scores BLEU pour le système Moses et le prototype de traduction à base d'exemples

Afin d'évaluer la qualité des traductions produites par les deux systèmes EBMT et Moses pour des domaines spécialisés ou non, nous avons choisi deux exemples de traductions extraites de textes relatifs à l'Agence Européenne des Médicaments et aux débats du Parlement Européen (Tables 6 et 7).

Exemple 1: our success must be measured by our capacity to *keep* growing while ensuring *solidarity and cohesion*.	
Traduction de référence	nous devons mesurer notre réussite à notre capacité à *poursuivre sur la voie* de la croissance tout en garantissant *la solidarité et la cohésion*.
Prototype EBMT: Run 1	notre succès doit être mesuré à notre capacité à *garder* la croissance en garantissant *la solidarité et la cohésion*.
Prototype EBMT: Run 6	notre succès doit être mesuré à notre capacité à *continuer* la croissance en garantissant *la solidarité et la cohésion*.
Moses: Run 1	notre succès doit être mesuré par notre capacité à *maintenir* la croissance tout en assurant *la solidarité et de cohésion*.
Moses: Run 6	notre succès doit être mesuré par notre capacité à *suivre* la croissance tout en assurant *la solidarité et de cohésion*.

TABLE 6 : Traductions produites par Moses et le prototype EBMT pour une phrase du domaine

Pour la phrase du domaine (Exemple 1), les deux systèmes EBMT et Moses produisent des traductions presque similaires et plus ou moins correctes. Dans le premier exemple, le mot anglais « keep » a été identifié par l'analyseur morpho-syntaxique utilisé dans le prototype EBMT comme un verbe et traduit par le lexique bilingue par les deux mots « garder » et « continuer ». Bien sûr, la traduction proposée dans le premier *run* (garder) est correct mais elle est moins expressive que celle proposée dans le sixième *run* (continuer). Le lexique anglais-français propose pour le mot « keep » plusieurs traductions (continuer, entretenir, garder, maintenir, observer, protéger, respecter, tenir, etc.), mais le prototype EBMT a choisi le mot « garder » dans le *run* 1 et le mot « continuer » dans le *run* 6. Par ailleurs, Moses a ajouté la préposition « de » (au lieu de l'article défini « la ») au mot « cohésion » quand il a traduit le mot « cohesion » dans l'expression « solidarity and cohesion ».

Exemple 2: there was also a small increase in *fasting blood glucose* and in *total cholesterol* in duloxetine-treated patients while those laboratory tests showed a slight decrease in the *routine care group*.	
Traduction de référence	il y a eu également une faible augmentation de la *glycémie à jeun* et du *cholestérol total* dans le groupe duloxétine alors que les tests en laboratoire montrent une légère diminution de ces paramètres dans le *groupe traitement usuel*.
Prototype EBMT: Run 4	il y avait aussi une petite augmentation dans la *glycémie à jeun* et du *cholesterol total* chez les patients traités par la duloxétine alors que les tests en laboratoire montraient une légère diminution dans le *groupe de soins de routine*.
Prototype EBMT: Run 8	il y avait aussi une faible augmentation dans la *glycémie à jeun* et du *cholesterol total* chez les patients traités par la duloxétine alors que les tests en laboratoire montraient une légère diminution dans le *groupe de soins de routine*.
Moses: Run 4	il était également une légère augmentation de répréhensible *glycémie artérielle* et en total de patients duloxetine-treated *cholesterol* laboratoire alors que ces tests, ont montré une diminution sensible dans les *soins standards groupe*.
Moses: Run 8	il y a aussi une légère augmentation de la *glycémie à jeun* et *cholestérol total* de patients duloxetine-treated alors que ces tests de laboratoire a montré une légère baisse dans les *soins de routine groupe*.

TABLE 7 : Traductions produites par Moses et EBMT pour une phrase hors domaine

Dans l'exemple de la phrase hors-domaine (Exemple 2), les résultats du prototype EBMT sont nettement meilleurs que la plupart des traductions produites par Moses qui elles, sont incompréhensibles et non grammaticales. Ce résultat peut être expliqué par le fait que le corpus de test contient un lexique qui ne correspond pas aux entrées de la table de traductions de Moses. Par exemple, le prototype EBMT traduit correctement les mots composés « fasting blood glucose » et « total cholesterol » (glycémie à jeun, cholestérol total), mais traduit le mot composé « routine care group » par « groupe de soins de routine » au lieu de « groupe de soins routiniers ». Comme nous pouvons le constater, cette traduction est soit produite par le reformulateur bilingue à l'aide de la septième règle du module de transfert syntaxique (Table 3), soit elle correspond à une traduction partielle fournie par le moteur de recherche interlingue. Par ailleurs, Moses ne parvient pas à traduire correctement les expressions composées comme « fasting blood glucose », « total cholesterol », « duloxetine-treated patients » et « routine care group » dans le *run* 4 mais il réussit la traduction des expressions « fasting blood glucose » et « total cholesterol » dans le *run* 8.

L'analyse des traductions montre que les principales faiblesses du prototype EBMT sont liées à trois problèmes : les erreurs produites par l'analyseur syntaxique de la langue source, la différence entre

les syntaxes des deux langues ainsi que la polysémie dans le lexique bilingue. Pour pallier les deux premiers problèmes, nous avons proposé de prendre en compte les traductions candidates retournées par le moteur de recherche interlingue même si ces traductions ne correspondent qu'à une partie de la phrase à traduire. Pour le problème de la polysémie dans le lexique bilingue, le prototype EBMT ne propose aucun traitement spécifique. Cela peut expliquer en partie pourquoi les performances de Moses sont meilleures que celles du prototype EBMT pour la traduction des phrases de l'intra-domaine. Il semble que les probabilités de la table de traductions qui sont calculées au cours du processus d'alignement de mots avec Giza++ (Och et Ney, 2002) contribuent à choisir la bonne traduction. Par ailleurs, dans le cas de Moses, nous avons constaté que pour les phrases hors-domaine, la plupart des erreurs de traduction sont liées au lexique. Par exemple, Moses propose le mot composé « glycémie artérielle » comme traduction de l'expression « fasting blood glucose » pour le *run* 4, ce qui est inexact. Dans les systèmes de traduction statistique tels que Moses, le décodeur consulte les tables de traductions qui sont construites automatiquement à l'aide de l'outil d'alignement de mots Giza++ qui peut produire des erreurs, en particulier lorsqu'il tente d'aligner des expressions multi-mots (Fraser, Marcu, 2007). (Ren et al, 2009) ont montré que l'intégration de ces expressions dans le modèle de traduction de Moses améliore la qualité de la traduction du système de manière significative. Il est difficile d'affirmer à ce stade que la qualité de traduction du prototype EBMT continuera à être meilleure que celle de Moses pour les textes du domaine de spécialité même si on augmente considérablement la taille des corpus parallèles du domaine. De même, on ne peut pas dire que la comparaison effectuée entre les deux systèmes est tout à fait adéquate puisque le prototype EBMT utilise plusieurs composants nécessitant des ressources linguistiques externes. Nous pourrons conclure tout de même, d'une part, que même si la taille du corpus spécialisé et celle du corpus monolingue n'est pas très grande, le prototype EBMT produit des traductions correctes à la fois pour les textes intra-domaine et hors-domaine, et d'autre part, que l'ajout d'un corpus parallèle spécialisé dans le corpus d'apprentissage du système Moses peut dégrader sa performance sur les textes du domaine général.

5 Conclusion

Dans cet article, nous avons présenté un prototype de traduction à base d'exemples utilisant la recherche d'information interlingue et ne nécessitant qu'un corpus de textes en langue cible. Nous avons, en particulier, montré que l'ajout d'un lexique bilingue spécialisé au dictionnaire général de ce prototype améliore significativement la qualité de traduction des textes du domaine de spécialité sans perte de qualité de traduction pour les textes du domaine général. Deux types de corpus de test ont été utilisés pour les expérimentations: des textes correspondant au domaine extraits du corpus Europarl et des textes hors-domaine extraits du corpus Emea. Nous avons constaté que le prototype de traduction à base d'exemples et Moses réalisent un score BLEU élevé pour les textes intra-domaine. Nos expériences sur les textes hors-domaine ont montré que le prototype EBMT est plus performant que Moses. Les résultats encourageants obtenus par le moteur de traduction à base d'exemples pour le couple de langues anglais-français montrent l'intérêt de l'approche utilisée et nous incite à ajouter et évaluer d'autres couples de langues peu dotées en corpus parallèles. Les approches d'alignement de mots à partir de corpus comparables constituent une piste pour la production de ressources linguistiques multilingues pour de tels couples de langues. Nos travaux futurs s'orientent, d'une part, vers une évaluation à une large échelle de notre prototype de traduction en récupérant à partir du Web de gros volumes de corpus et en les indexant dans le but d'obtenir un modèle de langue représentatif de la langue cible, et d'autre part, vers l'amélioration de la qualité de la traduction en apprenant automatiquement d'autres règles de projection syntaxique pour le reformulateur bilingue et en enrichissant d'avantage le dictionnaire bilingue.

Références

BANERJEE P., DU J., LI B., NASKAR S. K., WAY A., GENABITH J. (2010). Combining Multi-Domain Statistical Machine Translation Models using Automatic Classifiers. Actes de *the Ninth Conference of the Association for MT in the Americas*.

BERTOLDI N., FEDERICO M. (2009). Domain adaptation for statistical machine translation with monolingual resources. Actes de *the 4th Workshop on Statistical Machine Translation*.

BESANÇON R., DE CHALENDAR G., FERRET O., GARA F., LAIB M., MESNARD O., SEMMAR N. (2010). LIMA: A multilingual framework for linguistic analysis and linguistic resources development and evaluation. Actes de *LREC 2010.*

BOUAMOR D., SEMMAR N., ZWEIGENBAUM P. (2012). Identifying bilingual Multi-Word Expressions for Statistical Machine Translation. Actes de *LREC 2012.*

BROWN P. F., COCKE J., DELLA PIETRA S., DELLA PIETRA V. J., JELINEK F., LAFFERTY J. D., MERCER R. L., ROOSSIN, P. S. (2005). A statistical approach to machine translation. *Computational Linguistics, 16(2):79–85, 1990.*

DAUMÉ III H., JAGARLAMUDI J. (2011). Domain Adaptation for Machine Translation by Mining Unseen Words. Actes de *the 49th Annual Meeting of the Association for Computational Linguistics: short papers*.

FRASER A., MARCU D. (2007). Measuring word alignment quality for statistical machine translation. *Computational Linguistics, 33(3), 293–303.*

GREFENSTETTE G. (1998). Cross-Language Information Retrieval. *The Information Retrieval Series, Vol. 2,* Springer.

HILDEBRAND A. S., ECK M., VOGEL S., ALEX W. (2005). Adaptation of the translation model for statistical machine translation based on information retrieval. Actes de *the European Association for Machine Translation Conference*.

HUTCHINS J. (2003). Machine Translation: General Overview. *The Oxford Handbook of Computational Linguistics,* Oxford: University Press.

LAFFERTY J., MCCALLUM A., PEREIRA F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Actes de *the Eighteenth International Conference on Machine Learning*.

LANGLAIS P. (2002). Improving a general-purpose statistical translation engine by terminological lexicons. Actes de *COLING: Second international workshop on computational terminology.*

LEWIS W. D., WENDT C., BULLOCK D. (2010). Achieving Domain Specificity in SMT without Overt Siloing. Actes de *the seventh international conference on Language Resources and Evaluation*.

MOHRI M., PEREIRA F., RILEY M. (2002). Weighted Finite-State Transducers in Speech Recognition. *Computer Speech and Language, 16(1).*

OCH F. J, NEY H. (2002). Discriminative training and maximum entropy models for statistical machine translation. Actes de *the 40th meeting of the Association for Computational Linguistics*.

PAPINENI K., ROUKOS S., WARD T., ZHU W. J. (2002). BLEU: a method for automatic evaluation of machine translation. Actes de *the 40th Annual meeting of the Association for Computational Linguistics*.

PECINA P., TORAL A., WAY A., PAPAVASSILIOU V., WAY A., PROKOPIDIS P., GIAGKOU M. (2011). Towards Using Web-Crawled Data for Domain Adaptation in Statistical Machine Translation. Actes de *the 15th Conference of the European Association for Machine Translation*.

REN Z., LU Y., CAO J., LIU Q., HUANG Y. (2009). Improving statistical machine translation using domain bilingual multiword expressions. Actes de *the Workshop on Multiword Expressions, ACL-IJCNLP*.

SALTON S., MCGILL M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc.

SEMMAR N., SERVAN C., DE CHALENDAR G., LE NY B., BOUZAGLOU J. (2010). A Hybrid Word Alignment Approach to Improve Translation Lexicons with Compound Words and Idiomatic Expressions. Actes de *32nd Translating and the Computer conference, London, England, 2010.*

SEMMAR N., ZENNAKI O., LAIB M. (2015). Improving the Performance of an Example-Based Machine Translation System Using a Domain-specific Bilingual Lexicon. Actes de *29th Pacific Asia Conference on Language, Information and Computation, Shanghai, China, 2015.*

SIMARD M., UEFFING N., ISABELLE P. (2007). Rule-based translation with statistical phrase-based post-editing. Actes de *the Workshop on Statistical Machine Translation, pages 203–206, Prague, Czech Republic, June 2007*.

SNOVER M., DORR B., SCHWARTZ R. (2008). Language and translation model adaptation using comparable corpora. Actes de *the Conference on Empirical Methods in Natural Language Processing*.

SOMERS H. (1999). Review article: Example-based machine translation. *Machine Translation, 14(2):113–157, June 1999.*

SOMERS H. (2003). Machine Translation: Latest Developments. *The Oxford Handbook of Computational Linguistics,* Oxford: University Press.

TIEDEMANN J. (2012). Parallel Data, Tools and Interfaces in OPUS. Actes de *the 8th International Conference on Language Resources and Evaluation.*

VERONIS J., HAMON O., AYACHE C., BELMOUHOUB R., KRAIF O., LAURENT D., NGUYEN T. M. H., SEMMAR N., STUCK F., WAJDI Z. (2008). L'évaluation des technologies de traitement de la langue: La campagne d'évaluation Arcade II. *Chapitre 2,* Editions Hermès, 2008.

WANG W., MACHEREY K., MACHEREY W., OCH F., XU P. (2012). Improved Domain Adaptation for Statistical Machine Translation. Actes de *the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.