Improving the Performance of an Example-Based Machine Translation System Using a Domain-specific Bilingual Lexicon

Nasredine Semmar, Othman Zennaki, Meriama Laib
CEA, LIST, Vision and Content Engineering Laboratory
F-91191, Gif-sur-Yvette, France
`{nasredine.semmar,othman.zennaki,meriama.laib}@cea.fr`

Abstract

In this paper, we study the impact of using a domain-specific bilingual lexicon on the performance of an Example-Based Machine Translation system. We conducted experiments for the English-French language pair on in-domain texts from Europarl (European Parliament Proceedings) and out-of-domain texts from Emea (European Medicines Agency Documents), and we compared the results of the Example-Based Machine Translation system against those of the Statistical Machine Translation system Moses. The obtained results revealed that adding a domain-specific bilingual lexicon (extracted from a parallel domain-specific corpus) to the general-purpose bilingual lexicon of the Example-Based Machine Translation system improves translation quality for both in-domain as well as out-of-domain texts, and the Example-Based Machine Translation system outperforms Moses when texts to translate are related to the specific domain.

1 Introduction

There are mainly two approaches for Machine Translation (MT): rule-based and corpus-based (Trujillo, 1999; Hutchins, 2003). Rule-Based MT (RBMT) approaches require manually made bilingual lexicons and linguistic rules, which can be costly, and not generalized to other languages. Corpus-based machine translation approaches are effective only when large amounts of parallel corpora are available. However, parallel corpora are only available for a limited number of language pairs and domains. In several fields, available corpora are not sufficient to make Statistical Machine Translation (SMT) approaches operational. Most previous works addressing domain adaptation in machine translation have proven that a SMT system, trained on general texts, has poor performance on specific domains. In this paper, we study the impact of using a domain-specific bilingual lexicon on the performance of an Example-Based Machine Translation (EBMT) system, and we compare the results of the EBMT system against those of the SMT system Moses on in-domain and out-of-domain texts.

The rest of the paper is organized as follows. In Section 2, we present previous research in the field of domain adaptation in SMT. Section 3 describes the translation process and the main components of the EBMT system. Section 4 presents the experimental setup and inspects the results of the EBMT system in qualitative and quantitative evaluations. Section 5 concludes our study and presents our future research directions.

2 Related Work

Domain adaptation consists in adapting MT systems designed for one domain to work in another. Several ideas have been explored and implemented in domain adaptation of SMT (Bungum and Gambäck, 2011). Langlais (2002) integrated domain-specific lexicons in the translation model of a SMT engine which yields a significant reduction in word error rate. Lewis et al. (2010) developed domain specific SMT by pooling all training data into one large data pool, including as much in-domain parallel data as possible. They trained highly specific language models on in-domain monolingual data in order to reduce the dampening effect of heterogeneous data on quality within the domain. Hildebrand et al.

(2005) used an approach which consisted essentially in performing test-set relativization (choosing training samples that look most like the test data) to improve the translation quality when changing the domain. Civera and Juan (2007), and Bertoldi and Federico (2009) used monolingual corpora and Snover et al. (2008) used comparable corpora to adapt MT systems designed for Parliament domain to work in News domain. The obtained results showed significant gains in performance. Banerjee et al. (2010) combined two separate domain models. Each model is trained from small amounts of domain-specific data. This data is gathered from a single corporate website. The authors used document filtering and classification techniques to realize the automatic domain detection. Daumé III and Jagarlamudi (2011) used dictionary mining techniques to find translations for unseen words from comparable corpora and they integrated these translations into a statistical phrase-based translation system. They reported improvements in translation quality (between 0.5 and 1.5 BLEU points) on four domains and two language pairs. Pecina et al. (2011) exploited domain-specific data acquired by domain-focused web-crawling to adapt general-domain SMT systems to new domains. They observed that even small amounts of in-domain parallel data are more important for translation quality than large amounts of in-domain monolingual data. Wang et al. (2012) used a single translation model and generalized a single-domain decoder to deal with different domains. They used this method to adapt large-scale generic SMT systems for 20 language pairs in order to translate patents. The authors reported a gain of 0.35 BLEU points for patent translation and a loss of only 0.18 BLEU points for generic translation.

3 The Translation Process of the Example-Based Machine Translation System

The translation process of the EBMT system consists of several steps: retrieving translation candidates from a monolingual corpus using a cross-language search engine, producing translation hypotheses using a transducer, using word lattices to represent the combination of translation candidates and translation hypotheses, and choosing the n-best translations according to a statistical language model learned from a target language corpus (Semmar and Bouamor 2011; Semmar et al., 2011; Semmar et al., 2015). This process uses a cross-language search engine, a bilingual reformulator (transducer) and a generator of translations. In order to illustrate the functioning of the EBMT system, we indexed a small textual database composed of 1127 French sentences extracted from the ARCADE II corpus (Veronis et al., 2008) and we considered the input source sentence "Social security funds in Greece encourage investment in innovation." as the sentence to translate.

3.1 The Cross-language Search Engine

The role of the cross-language search engine is to extract for each sentence to translate (user's query) sentences or sub-sentences from an indexed monolingual corpus in the target language (Davis and Ogden, 1997; Grefenstette, 1998; Baeza-Yates and Ribeiro-Neto, 1999). These sentences or sub-sentences correspond to a total or a partial translation of the sentence to translate. The cross-language search engine used in the EBMT system is based on a deep linguistic analysis of the query and the monolingual corpus to be indexed and uses a weighted vector space model (Salton and McGill, 1986; Besançon et al., 2003; Semmar et al., 2006). This cross-language search engine is composed of the following modules:

- A linguistic analyzer based on the open source multilingual platform LIMA [1] (Besançon et al., 2010) which includes a morphological analyzer, a Part-Of-Speech tagger and a syntactic analyzer. This analyzer processes both sentences to be indexed in the target language and the sentence to translate in order to produce a set of normalized lemmas with their linguistic information (Part-Of-Speech, gender, number, etc.). The syntactic analyzer implements a dependency grammar to produce syntactic dependencies relations (used to compute compound words) and works by identifying verbal and nominal chains. These syntactic dependencies are detected using finite-state automata defined by rules expressing possible successions of grammatical categories.

---

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

- A statistical analyzer that attributes to each word or a compound word of the sentences to be indexed a weight. For this purpose, we use the TF-IDF weighting. The weight w_{ij} of term j in document i is defined with the formula $w_{ij}=tf_{ij}logN/n_j$, where tf_{ij} is the frequency of term j in document i, N is the total number of documents in the collection, and n_j is the number of documents where term j appears.

- An indexer to build the textual database which contains the sentences of the target language.

- A query reformulator to expand queries during the interrogation of the textual database. The query terms are translated using a bilingual lexicon. Each term of the query is reformulated into its translations in target language using an English-French lexicon composed of 243539 entries[2].

- A comparator which measures the similarity between the sentence to translate (query) and the indexed sentences in order to retrieve the closest sentences to the reformulated query. The *Cosine* similarity is used to measure the distance between the sentence to translate and each sentence of the textual database. The retrieved sentences are classified by the comparator which groups in the same cluster the sentences that share the same words.

For example, from the sentence "Social security funds in Greece encourage investment in innovation.", two nominal chains are recognized: "Social security funds in Greece" and "investment in innovation". From the first nominal chain, the syntactic analyzer recognizes three compound words: Social security funds in Greece (Greece_fund_security_social), Social security funds (fund_security_social), and Social security (security_social). Table 1 illustrates the two first translation candidates provided by the cross-language search engine for the sentence to translate "Social security funds in Greece encourage investment in innovation.". These sentences share with the query the terms "fund_security_social, Greece, investment" for the first class and the term

"fund_security_social" for the second class. In addition, the cross-language search engine provides the linguistic information (lemma, Part-Of-Speech, gender, number and syntactic dependency relations) of all words included in the translation candidates (Table 2). The translation candidates are represented as graphs of words and encoded with Finite-State Machines (FSMs). Each transition of the automaton corresponds to the lemma and its linguistic information which is provided by the linguistic analyzer of the cross-language search engine (Figure 1).

Class n°.	Class query terms	Translation candidates
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. The two first translation candidates returned by the cross-language search engine for the sentence of to translate "Social security funds in Greece encourage investment in innovation.".

Les [*le, Plural determiner*] caisses [*caisse, Plural common noun*] de [*de, Singular preposition*] sécurité [*sécurité, Singular common noun*] sociale [*social, Singular adjective*] de [*de, Singular preposition*] Grèce [*Grèce, Singular proper noun*] revendiquent [*revendiquer, Third person plural verb*] l'[*le, Singular determiner*] indépendance [*indépendance, Singular common noun*] en [*en, Singular preposition*] matière [*matière, Singular common noun*] d'[*de, Singular preposition*] investissements [*investissement, Plural common noun*]. [*., Punctuation*]

Table 2: Linguistic information (lemma, grammatical category) of the words of the first translation candidate. This sentence is composed of two nominal chains linked by the word "revendiquent".
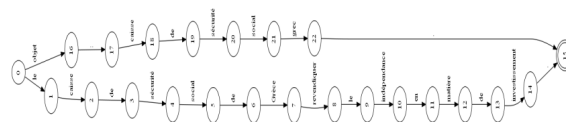

Figure 1: FSMs representing the retrieved sentences returned by the cross-language search engine.

---

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

3.2 The Bilingual Reformulator

The role of the bilingual reformulator is to produce a set of translation hypotheses from the sentence to translate. It consists, on the one hand, in transforming into the target language the syntactic structure of the sentence to translate, and, on the other hand, in translating its words. The reformulator uses a set of linguistic rules to transform syntactic structures from the source language to the target language (Syntactic transfer) and the bilingual lexicon of the cross-language search engine to translate words of the sentence to translate (Lexical transfer). The rules of the syntactic transfer are built manually and are based on morpho-syntactic patterns (Table 3). Expressions (phrases) corresponding to each pattern are identified by the syntactic analyzer during the step of recognition of verbal and nominal chains. These expressions can be seen as sentences accepted by a FSM transducer whose outputs are instances of these sentences in the target language (Figure 2).

Rule n°.	Tag pattern (English)	Tag pattern (French)
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 3: Frequent Part-Of-Speech tag patterns used to transform syntactic structures of the sentence to translate from English to French. In these patterns A refers to an Adjective, P to a Preposition, T to Past Participle, and N to a Noun.
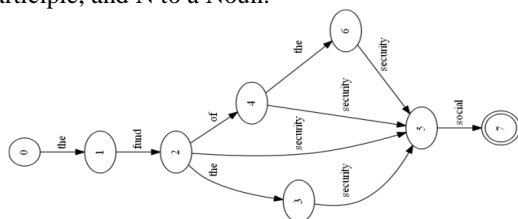

Figure 2: Example of a lattice of words corresponding to the syntactic transformation of the compound word "Social security funds".

For example, from the sentence to translate "Social security funds in Greece encourage investment in innovation.", two nominal chains are recognized: "Social security funds in Greece" and "investment in innovation". These nominal chains are linked with the verb "encourage". The expression "investment in innovation" is transformed using the sixth rule (Table 3) into the expression "the investment in the innovation". It is important to mention here that the linking word "the" (definite article) is added to the applied rule before each noun (investment, innovation) in order to complete the transformation. The FSM transducer of the syntactic transfer step produces a lattice of words in the source language (Figure 2). Each word is represented with its lemma in the lattice and is associated with its linguistic information (Part-Of-Speech, gender, number, etc.).

Lexical transfer translates in the target language the lemmas of the obtained syntactic structures words using the bilingual lexicon of the cross-language search engine. This English-French lexicon is composed of 243539 entries. These entries are represented in their normalized forms (lemmas). A lemmatization process provided by the linguistic analyzer LIMA is applied on the obtained syntactic structures words. This step produces an important number of translation hypotheses. This is due to the combination of the syntactic transfer rules and the polysemy in the bilingual lexicon. The result of the bilingual reformulator is a set of lattices in which words are in the target language.

3.3 The Generator of Translations

The role of the generator of translations consists in assembling the results returned by the cross-language search engine and the bilingual reformulator, and in choosing the n-best translations according to a statistical language model learned from the target language corpus. The assembling process consists in composing FSMs corresponding to the translation candidates with FSMs corresponding to the translation hypotheses. The FSM state where the composition is made is determined by words which link the nominal chains of the translation candidates and the translation hypotheses. All the operations applied on the FSMs are made with the AT&T

FSM Library[3] (Mohri et al., 2002). In order to find the best translation hypothesis from the set of word lattices (Dong et al., 2014), a statistical model is learned with the CRF++ toolkit[4] (Lafferty et al., 2001) on lemmas and Part-Of-Speech tags of the target language corpus. Therefore, the n-best translations words are in their normalized forms (lemmas). To generate the n-best translations with words in their surface (inflected) forms, we applied a morphological generator (flexor) which uses the linguistic information (Part-Of-Speech, gender, number, etc.). The word lattices corresponding to the translations are enriched with the results of the flexor. These lattices are then scored with another statistical language model learned from texts of the target language containing words in inflected forms. The CRF++ toolkit is used to select the n-best translations in inflected forms.

4   Experiments and Results

4.1   Data and Experimental Setup

In order to study the impact of using a domain-specific bilingual lexicon on the performance of the EBMT system, we conducted our experiments on two English-French parallel corpora (Table 4): Europarl (European Parliament Proceedings) and Emea (European Medicines Agency Documents). Both corpora were extracted from the open parallel corpus OPUS (Tiedemann, 2012). Evaluation consists in comparing translation results produced by the open source SMT system Moses (Khoen et al., 2007) and the EBMT system on in-domain and out-of-domain texts. The English-French training corpus is used to build Moses's translation and language models. The French sentences of this training corpus are used to create the indexed database of the cross-language search engine integrated in the EBMT system. We conducted eight runs and two test experiments for each run: In-Domain and Out-Of-Domain. For this, we randomly extracted 500 parallel sentences from Europarl as an In-Domain corpus and 500 pairs of sentences from Emea as an Out-Of-Domain corpus. These experiments are done to show the impact of the domain vocabulary on the translation results.

---

[3] FSM Library is available from AT&T for non-commercial use as executable binary programs.
[4] http://wing.comp.nus.edu.sg/~forecite/services/parscit-100401/crfpp/CRF++-0.51/doc/.

The domain vocabulary is represented in the case of Moses by the specialized parallel corpus (Emea) which is added to the training data (Europarl). In the case of the EBMT system, the domain vocabulary is identified by a bilingual lexicon which is extracted automatically from the specialized parallel corpus (Emea) using a word alignment tool (Semmar et al., 2010; Bouamor et al., 2012). This specialized bilingual lexicon is added to the English-French lexicon which is used jointly by the cross-language search engine and the bilingual reformulator. To evaluate the performance of the EBMT system and Moses, we used the BLEU score (Papineni et al; 2002).

Run n°.	Training (# sentences)	Tuning (# sentences)
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: Corpora details used to train Moses language and translation models, and to build database of the EBMT system. In this table, K refers to 1000.

4.2   Results and Discussion

The performance of Moses and the EBMT system is evaluated using the BLEU score on the two test sets for the eight runs described in the previous section. Note that we consider one reference per sentence. The obtained results are reported in Table 5.

Run n°.	In-Domain		Out-Of-Domain	
	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: BLEU scores of Moses and the EBMT system.

The first observation is that, when the test set is In-Domain, we achieve a relatively high BLEU score for both the two systems and the score of Moses is better in all the runs. For the Out-Of-Domain test corpus, the EBMT system performs better than Moses in all the runs and in particular Moses has obtained a very low BLEU score in the first and fifth runs (13.62 and 14.74). Furthermore, it seems that the English-French lexicon used in the cross-language search engine and the bilingual reformulator has had a significant impact on the result of the EBMT system. It improved regularly its BLEU score in all the runs. Likewise, these results show that small amounts of in-domain parallel data are more important for translation quality of Moses than large amounts of out-of-domain data. For example, adding a specialized parallel corpus composed of 30000 sentences to the 500000 sentences of Europarl reported a gain of 14.52 BLEU points. However, for the In-Domain test corpus, Moses's BLEU score in runs 7 and 8 (adding respectively 20000 and 30000 sentences to the 500000 sentences of Europarl) is little than Moses's BLEU score in run 6 (adding only 10000 sentences to the 500000 sentences of Europarl).

In order to evaluate qualitatively the EBMT system and Moses when translating specific and general-purpose texts, we take two examples of translations drawn from texts relating to the European Medicines Agency texts and the European Parliament proceedings (Tables 6 and 7). For the In-Domain sentence (Example 1), the EBMT system and Moses provide close translations and these translations are more or less correct. In the first example, the English word "keep" was identified by the morpho-syntactic analyzer used by the EBMT system as a verb and the bilingual lexicon proposed respectively the words "garder" and "continuer" as translations for this word. Of course, the translation proposed in the first run (garder) is correct but it is less expressive than the one proposed in the fifth run (continuer). The English-French lexicon proposes for the word "keep" several translations (continuer, entretenir, garder, maintenir, observer, protéger, respecter, tenir, etc.) but the EBMT system has chosen "garder" in run 1 and "continuer" in run 6. On the other hand, Moses added the preposition "de" (instead of the definite article "la") to the word "cohésion" when it translated the word "cohesion" in the expression "solidarity and cohesion".

Example 1 Input: our success must be measured by our capacity to *keep* growing while ensuring *solidarity and cohesion.*	
Reference	nous devons mesurer notre réussite à notre capacité à *poursuivre sur la voie* de la croissance tout en garantissant *la solidarité et la cohésion.*
EBMT system: Run 1	notre succès doit être mesuré à notre capacité à *garder* la croissance en garantissant *la solidarité et la cohésion.*
EBMT system: 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: Translations produced by the EBMT system and Moses for an In-Domain sentence.

For the Out-Of-Domain sentence (Example 2), the EBMT system results are clearly better and most of the translations produced by Moses are incomprehensible and ungrammatical. This result can be explained by the fact that the test corpus has a vocabulary which is different from the entries of Moses's translation table. For instance, the EBMT system translates correctly the compound words "fasting blood glucose" and "total cholesterol" (glycémie à jeun, cholesterol total) but it translates the compound word "routine care group" as "groupe de soins de routine" instead of "groupe de soins routiniers". As we can see, this translation could not be provided by the bilingual reformulator because there is no transfer rule implementing the tag pattern of this compound word which is NPNPN (Table 3). This expression corresponds to a partial translation provided by the cross-language search engine for the sentence to translate. On the other hand, Moses fails to translate correctly the multiword expressions "fasting blood glucose", "total cholesterol", "duloxetine-treated patients" and "routine care group" in run 4. However, it succeeds in the translation of the expressions "fasting blood glucose" and "total cholesterol" in run 8.

111

Example 2 Input: 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*.	
Reference	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*.
EBMT System: 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*.
EBMT System: 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: Translations produced by the EBMT system and Moses for an Out-Of-Domain sentence.

After analyzing some translations, we observed that the major issues of our EBMT system are related to errors from the source-language syntactic analyzer, the non-isomorphism between the syntax of the two languages and the polysemy in the bilingual lexicon. To handle the first two issues, we proposed to take into account translation candidates returned by the cross-language search engine even if these translations correspond only to a part of the sentence to translate. However, for the presence of the polysemy in the bilingual lexicon, the EBMT system has no specific treatment. This can explain partially why the EBMT system is outperformed by Moses when translating In-Domain sentences. It seems that translation table probabilities which are computed during the word alignment process with Giza++ (Och and Ney,

2002; Och and Ney, 2003) have contributed to choose the right translation. On the other hand, we noted that most of Moses's translation errors for Out-Of-Domain sentences are related to vocabulary. For example, Moses proposes the compound word "glycémie artérielle" as a translation for the expression "fasting blood glucose" in run 4 which is not correct. In SMT systems such as Moses, phrase tables are the main knowledge source for the machine translation decoder. The decoder consults these tables to figure out how to translate an input sentence from the source language into the target language. These tables are built automatically using the open source word alignment tool Giza++. However, Giza++ could produce errors in particular when it aligns multiword expressions. (Bouamor et al, 2012; Ren et al., 2009) showed that the integration of multiword expressions in Moses's translation model improves the translation quality. Multiword expressions include a large list of categories such as collocations, compound words, idiomatic expressions, named entities and domain-specific terms (Baldwin and Kim, 2010). To reduce word alignment errors with Giza++, we propose the following three methods to integrate into Moses the bilingual lexicon which is extracted automatically by our word alignment tool from the specialized parallel corpus (Emea):

- $Moses_{CORPUS}$: In this method, we add the extracted bilingual lexicon as a parallel corpus and retrain the translation model. By increasing the occurrences of the specialized words and their translations, we expect a modification of alignment and probability estimation.

- $Moses_{TABLE}$: This method consists in adding the extracted bilingual lexicon into Moses's phrase table. We use the *Cosine* similarity measure provided by our word alignment tool for each specialized word of the bilingual lexicon as a translation probability.

- $Moses_{FEATURE}$: In this method, we extend "$Moses_{TABLE}$" by adding a new feature indicating whether a word comes from the specialized bilingual lexicon or not (1 or 0 is introduced for each entry of the phrase table).

For these experiments, we used only English-French training corpora of runs 2, 3 and 4 to build Moses's translation and language models. We measure the translation quality on the same test sets of the previous experiments (500 parallel sentences extracted randomly from Europarl for the In-Domain test and 500 pairs of sentences extracted randomly from Emea for the Out-Of-Domain test). Because the bilingual lexicon which is extracted automatically from the specialized parallel corpus is composed of entries in their normalized forms (lemmas), we used the factored translation model of Moses (Koehn et al., 2010). This model accepts the use of additional annotations at the word level and operates on lemmas instead of surface forms. The translation process consists, first, in translating lemmas of words from the source language into the target language, and second, in generating the inflected forms for each lemma. Tables 8 and 9 present respectively the Moses's results for the In-Domain and the Out-Of-Domain sentences when using the three integration strategies.

The first important point to mention here is that there is improvement of the BLEU score in all the integration methods for the Out-Of-Domain sentences. The best improvement is achieved using the $\text{Moses}_{\text{FEATURE}}$ method which guides Moses to choose specialized words instead of those provided by the translation model built with Giza++. Compared to the baseline system (Moses without using integration strategies), this method reports a gain of 3.63 BLEU points for the fourth run. The obtained BLEU score (28.18) is not very far from the BLEU score obtained by the EMBT system in the same run (29.50). We think that this high score is due to the feature which guides Moses in choosing the best translation with a preference to the words of the specialized bilingual lexicon. In this case, Moses neglects the other translations found in the translation table. On the other hand, the $\text{Moses}_{\text{TABLE}}$ method has lower scores in all the runs. We assume that we obtain such lower scores because the content of the translation table is not coherent. Indeed, we considered the *Cosine* similarity measure provided by our word alignment tool for each specialized word of the bilingual lexicon as a translation probability. However, in actual fact, values of *Cosine* similarity measures are not similar to translation probabilities provided by Giza++.

Run n°.	In-Domain		
	$\text{Moses}_{\text{CORPUS}}$	$\text{Moses}_{\text{TABLE}}$	$\text{Moses}_{\text{FEATURE}}$
2	32.82	32.15	29.18
3	33.89	33.48	30.26
4	34.64	34.11	31.84

Table 8. Translation results in terms of BLEU scores corresponding to the three integration methods for the In-Domain sentences.

Run n°.	Out-Of-Domain		
	$\text{Moses}_{\text{CORPUS}}$	$\text{Moses}_{\text{TABLE}}$	$\text{Moses}_{\text{FEATURE}}$
2	23.45	23.11	24.69
3	24.09	23.76	25.68
4	25.43	25.05	28.18

Table 9. Translation results in terms of BLEU scores corresponding to the three integration methods for the Out-Of-Domain sentences.

As it can be seen, these results confirm that adding specialized parallel corpora to the training data improves the translation quality of Out-Of-Domain test corpus for the both MT systems in all cases but the improvement of the EBMT system is more significant. Likewise, even if the size of the specialized corpus and the size of the general-purpose monolingual corpus are not significant, the EBMT prototype produces correct translations for both in-domain and out-of-domain texts.

5 Conclusion

In this paper, we have studied the impact of using a domain-specific corpus on the performance of an EBMT system and Moses. Two kinds of texts are used in our experiments: in-domain texts from Europarl and out-of-domain texts from Emea. We have seen that both the two systems achieved a relatively high BLEU score for in-domain texts. Our experiments on out-of-domain texts have showed that the EBMT system performs better than Moses. Moreover, we have noticed that the method which guides Moses to choose specialized words instead of those provided by the translation model built with Giza++ achieves the best improvement. In the future, we plan, on the one hand, to use machine learning techniques to extract transfer rules for the bilingual reformulator from annotated parallel corpora, and on the other hand, to evaluate the EBMT system on other specific domains such as security, finance, etc.

Philippe Langlais. 2002. Improving a general-purpose statistical translation engine by terminological lexicons. In *Proceedings of COLING: Second international workshop on computational terminology.*

William D. Lewis, Chris Wendt, and David Bullock. 2010. Achieving Domain Specificity in SMT without Overt Siloing. In *Proceedings of the seventh international conference on Language Resources and Evaluation.*

Mehryar Mohri, Fernando Pereira, and Michael Riley. 2002. Weighted Finite-State Transducers in Speech Recognition. In *Computer Speech and Language, 16(1),* pages 69-88.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th meeting of the Association for Computational Linguistics.*

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics, volume 29, number 1, pages 19-51.*

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual meeting of the Association for Computational Linguistics,* pages 311–318.

Pavel Pecina, Antonio Toral, Andy Way, Vassilis Papavassiliou, Prokopis Prokopidis, and Maria Giagkou. 2011. Towards Using Web-Crawled Data for Domain Adaptation in Statistical Machine Translation. In *Proceedings of the 15th Conference of the European Association for Machine Translation.*

Zhixiang Ren, Yajuan Lu, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions, ACL-IJCNLP,* pages 47–57, Suntec, Singapore.

Gerard Salton and Michael J. McGill. 1986. Introduction to Modern Information Retrieval. In *McGraw-Hill, Inc.*

Nasredine Semmar, Meriama Laib and Christian Fluhr. 2006. A Deep Linguistic Analysis for Cross-language Information Retrieval. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC,* Italy.

Nasredine Semmar and Meriama Laib. 2010b. Using a Hybrid Word Alignment Approach for Automatic Construction and Updating of Arabic to French Lexicons. In *Proceedings of the Workshop on LR and HLT for Semitic Languages, LREC.*

Nasredine Semmar, Dhouha Bouamor. 2011. A New Hybrid Machine Translation Approach Using Cross-Language Information Retrieval and Only Target Text Corpora. In *Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation*, Spain.

Nasredine Semmar, Christophe Servan, Dhouha Bouamor, and Ali Joua. 2011. Using Cross-Language Information Retrieval for Machine Translation. In *Proceedings of the 5th Language & Technology Conference*, Poland.

Nasredine Semmar, Othman Zennaki, and Meriama Laib. 2015. Evaluating the Impact of Using a Domain-specific Bilingual Lexicon on the Performance of a Hybrid Machine Translation Approach. In *Proceedings of the 10th International Conference on Recent Advances in Natural Language Processing*, Bulgaria.

Matthew Snover, Bonnie Dorr, and Richard Schwartz. 2008. Language and translation model adaptation using comparable corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

Ankit Srivastava, Sergio Penkale, Declan Groves, and John Tinsley. 2009. Evaluating Syntax-Driven Approaches to Phrase Extraction for MT. In *Proceedings of the 3rd International Workshop on Example-Based Machine Translation,* Ireland.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation.*

Arturo Trujillo. 1999. Translation Engines: Techniques for Machine Translation. In *Applied Computing, Springer.*

Veronis Jean, Hamon Olivier, Ayache Christelle, Belmouhoub Rachid, Kraif Olivier, Laurent Dominique, Nguyen Thi Minh Huyen, Semmar Nasredine, Stuck François and Wajdi Zaghouani. 2008. L'évaluation des technologies de traitement de la langue: La campagne d'évaluation Arcade II. In *Chapitre 2, Editions Hermès, 2008.*

Wei Wang, Klaus Macherey, Wolfgang Macherey, Franz Och, and Peng Xu. 2012. Improved Domain Adaptation for Statistical Machine Translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*