Evaluating the Impact of Using a Domain-specific Bilingual Lexicon on the Performance of a Hybrid Machine Translation Approach

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

This paper describes an Example-Based Machine Translation prototype and presents an evaluation of the impact of using a domain-specific vocabulary on its performance. This prototype is based on a hybrid approach which needs only monolingual texts in the target language and consists to combine translation candidates returned by a cross-language search engine with translation hypotheses provided by a finite-state transducer. The results of this combination are evaluated against a statistical language model of the target language in order to obtain the n-best translations. To measure the performance of this hybrid approach, we achieved several experiments using corpora on two domains from the European Parliament proceedings (Europarl) and the European Medicines Agency documents (Emea). The obtained results show that the proposed approach outperforms the state-of-the-art Statistical Machine Translation system Moses when texts to translate are related to the specialized domain.

1 Introduction

Current Machine Translation (MT) technology has serious limitations: there are, on the one hand, the rule-based systems which require handcrafted linguistic rules and their manual construction is time consuming and expensive, and, on the other hand, the statistical systems which try to learn how to translate by analyzing the translation patterns found in large collections of human translations and these systems 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.

We present, in this paper, an Example-Based Machine Translation (EBMT) prototype and we study the impact of using a domain-specific lexicon on its performance. The EBMT prototype is based on a hybrid approach which uses only a monolingual corpus in the target language. This corpus is considered as a textual database of a cross-language search engine. For each sentence to translate (query in natural language), the cross-language search engine is used to provide a set of sentences in the target language. These sentences are combined with translation hypotheses provided by a finite-state transducer. The result of this combination is evaluated against a statistical language model learned from the target language corpus in order to produce the n-best translations. We believe that this is the first application of cross-language information retrieval in machine translation (Semmar and Bouamor 2011; Semmar et al., 2011; Semmar et al., 2014).

The remainder of this paper is organized as follows: Section 2 describes the main approaches used in machine translation and presents previous works addressing the task of domain adaptation in statistical machine translation. Section 3 introduces the hybrid approach used to implement the EBMT prototype and presents its architecture. In section 4 we discuss results obtained after translating two types of texts in general and specialized domains. Section 5 concludes our study and presents our future work.

2 Related Work

Machine translation systems are indispensable tools in a globalizing world. In the last years, several online MT systems have been proposed and are used by millions of people every day. However, there are serious limitations to current MT technology which mainly uses two approaches: rule-based and corpus-based (Trujillo, 1999; Hutchins, 2003). The rule-based approach-

Proceedings of Recent Advances in Natural Language Processing, pages 579–587,
Hissar, Bulgaria, Sep 7–9 2015.

es regroup word-to-word translation, syntactic translation with transfer rules and interlingua. The corpus-based machine translation approaches regroup Example-based MT and statistical-based MT techniques (Somers, 2003). These two techniques have in common the use of a database containing already translated sentences. EBMT uses a process which consists in matching a new sentence against this database to extract suitable sentences which are recombined in an analogical manner to determine the correct translation. The second corpus-based strategy is the statistical approach (Brown et al., 1993) which consists in searching for a target language string that maximizes the probability that this string is the translation of a source target string (translation model) and the probability that this target language string is a valid sentence (language model). This approach uses strings co-occurrence frequency in aligned texts in order to build the translation model and strings succession (n-grams) in order to build the language model. 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 MT approaches are effective only when large amounts of parallel corpora are available. Recently, several strategies have been proposed to combine the strengths of rule-based and corpus-based MT approaches or to add deep linguistic knowledge into statistical machine translation. Examples include Part-Of-Speech and morphological information (Koehn et al., 2010), word sense disambiguation models (Carpuat and Wu, 2007) and semantic role labels (Wu and Fung, 2009). Carbonell et al. (2006) described a new paradigm for corpus based translation that does not require parallel text. They called this paradigm Context-Based Machine Translation (CBMT) which relies on a lightweight translation model utilizing a full-form bilingual lexicon and a decoder using long-range context via long n-grams and cascaded overlapping. The authors evaluated their approach in Spanish-English translation using Spanish newswire text. This approach achieves a BLEU score of 0.64. The results showed that quality increases above the reported score as the target corpus size increases and as dictionary coverage of source words and phrases becomes more complete.

As regards domain adaptation in MT, most previous works addressing this task have proven that a statistical machine translation system trained on general texts, has poor performance on specific domains. In order to adapt MT systems designed for one domain to work in another, several ideas have been explored and implemented (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. However, this work did not report the impact of generic data on domain translation accuracy. 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.

The approach we propose for domain adaptation is close in spirit to the work of Langlais (2002), but assumes the integration of the domain-specific lexicon in the two components of the EBMT prototype: the cross-language search engine and the bilingual reformulator.

3 Machine Translation Based on Cross-language Information Retrieval

The hybrid approach used in the Example-Based Machine Translation prototype consists, on the one hand, in indexing a database of sentences in the target language and considering each sentence to translate as a query to that database, and on the other hand, in combining sentences returned by a cross-language search engine with translation hypotheses provided by a bilingual reformulator (Figure 1).

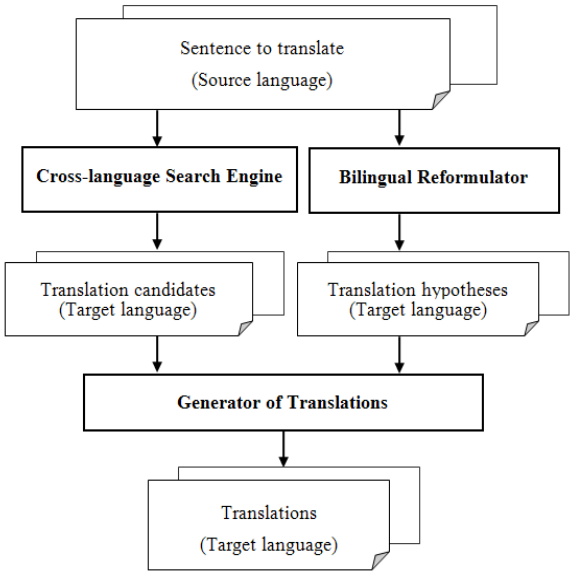

Figure 1: Architecture of the Example-Based Machine Translation prototype.

The EBMT prototype is composed of:

- A cross-language search engine to extract sentences or sub-sentences of the target language from the textual database which correspond to a total or a partial translation of the sentence to translate.
- A bilingual reformulator to transfer syntactic structures from the source language to the target language using transfer rules and bilingual lexicons.
- A generator of translations which 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.

In order to illustrate the translation process of the EBMT prototype, we indexed a textual database composed of 1127 French sentences extracted from the ARCADE II corpus[1] 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 purpose of Cross-Language Information Retrieval (CLIR) is to find similar or relevant documents for a given query where the documents and the query are written in different languages (Davis and Ogden, 1997; Grefenstette, 1998). In our use of CLIR in machine translation, a document corresponds to a sentence. The role of the cross-language search engine is to retrieve for each user's query (which is introduced as a sentence in natural language) translation candidates from an indexed monolingual corpus. The cross-language search engine used in the EBMT prototype is based on a deep linguistic analysis (Besançon et al., 2010) of the query and the monolingual corpus to be indexed and uses a weighted vector space model in which sentences to be indexed are grouped into classes characterized by the same set of words (Salton and McGill, 1986). This cross-language search engine (Besançon et al., 2003) is composed of a linguistic analyzer based on the open source multilingual platform LIMA[1], a statistical analyzer that attributes to each word or a compound word of the sentences to be indexed a weight by using the TF-IDF weighting, a comparator which measures the similarity between the sentence to translate (query) and the indexed sentences in the target language, a query reformulator to translate words of the query from the source language into the target language using a bilingual lexicon, and a indexer to build the inverted files of the sentences to be indexed on the basis of their linguistic analysis. The cross-language search engine provides the linguistic information (lemma, Part-Of-Speech, gender, number and syntactic dependen-

[1] http://www.technolangue.net/article.php3?id_article=201.

cy relations) of all words included both in the sentence to translate and the retrieved sentences (translation candidates). The result is a list of sentences classes ordered according to the weight of the intersection (similarity measure) between the sentence to translate and the indexed sentences. The translation candidates are represented as graphs of words and encoded with Finite-State Machines (FSMs). The nodes correspond to the states and the arcs refer to transitions. Each transition of the automaton indicates a lemma and its linguistic information which is provided by the linguistic analyzer of the cross-language search engine. 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.".

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 query "Social security funds in Greece encourage investment in innovation.".

3.2 The Bilingual Reformulator

Because the indexed monolingual corpus does not contain the entire translation of each sentence, we added a mechanism to extend translations returned by the cross-language search engine. This is achieved by a Finite-State Transducer (FST) which 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 transducer uses a set of linguistic rules to transform syntactic structures from the source language to the target language and the cross-language search engine bilingual lexicon to translate words of the sentence to translate. This reformulator produces translation hypotheses for the sentence to translate and proceeds in two phases: The first one (Syntactic transfer) consists in transforming syntactic structures from the source language to the target language using transfer rules. These rules built manually are based on morpho-syntactic patterns (Table 2). Expressions (phrases) corresponding to each pattern are identified by the LIMA's 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 2: 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.

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 2) 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.

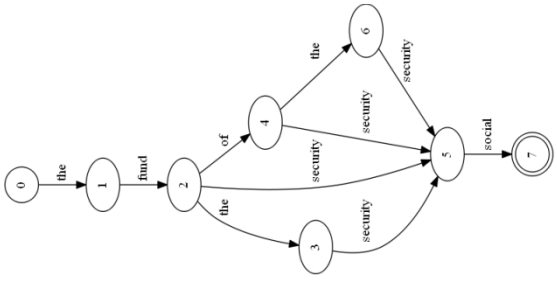

Figure 2: Example of syntactic transformation of the compound word "Social security funds".

The second phase of the bilingual reformulator (Lexical transfer) translates in the target language the lemmas of the obtained syntactic structures words using the cross-language search engine bilingual lexicon. This English-French lexicon is composed of 243539 entries[2]. These entries are represented in their normalized forms (lemmas). A lemmatization process provided by the linguistic analyzer is applied on the obtained syntactic structures words. This step could produce 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 bilingual transducer produces a lattice of words. Each word is represented with its lemma in the lattice and is associated with its linguistic information (Part-Of-Speech, gender, number, etc.).

3.3 The Generator of Translations

The generator of translations consists in producing correct sentences in the target language by using morphological information and syntactic structures of translation candidates. Its role is to assemble in a lattice of words translation hypotheses produced by the transducer with the translation candidates returned by the cross-language search engine. The assembling process consists in composing FSMs corresponding to the translation hypotheses with FSMs corresponding to the translation candidates. Syntactic dependency relations of the translation hypotheses and the translation candidates as well as transfer rules are used to determine the FSM state where the composition is made. In our example, the verb "encourager" (encourage) which links the two patterns involved in the syntactic transformation of the sentence to translate, and the word "revendiquer" (claim) which links the two nominal chains of the first translation candidate (Table 1) determine this state. 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 lattice, a statistical model is learned with the CRF++ toolkit[4] (Lafferty et al., 2001) on the lemmatized corpus of the target language. Therefore, the n-best translations words are in their normalized forms (lemmas). To generate the n-best translations with words in inflected forms, a morphological generator (flexor) is used to transform the lemmas of the translations words into their surface forms. This flexor uses the linguistic information (Part-Of-Speech, gender, number, etc.) provided by the linguistic analyzer of the cross-language search engine for each word of the sentence to translate and the retrieved sentences. The lattice of words corresponding to the translations is enriched with the results of the flexor. This lattice is 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. Table 3 shows the two first translations provided by the EBMT prototype for the input source sentence "Social security funds in Greece encourage investment in innovation.".

Rank	Translation
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: The two first translations for the English sentence "Social security funds in Greece encourage investment in innovation.".

4 Experimental Results

4.1 Data and Experimental Setup

We conducted our experiments on two English-French parallel corpora: Europarl (European Parliament Proceedings) and Emea (European Medicines Agency Documents). Both corpora were extracted from the open parallel corpus OPUS (Tiedemann, 2012). Table 4 lists corpora details.

Run n°.	Training (# sentences)	Tuning (# sentences)
1	150000 (Europarl)	3750 (Europarl)
2	150000+10000 (Europarl+Emea)	1500 (Europarl)
3	150000+20000 (Europarl+Emea)	1500 (Europarl)
4	150000+30000 (Europarl+Emea)	1500 (Europarl)

Table 4: Corpora details used to train Moses and to build the database of the cross-language search engine integrated in the EBMT prototype.

[2] http://catalog.elra.info/product_info.php?products_id=666.
[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 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 prototype. We conducted four 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. 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 prototype, 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 by the cross-language search engine and the bilingual reformulator. First, both corpora have been normalized through the following preprocessing tools provided by the open source SMT toolkit Moses (Khoen et al., 2007): Tokenization, True-casing (the initial words in each sentence are converted to their most probable casing) and Cleaning (long sentences –more than 80 characters- and empty sentences are removed). To evaluate the performance of our approach, we used Moses (Koehn et al., 2007) as a baseline, and the BLEU score as an automatic evaluation metric (Papineni et al; 2002).

4.2 Results and Discussion

We measure translation quality on the two test sets for the four runs described in the previous section and calculate the BLEU score. We also consider only one reference for each test sentence. 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

Table 5: BLEU scores of Moses and the EBMT prototype.

The first observation is that, when the test set is In-Domain, we achieve a relatively high score BLEU 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 prototype performs better than Moses in all the runs and in particular Moses has obtained a very low BLEU score in the first run. This result can be explained by the fact that the test corpus has a vocabulary which is different from the entries of the translation table. Furthermore, it seems that the English-French lexicon used by the cross-language search engine and the bilingual reformulator has had a significant impact on the result of the EBMT prototype. It improved regularly its BLEU score in all the runs. These results confirm that adding specialized parallel corpora to the training data improves the translation quality for the both MT systems in all cases but the improvement of the EBMT prototype is more significant. These results also show that the proportion of the specialized corpus in the training data has a strong impact on the performance of Moses. Indeed, in the fourth run, adding a specialized parallel corpus composed of 30000 sentences to the 150000 sentences of Europarl, reported a gain of 10.93 BLEU score. Tables 6 and 7 illustrate two examples of translations produced by our EBMT prototype and Moses drawn from texts relating to the European Parliament proceedings and the European Medicines Agency texts. Analysis of the translation results shows that for the In-Domain sentences (Example 1) the EBMT prototype and Moses provide close translations and these translations are more or less correct.

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 prototype	notre succès doit être mesuré à notre capacité à *garder* la croissance en garantissant la solidarité et la cohésion.
Moses	notre succès doit être mesuré par notre capacité à *maintenir* la croissance tout en assurant la solidarité et de cohésion.

Table 6: Translations produced by the EBMT prototype and Moses for an In-Domain sentence.

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 prototype	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.*
Moses	il y a également une légère augmentation de répréhensible *glycémie artérielle* et *cholesterol total* de patients duloxetine-treated laboratoire alors que ces tests, ont montré une diminution sensible dans les *soins standards groupe*.

Table 7: Translations produced by the EBMT prototype and Moses for an Out-Of-Domain sentence.

For the Out-Of-Domain sentences, the EBMT prototype results are clearly better and most of the translations produced by Moses are incomprehensible and ungrammatical (Example 2). This result could be due, on the one hand, to differences between the vocabulary of the test corpus and the entries of Moses's translation table, and, on the other hand, to their impact on the phrase reordering model. In the first example, the English word "keep" was identified by the morpho-syntactic analyzer as a verb and the bilingual lexicon of the EBMT prototype proposed the word "garder" as translation. Of course, this translation is correct but it is less expressive than "poursuivre sur la voie" of the translation reference. Likewise, the compound words "fasting blood glucose" and "total cholesterol" of the second example are translated correctly (glycémie à jeun, cholesterol total). On the other hand, the compound word "routine care group" is translated 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 2). This expression corresponds to a partial translation provided by the cross-language search engine for the sentence to translate. We observed that the major issues of our EBMT prototype 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. For the presence of the polysemy in the bilingual lexicon, the EBMT prototype has no specific treatment.

Concerning Moses's translation results for Out-Of-Domain sentences, we noted that most of errors are related to vocabulary. For example, Moses proposes the compound word "glycémie artérielle" as a translation for the expression "fasting blood glucose" 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++[5] (Och and Ney, 2003). However, this tool could produce errors in particular when it aligns multiword expressions.

As a conclusion to this study, even if the comparison between the results of the two MT systems is not completely adequate since the EBMT prototype includes several components that require additional training data (Part-Of-Speech tagger), handwritten rules (Syntactic analyzer, Bilingual reformulator), monolingual and bilingual lexicons (Morphological analyzer, Bilingual reformulator), and Moses is trained on a small amount of the Emea corpus, the experiments show that the EBMT prototype performs better than Moses when texts to translate are related to the specialized domain in all configurations. Our preliminary results also show that the EBMT prototype continues to perform better than Moses when we increase the size of the training corpus of the specialized domain. Likewise, after analyzing qualitatively translations produced by Moses and the EBMT prototype, we observed that the good quality translation of the EBMT prototype is due to its linguistic components and in particular to the syntactic parser and the bilin-

[5] http://www.statmt.org/moses/giza/GIZA++.html.

gual lexicon which contains correct translations of most of the multiword expressions present in the Emea corpus. On the other hand, we noted that Moses fails to translate correctly several multiword expressions (which are very frequent in this corpus) as those of the Example 2, and we are not sure that increasing the training corpus size would limit these incomprehensible and ungrammatical translations.

5 Conclusion

We presented in this paper an EBMT prototype and we compared its performance to the SMT system Moses on domain-specific translation. The first results of our experiments show that, on the one hand, the EBMT prototype performs better than Moses when texts to translate are related to the specialized domain, and, on the other hand, large amounts of in-domain parallel data are necessary for Moses to obtain an acceptable translation quality. These experiments reveal the ability of the EBMT prototype to adapt better to out-domain material. In order to consolidate and improve these encouraging results, we expect to explore a number of ways. First, we will focus on using machine learning techniques to automatically extract transfer rules for the finite-state transducer from a bi-parsed and a word-aligned parallel corpus. Second, we will develop filtering techniques to be applied on these rules in order to reduce the number of translation hypotheses proposed by the bilingual reformulator. Third, we will use word sense disambiguation approaches to deal with polysemy in the extracted bilingual lexicon. In the final line of our future work, we will continue experimenting our machine translation approach on other specific domains and comparing its performance to other domain adaptation techniques.

Acknowledgments

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 312651 – ePOOLICE.

References

Pratyush Banerjee, Jinhua Du, Baoli Li, Sudip Kr. Naskar, Andy Way, and Josef van Genabith. 2010. Combining Multi-Domain Statistical Machine Translation Models using Automatic Classifiers. In *Proceedings of the Ninth Conference of the Association for MT in the Americas, pages 141–150*.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the 4th Workshop on Statistical Machine Translation*.

Romaric Besançon, Gaël De Chalendar, Olivier Ferret, Christian Fluhr, Olivier Mesnard, and Hubert Naets. 2003. Concept-Based Searching and Merging for Multilingual Information Retrieval: First Experiments at CLEF 2003. In *C. Peters et al. (Ed.): CLEF 2003, Springer Verlag, Berlin, 2004.*

Romaric Besançon, Gaël De Chalendar, Olivier Ferret, Faïza Gara, Meriama Laib, Olivier Mesnard, and Nasredine Semmar. 2010. LIMA: A multilingual framework for linguistic analysis and linguistic resources development and evaluation. In *Proceedings of the seventh international conference on Language Resources and Evaluation*.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2012. Automatic Construction of a Multiword Expressions Bilingual Lexicon: A Statistical Machine Translation Evaluation Perspective. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon, COLING 2012,* India.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. In *Computational Linguistics - Special issue on using large corpora: II, Volume 19 Issue 2, MIT Press Cambridge, pages 263-311.*

Lars Bungum and Bjorn Gambäck. 2011. A Survey of Domain Adaptation in Machine Translation Towards a refinement of domain space. In *Proceedings of the India-Norway Workshop on Web Concepts and Technologies*.

Jaime Carbonell, Steve Klein, David Miller, Michael Steinbaum, Tomer Grassiany, and Jochen Frey. 2006. Context-Based Machine Translation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the Joint Conference EMNLP-CoNLL*.

Jorge Civera and Alfons Juan. 2007. Domain adaptation in statistical machine translation with mixture modelling. In *Proceedings of the Second Workshop on Statistical Machine Translation*.

Hal Daumé III and Jagadeesh Jagarlamudi. 2011. Domain Adaptation for Machine Translation by Mining Unseen Words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: short papers, pages 407–412.*

Mark W. Davis and William C. Ogden. 1997. QUILT: Implementing a large-scale cross-language text retrieval system. In *Proceedings of SIGIR*.

Gregory Grefenstette. 1998. Cross-Language Information Retrieval. In *The Information Retrieval Series, Vol. 2, Springer.*

Almut Silja Hildebrand, Matthias Eck, Stephan Vogel, and Waibel Alex. 2005. Adaptation of the translation model for statistical machine translation based on information retrieval. In *Proceedings of the European Association for Machine Translation Conference.*

John Hutchins. 2003. Machine Translation: General Overview. In *Ruslan (Ed.), The Oxford Handbook of Computational Linguistics, Oxford: University Press, pages 501-511.*

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicolas Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Conference ACL 2007, demo session, Prague, Czech Republic.*

Philipp Koehn, Barry Haddow, Philip Williams, and Hieu Hoang. 2010. More Linguistic Annotation for Statistical Machine Translation. In *Proceedings of the Fifth Workshop on Statistical Machine Translation and Metrics.*

John Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning.*

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 7th 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. 2003. A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics, Vol. 29(1).*

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.*

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

Nasredine Semmar and Meriama Laib. 2010. 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. 2014. Using Cross-Language Information Retrieval and Statistical Language Modelling in Example-Based Machine Translation. In *Proceedings of the 36th Translating and the Computer conference*, England.

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.*

Harold Somers. 2003. Machine Translation: Latest Developments. In *Ruslan (ed.), The Oxford Handbook of Computational Linguistics, Oxford: University Press, pages 513-527.*

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.*

Wei Wang, Klaus Macherey, Wolfgang Macherey, Franz Och, and Peng Xu. 2012. Improved Domain Adaptation for Statistical Machine Translation. In *Proceedings of the Association for Machine Translation in the Americas Conference.*

Dekai Wu and Pascale Fung. 2009. Semantic Roles for SMT: A Hybrid Two-Pass Model. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*