Building Multiword Expressions Bilingual Lexicons for Domain Adaptation of an Example-Based Machine Translation System

Nasredine Semmar, Meriama Laib
CEA, LIST, Vision and Content Engineering Laboratory, F-91191, Gif-sur-Yvette, France
nasredine.semmar@cea.fr, meriama.laib@cea.fr

Abstract

We describe in this paper a hybrid approach to build automatically bilingual lexicons of Multiword Expressions (MWEs) from parallel corpora. We more specifically investigate the impact of using a domain-specific bilingual lexicon of MWEs on domain adaptation of an Example-Based Machine Translation (EBMT) system. We conducted experiments on the English-French language pair and two kinds of texts: in-domain texts from Europarl (European Parliament proceedings) and out-of-domain texts from Emea (European Medicines Agency documents) and Ecb (European Central Bank corpus). The obtained results indicate that integrating domain-specific bilingual lexicons of MWEs improves translation quality of the EBMT system when texts to translate are related to the specific domain and induces a relatively slight deterioration of translation quality when translating general-purpose texts.

1 Introduction

Multiword Expressions (MWEs) play a major role in several natural language processing applications such as Machine Translation (MT) and Cross-Language Information Retrieval (CLIR) because they often characterize specific-domains vocabularies. The identification and the alignment of MWEs from parallel texts is a complex task (Sag et al., 2002; Hurskainen, 2008; DeNero and Klein, 2008; Bouamor et al., 2012; Ramisch, 2014; Semmar and Laib, 2017). Statistical approaches for word alignment (Brown et al., 1993) are unable to handle many-to-many alignments and as a result they cannot take into account correctly MWEs present in parallel corpora. For instance, the automatic word alignment tool Giza++ (Och and Ney, 2002) which implements IBM models can produce noisy (non perfect) outputs in particular when it aligns MWEs (Fraser and Marcu, 2007). As Statistical Machine Translation (SMT) systems use the translation table probabilities produced by this word alignment tool, the translation quality of these systems is highly impacted by the performance of Giza++.

In this paper, we discuss the application of domain adaptation to an Example-Based Machine Translation (EBMT) system. In particular, our investigation focuses on the impact of using a domain-specific bilingual lexicon of MWEs on the performance of this system. Two kinds of texts corpora are used in our investigation: in-domain texts from Europarl (European Parliament proceedings) and out-of-domain texts from Emea (European Medicines Agency documents) and Ecb (European Central Bank corpus).

The remainder of the paper is organized as follows. In section 2, we first present a state-of-the-art on domain adaptation for SMT, and then, we survey previous work on the use of multiword expressions in MT. Section 3 describes a hybrid approach to build bilingual lexicons of multiword expressions from parallel corpora. Section 4 presents briefly the EBMT system. In section 5, the experimental results are reported and discussed. Finally, the conclusions and future work are presented in section 6.

2 Related Work

In the last few years, a number of approaches have been explored to deal with domain adaptation in SMT (Hildebrand et al., 2005; Lewis et al., 2010; Banerjee et al., 2010; Bungum and Gambäck, 2011; Axelrod et al., 2011; Pecina et al., 2011; Wang et al., 2012; Mathur et al., 2015; Semmar et al., 2015). These approaches can be classified into three distinct categories: supervised (Daumé III, 2007; Foster and Kuhn, 2007; Koehn and Schroeder, 2007; Civera and Juan, 2007), semi-supervised (Ueffing, 2006; Ueffing et al., 2007; Ueffing et al., 2008) and unsupervised (Wu et al., 2008; Bertoldi and Federico, 2009). The first category (supervised approaches) consists in manipu-

Proceedings of Recent Advances in Natural Language Processing, pages 661–670,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_085

lating in-domain and out-of-domain data in order to adapt language and translation models (Eck et al. 2004; Daumé III, 2007; Foster and Kuhn, 2007; Koehn and Schroeder, 2007; Civera and Juan, 2007; Daumé III and Jagarlamudi, 2011). Daumé III (2007) used dictionary mining techniques to find translations for unseen words from comparable corpora and integrated these translations into a statistical phrase-based translation system. Likewise, Civera and Juan (2007) used monolingual corpora to adapt MT systems designed for Parliament domain to work in News domain. The obtained results showed significant gains in performance. On the other hand, Eck et al. (2004) limited adaptation to the target language model. The general-purpose language model is combined with one estimated on documents retrieved from the Web. These documents are obtained by using cross-language information retrieval techniques. The objective of the second category (semi-supervised approaches) is to train an SMT system on a small amount of data and then iteratively improve its performance by translating additional monolingual source language corpora and adding the reliable translations to the training corpus (Ueffing, 2006; Ueffing et al., 2007; Ueffing et al., 2008). Ueffing et al. (2007) explored monolingual texts in the source language to improve the MT system performance. They used an initial version of the translation system to translate texts in the source language. The generated translations and their sources are then used as a new parallel corpus for training an additional translation model. In this way, the translation system is adapted to the new source texts even if no bilingual corpus in this domain is available. The third category includes (unsupervised) approaches where in-domain bilingual corpora do not exist (Langlais, 2002; Wu et al., 2008; Bertoldi and Federico, 2009). Langlais (2002) added the content of a domain-specific lexicon into the training corpus used to generate the translation model of a SMT system. Wu et al. (2008) described a method which first uses out-of-domain corpora to train a baseline system and then uses in-domain translation dictionaries and in-domain monolingual corpora to improve the in-domain performance. Bertoldi and Federico (2009) investigated cross-domain adaptation of the state-of-the-art SMT system Moses (Koehn et al., 2007) by exploiting large monolingual corpora. They generated a synthetic parallel corpus by translating a monolingual adaptation corpus with an existing machine translation system, and they trained statistical models from the synthetic corpus. They reported that the most important improvement is provided by adapting the language model. The adaptation of the translation model and the reordering model produces small improvement.

As regards exploiting multiword expressions in domain adaptation, several works attempted to integrate these units in machine translation systems (Hurskainen, 2008; Ren et al., 2009; Carpuat and Diab, 2010; Pal et al., 2011; Bouamor et al., 2012; Semmar and Laib, 2017). Hurskainen (2008) described different ways to identify and isolate MWEs, and presented a method to mark MWEs in order to integrate this resource in a rule-based MT system. The author observed that when MWEs have been described in the linguistic analyzer of the MT system, they can be automatically included as part of the bilingual lexicon of the system. Ren et al. (2009) integrated bilingual MWEs into the decoder of the SMT system Moses. They observed a high improvement when they added a feature that identifies whether or not a bilingual phrase contains bilingual MWEs. Carpuat and Diab (2010) generalized this approach by replacing the binary feature by a count feature representing the number of MWEs in the source language phrase. Pal et al. (2011) converted the MWEs present in the parallel training corpus into single tokens in order to improve the phrase alignment quality. They reported that this preprocessing step improved the translation quality of a phrase-based statistical machine translation system. Recently, Semmar and Laib (2017) described, on the one hand, a hybrid approach to identify and find bilingual MWEs correspondences from a parallel corpus, and on the other hand, three strategies to integrate a bilingual lexicon of MWEs into the SMT system Moses.

The approach we implemented to deal with domain adaptation in our Example-Based Machine Translation system is close to the work of (Langlais, 2002) and (Hurskainen, 2008). It consists in adding a domain-specific bilingual lexicon of MWEs to the general-purpose bilingual dictionary of the EBMT system.

3 Building Bilingual Lexicons of Multiword Expressions

There are mainly two strategies to extract bilingual MWEs from parallel corpora. The first strategy consists to acquire translations of phrases from parallel corpora in one step. Phrases are not necessary MWEs, they are contiguous sequences of a few words that encapsulate enough context to be translatable (DeNero and Klein, 2008; Marchand and Semmar, 2011). The second strategy firstly, identifies monolingual MWEs candidates and then applies alignment approaches to find bilingual correspondences (Daille et al., 1994; Blank, 2000; Barbu, 2004). In the second strategy, the MWEs extraction can be processed by using symbolic methods founded on morphosyntactic patterns, or, through statistical approaches, which use automatic measures to rank MWEs candidates. Finally, MWEs extraction can be done by using hybrid approaches, which combine the two first strategies.

Our hybrid approach for MWEs alignment performs terminology extraction and alignment of MWEs from parallel texts in one step (Marchand and Semmar, 2011). The main idea of this approach is to consider the global task of identification and alignment of MWEs as an optimization problem. In order to linearize this optimization problem, we made the hypothesis that a MWE is composed of contiguous units. We use, then, integer linear programming to find an approximated optimal solution (DeNero and Klein, 2008). In this model, a sentence pair consists of two word sequences *e* and *f*, e_{ij} is the MWE from between-word positions *i* to *j* of *e*, and f_{kl} is the MWE from between-word positions *k* to *l* for *f*. A link is an aligned pair of MWEs, denoted (e_{ij}, f_{kl}). Each e_{ij} is allowed to be linked with several f_{kl} and each f_{kl} with several e_{ij}. An alignment *a* of the sentence pair (*e*, *f*) is a segmentation of the two sentences in MWEs with the set of links between these MWEs. We use a real-valued function ϕ to score links.

$$\phi : \{e_{ij}\} \times \{f_{kl}\} \rightarrow R$$

The score of an alignment *a* is the product of all the links inside it:

$$\phi(a) = \prod_{(e_{ij}, f_{kl}) \in a} \phi(e_{ij}, f_{kl})$$

In order to find the alignment (segmentation + links) that maximizes this score, we, first, introduce binary variables A_{ijkl} denoting whether a link exists between e_{ij} and f_{kl}. Furthermore, we introduce binary indicators E_{ij} and F_{kl} that denote whether some (e_{ij}, .) and (., f_{kl}) appear in *a*, respectively. Finally, we use $W_{ijkl} = log(\Phi(e_{ij}, f_{kl}))$ to transform the product into a sum. When optimized, the integer program yields the optimal alignment:

$$\begin{cases} \max \sum_{i,j,k,l} W_{i,j,k,l} A_{i,j,k,l} & & \\ \forall\, x : 1 \leq x \leq |e| & \sum_{i,j:i<x\leq j} E_{i,j} = 1 & (1) \\ \forall\, y : 1 \leq y \leq |f| & \sum_{k,l:k<y\leq l} F_{k,l} = 1 & (2) \\ \forall\, i,j & \sum_{k,l} A_{i,j,k,l} \geq E_{i,j} & (3) \\ \forall\, k,l & \sum_{i,j} A_{i,j,k,l} \geq F_{k,l} & (4) \\ \forall\, i,j,k,l & 2 \cdot A_{i,j,k,l} \leq E_{i,j} + F_{k,l} & (5) \end{cases}$$

Under the following constraints:

$$\begin{cases} 0 \leq i < |e|, & 0 < j \leq |e|, & i < j \\ 0 \leq k < |f|, & 0 < l \leq |f|, & k < l \end{cases}$$

Constraints (1) and (2) indicate that a word is inside exactly one phrase. Constraint (3) ensures that each phrase in the selected partition of *e* appears in at least one link (and likewise constraint (4) for *f*). Finally, constraint (5) ensures that if a link exists between e_{ij} and f_{kl} ($A_{ijkl} = 1$) then e_{ij} and f_{kl} are in the selected partitions of *e* and *f*. This constraint allows a phrase to be aligned with several other phrases. This integer program can work with any real-valued scoring function.

3.1 Scoring Based on Co-occurrence of MWEs

We use a sentence aligned corpus to compute the co-occurrence score. For each MWE, we consider its presence or absence in each sentence, and thus, the score between two MWEs e_{ij} and f_{kl} is computed as follows:

$$\phi_c(e_{ij}, f_{kl}) = \frac{\sum_{s' \in S} N_{s'}(e_{ij}) \times N_{s'}(f_{kl})}{\sum_{s \in S} N_s(e_{ij}) + N_s(f_{kl}) - N_s(e_{ij}) \times N_s(f_{kl})}$$

Where $N_s(e_{ij})$ is *1* if the phrase e_{ij} of the first language is present in the sentence *s* of the corpus *S* and *0* otherwise. $N_s(f_{kl})$ is similar for the other language. This score calculates the number of common presence of both MWEs divided by the number of total presence of either MWE. Note that if none of e_{ij} or f_{kl} appears in the whole corpus, the score is set to *0*. Indeed, if two MWEs appear exactly in the same bi-sentences, they are probably translation of each other and the score will be *1*.

3.2 Filtering MWEs Candidates

After obtaining an ordered list of bilingual MWEs, we filter the results, on the one hand, by removing the longer MWEs if a shorter MWE occurs in these candidates, and on the other hand, by keeping only MWEs which match with a list of morpho-syntactic patterns built manually (Bouamor et al., 2012).

4 Example-Based Machine Translation System

The translation process of the proposed EBMT system consists of several steps (Figure 1): retrieving translation candidates from a monolingual corpus using a cross-language search engine, producing translation hypotheses using a bilingual reformular, and generating the n-best translations from the combination of translation candidates and translation hypotheses (Semmar et al., 2015).

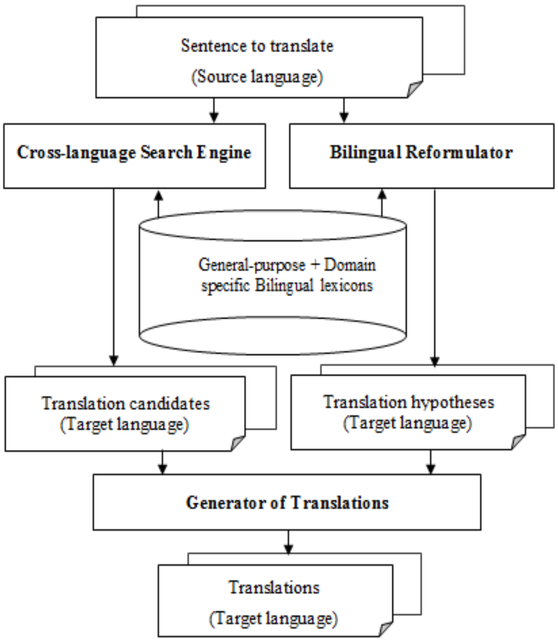

Figure 1: Main components of the Example-Based Machine Translation system.

The Cross-language Search Engine extracts for each sentence to translate (user's query) sentences or sub-sentences from an indexed monolingual corpus in the target language. These sentences or sub-sentences correspond to a total or a partial translation of the sentence to translate. This cross-language search engine is based on a deep linguistic analysis of the query and the monolingual corpus to be indexed, a bilingual lexicon and a weighted vector space model (Besançon et al., 2003). The deep linguistic analysis is achieved by means of the multilingual analyzer LIMA (Besançon et al., 2010) and the English-French lexicon is composed of 243539 entries[1]. The cross-language search engine returns translation candidates 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 (Part-Of-Speech, gender, number, etc.) which is provided by LIMA.

The role of the Bilingual Reformulator 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). These rules are built manually and are based on morpho-syntactic patterns (Table 1). Phrases corresponding to each pattern are identified by the syntactic analyzer of LIMA during the step of recognition of verbal and nominal chains. These phrases are extracted from the sentence to translate and are accepted by a FSM transducer whose outputs are instances of these phrases in the target language.

English pattern	French pattern
Adj-Noun	Noun-Adj
Adj-Adj-Noun	Noun-Adj-Adj
Noun-Prep-Noun	Noun-Prep-Noun
Noun-Prep-Adj-Noun	Noun-Prep-Adj-Noun
Noun-Noun	Noun-Noun

Table 1: Some frequent patterns to transform syntactic structures from English to French.

The Generator of Translations produces the n-best translations from a set of word lattices. These word lattices correspond to the combination of the results returned by the cross-language search engine and the bilingual reformulator. The combination 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. In order to find the best translation hypothesis from the set of word lattices, we first

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

use a statistical language model learned on lemmas and Part-Of-Speech tags of the target language corpus, and then, we apply a morphological generator (flexor) associated with linguistic information provided by LIMA to generate the n-best translations with words in their surface (inflected) forms.

5 Experimental Results

In this section, we first describe the corpora and the experimental setup used to train the Example-Based Machine Translation system and our baseline which is the state-of-the-art SMT system Moses[2]. Then, we give some details concerning the integration of the bilingual lexicons extracted from domain-specific corpora in these two systems, and we present the different sets of experiments that we carried out with a brief discussion.

5.1 Data and Experimental Setup

In order to study the impact of using a domain-specific bilingual lexicon on the performance of the EBMT system and Moses, we conducted our experiments on three English-French parallel corpora (Table 2): Europarl (European Parliament proceedings), Emea (European Medicines Agency documents) and Ecb (European Central Bank corpus). These corpora were extracted from the open parallel corpus OPUS (Tiedemann, 2012). We use the factored translation model of Moses. It is an extension of the phrase-based models which are limited to the mappings of phrases without any explicit use of linguistic information. The factored model enables the use of additional markup at the word level. Our model operates on lemmas instead of surface forms because the entries of the general-purpose dictionary and the bilingual lexicon of MWEs are in lemma forms. Therefore, training corpora are lemmatized using the multilingual analyzer LIMA.

Evaluation consists in comparing translation results produced by Moses 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 six 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, 500 pairs of sentences from Emea and 500 pairs of sentences from Ecb as Out-Of-Domain corpora.

Run n°.	**Training (# sentences)**	**Tuning (# sentences)**
1	150K+10K (Europarl+Emea)	2K+0.5K (Europarl+Emea)
2	150K+20K (Europarl+Emea)	2K+0.5K (Europarl+Emea)
3	150K+30K (Europarl+Emea)	2K+0.5K (Europarl+Emea)
4	500K+10K (Europarl+Ecb)	2K+0.5K (Europarl+Ecb)
5	500K+20K (Europarl+Ecb)	2K+0.5K (Europarl+Ecb)
6	500K+30K (Europarl+Ecb)	2K+0.5K (Europarl+Ecb)

Table 2: Corpora details used to train Moses language and translation models, and to build the database of the EBMT system (K refers to 1000).

The goal of our experiments is to show the impact of the domain vocabulary on the translation results. The domain vocabulary is identified by a bilingual lexicon of MWEs which is extracted automatically from the specialized parallel corpus (Emea or Ecb) using our MWEs alignment approach. In the case of the EBMT system, the specialized bilingual lexicon is added to the general-purpose English-French lexicon which is used jointly by the cross-language search engine and the bilingual reformulator. In the case of Moses, on the one hand, we added the English-French lexicon of the cross-language search engine to the training data (Europarl), and on the other hand, we integrated the domain-specific bilingual lexicon of MWEs using the following three methods (Bouamor et al., 2012):

- Moses$_{\text{CORPUS}}$: In this method, we include the extracted MWE pairs of the bilingual lexicon to the training data of Moses.

- Moses$_{\text{TABLE}}$: This method consists to insert the extracted MWE pairs in the phrase table which is generated while training Moses.

- Moses$_{\text{FEATURE}}$: In this method, we extend Moses$_{\text{TABLE}}$ by adding a new feature indicating whether a phrase is a MWE or not.

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

5.2 Results and Discussion

The performance of Moses and the EBMT system is evaluated using the BLEU score (Papineni et al., 2002) on the two test sets for the six runs described in the previous section. Note that we consider one reference per sentence. Table 3 illustrates the performance of the EBMT system for In-Domain and Out-Of-Domain texts. Table 4 and Table 5 report the BLEU scores of the different integration strategies of the specialized bilingual lexicon in Moses respectively for In-Domain and Out-Of-Domain texts.

Run n°.	In-Domain (Europal)	Out-Of-Domain (Emea and Ecb)
1	32.05	29.02
2	31.03	30.26
3	29.92	31.84
4	33.82	32.64
5	33.31	33.71
6	32.93	37.74

Table 3: BLEU scores of the EBMT system for In-Domain and Out-Of-Domain texts.

Run n°.	In-Domain (Europal)		
	Moses$_{CORPUS}$	Moses$_{TABLE}$	Moses$_{FEATURE}$
1	32.79	32.10	32.91
2	34.06	33.51	34.12
3	34.61	34.19	34.68
4	37.58	37.44	37.63
5	37.61	37.52	37.72
6	37.79	37.76	37.84

Table 4: BLEU scores of Moses for In-Domain texts.

Run n°.	Out-Of-Domain (Emea and Ecb)		
	Moses$_{CORPUS}$	Moses$_{TABLE}$	Moses$_{FEATURE}$
1	23.38	23.05	23.59
2	23.95	23.71	24.12
3	25.37	24.87	25.40
4	25.92	25.67	26.15
5	26.85	26.73	27.07
6	31.04	30.94	31.39

Table 5: BLEU scores of Moses for Out-Of-Domain texts.

As shown in Table 3 and Table 4, for In-Domain texts, Moses and the EBMT system achieve a relatively high BLEU score and the score of Moses is better in all the runs. For the Out-Of-Domain test corpora, the EBMT system performs better than Moses. This may be due to the bilingual lexicon of MWEs which is built automatically from the specialized parallel corpus (Emea or Ecb). It seems that it has had a significant impact on the result of the EBMT system, it improved regularly its BLEU score in all the runs. These results also show that adding a specialized lexicon of MWEs to the translation model of Moses improves translation quality of Out-Of-Domain texts without loss of translation quality when translating In-Domain texts, and confirm the results obtained by Ren et al. (2009) and Bouamor et al. (2012). However, in the case of the EBMT system, adding the specialized lexicon to the general-purpose bilingual dictionary of the search engine translation has had a negative impact on the translation quality of In-Domain texts. For example, adding a bilingual lexicon of MWEs built from the Emea specialized parallel corpus composed of 30K sentences to the 150K sentences of Europarl reported a gain of 2.82 BLEU points when translating Out-Of-Domain texts but had led to a loss of 2.13 BLEU points when translating In-Domain texts (Table 3: runs 1 and 3). Likewise, adding a bilingual lexicon of MWEs built from the Ecb specialized parallel corpus composed of 30K sentences to the 150K sentences of Europarl reported a gain of 5.10 BLEU points when translating Out-Of-Domain texts but had led to a loss of 0.89 BLEU points when translating In-Domain texts (Table 3: runs 4 and 6). The other important point to mention here is that the integration method Moses$_{FEATURE}$ provides the best BLEU score in all the runs for In-Domain texts (Table 4) and Out-Of-Domain texts (Table 5).

In order to evaluate and analyze the translation quality of the EBMT system and Moses when translating specific-domain texts, we take an example of translations drawn from the Ecb test corpus (Table 6). For this sentence, the EBMT system and Moses provide close translations and these translations are more or less correct, but translations provided by Moses contain many spelling and grammatical errors. In this example, the English word “shares” was identified by the morpho-syntactic analyzer used by the EBMT system as a noun and is translated as “actions”. This translation is the right one despite the fact that the English-French lexicon contains for the word “share” several translations (avoir part à, contingent, diviser, lot, lotir, part, partager, participer, etc.). On the other hand, Moses translates the word “shares” with the word “partage” which is not a correct translation. It seems that Moses has taken this translation from Europal training corpus instead of the Ecb corpus. Likewise, Moses fails to translate correctly the word “investors” (plural of the word “investor”) even if it has translated correctly the singular form of this word (investor).

Example Input (Ecb): for example, in France, statistics show that a non resident *investor* holds *shares* on average only for five months (compared to eleven months for resident *investors*).	
Reference	en France, par exemple, les statistiques montrent qu’un *investisseur* non résident ne conserve, en moyenne, ses *actions* que pour une période de cinq mois (contre onze mois pour un *investisseur* résident).
EBMT system: Run 6	par exemple, en France, les statistiques montrent qu'un *investisseur* non résident garde des *actions* en moyenne seulement pour cinq mois (par comparaison avec onze mois pour les *investisseurs* résidents).
Moses$_{\text{FEATURE}}$: Run 6	par exemple, en France, les statistiques montrent qu’un *investisseur* non résidente détient *partage* en moyenne seulement pour cinq mois (par rapport à onze mois pour résidant *investors*).

Table 6: Translations produced by the EBMT system and Moses for a sentence from Ecb corpus.

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. We think that taking 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 could reduce the impact of the issues related to syntactic parsing. However, for the presence of the polysemy in the bilingual lexicon, the EBMT system has no specific treatment. On the other hand, we noted that most of Moses’s translation errors for Out-Of-Domain sentences are related to vocabulary. In another example of translation, Moses proposes the compound word “capitalisation du marché” as a translation for the expression “market capitalisation” instead of the compound word “capitalisation boursière” which is more precise. In SMT systems such as Moses, phrase tables are the main knowledge source for the machine translation decoder which consults these tables to figure out how to translate an input sentence. These tables are built using Giza++ but, as we mentioned before, this tool could produce errors in particular when it aligns multiword expressions (Fraser and Marcu, 2007). Finally on this point, we can observe that the major issues of Moses concern errors produced by Giza++ when aligning MWEs (translation model), incorrect spelling and poor grammar generated by the decoder (language model). To handle the first issue, we proposed to take into account the specialized bilingual lexicon extracted with the MWEs aligner into Moses’s phrase table and we added a new feature indicating whether a word comes from this lexicon or not (Moses$_{\text{FEATURE}}$ method). However, for spelling and poor errors, statistical machine translation toolkits like Moses have no specific treatment because they have not been designed with grammatical error correction in mind.

6 Conclusion and Future Work

We have presented in this paper a hybrid approach to build automatically bilingual lexicons of Multiword Expressions (MWEs) from parallel corpora. We have also investigated the impact of using a domain-specific bilingual lexicon of MWEs on domain adaptation for two MT systems: the state-of-the-art SMT system Moses and an Example-Based Machine Translation (EBMT) system. The obtained results have showed that adding a specialized bilingual lexicon of MWEs to the general-purpose dictionary of the EBMT system improves significantly its performance for Out-Of-Domain texts. We have compared the results of the EBMT system with those of Moses, and the results of the EBMT system are always better for Out-Of-Domain texts and almost comparable for In-Domain texts. This study offers several open issues for future work. First, we plan to analyze the obtained translations for both the EBMT and the SMT systems in terms of perplexities in order to improve the modeling of Out-Of-Vocabulary words. The second perspective is to adapt and evaluate the MWEs alignment approach and the EBMT system to new language pairs and new domains. We also expect to explore the use of recurrent neural networks language models for rescoring the n-best translations produced by the EBMT system.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 700381.

References

A. Axelrod, X. He, and J. Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of EMNLP 2011*.

P. Banerjee, J. Du, B. Li, S.K. Naskar, A. Way, and J. van Genabith. 2010. Combining Multi-Domain Statistical Machine Translation Models using Automatic Classifiers. In *Proceedings of AMTA 2010*.

A. M. Barbu. 2004. Simple linguistic methods for improving a word alignment algorithm. In *Proceedings of the 7th International Conference on the Statistical Analysis of Textual Data*.

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

R. Besançon, G. De Chalendar, O. Ferret, F. Gara, M. Laib, O. Mesnard, and N. Semmar. 2010. LIMA: A multilingual framework for linguistic analysis and linguistic resources development and evaluation. In *Proceedings of LREC 2010*.

R. Besançon, G. De Chalendar, O. Ferret, C. Fluhr, O. Mesnard, and H. Naets. 2003. Concept-Based Searching and Merging for Multilingual Information Retrieval: First Experiments at CLEF 2003. *C. Peters et al. (Ed.): CLEF 2003,* Springer Verlag.

I. Blank. 2000. Terminology extraction from parallel technical texts. *Parallel text processing,* Springer.

D. Bouamor, N. Semmar, and P. Zweigenbaum. 2012. Automatic Construction of a MultiWord Expressions Bilingual Lexicon: A Statistical Machine Translation. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III), COLING 2012.*

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics, 19(2).*

L. Bungum and B. 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*.

M. Carpuat and M. Diab. 2010. Task-based Evaluation of Multiword Expressions: a Pilot Study in Statistical Machine Translation. In *Proceedings of Human Language Technology conference and the North American Chapter of the Association for Computational Linguistics conference*.

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

B. Daille, E. Gaussier, and J. M. Langé. 1994. Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th conference on Computational linguistics ACL 1994*.

H. Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings Conference of the Association for Computational Linguistics. ACL 2007*.

H. Daumé III and J. Jagarlamudi. 2011. Domain Adaptation for Machine Translation by Mining Unseen Words. In *Proceedings of ACL 2011*.

J. DeNero and D. Klein. 2008. The complexity of phrase alignment problems. In *Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies*.

M. Eck, S. Vogel, and A. Waibel. 2004. Language model adaptation for statistical machine translation based on information retrieval In *Proceedings of the 4th International Conference on language resources and evaluation LREC 2004*.

G. Foster and R. Kuhn. 2007. Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation.*

A. Fraser and D. Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics, 33(3).*

A. S. Hildebrand, M. Eck, S. Vogel, and W. Alex. 2005. Adaptation of the translation model for statistical machine translation based on information retrieval. In *Proceedings of the EAMT 2005*.

A. Hurskainen. 2008. Multiword Expressions and Machine Translation. In *Technical Report No 1 in Language Technology, Institute for Asian and African Studies*, University of Helsinki, Finland.

P. Koehn and J. Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the second workshop on statistical machine translation*.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual meeting of the Association for Computational Linguistics ACL 2007*.

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

W. D. Lewis, C. Wendt, and D. Bullock. 2010. Achieving Domain Specificity in SMT without Overt Siloing. In *Proceedings of LREC 2010*.

M. Marchand and N. Semmar. 2011. A Hybrid Multi-Word Terms Alignment Approach Using Word Co-occurrence with a Bilingual Lexicon. In *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics LTC'11*.

P. Mathur, M. Federico, S. Köprü, S. Khadivi, and H. Sawaf. 2015. Topic Adaptation for Machine Translation of E-commerce Content. In *Proceedings of MT Summit XV*.

F. J. Och and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL 2002*.

S. Pal, T. Chakraborty, and S. Bandyo-padhyay. 2011. Handling Multiword Expressions in Phrase-Based Statistical Machine Translation. In *Proceedings of the XIII Machine Translation Summit*.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*.

P. Pecina, A. Toral, A. Way, V. Papa-vassiliou, P. Prokopidis, and M. Giagkou. 2011. Towards Using Web-Crawled Data for Domain Adaptation in Statistical Machine Translation. In *Proceedings of EAMT 2011*.

C. Ramisch. 2014. Multiword Expressions Acquisition: A Generic and Open Framework. *Theory and Applications of Natural Language Processing Monographs*, Springer.

Z. Ren, Y. Lu, J. Cao, Q. Liu, and Y. Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions, ACL-IJCNLP 2009*.

I. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of CICLing '02*.

N. Semmar, O. Zennaki and M. Laib. 2015. Evaluating the Impact of Using a Domain-specific Bilingual Lexicon on the Performance of a Hybrid Machine Translation Approach. In *Proceedings of Recent Advances in Natural Language Processing RANLP 2015*.

N. Semmar and M. Laib. 2017. Integrating Specialized Bilingual Lexicons of Multiword Expressions for Domain Adaptation in Statistical Machine Translation. In *Proceedings of PACLING 2017*.

J. Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of LREC 2012*.

N. Ueffing. 2006. Using monolingual source-language data to improve MT performance. In *Proceedings of the international workshop on spoken language translation IWSLT 2006*.

N. Ueffing, G. Haffari, and A. Sarkar. 2007. Semi-supervised model adaptation for statistical machine translation. *Machine Translation*, 21(2).

N. Ueffing, G. Haffari, and A. Sarkar. 2008. Semi-supervised learning for machine translation. *Learning machine translation*, NIPS Series, MIT Press.

W. Wang, K. Macherey, W. Macherey, F. Och, and P. Xu. 2012. Improved Domain Adaptation for Statistical Machine Translation. In *Proceedings of AMTA 2012*.

H. Wu, H. Wang, and C. Zong. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd International Conference on Computational Linguistics COLING'08.*