A hybrid approach for automatic extraction of bilingual Multiword Expressions from parallel corpora

Nasredine Semmar

A Hybrid Approach for Automatic Extraction of Bilingual Multiword Expressions from Parallel Corpora

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

Abstract

Specific-domain bilingual lexicons play an important role for domain adaptation in machine translation. The entries of these types of lexicons are mostly composed of MultiWord Expressions (MWEs). The manual construction of MWEs bilingual lexicons is costly and time-consuming. We often use word alignment approaches to automatically construct bilingual lexicons of MWEs from parallel corpora. We present in this paper a hybrid approach to extract and align MWEs from parallel corpora in a one-step process. We formalize the alignment process as an integer linear programming problem in order to find an approximated optimal solution. This process generates lists of MWEs with their translations, which are then filtered using linguistic patterns for the construction of the bilingual lexicons of MWEs. We evaluate the bilingual lexicons of MWEs produced by this approach using two methods: a manual evaluation of the alignment quality and an evaluation of the impact of this alignment on the translation quality of the phrase-based statistical machine translation system Moses. We experimentally show that the integration of the bilingual MWEs and their linguistic information into the translation model improves the performance of Moses.

Keywords: Bilingual lexicon, Multiword expression, Terminology extraction, Domain adaptation, Statistical machine translation

1. Introduction

A MultiWord Expression (MWE) is a combination of words for which syntactic or semantic properties of the whole expression cannot be obtained from its components (Sag et al., 2002). Such units could be collocations, compound words, named entities, idioms, etc. They constitute an important part of the lexicon of any natural language (Jackendoff, 1997). Bilingual lexicons of MWEs play a vital role in Machine Translation (MT) and Cross-Language Information Retrieval (CLIR) because for a specific domain the specialized vocabulary is largely dominated by MWEs. The manual construction of these lexicons is costly and time-consuming. Word alignment approaches are often used to automatically construct bilingual lexicons from parallel corpora. Several word alignment approaches have been explored (Daille et al., 1994; Barbu, 2004) and many automatic word alignment tools are available, such as Giza++ (Och and Ney, 2000). However, most of these tools are efficient only to align single words (Fraser and Marcu, 2007). In this paper, we describe and evaluate a hybrid approach to automatically extract and align MWEs from an English-French parallel corpus. In contrast to traditional approaches for MWEs alignment which consist in firstly identifying monolingual MWEs candidates and secondly applying alignment to find bilingual correspondences, our approach extracts and aligns MWEs in a one-step process.

The remainder of the paper is organized as follows. We define in Section 2 the notion of Multiword Expression and describe different types of MWEs with examples. In Section 3, we survey previous works addressing the tasks of extracting and aligning MWEs from parallel corpora. Section 4 introduces our hybrid approach to build bilingual lexicons of MWEs from sentence aligned parallel corpora. The experimental results are reported and discussed in Section 5. Finally, we present in Section 6 the conclusion and future work.

2. Multiword Expressions

In Natural Language Processing (NLP), a multiword expression refers to a non-compositional sequence of words whose exact and unambiguous meaning, connotation and syntactic properties cannot be derived from the meaning or connotation of its components (Sag et al., 2002). MWEs are frequently used in written texts and constitute a significant part of the language lexicon. Sag et al. (2002) classify multiword expressions into two main categories: lexicalized phrases and institutionalized phrases. Lexicalized phrases "have at least partially idiosyncratic syntax or semantics, or contain "words" which do not occur in isolation". Institutionalized phrases are "semantically and syntactically compositional, but statistically idiosyncratic".

2.1 Lexicalized Phrases

In a decreasing order of lexical rigidity, lexicalized phrases are broken down into three classes: fixed expressions, semi-fixed expressions and syntactically-flexible expressions.

Fixed expressions are non-compositional sequences of words. They are syntactically and morphologically rigid and undergo neither internal modification nor morphological and syntactical variations (e.g. "*nest of vipers*" in English or "*pomme de terre*" in French). To determine whether or not a sequence of words is a fixed expression, we can use linguistic criteria such as using synonyms or adding words between its components (e.g. "*nest of many black vipers*" in English or "*pomme de jolie terre lointaine*" in French). Fixed expressions can be considered as single entries in the dictionary.

A semi-fixed expression is a non-compositional sequence of words whose components do not contribute to its figurative meaning. Semi-fixed expressions should respect a strict word order and some of them undergo limited lexical and morphological variability such as inflection and some variation in the reflexive form. According to their

characteristics, they can be broken down into three basic categories: non-decomposable idioms, proper names and some compound nominals (Sag et al., 2002).

Non-decomposable idioms do not undergo syntax variability but their components accept lexical changes such as pronominal reflexivity form (e.g. "*wet him-self*", "*wet themselves*"), verbal inflection ("*kick the bucket*", "*kicked the bucket*") or passivization (e.g. "*briser le silence*" or "*le silence est brisé*" in French). Proper Names "are syntactically highly idiosyncratic" (Sag et al., 2002). They can be complex with two or three proper names as components, including person, place and organization names. Compound nominals are syntactically unalterable and undergo number inflection (e.g. "*car park(s)*" in English or "*pomme(s) de terre*" in French).

Unlike semi-fixed expressions, syntactically-flexible expressions undergo a wide degree of syntactic variation such as passivization (e.g. "*The cat was let out of the bag*") and allow external elements to intervene between their components (e.g. "*slow the car down*"). This type of expressions includes verb-particle constructions, decomposable idioms. Particle verbs constructions are made up of a verb whose meaning is modified by one or more particles. They can be either semantically idiosyncratic such as "*brush up on*" or compositional such as "*take after*", "*look out*", "*go back*" and "*run over*". Decomposable idioms tend to be syntactically flexible to some degree that is unpredictable. Semantically, they behave as if their components were linked parts contributing independently to the figurative interpretation of the expression as a whole.

2.2 Institutionalized Phrases

Institutionalized phrases are semantically and syntactically fully compositional, but statistically idiosyncratic (Sag et al., 2002). They occur in a high frequency and their idiosyncrasy is statistical rather than linguistic. They generally allow one available meaning. Institutionalized phrases often refer to "collocations", described as sequences of words that statistically have a high probability to appear together whether they are contiguous or not (e.g. "*make a difference*").

3. Related Work

Automatic identification of MWEs from texts is a real challenge in Natural Language Processing. This is due to the diversity and the complexity of their lexical, syntactic and semantic characteristics (Moon, 1998; Riehemann 2001; Sag et al. 2002). Two approaches have emerged to extract bilingual MWEs from parallel corpora. The first approach consists of acquiring translations of MWEs from parallel corpora in one-step (DeNero and Klein, 2008; Marchand and Semmar 2011). DeNero and Klein (2008) consider, on the one hand, MWEs as phrases composed of contiguous sequences of words that encapsulate enough context to be translatable, and on the other hand, that the problem of finding an optimal alignment between bilingual MWEs can be cast as an integer linear program. Marchand and Semmar (2011) used an approach which followed to some extent that of DeNero and Klein (2008) while they added two scoring functions based on co-occurrence and a seed single word bilingual dictionary. The second approach for extracting bilingual MWEs from parallel corpora, firstly, identifies monolingual MWEs candidates and then applies alignment techniques to find bilingual correspondences (Daille et al., 1994; Blank 2000; Barbu 2004; Deng et al., 2005; Samuelsson et al., 2007; MacCartney et al., 2008; Lefever et al., 2009; Semmar et al., 2011; Bouamor et al., 2012). In the second approach, MWEs extraction can be processed by using symbolic methods based on linguistic patterns (Dagan et al., 1994; Okita et al., 2010; Bouamor et al., 2012), or, through statistical approaches which use automatic measures to rank MWEs candidates (Pearce 2002; Evert and Krenn 2005; Zhang et al. 2006; Villavicencio et al. 2007; Vintar et al., 2008). Finally, MWEs extraction can be done by using hybrid approaches, which combine statistical information with some kinds of linguistic information such as syntactic and semantic properties (Baldwin and Villavicencio 2002; Van de Cruys and Villada Moiron 2007; Caseli et al., 2010). Dagan and Church (1994) proposed to use syntactic analysis to extract terminology. MWEs are then extracted by grouping linguistically related terms. In the same way, Okita et al. (2010) proposed to link across two languages MWEs according to their syntactic and lexical information. Tufis and Ion (2007) introduce a linguistic approach in which they claim that MWEs keep in most cases the same morpho-syntactic structure in the source and target languages. Statistical approaches also have proven to be useful in collecting bilingual MWEs from parallel corpora. Kupiec (1993) introduced the use of machine learning algorithms such as the Expectation Maximization (EM) to extract MWEs. Similarly, Vintar and Fiser (2008) proposed to extract bilingual MWEs by translating MWEs from a well-known language (English) to a low resource language (Slovene) by using machine translation. They have shown that their translation-based approach performs better than using linguistic approaches. However, they did not combine these two kind of approaches. The combination of such approaches enables to extract finer MWEs. In this way, Wu and Chang (2004) and later Boulaknadel et al. (2008), proposed to use syntactic and statistical analysis to extract bilingual MWEs from a parallel corpus. The main aspect of their approach is a monolingual parsing to extract MWEs combined with statistical detection in each language, then, they confront candidates from each side to find bilingual MWEs. Other approaches proposed to use machine translation to translate MWEs candidates found with a syntactic analysis (Seretan and Wehrli, 2007).

4. Building Bilingual Lexicons of MWEs

The process of building MWEs bilingual lexicons from parallel corpora is composed of the following two steps:

1. MWEs extraction and alignment using scoring functions.
2. MWEs candidates filtering using morpho-syntactic patterns.

4.1 Extraction and Alignment of MWEs

In this section, we describe our approach to extract and align MWEs from an English-French parallel corpus in a one-step process (Marchand and Semmar, 2011; Semmar and Marchand, 2017; Semmar and Laib, 2017). This approach is hybrid because it considers the global task of identification and alignment of MWEs as an optimization

problem and it uses external linguistic resources: a seed single word bilingual dictionary and morpho-syntactic patterns. It handles MWEs which are composed of contiguous units. As the only restriction we made is the contiguity of MWEs, the alignment task is a NP-hard problem. We formalize, then, the alignment task as an integer linear programming problem to find an approximated optimal solution (DeNero and Klein, 2008; Marchand and Semmar, 2011).

In this formalization, 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 (objective function) 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[1]:

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

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 MWE. Constraint (3) ensures that each MWE 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 MWE to be aligned with several other MWEs. This integer program can work with any real-valued scoring function.

Because the only restriction we made on MWEs is their contiguity, the alignment task model can handle the following MWEs:

- Compound nouns: A sequence of words acting as a single noun. These compounds could be proper nouns or common nouns.
- Phrasal verbs: Collocations containing a verb followed by a preposition.
- Verb constructions: Concatenations of a verb and a noun collocation.
- Verb phrase idioms: Verb phrases whose semantics are non-compositional.
- Verb-prepositional phrase constructions: Verbs attached to prepositional phrases without compositional semantics.

The integer linear program describing the alignment task can work with any scoring function. To solve this program, we used two scoring functions.

4.1.1 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. 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*.

As expected with this scoring function, if the program finds an unknown word or if the word co-occurs with no other word in the translated sentence, all the links containing this word will obtain a score equal to *0*. Therefore, the global score of the alignment will be also equal to *0* whatever the other links because the scoring function is multiplicative. In order to overcome this limit, we used an external linguistic resource: a seed bilingual dictionary.

4.1.2 Scoring Based on a Bilingual Dictionary

The bilingual dictionary provides several word-to-word alignments. We want to comply with these alignments as often as possible as we infer that they are mostly correct. The dictionary also gives negative alignment information. Of course, if two words are not aligned by the dictionary we can't take for sure that they shouldn't, and we have to take that into account. The dictionary score is calculated with the following formula:

[1] We used the open source solver GLPK (www.gnu.org/s/glpk/).

$$\phi(e_{ij}, f_{kl}) = \frac{a \times R_1 + b \times R_0}{a \times R_1 + b \times R_0 + c \times N_1 + d \times N_0}$$

R_1 is the number of respected links, R_0 is the number of respected non-links, N_1 is the number of non-respected links, and N_0 is the number of non-respected non-links. The coefficients *a*, *b*, *c* and *d* can be adapted to balance the relative influence of the four terms. We analyzed a small corpus that allowed us to empirically choose the use of the following values: $a = b = c = 1$ and $d = 0.5$. The score is calculated for each part of the bilingual MWEs and then the two of them are multiplied. We have to take into account R_0 and N_0 because otherwise the whole bi-sentence would be the optimal segmentation.

As we can see, this score has a double effect. First, it gives a high score if the bilingual MWEs respect dictionary word-to-word alignment. Second, due to R_0, it sets a threshold score for unknown couples. Both effects can have a positive role in alignment task as we will see in the examples below. The dictionary-based score is not intended to be used separately. It is mixed with co-occurrence score. We used an English-French bilingual dictionary containing 243539 entries with doubles[2]. It is important to point out here that the entries of the English-French bilingual dictionary are in lemmas forms. Therefore, to take full advantage of this dictionary, it is preferable to lemmatize the parallel corpus before extracting and aligning MWEs. However, as some surface forms are similar to lemmas in English and French languages, we experimented the two possibilities. The parallel corpus has been lemmatized using the multilingual analyzer LIMA (Besançon et al., 2010).

4.2 Filtering MWEs Candidates

The result of the previous step (Extraction and alignment of MWEs) is a list of alignment links candidates. Each link is composed of a MWE in the source language and its translation candidate in the target language. This step covers all the categories of MWEs (Compound nouns, Phrasal verbs, etc.).

In order to increase the accuracy of this step, we filter the results, on the one hand, by removing the longer MWEs if shorter MWEs occur in these candidates, and on the other hand, by selecting only MWEs which match with a list of morpho-syntactic patterns (Table 1). The MWEs candidates are composed of sequences of words of size $n \geq 2$ that follow the most frequent Part-Of-Speech patterns. Part-Of-Speech tags of the components of each MWE are provided by the multilingual analyzer LIMA after processing the parallel corpus. We have built manually the list of morpho-syntactic patterns by analyzing the sequences of Part-Of-Speech tags corresponding to the MWEs candidates provided by the first step. We have also used the patterns derived by other research works (Bouamor et al., 2012). At the end, we obtained a set of 25 patterns, most of which are related to noun phrases. However, it is important to note that a same pattern in a source language could have several patterns in the target language. It is for instance the case of the English pattern "Adj-Noun-Noun" which have three equivalent patterns in French "Noun-Adj-Prep-Noun", "Noun-Noun-Adj" and "Noun-Prep-Noun-Adj".

Contrary to the work of Bouamor et al. (2012), we consider a MWE in the target language as a translation of a MWE in the source language only if the morpho-syntactic pattern of the source MWE has an equivalent morpho-syntactic pattern in the target language. This led us to take a decision on the set of MWEs which are most probable to be entries of the bilingual lexicon. Indeed, the objective of using filtering morpho-syntactic is to identify and separate only the strongest possible MWEs from among the list of all possible MWEs candidates. Naturally, this step increases the precision of the alignment but at the same time it decreases the recall.

English Pattern	Equivalent French Pattern
Adj-Noun	Adj-Noun
Adj-Noun	Noun-Adj
Noun-Noun	Noun-Prep-Noun
Noun-Noun	Noun-Noun
Adj-Noun-Noun	Noun-Adj-Prep-Noun
Adj-Noun-Noun	Noun-Noun-Adj
Adj-Noun-Noun	Noun-Prep-Noun-Adj
Noun-Prep-Noun	Noun-Prep-Noun
Noun-Noun-Noun	Noun-Prep-Noun-Det-Noun
Noun-Noun-Noun	Noun-Prep-Noun-Noun
Adj-Adj-Noun	Noun-Adj-Adj
Noun-Noun-Noun-Noun	Noun-Prep-Noun-Noun-Prep-Det-Noun
Adj-Noun-Noun-Noun	Noun-Prep-Det-Noun-Prep-Noun-Adj
Adj-Adj-Noun-Noun	Noun-Noun-Adj-Adj

Table 1: Some English and French filtering morpho-syntactic patterns (Adj refers to an Adjective, Prep to a Preposition, and Det to a Determiner).

5. Experimental Results

The quality of alignment of MWEs and the impact of using MWEs on machine translation have been evaluated, firstly, manually, by comparing the results of our approach with a reference alignment; and secondly automatically by using the results of our MWEs alignment approach to build the translation model of the state-of-the-art statistical machine translation system Moses (Koehn et al., 2007).

5.1 Manual Evaluation

Our hybrid approach for MWEs alignment and the baseline Giza++ (Och and Ney, 2000) have been evaluated using the evaluation metrics defined in (Mihalcea et al., 2003). The

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

corpus used to evaluate the performance of the English-French MWE aligners is composed of a set of 1992 parallel sentences extracted from Europarl (European Parliament Proceedings). This parallel corpus is composed of 46265 English words and 49332 French words and has been used to build manually the reference alignment by the Yawat tool (Germann, 2008). Alignment with Giza++ was achieved in source–target and target–source directions and the results were merged using the union heuristic.

At first glance, we can see that the combination of the scoring using co-occurrence and the scoring based on the bilingual dictionary with the filtering patterns provides the best performance of our MWEs alignment approach. It clearly appears that keeping only MWEs candidates that have equivalent morpho-syntactic patterns in source and target languages has had a significant impact on the precision of the alignment. This filtering step certainly has improved the precision but the recall has dropped.

MWEs Aligner	Precision	Recall	F-measure
Baseline (Giza++)	0.83	0.37	0.51
Co-occurrence	0.61	0.63	0.61
Co-occurrence + Bilingual dictionary	0.85	0.54	0.66
Co-occurrence + Bilingual dictionary + Filtering patterns	0.95	0.52	0.67

Table 2: Performance of Giza++ and our MWEs alignment approach.

We observed after aligning some sentences that when both sentence structures are similar, our MWEs aligner performs well. The segmentation is word to word or MWE to MWE depending on what is more frequent in the corpus. Moreover, the surjective formulation of the problem allows our approach to detect expressions in two parts. We can see in the following example that both the English words "*role*" and "*play*" are linked to the French word "*rôle*" (Figure 1).

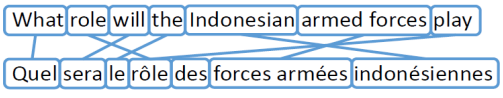

Figure 1: Example of a correct alignment with only the co-occurrence socre.

We have also observed some improvements due the information provided by the bilingual dictionary, as presented in Figure 2. In this example, the bilingual dictionary provides the alignments: "*be/être*", "*decided/décidé*" and "*there/y*". Therefore, our MWEs aligner reconstructs the whole expression "*is to be decided on there/doit y être décidé*". Moreover, the links "*concrete/concret*" and "*programme/programme*" are consolidated by the bilingual dictionary.

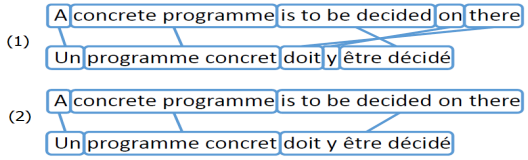

Figure 2: Improvement of alignments (1) Alignment without the bilingual dictionary and (2) Alignment with the bilingual dictionary.

Finally, it should be noted that the link "*is to be decided on there/doit y être décidé*" is abandonned after the step of filtering because no morpho-syntactic pattern matches this expression.

5.2 Evaluation through a Translation Task

The unavailability of a reference alignment of a significant size for MWEs does not allow us to achieve a large scale evaluation. That's why we considered evaluating the impact of MWEs on the quality of translation by integrating the results of our MWEs alignment approach in the training corpus used to extract the translation model of the phrase based statistical machine translation system Moses. We used the factored translation model (Koehn and Hoang, 2007) as our baseline system. It is an extension of the phrase-based model which enables the use of additional linguistic information at the word level such as morphology and Part-Of-Speech. Note that in Moses translation models are produced by the word alignment tool Giza++.

The factored translation model operates on lemmas instead of surface forms. The translation process is then broken up into the following mapping steps:

1. Translate the lemmas of the source language into lemmas in the target language.
2. Generate surface forms given the lemma and linguistic information (Morphology and Part-Of-Speech).

The goal of these experiments is to study in what respect bilingual MWEs are useful to improve the performance of Moses. In 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 into the target language. These tables are built automatically using Giza++. In order to integrate into Moses the bilingual lexicon which is extracted automatically by our MWEs alignment approach, we add the extracted bilingual lexicon as a parallel corpus and retrain the translation model.

5.2.1 Data and Experimental Setup

In order to study the impact of the bilingual lexicon of MWEs on the performance of Moses, we conducted our experiments on two English-French parallel corpora (Table 3): Europarl (European Parliament Proceedings) and Emea (European Medicines Agency Documents). These corpora were extracted from the open parallel corpus OPUS (Tiedemann, 2012). We achieved three runs and two test experiments for each run: In-Domain and Out-Of-Domain. For this, we randomly extracted 500 parallel sentences rom

Europarl as an In-Domain corpus and 500 pairs of sentences from Emea as Out-Of-Domain corpus. The domain vocabulary is represented in the case of the baseline (Giza++) by the specialized parallel corpus Emea which is added to the training data (Europarl). For our MWEs alignment approach, the domain vocabulary corresponds to the bilingual lexicon of MWEs extracted from the specialized corpus. This bilingual lexicon of MWEs is added to the training corpus (Europarl). It is important to note here that the word alignment tool Giza++ is used to generate the translation tables for both methods (baseline and our approach). In other words, for the baseline (Giza++), the translation table is generated from the parallel corpus which is the concatenation of the general-purpose training data (Europarl) and the domain-specific data (Emea). For our MWEs aligner, the translation table is generated from the parallel corpus which is the concatenation of the general-purpose training data (Europarl) and the bilingual lexicon of MWEs extracted from the domain-specific data (Emea).

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)

Table 3: Corpora details used to train Moses language and translation models (K refers to 1000)

5.2.2 Results and Discussion

The performance of the SMT system Moses is evaluated using the BLEU score (Papineni et al., 2002) on the two test sets for the three runs described in the previous section. Note that we consider only one reference per sentence. The obtained results are reported in tables 4 and 5. As shown in tables 4 and 5, for In-Domain texts, Moses achieves a relatively high BLEU score and the scores of Moses when using the results of our MWEs alignment approach are better than those when we use the baseline (Giza++) in all the runs. Again, the best performance for both In-Domain and Out-Of-Domain texts is achieved using the combination of the scoring using co-occurrence and the scoring based on the bilingual dictionary with the filtering morpho-syntactic patterns.

In addition, we explored the use of LSTM (Long Short-Term Memory) recurrent neural network language models (Hochreiter and Schmidhuber, 1997) for rescoring the 100-best translations proposed by the SMT system Moses. This has been limited to only the third run for both In-Domain texts and Out-Of-Domain texts. When we experimented the LSTM to rerank the 100 hypotheses, the BLEU score (corresponding to the combination of the scoring using co-occurrence and the scoring based on the bilingual dictionary with the filtering morpho-syntactic patterns) increases to 35.82 (+1.49 BLEU points) for In-Domain texts and to 25.53 for Out-Of-Domain texts (+0.9 BLEU points).

Run n°.	In-Domain (Europal)			
	Baseline (Giza++)	Co-occurrence	Co-occurrence + Bilingual dictionary	Co-occurrence + Bilingual dictionary + Filtering patterns
1	32.62	32.69	32.71	32.72
2	33.81	33.88	33.89	33.91
3	34.25	34.30	34.32	34.33

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

Run n°.	Out-Of-Domain (Emea)			
	Baseline (Giza++)	Co-occurrence	Co-occurrence + Bilingual dictionary	Co-occurrence + Bilingual dictionary + Filtering patterns
1	22.96	23.03	23.06	23.07
2	23.30	23.37	23.39	23.41
3	24.55	24.59	24.62	24.63

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

Because the BLEU score reports only global improvements and does not necessarily reveal the impact of the domain vocabulary (represented by the bilingual lexicon of MWEs extracted with our word alignment approach) on the translation quality of Moses, we manually analyzed some examples of translations drawn from the Out-Of-Domain test corpus (Table 6). We noted after analyzing the translation results of the specialized test corpus (Emea) that in some cases errors come from the training parallel corpus. For instance, the English word "*hypertension*" is sometimes translated as the uniterm "*hypertension*" such as in the bilingual sentence "*Cases of hypertensive crisis have been reported with duloxetine, especially in patients with pre-existing hypertension./Des cas de crise hypertensive ont été rapportés avec la duloxétine, en particulier chez des patients présentant une hypertension préexistante.*", and sometimes translated as the multiterm "*hypertension artérielle*" such as in the bilingual sentence "*The initiation of treatment with XERISTAR is contraindicated in patients with uncontrolled hypertension that could expose patients to a potential risk of hypertensive crisis./L'instauration du traitement par XERISTAR est contre-indiquée chez les patients présentant une hypertension artérielle non équilibrée qui pourrait les exposer à un risque potentiel de crise hypertensive.*". In the example of Table 6, the baseline system provides for the word "*hypertension*" the translation "*hypertension*" and our MWE alignment approach provides for this word the translation "*hypertension artérielle"*. Of course, both translations are correct.

Similarly, both the baseline and our MWE alignment approach translate correctly the multiword expression "*increase in blood pressure/augmentation de la pression artérielle*". On the other hand, as we can see, some translations provided when using the baseline and when we use our approach have many spelling and grammatical errors and are very approximate. As examples, we may mention the translations of the expressions "*has been*

associated/a été associé" and "*in some patients/dans certains patients*". These results can be explained by the fact that, on the one hand, statistical machine translation toolkits like Moses have not been designed with grammatical error correction in mind, and on the other hand, these two expressions have not been considered by our alignment approach as being MWEs. Indeed, even if after the scoring function based on co-occurrence, these two expressions have been identified as MWEs, but the filtering step based on morpho-syntax patterns takes that possibility away (no patterns for these expressions). This is one of the major weaknesses of MWEs alignment approaches based on patterns. Applying morpho-syntax patterns to filter the list of MWEs candidates increases the precision of the alignment but at the same time it decreases the recall.

Example Input (Emea): Duloxetine has been associated with an *increase in blood pressure* and clinically significant *hypertension* in some patients.	
Translation reference	La duloxétine a été associée à une *augmentation de la pression artérielle* et à une *hypertension artérielle* cliniquement significative chez certains patients.
Translation when using the Baseline (Giza++)	Duloxetine a été associé à une *augmentation de la pression artérielle* et de différence cliniquement significative *hypertension* dans certains patients.
Translation when using our MWE aligner (Co-occurrence + Bilingual dictionary + Filtering patterns)	Duloxetine a été associé à une *augmentation de la pression artérielle* et de *hypertension artérielle* cliniquement significative dans certains patients.

Table 6: Translations produced by Moses for a sentence from the Emea corpus.

For the multiword expression "*clinically significant hypertension*", the translation proposed by Moses when using the baseline provides an ungrammatical and meaningless translation (*clinically significant hypertension/différence cliniquement significative hypertension*).

6. Conclusion and Future Work

This paper presented, on the one hand, a hybrid approach to extract and align MWEs from a parallel corpus in a one-step process, and on the other hand, an experimental evaluation of the impact of integrating the results of this MWEs alignment approach on the performance of the statistical machine translation system Moses. We have more specifically shown that adding external knowledge (bilingual lexicons and filtering linguistic patterns) to the co-occurrence scoring function improves significantly the precision of the MWEs alignment approach. We have also showed that the results of the SMT system Moses can be improved by rescoring its n-best translations using a LSTM language model. This study offers several open issues for future work. First, we expect to use machine learning approaches to extend the morpho-syntactic patterns to take into account other forms of MWEs. The second perspective is to explore the integration of bilingual MWEs into other machine translation systems such as neural machine translation ones. We also expect to adapt our MWEs alignment approach to new language pairs such as English-Arabic and French-Arabic.

7. Acknowledgements

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

8. Bibliographical References

Baldwin, T., Villavicencio, A. (2002). Extracting the unextractable: A case study on verb–particles. In Proceedings of the 6th conference on natural language learning (CoNLL).

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

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

Bouamor, D., Semmar, N., and Zweigenbaum, P. (2012). Identifying bilingual Multiword expressions for statistical machine translation. In Proceedings of the eighth international conference on Language Resources and Evaluation (LREC).

Boulaknadel, S., Daille, B., and Aboutajdine, D. (2008). A multi-word term extraction program for arabic language. In Proceedings of the sixth international conference on Language Resources and Evaluation (LREC).

Caseli, H., Ramisch, C., Nunes, M., and Villavicencio, A. (2010). Alignment-based extraction of multiword expressions. *Language Resources & Evaluation*, 44.

Dagan, I. and Church, K. (1994). Termight: Identifying and translating technical terminology. In Proceedings of the fourth conference on applied natural language processing.

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

Deng Y. and Byrne W. (2005). HMM Word and Phrase Alignment for Statistical Machine Translation. In Proceedings of Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP).

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

Evert, S. and Krenn, B. (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language* 19(4).

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

Germann, U. (2008). Yawat: Yet Another Word Alignment Tool. In Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies (ACL HLT).

Hochreiter, S. and Schmidhuber J. (1997). Long short-term memory. *Neural Computation*, Volume 9 Issue 8, MIT Press Cambridge, MA, USA.

Jackendoff, R. (1997). The architecture of the language faculty. *The MIT Press*.

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

Koehn, P. and Hoang, H. 2007. Factored translation models. In Proceedings of the 7th Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.

Kupiec, J. (1993). An algorithm for finding noun phrase correspondences in bilingual corpora. In Proceedings of the 31st annual meeting on Association for Computational Linguistics (ACL).

Lefever E., Macken L., Hoste V. (2009). Language-independent bilingual terminology extraction from a multilingual parallel corpus. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (ACL).

MacCartney B., Galley M., and Manning C. D. (2008). A Phrase-Based Alignment Model for Natural Language Inference. In Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Marchand, M. and Semmar, N. (2011). A Hybrid Multi-Word Terms Alignment Approach Using Word Co-occurrence with a Bilingual Lexicon. In Proceedings of the fifth language and technology conference: human language technologies as a challenge for computer science and linguistics.

Mihalcea, R. and Pedersen, T. (2003). An evaluation exercise for word alignment. In: Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond - Volume 3.

Moon, R. (1998). Fixed Expressions and Idioms in English. *Oxford: Clarendon Press*.

Och, F. J. and Hermann, N. (2000). Improved statistical alignment models. In Proceedings of the 38th annual meeting on Association for Computational Linguistics (ACL).

Okita, T., Guerra, A. M., Graham Y., and Way, A. (2010). Multi-word expression-sensitive word alignment. In Proceedings of the 4th International Workshop on Cross Lingual Information Access at the 23rd International Conference on Computational Linguistics (COLING).

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL).

Pearce, D. (2002). A comparative evaluation of collocation extraction techniques. In Proceedings of the third international conference on language resources and evaluation.

Riehemann, S. Z. 2001. A Constructional Approach to Idioms and Word Formation. *Ph.D. Thesis, Stanford University*.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In Proceedings of third CICLing: International Conference on Computational Linguistics and Intelligent Text Processing (CICLing).

Samuelsson Y. and Volk M. (2007). Automatic Phrase Alignment: Using Statistical N-Gram Alignment for Syntactic Phrase Alignment. In Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories.

Semmar N., Servan C., de Chalendar G., and Le NY B. (2010). A Hybrid Word Alignment Approach to Improve Translation Lexicons with Compound Words and Idiomatic Expressions. In Proceedings of the 32nd Translating and the Computer conference.

Semmar, N. and Marchand, M. (2017). Une approche hybride pour la construction de lexiques bilingues d'expressions multi-mots à partir de corpus parallèles. In Proceedings of the 24ème conférence sur le Traitement Automatique des Langues Naturelles (TALN).

Semmar, N. and Laib, M. (2017). Building Multiword Expressions Bilingual Lexicons for Domain Adaptation of an Example-Based Machine Translation System. In Proceedings of the Recent Advances in Natural Language Processing (RANLP).

Seretan, V. and Wehrli, E. (2007). Collocation translation based on sentence alignment and parsing. In Proceedings of the 14ème conférence sur le Traitement Automatique des Langues Naturelles (TALN).

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the eighth international conference on Language Resources and Evaluation (LREC).

Tufis, D. and Ion, R. (2007). Parallel corpora, alignment technologies and further prospects in multilingual resources and technology infrastructure. In Proceedings of the 4th International Conference on Speech and Dialogue Systems.

Van de Cruys, T. and Villada Moiron, B. (2007). Semantics-based multiword expression extraction. In Proceedings of the workshop on a broader perspective on multiword expressions.

Villavicencio, A., Kordoni, V., Zhang, Y., Idiart, M., and Ramisch, C. (2007). Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In Proceedings of the joint conference on empirical methods in natural language processing and computational natural language learning.

Vintar, S. and Fiser D. (2008). Harvesting multi-word expressions from parallel corpora. In Proceedings of the sixth international conference on Language Resources and Evaluation (LREC).

Wu, C. and Chang, J. S. (2004). Bilingual collocation extraction based on syntactic and statistical analyses. Computational Linguistics and Chinese Language Processing, Vol. 9, No. 1, Association for Computational Linguistics and Chinese Language Processing.

Zhang, Y., Kordoni, V., Villavicencio, A., and Idiart, M. (2006). Automated multiword expression prediction for grammar engineering. In Proceedings of the workshop on multiword expressions: Identifying and exploiting underlying properties.