Chapter 9

Extracting and aligning multiword expressions from parallel corpora

Nasredine Semmar
CEA LIST, Vision and Content Engineering Laboratory

Christophe Servan
University of Grenoble Alpes – Grenoble Informatics Laboratory
SYSTRAN

Meriama Laib
CEA LIST, Vision and Content Engineering Laboratory

Dhouha Bouamor
Actimos, Groupe Accord

Morgane Marchand
eXenSa

Bilingual lexicons of multiword expressions play a vital role in several natural language processing applications such as machine translation and cross-language information retrieval because they often characterize domain-specific vocabularies. Word alignment approaches are generally used to construct bilingual lexicons automatically from parallel corpora. We present in this chapter three approaches to align multiword expressions from parallel corpora. We evaluate the bilingual lexicons produced by these approaches 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 lexicons of multiword expressions in the translation model improves the performance of Moses.

Nasredine Semmar, Christophe Servan, Meriama Laib, Dhouha Bouamor & Morgane Marchand. 2019. Extracting and aligning multiword expressions from parallel corpora. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 239–268. Berlin: Language Science Press. DOI:10.5281/zenodo.3264764

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 parts (Sag et al. 2002). Such units could be collocations, compound words, named entities, etc. They constitute an important part of the lexicon of any natural language (Jackendoff 1997). Bilingual lexicons of MWEs play a vital role in several Natural Language Processing (NLP) applications such as Machine Translation (MT) and Cross-Language Information Retrieval (CLIR) because they generally characterize domain-specific vocabularies. The manual construction of these lexicons is often costly and time consuming. Word alignment approaches are generally used to automatically construct bilingual lexicons from parallel or comparable corpora. Several word alignment approaches have been explored (Daille et al. 1994; Blank 2000; Barbu 2004) and many automatic word alignment tools are available, such as GIZA++ (Och & Ney 2000). However, most of these tools are efficient only to align single words (Fraser & Marcu 2007).

The chapter is organized as follows. We survey in Section 2 previous works addressing the tasks of extracting and aligning MWEs from parallel corpora. We define in Section 3 the notion of MultiWord Expression and describe different types of MWEs with examples. In Section 4, we introduce three approaches 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 Related work

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 necessarily MWEs, they can be contiguous sequences of a few words that encapsulate enough context to be translatable (DeNero & Klein 2008). The second strategy firstly, identifies monolingual MWE candidates and then applies alignment approaches to find bilingual correspondences (Daille et al. 1994; Blank 2000; Gaussier & Yvon 2011; Barbu 2004).

In the second strategy, MWEs extraction can be processed by using symbolic methods based on morpho-syntactic patterns, or, through statistical approaches, which use automatic measures to rank MWE candidates. Finally, MWEs extraction can be done by using hybrid approaches, which combine the two first strategies.

Dagan & Church (1994) proposed to use syntactic analysis to extract terminology. MWEs are 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. Tufiş & Ion (2007) and Seretan & Wehrli (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 & Fišer (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. But they did not combine these two kind of approaches. The combination of such approaches enables to extract finer MWEs (Daille 2001). In this way Wu & Chang (2003) 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 & Wehrli 2007). Again, the first step is done on each language independently and then, a second step aims to match candidates across languages.

3 Multiword expressions

3.1 Definition

In 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 (Choueka 1988; Sag et al. 2002). MWEs are frequently used in written texts and constitute a significant part of the language lexicon.

Jackendoff (1997) considers that the frequency of their use is equivalent to that of single words. Although MWEs are easily computed, stored and used by humans, their identification is a major issue for different type of NLP applications, namely for syntactic analysis (Nivre & Nilsson 2004; Constant et al. 2011), automatic summarization (Hogan et al. 2007), information extraction (Vechtomova 2005) and especially for machine translation and cross-language information retrieval (Carpuat & Diab 2010; Ren et al. 2009).

3.2 Multiword expressions typology

In the literature, MWEs are presented under different names or classifications such as idioms, lexicalized phrases or collocations and several authors (Ramisch et al. 2013) give a list of examples instead of giving an exact description of them. According to Calzolari et al. (2002), MWEs are "different but related phenomena" and "At the level of greatest generality, all of these phenomena can be described as a sequence of words that acts as a single unit at some level of linguistic analysis".

Sag et al. (2002) classify them into two main categories: lexicalized phrases and institutionalized phrases (Figure 1). 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".

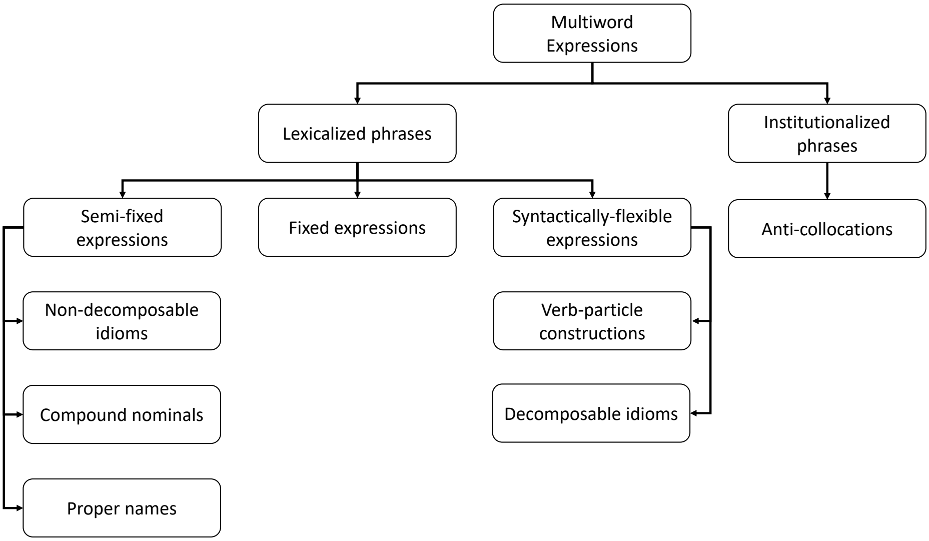

Figure 1: Typology of multiword expressions by Sag et al. (2002)

3.2.1 Lexicalized phrases

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

3.2.1.1 Fixed 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 (cf. *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.

3.2.1.2 Semi-fixed expressions

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 himself, 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, places 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).

3.2.1.3 Syntactically-flexible expressions

Unlike semi-fixed expressions, syntactically-flexible expressions undergo a wide degree of syntactic variation such as passivation (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 (Riehemann 2001). Semantically, they behave as if their components were linked parts contributing independently to the figurative interpretation of the expression as a whole.

3.2.2 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" (Barz 1996; Riehemann 2001; Burger 2010), described as sequences of words that statistically have a high probability to appear together whether they are contiguous or not (e.g., *make love* or *make a difference*).

4 Construction of bilingual lexicons of MWEs from parallel corpora

In this section, we describe three approaches to build bilingual lexicons of MWEs from a sentence aligned parallel corpus. The first two approaches are composed of two steps. The first step identifies MWEs present in the parallel corpus, and the second step establishes correspondence relations between the MWEs of the source text and their translations in the target text. The third approach performs the terminology extraction and alignment tasks in one step.

4.1 Statistical approach for MWEs alignment

The statistical approach for MWEs alignment consists first in identifying the relevant word groups through the use of *n*-gram statistics in both the source and target languages. Then for each source MWE extracted we compile a list of candidate translations through the use of two distance metrics. The list of candidates is then pruned through the use of heuristics like the length of each MWE, and a translation is "found" if it satisfies confidence threshold on the distance metric and the heuristics.

The alignment process has the following four steps (Semmar et al. 2010):

1. Monolingual extraction of MWEs: The role of this step consists to identify all the *n*-grams (up to 6-grams) that may represent a MWE. This is done through frequency analysis and heuristic scoring. This step outputs two lists of terms, which we will refer to as SC (MWE in the Source Language) and TC (MWE in the Target Language).

2. Frequency distance calculation: This step calculates for all source MWEs in SC the distance to each of the target MWEs in TC. The main idea of this metric is that if two MWEs are translations of each other then they must appear together in the corpus segments, and only together. Their frequency distance is then calculated as follows:

   (1) $$\mathrm{FD}(s,t) = \frac{|f(s) - f(t)|}{\max(f(s), f(t))}$$

   Where $f(s)$ is the frequency of the source MWE and $f(t)$ is the frequency of the target MWE under consideration.

   We observe that if t is the translation of s, $f(s) = f(t)$ then we have distance equal to 0. Also, if two MWEs always occur together but one is much more frequent than the other, the distance could have a value other than 0 and they would not be considered translations of each other. Here we chose to apply a threshold of 0.25 as the maximum allowable distance. This threshold is calculated empirically and can be tuned to achieve better precision.

3. Co-occurrence distance (CD): The previous step only considers frequencies so it may be possible for two completely unrelated MWEs to achieve a low distance score. To refine extraction results, we also check for a co-occurrence score as follows:

   (2) $$\mathrm{CD}(X,Y) = \frac{\sqrt{\sum(X_i - Y_i)^2}}{N}$$

   Where, X_i is the number of occurrences of s in the i^{th} segment of the SL, Y_i is the number of occurrences of t in the i^{th} segment of the TL and N is the number of segments.

   This check allows the rejection of the MWEs that fortuitously have similar frequency. Since they would not appear in the same segments, the terms $X_i - Y_i$ would increase. The candidate list can be ordered through CD.

4. Pruning MWEs candidates: After obtaining an ordered list of target MWEs candidates, we remove:
   - The candidates which have a length different from the source MWE;
   - The candidates which have been previously aligned with another source MWE and where the co-occurrence score was better.

Because of the statistical nature of this approach, it performs much better for MWEs that occur often in the corpus. Table 1 illustrates some MWEs and their translations extracted from the bi-sentence *Approval of the Minutes of the previous sitting/Approbation du procès-verbal de la séance précédente*. It should be noted that before applying the MWEs alignment approach, we lemmatize the parallel corpus. This lemmatization is achieved using the CEA LIST Multilingual Analyzer LIMA (Besançon et al. 2010).

Table 1: Some examples of aligned MWEs with the statistical approach

English MWE	French MWE
minute	*procès-verbal*
approval of the minute	*approbation du procès-verbal*
previous sitting	*séance précédent*

4.2 Hybrid approach for MWEs alignment based on morpho-syntactic patterns

The hybrid approach for MWEs alignment is composed of the following two steps (Bouamor et al. 2012a,c,b):

1. MWEs identification: The method used to extract MWEs is based on a symbolic approach relying on morpho-syntactic patterns.
2. MWEs alignment: After extracting MWE candidates, context vectors from the parallel corpus are separately built and similarity scores between one MWE and all target MWEs are computed.

4.2.1 MWEs extraction

The method to extract monolingual MWEs from a parallel corpus is based on a symbolic approach relying on morpho-syntactic patterns. It handles both frequent and infrequent expressions and do not use any lexicon. This method involves a full morpho-syntactic analysis of source and target texts. The analysis is done using the CEA LIST Multilingual Analysis platform LIMA (Besançon et al. 2010), which produces Part-of-Speech (POS) tags and lemmas associated to each word. Since most MWEs consist of noun, adjectives and prepositions, we adopted a linguistic filter. It consists in keeping only *n*-gram (*n* from 2 to 4) units, which match with a list of a hand created morpho-syntactic patterns. Such process is used to keep only specific strings and filter out undesirable ones such as candidates composed mainly of stop words (*of a*, *is a*, *that was*). The algorithm operates on lemmas instead of surface forms which can draw on richer statistics and overcome the data sparseness problems.

In Table 2, we give an example of MWEs produced for each pattern. There exists extraction patterns (or configurations) for which no MWE has been generated (i.e., Noun-Adj). To this list are added some prepositional idiomatic expressions (*in particular*, *in the light of*, *as regards*, etc.) and named entities (*Middle East*, *South Africa*, *United States of America*, etc.) recognized by the morpho-syntactic analyzer LIMA. Then, we scored all extracted MWEs with their total frequency of occurrence in the corpus. To avoid an over-generation of MWEs and remove irrelevant candidates from the process, a redundancy cleaning approach is introduced. In this approach, if a MWE is nested in another, and they both have the same frequency, we discard the smaller one. Otherwise we keep both of them. We consider also the case in which a MWE appears in a high number of terms and discard all longer ones.

Our approach does not use any additional correlations statistics such as Mutual Information or Log Likelihood Ratio. It finds translations for all extracted MWEs (both frequent and infrequent ones).

Table 2: Example of morpho-syntactic patterns used to detect MWEs in each language independently

pattern	English MWE	French MWE
Adj-Noun	*plenary meeting*	*libre circulation*
Noun-Noun	*member state*	*état membre*
Noun-Prep-Noun	*point of view*	*point de vue*
Noun-Prep-Adj-Noun	*court of first instance*	*court de première instance*

4.2.2 MWEs alignment

MWEs alignment aims to find for each MWE in a source language its adequate translation in the target one. This task used to be handled through an external linguistic resource such as bilingual lexicons or single words alignment tools. Our approach for MWEs alignment is resource-independent and uses a parallel corpus and a list of input MWEs candidates to translate. It associates a specific representation to each expression (source and target).

We associate to each MWE an N sized vector, where N is the number of sentences in the corpus, indicating whether it appears or not in each sentence of the corpus. Our algorithm is based on the Vector Space Model (Salton et al. 1975). This vector space representation will serve, eventually, as a basis to establish a translation relation between each pair of MWEs.

To extract translation pairs of MWEs, we propose an iterative alignment algorithm operating as follows:

1. Find the most frequent MWE *exp* in each source sentence;
2. Extract all target translation candidates, appearing in all parallel sentences to those containing *exp*;
3. Compute a confidence value V_{Conf} for each translation relation between *exp* and all target translation candidates;
4. Consider that the target MWE maximizing V_{Conf} is the best translation;
5. Discard the translation pair from the process and go back to 1.

To compute the confidence value V_{Conf}, we adopted the *Jaccard* Index. This measure is based on the number I_{st} of sentences shared by each target and a source MWE. I_{st} is normalized by the sum of the number of sentences where the source and target MWEs appear independently of each other (respectively V_s and V_t) decreased by I_{st}.

(3) $$Jaccard = \frac{I_{st}}{V_s + V_t - I_{st}}$$

We illustrate in Table 3, a sample of aligned MWEs by means of the algorithm described above. When we observe MWE pairs, we noticed that our method has two advantages. On the one hand, it allows the translation of MWEs aligned in most previous work (Dagan & Church 1994; Ren et al. 2009) using single words alignment tools to establish word-to-word alignment relations. The approach can capture the semantic equivalence between expressions such as *insulaire en développement* and *small island developing* in a different way. On the other hand, the approach enables the alignment of idioms such as *à nouveau* ('once more').

Table 3: Some examples of aligned MWEs with the hybrid approach based on morpho-syntactic patterns

English MWE	French MWE
european parliament	*parlement européen*
military coup	*coup d'état*
in favour of	*en faveur de*
no smoking area	*zone non fumeur*
small island developing	*insulaire en développement*
good faith	*de bonne foi*
competition policy	*politique de concurrence*
process of consultation	*processus de consultation*
railway sector	*chemin de fer*
with regard to	*en ce qui concerne*
once more	*à nouveau*
cut in forestation	*coupe forestière*

4.3 Hybrid approach for MWEs alignment based on linear programming

This section describes a hybrid approach combining linguistic and statistical information which performs terminology extraction and alignment of MWEs from parallel texts in one step (Marchand & Semmar 2011).

Most of works on MWEs alignment are divided in two tasks: a monolingual step in which candidate terms are extracted and a bilingual step in which these terms are aligned with their translations (Gaussier & Yvon 2011). Word alignment techniques are generally used to achieve the bilingual step. These approaches in multiple steps have the disadvantage to potentially propagate errors.

The main idea of the hybrid approach for MWEs alignment based on linear programming is to consider the global task of selection and alignment as an optimization problem. The challenge when we deal with alignment of MWEs is the exponential complexity of such a task. The possible number of fragments in a sentence improves exponentially according to the number of the words of the sentence. Several works impose some constraints on the number of fragments of a MWE. In our approach, the only restriction we made on MWEs is contiguity. The advantage to assume the continuity is to enable a linearized formulation of the optimization problem to solve. We use an integer linear programming approach inspired by the work described in DeNero & Klein (2008) to quickly find an approximated optimal solution.

4.3.1 Linear programming 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. f_{kl} is the same 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 $\phi : \{e_{ij}\} \times \{f_{kl}\} \longrightarrow R$ to score links. The score of an alignment is then the product of all the links inside it:

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

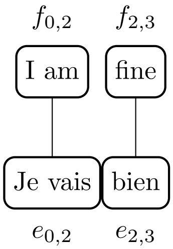

Figure 2: Example of alignment

In the example shown in Figure 2, the score of the alignment is computed as follows:

(5) $$\phi(a) = \phi(e_{0,2}, f_{0,2}) \times \phi(e_{2,3}, f_{2,3})$$

Formally this function has no constraints other than that of being real. In practice, we choose a function that gives an idea about the relevance to align such fragments. The higher the score, the higher the relevance of alignment is important. Therefore, we look for the alignment (segmentation + links) that maximizes the score described above.

First, we introduce binary variables $A_{i,j,k,l}$ denoting whether $(e_{ij}, f_{kl}) \in a$. Furthermore, we introduce binary indicators $E_{i,j}$ and $F_{k,l}$ that denote whether some $(e_{ij}, \cdot)$ or $(\cdot, f_{kl})$ appears in a, respectively. Finally, we will use $W_{i,j,k,l} = \log \phi(e_{ij}, f_{kl})$ to transform the product into a sum. When optimized[1], the integer program yields the optimal alignment:

[1] We used the open source solver GLPK (GNU Linear Programming Kit), available at http://www.gnu.org/s/glpk/.

(6)
$$\begin{cases} \max \sum\limits_{i,j,k,l} W_{i,j,k,l} A_{i,j,k,l} & & \\ \forall x : 1 \leq x \leq |e| & \sum\limits_{i,j : i<x\leq j} E_{i,j} = 1 & (1) \\ \forall y : 1 \leq y \leq |f| & \sum\limits_{k,l : k<y\leq l} F_{k,l} = 1 & (2) \\ \forall i, j & \sum\limits_{k,l} A_{i,j,k,l} \geq E_{i,j} & (3) \\ \forall k, l & \sum\limits_{i,j} A_{i,j,k,l} \geq F_{k,l} & (4) \\ \forall i, j, k, l & 2 \cdot A_{i,j,k,l} \leq E_{i,j} + F_{k,l} & (5) \end{cases}$$

With the following constraints:

(7)
$$\begin{cases} 0 \leq i < |e|, & 0 < j \leq |e|, & i < j \\ 0 \leq k < |f|, & 0 < l \leq |f|, & k < l \end{cases}$$

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} (i.e. $A_{i,j,k,l} = 1$) then e_{ij} and f_{kl} are in the selected partitions of e and f.

In that way, our approach differs from the one proposed in DeNero & Klein (2008). Their work focuses on bijective alignments while we consider surjective alignments. We have also modified constraints (3) and (4) and added constraint (5) to allow a phrase to be aligned with several other phrases. We have chosen this formalism because phrases are not necessarily composed of contiguous words.

This integer program can work with any real-valued scoring function.

4.3.2 Co-occurrence based metric

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

(8)
$$\phi_c(e_{ij}, f_{kl}) = \frac{\sum\limits_{s' \in S} N_{s'}(e_{ij}) \times N_{s'}(f_{kl})}{\sum\limits_{s \in S} N_s(e_{ij}) + N_s(f_{kl}) - N_s(e_{ij}) \times N_s(f_{kl})}$$

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 phrases divided by the number of total presence of either phrase. 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. The example in Table 4 illustrates this score.

Table 4: Example of ambiguous translation of MWEs

Je mange un *avocat*	–	I'm eating an *avocado*
L'*avocat* prend la parole	–	The *lawyer* takes the floor

In this small corpus, $N_1(avocat) = 1$, $N_1(avocado) = 1$, $N_2(avocat) = 1$ and $N_2(avocado) = 0$. Thus, the co-occurence score for the bi-gram *avocat/avocado* has the value:

$$\phi_c(avocado, avocat) = \frac{(1 \times 1) + (1 \times 0)}{(1 + 1 - 1 \times 1) + (1 + 0 - 1 \times 0)} = \frac{1}{2} \tag{9}$$

We observed after aligning some sentences that when both sentence structures are similar, the aligner performs well as shown in Figure 3. 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 us to begin to detect expressions in two parts. We can see that *rôle* is linked to both *role* and *play* (Figure 3, Alignment 3).

This would have been impossible with the bijective formulation of DeNero & Klein (2008). This result is encouraging but not yet sufficient. Actually this expression is partially recognized because it includes two plain words. Expressions with postponed prepositions would not be recovered this way because the prepositions are too common to be statistically relevant. If the structure is different we have more difficulties (as shown in Figure 4). Some sentences are also difficult to align because they are not perfect translation: *They/la population* or adverbs like *also* or *very* which are not translated.

We also observe that, for common words, the distribution of apparition is meaningless: *to* is linked with *de* and *a*. We should use a measure of information as suggested in Gao (1998). In addition, the program is powerless if it finds an unknown word or if a word co-occurs with no other word of the translated sentence. In that case, all links containing this word will obtain the score of 0 as

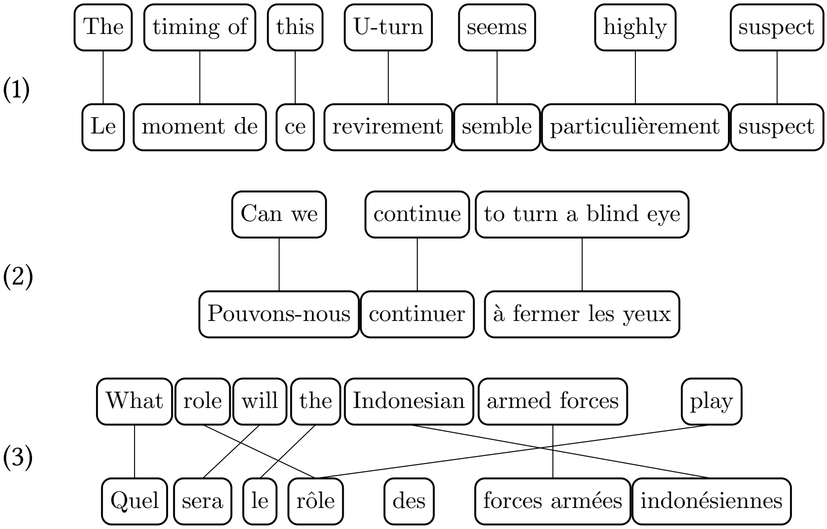

Figure 3: Good alignments with co-occurrence based metric

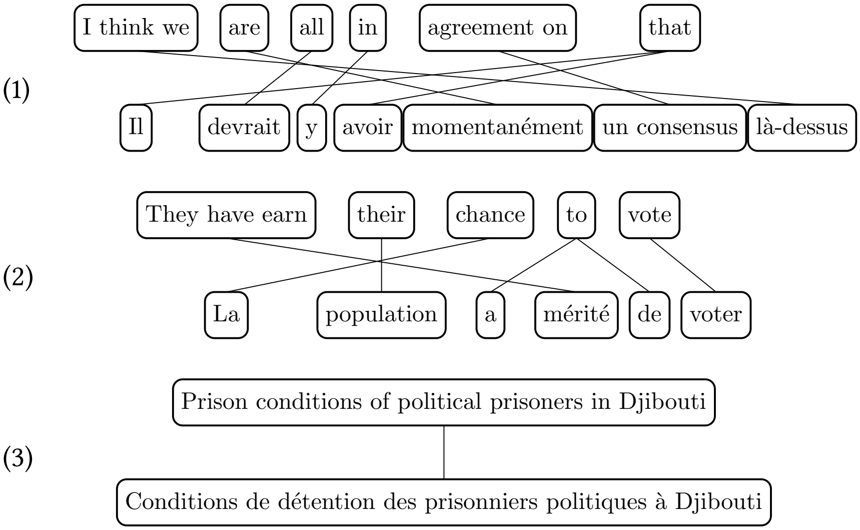

Figure 4: Bad alignments with co-occurrence based metric

they never occur. And as we use a multiplicative metric, the global score of the alignment will be 0 whatever the other links of the alignment. Unknown links should have a small, non-null score to allow the discovery of new links. Moreover, we can use an external resource such as a bilingual lexicon of single words which can improve the alignment of phrases.

4.3.3 Bilingual dictionary based metric

The bilingual dictionary gives us 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 cannot take for sure that they should not. But we have to take that into account.

The bilingual dictionary score is calculated as follows:

$$\phi_c(e_{ij}, f_{kl}) = \frac{a \times R_1 + b \times R_0}{a \times R_1 + b \times R_0 + c \times N_1 + d \times N_0} \tag{10}$$

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 bi-phrase 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 metric has a double effect. First, it gives a high score if bi-phrases respect dictionary word to word alignment. And 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 following examples. The dictionary-based metric is not intended to be used separately. It is mixed with co-occurrence score. We used an English-French bilingual dictionary containing 243,539 entries with doubles.[2]

In Figure 5, we observe some degradation of alignments. For these sentences, the threshold for unknown couples is too high relatively to the statistical score.

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

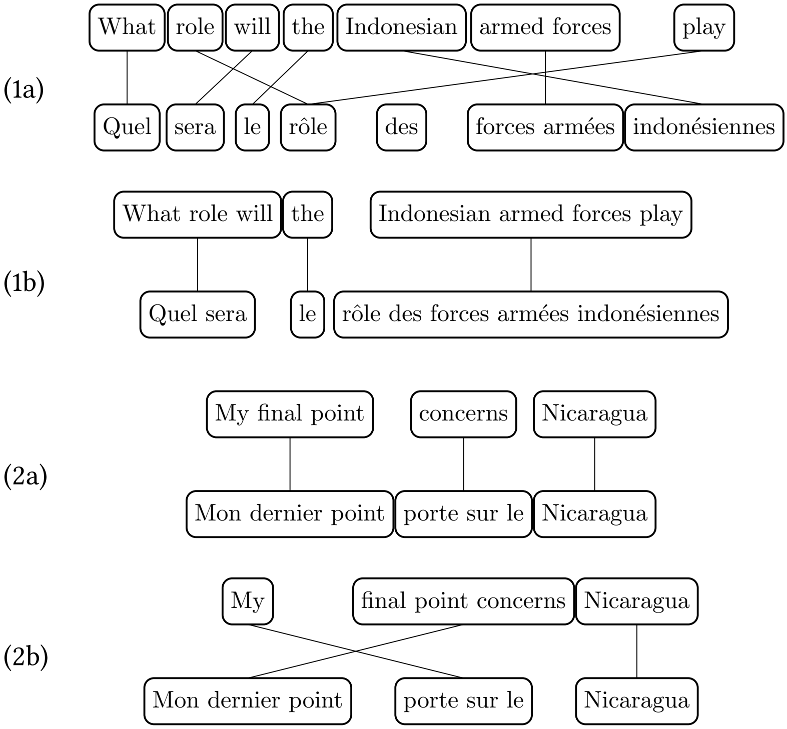

Figure 5: Degradation of alignments – (a) Alignments without the bilingual dictionary and (b) Alignments with the bilingual dictionary

So we lose the benefit of the co-occurrence metric. This problem should be partially solved by scaling the two metrics. However we have already observed some improvements, as presented in Figure 6. In the first example, the bilingual dictionary gives the alignments: *be/être*, *decided/décidé* and *there/y*. So the program manages to reconstruct the whole expression *is to be decided on there/doit y être décidé*. Moreover the links *concrete/concret* and *programme/programme* are strengthened. The second example is difficult to align due to the difference of structure. The alignment with dictionary is not perfect but is far more better. In this case the dictionary only gives links *verdict/jugement* and *request/requête* which were already aligned. However they are strengthened and others links are weakened. That is why we can observe an improvement.

Finally in the last example, the dictionary gives no links because the words are not lemmatized. The good result is here exclusively due to the threshold effect. The programme is allowed to consider links with no co-occurrence as long as others links have a good co-occurrence score.

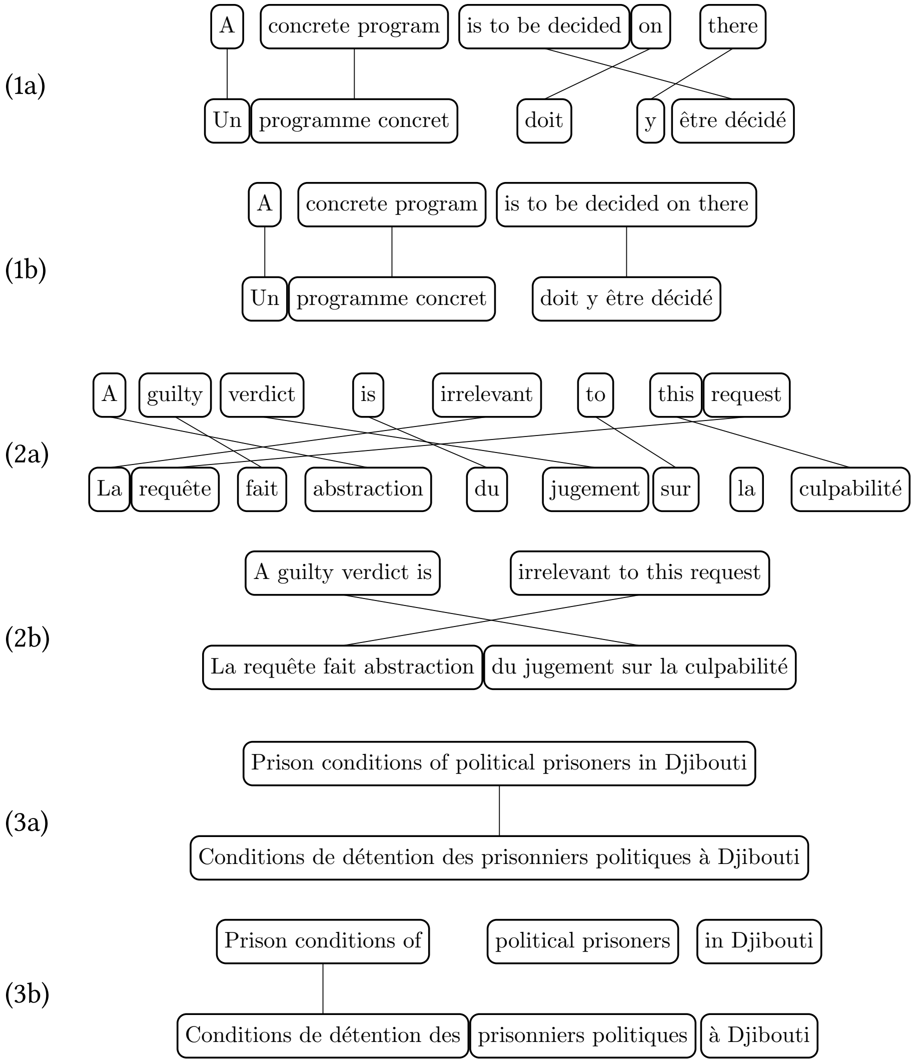

Figure 6: Amelioration of alignments – (a) Alignments without the bilingual dictionary and (b) Alignments with the bilingual dictionary

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 the three MWEs aligners with a reference alignment; and secondly automatically by using the results of the three MWEs aligners to build the translation model of the state-of-the-art statistical machine translation system Moses (Koehn et al. 2007).

5.1 Manual evaluation

The three approaches for MWEs alignment and the baseline Giza++ (Och & Ney 2000) have been evaluated using the following evaluation metrics. Given an alignment A, and a gold standard alignment (reference alignment) G, each such alignment set eventually consisting of two sets (A_s, A_p), and (G_s, G_p) where “s” and “p” correspond respectively to “sure” and “probable” alignments. The following measures are defined (where T is the alignment type, and can be set to either S or P). Each word aligner was evaluated in terms of Precision (P_T), Recall (R_T) and F-Measure (F_T).

(11) $$P_T = \frac{A_T \cap G_T}{A_T};\; R_T = \frac{A_T \cap G_T}{G_T};\; F_T = \frac{2 \times P_T \times R_T}{P_T + R_T}$$

The 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).

Table 5 summarizes the results of the three approaches for English–French MWEs alignments and the baseline (Giza++) in terms of precision, recall and F-measure.

The first observation is that, the hybrid approach based on morpho-syntactic patterns performs better than all the other methods. It clearly appears that the morpho-syntactic patterns used to extract the MWEs present in source and target texts has had a significant impact on the precision of the alignment. On the other hand, the statistical approach has the lower recall but it is better than the recall of the baseline (Giza++). And as a second observation, adding information coming from a bilingual lexicon to the co-occurrence metric used in the hybrid approach based on linear programming, certainly has improved the precision but the recall has dropped.

Table 5: Performance of the different English–French MWE aligners

MWE aligner	precision	recall	f-measure
Baseline (Giza++)	0.83	0.37	0.51
Statistical	0.81	0.39	0.52
Hybrid using morpho-syntactic patterns	0.87	0.55	0.67
Hybrid using co-occurrence	0.61	0.63	0.61
Hybrid using co-occurrence + lexicon	0.85	0.54	0.66

5.2 Alignment 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 evaluation and to compare our approaches with the state-of-the-art work. That's why we decided to study the impact of MWEs on the quality of translation by integrating the results of our word aligners in the training corpus used to extract the translation model of the phrase based statistical machine translation system Moses. We use the factored translation model (Koehn & Hoang 2007) as our baseline system. 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 (Figure 7).

Our model operates on lemmas instead of surface forms, in which the translation process is broken up into a sequence of mapping steps that either:

- Translate source lemmas into target's ones.
- Generate surface forms given the lemma.

The features used in the baseline system include: (1) four translation probability features, (2) two language models, (3) one generation model and (4) word penalty.

The goal of these experiments is to study in what respect 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 the open source word alignment tool Giza++ (Och & Ney 2000). However, Giza++ could produce errors in particular when it aligns multiword expressions (Fraser & Marcu 2007). In order to integrate into Moses the bilingual lexicon which is extracted automatically by the MWE alignment approaches, we propose the following three methods:

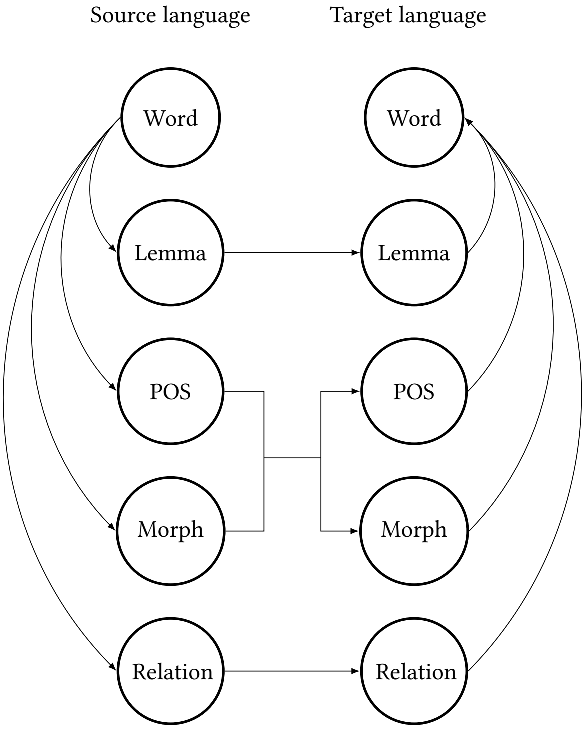

Figure 7: Factored model used in the SMT baseline system

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

TABLE: This method consists in adding the extracted bilingual lexicon into Moses's phrase table. We use a smoothed probability estimator to construct a translation probability for each MWE of the bilingual lexicon. This estimator is based on the similarity measure provided by each word alignment approach.

FEATURE: In this method, we extend the “TABLE” method by adding a new feature indicating whether a MWE comes from the bilingual lexicon or not (1 or 0 is introduced for each entry of the phrase table).

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 6): Europarl (European Parliament Proceedings) and Emea (European Medicines Agency Documents). These corpora were extracted from the open parallel corpus OPUS (Tiedemann 2012). For each MWE alignment approach, 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 from 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 our baseline (Moses) respectively by the specialized parallel corpus Emea which is added to the training data (Europarl). Afterwards, we extracted bilingual MWEs from the training corpus and applied the three methods described above. For the three integration methods (CORPUS, TABLE, FEATURE), the domain vocabulary is identified by a bilingual lexicon which is extracted automatically from the specialized parallel corpus Emea using the different MWEs alignment approaches.

Table 6: Europarl and Emea corpora details used to train language and translation models of Moses (K refers to 10^3)

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)

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 one reference per sentence. The obtained results are reported in Tables 7, 8, 9 and 10.

As shown in Tables 6, 7, 8 and 9, for In-Domain texts, Moses achieve a relatively high BLEU score and the scores of Moses when using the results of the hybrid approach based on morpho-syntactic patterns are better in all the runs. The best improvement is achieved using the “FEATURE” method. The “CORPUS” method (when compared to the baseline system) comes next with a slightly higher BLEU score with an improvement for In-Domain sentences and Out-Of-Domain texts.

Table 7: BLEU scores of Moses when using the results of the statistical approach

Run n°.	In-Domain (Europarl)				Out-Of-Domain (Emea)			
	Baseline	CORPUS	TABLE	FEATURE	Baseline	CORPUS	TABLE	FEATURE
1	32.62	32.41	32.36	32.55	22.96	22.82	22.75	22.91
2	33.81	33.76	33.71	33.79	23.30	23.09	23.04	23.27
3	34.25	34.23	34.21	34.24	24.55	24.49	24.45	24.52

Table 8: BLEU scores of Moses when using the results of the hybrid approach based on morpho-syntactic patterns

Run n°.	In-Domain (Europarl)				Out-Of-Domain (Emea)			
	Baseline	CORPUS	TABLE	FEATURE	Baseline	CORPUS	TABLE	FEATURE
1	32.62	32.82	32.15	32.88	22.96	23.45	23.11	23.69
2	33.81	34.05	33.48	34.09	23.30	24.09	23.76	24.18
3	34.25	34.64	34.11	34.67	24.55	25.43	25.05	25.48

Table 9: BLEU scores of Moses when using the results of the hybrid approach based on linear programming

Run n°.	In-Domain (Europarl)				Out-Of-Domain (Emea)			
	Baseline	CORPUS	TABLE	FEATURE	Baseline	CORPUS	TABLE	FEATURE
1	32.62	32.69	32.64	32.72	22.96	23.03	22.97	23.06
2	33.81	33.88	33.85	33.91	23.30	23.37	23.34	23.40
3	34.25	34.30	34.27	34.33	24.55	24.59	24.56	24.62

Table 10: BLEU scores of Moses when using the results of the hybrid approach based on linear programming and using a bilingual dictionary

Run n°.	In-Domain (Europarl)				Out-Of-Domain (Emea)			
	Baseline	CORPUS	TABLE	FEATURE	Baseline	CORPUS	TABLE	FEATURE
1	32.62	32.71	32.68	32.73	22.96	23.06	22.97	23.07
2	33.81	33.89	33.87	33.92	23.30	23.39	23.32	23.39
3	34.25	34.32	34.29	34.32	24.55	24.62	24.56	24.63

In order to show the impact of the domain vocabulary (represented by the bilingual MWEs extracted with the aligner based on the hybrid approach with morpho-syntactic patterns), on the translation quality of Moses, we manually analyzed an example of translations drawn from the Out-Of-Domain test corpus (Table 11).

Table 11: Translations produced by Moses for an Out-Of-Domain sentence

Input sentence	in the 12 week acute phase of three clinical trials of duloxetine in patients with diabetic neuropathic pain, small but statistically significant increases in fasting blood glucose were observed in duloxetine-treated patients.
Reference translation	lors de la phase aiguë de 12 semaines de trois essais cliniques étudiant la duloxétine chez les patients souffrant de douleur neuropathique diabétique, des augmentations faibles, mais statistiquement significatives de la glycémie à jeun ont été observées chez les patients sous duloxétine.
Moses translation with the "CORPUS" integration method	dans le 12 semaines de la phase aiguë trois études cliniques de duloxetine chez les patients avec douleur neuropathique diabétique, petites mais statistiquement significatif augmentations de répréhensible glycémie artérielle a été observée chez les patients traités duloxetine.
Moses translation with the "TABLE" integration method	dans le 12 semaine de la phase aiguë de trois essais cliniques de duloxetine dans les patients avec douleur neuropathique diabétique, petites mais statistiquement augmentations considérables dans le sang répréhensible glucose ont été constatées dans les patients duloxetine traités.
Moses translation with the "FEATURE" integration method	dans le 12 semaines de la phase aiguë de trois essais cliniques chez les patients avec douleur neuropathique diabétique, petites mais des augmentations statistiquement significatives de la glycémie à jeun ont été observées chez les patients traités duloxétine.

After analyzing the translations of this example, it is clear that in some cases, it is just impossible to perform a word-to-word alignment between two MWEs that are translation of each other. For example, the "FEATURE" method proposes the compound word *glycémie à jeun* as a translation for the expression *fasting blood glucose* which is correct, but, "CORPUS" and "TABLE" methods propose respectively the translations *répréhensible glycémie artérielle* and *sang répréhensible glucose* which are completely wrong. However, all the integration methods translate correctly the multiword expressions *diabetic neuropathic pain/douleur neuropathique diabétique* and *acute phase/phase aiguë*. The multiword expression *clinical trials/essais cliniques* is translated correctly by "TABLE" and "FEATURE" methods. Likewise, the translation provided by the "CORPUS" method for this expression is also correct *clinical trials/études cliniques* but it is different from the translation of the reference. It seems that the probabilities of the alignments proposed by Giza++ for these multiword expressions were very high and helped Moses decoder to choose these alignments. On the other hand, as we can see, all the translations have many spelling and grammatical errors, and in particular, the translations of some multiword expressions ('statistically significant increases'/*statistiquement significatif augmentations*, 'statistically significant increases'/*statistiquement augmentations considérables*) produced by the "CORPUS" and "TABLE" methods are very approximate. This result 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, Giza++ could produce errors in particular when it aligns multiword expressions (Fraser & Marcu 2007). For the multiword expression *duloxetine-treated patients*, the methods "FEATURE" and "CORPUS" provide a same translation which is more or less correct (*patients traités duloxetine*). However, the method "TABLE" provides a translation in a poor grammar (*patients duloxetine traités*).

Finally on this point, we can observe that the major issues of Moses concern errors produced by Giza++ when aligning multiword expressions (translation model), and 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 ("FEATURE" method). However, for spelling and grammar errors, Moses has no specific treatment.

6 Conclusion and future work

We have described, in this chapter, three approaches aiming to extract and align MWEs in English-French parallel corpora. We have also presented an experimental evaluation of the impact of integrating the results of these MWEs alignment approaches on the performance of the statistical machine translation system Moses. We have more specifically shown that, on the one hand, the hybrid approach based on morpho-syntactic patterns performs better than the other approaches and the "FEATURE" integration method achieves the best improvement, and on the other hand, using MWEs as additional parallel sentences to train the translation model of Moses improves its BLEU score.

This study offers several open issues for future work. First, we should explore machine learning approaches to extend the morphosyntactic patterns to take into account other forms of MWEs. The second perspective is to explore the integration of bilingual MWEs into other machine translation models such as rule-based translation ones. We also expect to explore the use of LSTM (Long Short-Term Memory) recurrent neural network language models for rescoring the *n*-best translations produced by Moses in order to reduce grammar errors.

References

Barbu, Ana Maria. 2004. Simple linguistic methods for improving a word alignment algorithm. In *Proceedings of the 7th international conference on the Statistical Analysis of Textual Data (JADT 2004)*, 88–98. Louvain-la-Neuve, Belgium.

Barz, Irmhild. 1996. Komposition und Kollokation. In Clemens Knobloch & Burkhard Schaeder (eds.), *Nomination – fachsprachlich und gemeinsprachlich*, 127–146. Springer.

Besançon, R., G. De Chalendar, O. Ferret, F. Gara, M. Laib, O. Mesnard & N. Semmar. 2010. LIMA: A multilingual framework of linguistic analysis and linguistic resources development and evaluation. In *Proceedings of the seventh international conference on Language Resources and Evaluation (LREC 2010)*, 3697–3704.

Blank, Ingeborg. 2000. Terminology extraction from parallel technical texts. In *Parallel text processing*, 237–252. Springer.

Bouamor, Dhouha, Nasredine Semmar & Pierre Zweigenbaum. 2012a. A study in using English-Arabic multiword expressions for statistical machine translation. In *Proceedings of the fourth international conference on Arabic Language Processing (CITALA-2012)*, 71–76. Rabat, Morocco.

Bouamor, Dhouha, Nasredine Semmar & Pierre Zweigenbaum. 2012b. 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 (CogALex-III)*, 95–108. Mumbai, India.

Bouamor, Dhouha, Nasredine Semmar & Pierre Zweigenbaum. 2012c. Identifying bilingual multi-word expressions for statistical machine translation. In *Proceedings of the eigth international conference on Language Resources and Evaluation (LREC 2012)*, 674–679. Istanbul, Turkey.

Boulaknadel, Siham, Béatrice Daille & Driss Aboutajdine. 2008. A multiterm extraction program for Arabic language. In *Proceedings of the international conference on Language Resources and Evaluation (LREC 2008)*, 1485–1488. Marrakech, Morocco.

Burger, H. 2010. Phraseologie: Eine Einführung am Beispiel des Deutschen [Phraseology: An introduction for German]. *Grundlagen der Germanistik* 36(4).

Calzolari, Nicoletta, Alessandro Lenci, Francesca Bertagna & Antonio Zampolli. 2002. Broadening the scope of the EAGLES/ISLE lexical standardization initiative. In *COLING '02: proceedings of the 3rd workshop on Asian Language Resources and International Standardization*, 1–8. Morristown, NJ, USA: Association for Computational Linguistics.

Carpuat, Marine & Mona Diab. 2010. Task-based evaluation of multiword expressions: A pilot study in statistical machine translation. In *Human Language Technologies: the 2010 annual conference of the North American chapter of the Association for Computational Linguistics (NAACL/HLT 2010)*, 242–245.

Choueka, Yaacov. 1988. Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *Proceedings of the 2nd international conference on Computer-Assisted Information Retrieval (recherche d'information assistée par ordinateur) - RIAO'88*, 609–624. Cambridge, MA, USA.

Constant, Mathieu, Isabelle Tellier, Denys Duchier, Yoann Dupont, Anthony Sigogne & Sylvie Billot. 2011. Intégrer des connaissances linguistiques dans un CRF: application à l'apprentissage d'un segmenteur-étiqueteur du français. In *Taln*, vol. 1, 321.

Dagan, Ido & Ken Church. 1994. Termight: Identifying and translating technical terminology. In *Proceedings of the fourth conference on Applied Natural Language Processing (ANLP'94)*, 34–40.

Daille, Béatrice. 2001. Extraction de collocation à partir de textes. In *Actes de la 8ème conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, 3–8. Tours, France: Association pour le Traitement Automatique des Langues.

Daille, Béatrice, Eric Gaussier & Jean-Marc Langé. 1994. Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th international conference on Computational linguistics (COLING 1994)*, 515–521.

DeNero, John & Dan 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: Short papers*, 25–28.

Fraser, Alexander & Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics* 33(3). 293–303.

Gao, Zhao-Ming. 1998. Automatic acquisition of a high-precision translation lexicon from parallel Chinese-English corpora. *Language* 248. 254.

Gaussier, Éric & François Yvon. 2011. *Modèles statistiques pour l'accès à l'information textuelle.* Lavoisier.

Germann, Ulrich. 2008. Yawat: Yet another word alignment tool. In *Proceedings of the ACL-08: HLT demo session*, 20–23. Columbus, Ohio, USA: Association for Computational Linguistics. https://www.aclweb.org/anthology/P08-4006.

Hogan, Deirdre, Conor Cafferkey, Aoife Cahill & Josef Van Genabith. 2007. Exploiting multi-word units in history-based probabilistic generation. In *Proceedings of the joint conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL)*, 267–276. Prague, Czech Republic.

Jackendoff, Ray. 1997. *The architecture of the language faculty.* Cambridge, MA, USA: MIT Press.

Koehn, Philipp & Hieu Hoang. 2007. Factored translation models. In *Proceedings of the 2007 joint conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 868–876.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin & Eva Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 177–180. Prague, Czech Republic.

Kupiec, Julian. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, 17–22.

Marchand, Morgane & Nasredine Semmar. 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*, 430–434. Poznań, Poland.

Nivre, Joakim & Jens Nilsson. 2004. Multiword units in syntactic parsing. In *Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications*, 39–46.

Och, F. J. & H. Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th annual meeting on Association for Computational Linguistics (acl 2000)*, 440–447.

Okita, Tsuyoshi, Alfredo Maldonado Guerra, Yvette Graham & Andy Way. 2010. Multi-word expression-sensitive word alignment. In *4th international workshop on Cross Lingual Information Access at COLING 2010*, 26–34. Beijing, China.

Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on Association for Computational Linguistics (ACL 2002)*, 311–318.

Ramisch, Carlos, Aline Villavicencio & Valia Kordoni. 2013. Introduction to the special issue on multiword expressions: From theory to practice and use. *ACM Transactions on Speech and Language Processing (TSLP)* 10(2). 3.

Ren, Z., Y. Lü, J. Cao, Q. Liu & Y. Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, 47–54.

Riehemann, Susanne. 2001. *A constructional approach to idioms and word formation.* Stanford University dissertation.

Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd international conference on Computational Linguistics and Intelligent Text Processing* (Lecture Notes in Computer Science 2276), 1–15. Springer.

Salton, G., A. Wong & C. S. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM* 18(11). 613–620.

Semmar, Nasredine, Christophe Servan, Gaël De Chalendar, Benoît Le Ny & Jean-Jacques Bouzaglou. 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 (ASLIB 2010).* London, UK.

Seretan, Violeta & Eric Wehrli. 2007. Collocation translation based on sentence alignment and parsing. In *Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles (TALN 2007)*, 401–410. Toulouse, France.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC-2012)*, 2214–2218. Istanbul, Turkey.

Tufiş, Dan & Radu Ion. 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*, 183–195.

Vechtomova, Olga. 2005. The role of multi-word units in interactive information retrieval. In *European conference on Information Retrieval*, 403–420.

Vintar, Špela & Darja Fišer. 2008. Harvesting multi-word expressions from parallel corpora. In *Proceedings of the 6th edition of the Language Resources and Evaluation conference (LREC 2008)*, 1091–1096.

Wu, Chien-Cheng & Jason S Chang. 2003. Bilingual collocation extraction based on syntactic and statistical analyses. In *Proceedings of Research on Computational Linguistics conference XV (ROCLING'2003)*, 33–55. Hsinchu, Taiwan: The Association for Computational Linguistics & Chinese Language Processing (ACLCLP). http://aclweb.org/anthology/O03-1003.