Automatic Construction of a MultiWord Expressions Bilingual Lexicon: A Statistical Machine Translation Evaluation Perspective

Dhouha Bouamor[1,2] *Nasredine Semmar*[1] *Pierre Zweigenbeaum*[2]
(1) CEA-LIST, Vision and Content Engineering Laboratory 91191 Gif-sur-Yvette Cedex, France
(2) LIMSI-CNRS, F-91403 Orsay, France
dhouha.bouamor@cea.fr, nasredine.semmar@cea.fr, pz@limsi.fr

ABSTRACT
Identifying and translating MultiWord Expressions (MWEs) in a text represent a key issue for numerous applications of Natural Language Processing (NLP), especially for Machine Translation (MT). In this paper, we present a method aiming to construct a bilingual lexicon of MWEs from a French-English parallel corpus. In order to assess the quality of the mined lexicon, a Statistical Machine Translation (SMT) task-based evaluation is conducted. We investigate the performance of three dynamic strategies and of one static strategy to integrate the mined bilingual MWEs lexicon in a SMT system. Experimental results shows that such a lexicon improves the quality of translation.

Construction Automatique d'un Lexique Bilingue d'Expressions Multi-Mots: Une Perspective d'Évaluation par un Système de Traduction Statistique

Identifier et traduire correctement les Expressions Multi-Mots (EMMs) dans un texte constituent un défi majeur pour différentes applications du Traitement Automatique des Langues Naturelles, et surtout en Traduction Automatique. Ce présent travail présente une méthode permettant de construire un lexique bilingue d'EMMs à partir d'un corpus parallèle Français-Anglais. Afin d'évaluer la qualité du lexique acquis, une évaluation axée sur la tâche de Traduction Automatique Statistique (TAS) est menée. Nous étudions les performances de trois stratégies dynamiques et d'une stratégie statique pour intégrer le lexique bilingue d'EMMs dans un système de TAS. Les expériences menées dans ce cadre montrent que ces unités améliorent la qualité de traduction.

KEYWORDS: MultiWord Expressions, Bilingual Alignment, Statistical Machine Translation.

KEYWORDS IN FRENCH: Expressions Multi-Mots, Alignement Bilingue, Traduction Automatique Statistique.

Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III), pages 95–108,
COLING 2012, Mumbai, December 2012.

1 Introduction

A MultiWord Expression (MWE) can be defined as 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 are made up of collocations (*"cordon bleu"*), frozen expressions (*"kick the bucket"*), named entities (*"New York"*) etc. (Sag et al., 2002; Constant et al., 2011). These units are numerous and constitute a significant portion of the lexicon of any natural language. (Jackendoff, 1997) claims that the frequency of MWEs in a speaker's lexicon is almost equivalent to the frequency of single words. While easily mastered by native speakers, their interpretation poses a major challenge for Natural Language Processing (NLP) applications, especially for those addressing semantic aspects of language.

For Statistical Machine Translation (SMT) systems, various improvements of translation quality were achieved with the emergence of phrase based approaches (Koehn et al., 2003). Phrases are usually defined as simply arbitrary n-grams with no sophisticated linguistic motivation consistently translated in a parallel corpus. In such systems, the lack of an adequate processing of MWEs could affect the translation quality. In fact, the literal translation of an unrecognized expression is the source of an erroneous and incomprehensible translation. For example, these systems would suggest *"way of iron"* as a translation of *"chemin de fer"* instead of *"railway"*. It is therefore important to use a lexicon in which MWEs are handled. But such a resource is not readily available in all languages, and if it exists, as described by (Sagot et al., 2005), it does not cover all MWEs of a given language.

In this paper, we propose a method aiming to acquire a bilingual lexicon of MWEs from a French-English parallel corpus. We consider any compositional and non-compositional contiguous sequence, belonging to one of the three classes defined by (Luka et al., 2006), as a MWE. Classes of MWEs were distinguished on the basis of their categorical properties and their syntactic and semantic fixedness degrees and consist of *compounds*, *idiomatic expressions* and *collocations*. Intuitively, bilingual MWEs are useful to improve the performance of SMT. However, further research is still needed to find the best way to bring such external knowledge to the decoder. In this study, we view SMT as an extrinsic evaluation of the usefulness of MWEs and explore strategies for integrating such textual units in an SMT system. Given a constructed bilingual MWEs lexicon, we propose (1) three *dynamic integration* strategies in which we attempt to change the translation model in several ways to handle MWEs and (2) a *static integration* strategy in which we would like to plug these translations into the decoder without changing the model.

This paper is organized as follows: the next section (section 2) describes in some details previous works addressing the task of bilingual extraction of MWEs and its applications. In section 3, we present the method we used to build the bilingual lexicon of MWEs and then introduce in section 4 four strategies aiming to integrate MWEs in an SMT system. In section 5, we report and discuss the obtained results. We finally conclude and present our future work in section 6.

2 Related Work

In recent years, a number of techniques have been introduced to tackle the task of bilingual MWEs extraction from parallel corpora. Most works start by identifying monolingual MWE candidates then, apply different alignment methods to acquire bilingual correspondences. Monolingual extraction of MWEs techniques revolve around three approaches: (1) symbolic

methods relying on morphosyntactic patterns (Okita et al., 2010; Dagan and Church, 1994); (2) statistical methods which use association measures to rank MWE candidates (Vintar and Fisier, 2008) and (3) Hybrid approaches combining (1) and (2) (Wu and Chang, 2004; Seretan and Wehrli, 2007; Daille, 2001; Boulaknadel et al., 2008). Each approach shows several limitations. It is, for example, difficult to apply symbolic methods to data without syntactic annotations. Furthermore, due to corpus size, statistical measures have mostly been applied to bigrams and trigrams, and it becomes more problematic to extract MWEs of more than three words. Concerning the alignment task, numerous approaches have already been introduced to deal with this issue. Some works make use of simple-word alignment tools (Dagan and Church, 1994; Lefever et al., 2009). Others rely on machine learning algorithms such as the *Expectation Maximisation (EM)* algorithm (Kupiec, 1993; Okita et al., 2010). In another direction, (Tufis and Ion, 2007; Seretan and Wehrli, 2007) introduce a linguistic approach in which they claim that MWEs keep in most cases the same morphosyntactic structure in the source and target language, which is not universal. For example the French MWE *"insulaire en développement"*, aligned with the English MWE *"small island developing"* do not share the same morphosyntactic structure.

Most of the methods described above aims at identifying MWEs in a corpus to construct or extend a bilingual lexicon without any application perspective. However, few works have focused on the extraction of bilingual MWEs lexicons in order to improve the performance of MT systems by reporting improved BLEU (Papineni et al., 2002) scores. This measure calculates the n-grams precision against a reference translation. In (Lambert and Banchs, 2005), authors introduce a method in which a bilingual MWEs lexicon was used to modify the word alignment in order to improve the translation quality. In their work, bilingual MWEs were grouped as one unique token before training alignment models. They showed on a small corpus, that both alignment quality and translation accuracy were improved. However, in a further study, a lower BLEU score is reported after grouping MWEs by part-of-speech on a large corpus (Lambert and Banchs, 2006). Some works have however focused on automatically learning translations of very specific MWEs categories, such as, for instance, idiomatic four character expressions in Chinese (Bai et al., 2009) or domain specific MWEs (Ren et al., 2009). (Carpuat and Diab, 2010) introduced a framework of two complementary integration strategies for monolingual MWEs in SMT. The first strategy segments training and test sentences according to the monolingual MWEs vocabulary of *Wordnet*. In the second strategy, they add a new MWE-based feature in SMT translation lexicons representing the number of MWEs in the source sentence. More recently, In (Bouamor et al., 2011), we proposed a method to enrich an SMT system's phrase table by a bilingual lexicon handling MWEs. On a small corpus (10k sentences), this method yeilds an improvement of 0.24 points in BLEU score. This study is an extention of the approach we presented in (Bouamor et al., 2011). We propose a method aiming to extract and align such units and study different strategies to integrate them into MOSES (Koehn, 2005), the state-of-the art SMT system.

3 Bilingual MWEs lexicon

In this section, we describe the approach we used to mine the bilingual lexicon of MWEs from a sentence aligned French-English parallel corpus. This approach is conducted in two steps. We first extract monolingual MWEs from each part of the parallel corpus. The second step consists in acquiring bilingual correspondences of MWEs.

Pattern	English/ French MWEs
Adj-Noun	Plenary meeting / Libre circulation
Noun-Adj	... / Parlement européen
Noun-Noun	Member state / Etat membre
Past_Participle -Noun	Developped country/ ...
Noun-Past_Participle	Parliament adopted/ Pays developpé
Adj-Adj-Noun	European public prosecutor /...
Adj-Noun-Adj	Social market economy / Bon conduite administratif
Adj-Noun-Noun	Renewable energy source / ...
Noun-Noun-Adj	... / Industrie automobile allemand
Noun-Adj-Adj	... / Ministère public européen
Adj-Noun-Adj	... / Important débat politique
Noun-Prep-Noun	Point of view / Chemin de fer
Noun-Prep-Adj-Noun	Court of first instance/ Court de première instance
Noun-Prep-Noun-Adj	.../ Source d'énergie renouvelable
Adj-Noun-Prep-Noun	European court of justice/ ...
Noun-Adj-Prep-Noun	.../ Politique européen de concurrence

Table 1: French and English MWE's morphosyntactic patterns

3.1 Monolingual Extraction of MWEs

The method we propose to identify monolingual MWEs in a text is based on a symbolic approach. This method is quite similar to the one used by (Okita et al., 2010). If they define patterns to handle only noun phrases, our approach takes into account both noun phrases, fixed expressions and named entities. Relatively simple, it does not use additional correlations statistics such as Mutual Information or Log Likelihood Ratio and attempts to find translations for all extracted MWEs (both highly and weakly correlated MWEs), to our knowledge, none of other approaches can make this claim. This method involves only a full morphosyntactic analysis of source and target texts. This morphosyntactic analysis is achieved using the CEA LIST Multilingual Analysis platform (LIMA) (Besançon et al., 2010) which produces a set of part of speech tagged normalized lemmas. Our algorithm operates on lemmas instead of surface forms which can draw on richer statistics and overcome the data sparseness problems. Since most MWEs consist of noun, adjectives and prepositions, we adopted a linguistic filter keeping only n-gram units ($2 \leq n \leq 4$) which match a list of 16 hand created morphosyntactic patterns. Such a 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"*). In Table 1 we give an example of MWE produced for each pattern. There exists extraction patterns (or configuration) for which no MWE has been generated (i.e. Noun-Adj).

Some of the fixed expressions such as (*in particular, in the light of, as regards...*) and named entities (*Midle East, South Africa, El-Salvador...*) recognized by the morphosyntactic analyzer are added to the candidate list. Then, all extracted MWEs are stored with their total frequency of occurrence. 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 them both. Additionally, if a MWE occurs in a high number of longer terms, we discard all such longer terms.

French	→	English
parlement européen	→	european parliament
état par état	→	amount of state
coup d'état	→	military coup
zone non fumeur	→	no smoking area
insulaire en développement	→	small island developing
de bonne foi	→	good faith
politique de concurrence	→	competition policy
chemin de fer	→	railway sector
en ce qui concerne	→	in regard to
en ce qui concerne	→	as regards
en ce qui concerne	→	with reference to
en ce qui concerne	→	with respect to
coupe forestier	→	cut in forestation

Table 2: Sample of aligned MWEs

3.2 Bilingual Alignment

Bilingual alignment is achieved by a method which consists in finding for each MWE in a source language its adequate translation in the target language. Traditionally, this task was handled through the use of external linguistic resources such as bilingual dictionaries or simple-word alignment tools. We propose a *resource-independant* method which simply requires a parallel corpus and a list of input MWE candidates to translate. Our approach is based on aspects of distributional semantics (Harris, 1954), where a specific representation is associated 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 or not it occurs in each sentence of the corpus. Our algorithm is based on the Vector Space Model (VSM). VSM (Salton et al., 1975) is a well-known algebraic model used in information retrieval, indexing and relevance ranking. 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, greedy alignment algorithm which operates as follows:

1. Find the most frequent MWE exp in each source sentence.
2. Extract all target translation candidates, occurring in all sentences parallel to those containing exp.
3. Compute a confidence value V_{Conf} for each translation relation between exp and each target translation candidate.
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.

The confidence value V_{Conf} is computed on the basis of the *Jaccard Index* (1).

$$Jaccard = \frac{I_{st}}{V_s + V_t + I_{st}} \tag{1}$$

This measure is based on the number I_{st} of sentences shared by each target and a source MWE. This is normalized by the sum of the number of sentences where the source and target MWEs

appear independently of each other (V_s and V_t) increased by I_{st}. In table 2, a sample of MWEs aligned by means of the algorithm described above.

From observing some pairs, we notice that our method presents several advantages: In order to find the adequate translation of a MWE and contrary to most previous works (Dagan and Church, 1994; Ren et al., 2009) using simple-word alignment tools to establish word-to-word alignment relations, our method captures the semantic equivalence between expressions such as *"insulaire en dévelopement"* and *"small island developing"* without any prior information about word alignment. It also permits the alignment of idioms such as *à nouveau* → *once more* or even *état par état* → *amount of state* and works for MWEs for which multiple correct target MWEs exist. For instance it captures that the *MWE "en ce qui concerne"* could be translated by *"in regard to", "with reference to", "with respect to"* and even by *"as regards"*.

4 Integration strategies

In the previous section, we described the approach we followed to mine the bilingual lexicon of MWEs. In order to assess the lexicon's quality, we carried out in a previous work (Bouamor et al., 2011) an intrinsic evaluation in which we compared the obtained pairs of bilingual MWEs to a manual alignment reference. On a small set of 100 French-English parallel sentences derived from the Europarl corpus, our approach yielded a precision of 63,93% , a recall of 62,46% and an F-measure of 63,19%. As it lacks a common benchmark data set for evaluation in MWE extraction and alignment researches, we carry out an extrinsic evaluation based on an SMT application and use MOSES (Koehn, 2005) as our BASELINE system. However, as we mentioned in section 1, the difficulty lies in how to integrate MWEs into such systems. To do so, we propose three dynamic integration strategies in which the translation model is amended in severals ways, and a static integration strategy in which we plug MWEs into the decoder without changing the model. We compare their performance in section 5.

4.1 Dynamic integration strategies

4.1.1 New Translation model with MWEs

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 candidate in a source language into the target one. However, due to the errors in automatic word alignment, extracted phrases might be meaningless. To alleviate this problem, we add the extracted bilingual MWE as a parallel corpus and retrain the translation model. In this method (TRAIN), we expect that by increasing the occurrences of bilingual MWEs, considered as good phrases, a modification of the alignment and the translation probability will be noticed.

4.1.2 Extention of the phrase table

In this method, we attempt to extend the BASELINE system's phrase table by integrating the found bilingual MWEs candidates. We use the Jaccard Index (proposed for each pair of MWEs) to define the translation probabilities in the two directions and set the lexical probabilities to 1 for simplicity. So, for each phrase in a given input sentence, the decoder will take into account bilingual MWEs when searching for all candidate translation phrases. This method is denoted TABLE in the remaining part of this paper.

4.1.3 New feature for MWEs

(Lopez and Resnik, 2006) pointed out that better feature mining can lead to substantial gain in translation quality. We followed this claim and extended TABLE by adding a new feature indicating whether a phrase is a MWE or not. The aim of this method (FEAT) is to guide the system to choose bilingual MWEs returned by our aligner instead of the BASELINE's system phrases.

4.2 Static Integration strategy

In this method, noted FORCED, we want to bring the bilingual MWEs lexicon to the decoder without changing the translation model. For this claim, we used the *forced decoding mode* of the Moses system. The decoder has an *XML markup scheme* that allows the specification of translations for parts of the sentence. In its simplest form, we can indicate to the decoder what to use to translate certain words or phrases in the sentence. So we represented each MWE occuring in the test set by its adequate XML markup scheme, using the translation pair of the lexicon. Below is an example of representing the MWE *à nouveau* in the test set.

... sembler être **à nouveau** mis en accusation, le ministère public ...

⇓

... sembler être < ***mwe*** **translation=*"once more"*** >**à nouveau**< /***mwe***> mis en accusation, le ministère public ...

5 Experiments

5.1 Data and tools

We used the French-English Europarl (Koehn, 2005) corpus of parliamentary debates as a source of the parallel corpus. To train the BASELINE system's translation model, we extracted 100000 pairs of sentences from the corpus. First, we tokenized, cleaned up the training corpus and kept only sentences containing at most 50 words. We mined the bilingual lexicon of MWEs from the same training corpus. Because the lexicon contains only lemmas of MWEs and the forced decoding mode of Moses is not currently compatible with factored models, the translation model was trained on lemmas instead of their surface forms. Training data were annotated with lemmas by means of the TreeTagger Toolkit[1]. Next, word-alignment for all the sentences in the parallel training corpus is established and uses the same methodology as in phrase-based models (symmetrized GIZA++ alignments) to create the phrase table. We also specified a language model using the IRST Language Modeling Toolkit [2] to train a lemma based tri-gram model on the total size of the Europarl corpus (1.8M sentences). Afterwards, we applied the above-described integration strategies.

The features used in the BASELINE system include: (1) four translation probability features, (2) one language model and (3) word penalty. For the "TRAIN" method, bilingual MWEs are added into the training corpus, as results, new alignments and phrase table are obtained. For the "TABLE" method, bilingual units are incorporated into the BASELINE system's phrase table. In "FEAT", an additional 1/0 feature is introduced for each entry of the phrase table. Concerning the FORCED method, it keeps the same models as BASELINE. Afterwards, the obtained models

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[2] http://hlt.fbk.eu/en/irstlm

Method	BLEU		TER	
	All_Test	*MWEs_Test*	*All_Test*	*MWEs_Test*
BASELINE	28.85	30.83	55.44	53.59
Dynamic				
TRAIN	**28.87**	**31.06**	**55.38**	**53.32**
TABLE	28.82	**30.88**	**55.42**	**53.46**
FEAT	**28.95**	**31.06**	55.48	**53.56**
Static				
FORCED	28.20	29.19	56.01	55.05

Table 3: Translation results in term of BLEU and TER

were tuned by Minimum Error Rate Training (Och, 2003) on a development set of 4000 pairs of sentences.

5.2 Results and discussion

We conducted two test experiments: *All_Test* and *MWEs_Test*. For this, we *randomly* extracted 1000 parallel sentences from the corpus described above to construct the *All_Test* test corpus. In order to measure the real contribution of bilingual MWEs handled by different translation models, we constituted the *MWEs_Test* corpus, in which we kept only sentences of the *All_Test* corpus containing at least one MWE of the lexicon. This corpus contains 323 pairs of sentences. We evaluate the translation quality of the described dynamic and static strategies on the two test sets with respect to BLEU (Papineni et al., 2002) score, which is based on *n-gram* precision, and Translation Error Rate (TER) (Snover et al., 2006), which generalizes edit distance beyond single-word edits. For this evaluation, we consider one reference per sentence. Table 3 reports the obtained results.

The first substantial observation, as can be seen, is related to the BLEU scores which vary according to the test set type. Concerning the *All_test* corpus, the best improvement is achieved by the FEAT dynamic strategy, in which we add a new feature indicating wether a phrase in the phrase table is a MWE or not. Compared to the BASELINE, this method reports a gain of +0.1 point in BLEU score. The first translation example in Table 4 points out the contribution of the introduced feature to the performance of the translation approach. Contrary to the BASELINE system, which translates the unit *"initiative communautaire"* as simply *"initiative"*, the FEAT strategy adequately translates both the MWE *"initiative communautaire"*→*"community initiative"* and its immediate right context(*"for africa"*). Lower BLEU scores are achieved by TABLE and FORCED wrt. the BASELINE system. For the *MWEs_Test* corpus, which considers only sentences containing MWEs of the lexicon, we notice that all dynamic integration strategies report increased BLEU scores compared to the BASELINE and the static integration strategy (FORCED). The FEAT and TRAIN methods achieve a gain of +0.23 BLEU points over the BASELINE system. The TABLE strategy comes next with a slighly improved BLEU score showing a gain of +0.05 BLEU points. However, the FORCED static strategy reports lower scores on both *All_Test* and *MWEs_Test* test corpora. This can certainly be explained as follows: while forcing the decoder to translate a MWE with a given MWE candidate, even if it is a good translation, it fails to adequately translate the immediate left or right context of the MWEs which consequently lowers the BLEU score. For example, in the second example of Table 4, both systems suggest a correct translation for the MWE *"aide international"* but FORCED fails to adequately translate the

SOURCE SENTENCE	je entendre en effet lancer un initiative communautaire pour le afrique en étendre le ligne nepad ...
REFERENCE	indeed , i intend to launch a *community initiative* **for africa** , develop the nepad line...
BASELINE	i hear be indeed launch an *initiative* **for the eu africa** by extend the nepad line ...
FEAT	i hear in fact launch a *community initiative* **for africa** by extend the nepad line ...
SOURCE SENTENCE	le deuxième groupe de problème relever de le aide international et du prochain engagement de johannesburg.
REFERENCE	another series of problem **mention be a matter of** *international aid* and the forthcoming johannesburg summit.
BASELINE	the second group of the problem **be a matter of** *international aid* and the forthcoming johannesburg commitment.
FORCED	the second group of the problem **relate to the** *international aid* and the forthcoming johannesburg commitment.

Table 4: Translation examples. Note that the text is lemmatized. We underline MWEs and put in bold different suggestion of immediate left or right context.

Method	*p-value* (95%*CI*)	
	All_Test	*MWE_Test*
BASELINE	-	-
TRAIN	0.1	0.05
TABLE	-	0.3
FEAT	0.01	0.01

Table 5: Statistical significance test of BLEU improvements in term of *p-value*

phrase *"relever de"*. It is important to note that this translation could be supported if we have for each source sentence multiple references. In an earlier study, (Ren et al., 2009) proposed a strategy quite similar to the FEAT method in which they indicate for each entry in the phrase table wether a phrase contains a *domain specific bilingual* MWE. For the medical domain, their method gained +0.17 of BLEU score compared to the baseline system, a lower improvement than the one reported by the FEAT method. The question that arises based on these different results is: Is it possible to claim that the system having the best score is the best one? In other words, are the obtained results for the different experimental settings statistically significant?

In order to assess statistical significance of previously obtained test results, we use the *paired bootstrap resampling* method (Koehn, 2004). This method estimates the probability (*p-value*) that a measured difference in BLEU scores arose by chance by repeatedly (10 times) creating new virtual test sets by drawing sentences with replacement from a given collection of translated sentences. If there is no significant difference between the systems (*i.e., the null hypothesis is true*), then this shuffling should not change the computed metric score. We carry out experiments using this method to compare each of the methods TRAIN, TABLE and FEAT, yielding improvements in BLEU scores (Table 3) over the BASELINE system on the two test set results *All_Test* and *MWE_Test*.

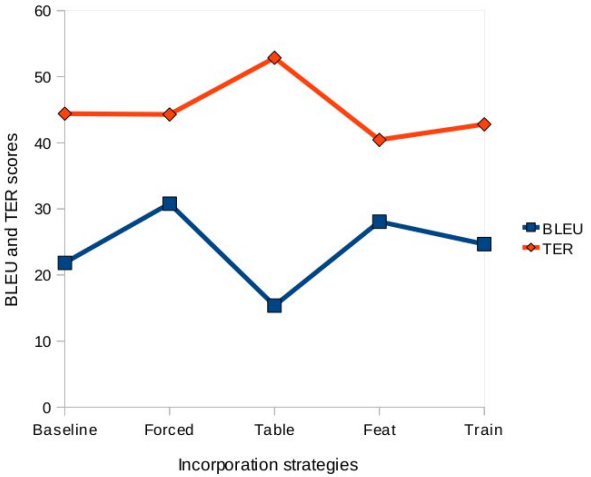

Figure 1: Lexical evaluation of MWEs in term of BLEU and TER

Table 5 displays reported *p-values* at the edge of the 95% *confidence interval* (CI). As can be observed, the results vary from insignificant (at $p > 0.05$) to highly significant. On both test set results, we notice that improvements achieved by the FEAT integration startegy are statistically significant. However, the small improvement of BLEU score yielded by the TABLE method (having a p-value of 0.3) is non significant. The reason being that we used the *Jaccard Index*, a measure defined for comparing similarity and diversity of sample sets, to define a translation probability. This could be adjusted by transforming obtained Jaccard Index for each pair of MWE to a translation probability in order to ensure the uniformity and consistency of translation probabilities in the phrase table.

The BLEU metric reports only global improvements and does not show significant differences that can be revealed by human evaluation. This observation motivated us to set up a fine-grained *lexical evaluation* of MWEs in the *MWEs_test* corpus. We kept only MWEs on the test corpus and manually created the gold standard from the reference. We translated the new test corpus according to dynamic and static integration strategies and computed BLEU and TER scores. Figure 1 illustrates obtained results. As one can note, a gain of almost +9.8 BLEU, −0.2 TER points are achieved by the FORCED strategy. This confirms that the worsening of BLEU scores in previous experiments are not affected by the quality of the bilingual MWEs lexicon. We notice also that both TRAIN and FEAT strategies report higher scores (respectively 24.67 and 28.06 points BLEU) compared to the BASELINE which comes with 21.84 points BLEU.

6 Conclusion

We proposed in this paper a hybrid approach to identify and find bilingual MWEs correspondences in a French-English parallel corpus. The alignment algorithm we propose works only on many to many correspondences and deals with highly and weakly correlated MWEs in a given sentence pair. In order to assess the lexicon's quality, we investigated the performance of three dynamic strategies and of one static strategy to integrate the mined bilingual MWE lexicon in the MOSES SMT system. We showed that the FEAT method, in which we add a new feature indicating whether a phrase is a MWE or not, brings a small but statistically significant improvement to the translation quality of the test sets. We also introduced a lexical evaluation of MWEs units based on the measure of BLEU score.

Although our initial experiments are positive, we believe that they can be improved in a number of ways. We first plan to set up a large scale evaluation by enlarging the size of the training corpus. In all experiments, we trained a translation model on lemmas instead of surface forms. We will make use of a generation model to generate adequate surface forms from lemmas. In addition to their application in a phrase based SMT system, we plan to evaluate the impact of the mined lexicon on the relevance of a cross-language search engine results. We also expect to extract such textual units from more available but less parallel data sources: *comparable corpora*.

References

Bai, M., Y., J.-M., C., K.-J., and Chang, J. S. (2009). Acquiring translation equivalences of multiword expressions by normalized correlation frequencies. In *Proceedings of EMNLP*.

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

Bouamor, D., Semmar, N., and Zweigenbaum, P. (2011). Improved statistical machine translation using multi-word expressions. In *Proceedings of MT-LIHMT*, Barcelona, Spain.

Boulaknadel, S., Daille, B., and Driss, A. (2008). A multi-term extraction program for arabic language. In *Proceedings of LREC*, Marrakech, Morocco.

Carpuat, M. and Diab, M. (2010). Task-based evaluation of multiword expressions:a pilot study in statistical machine translation. In *Proceedings of HLT-NAACL*.

Constant, M., Tellier, I., Duchier, D., Dupont, Y., Sigogne, A., Billot, S., et al. (2011). Intégrer des connaissances linguistiques dans un crf: application à l'apprentissage d'un segmenteur-étiqueteur du français. In *Actes de TALN*, Montpellier, France.

Dagan, I. and Church, K. (1994). Termight: Identifying and translating technical terminology. In *Proceedings of the 4th Conference on ANLP*, pages 34–40, Stuttgart, Germany.

Daille, B. (2001). Extraction de collocation à partir de textes. In Maurel, D., editor, *Actes de TALN 2001 (Traitement automatique des langues naturelles)*, Tours. ATALA, Université de Tours.

Harris, Z. (1954). Distributional structure. *Word*.

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

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT-SUMMIT*.

Koehn, P., Och, F., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 115–124, Edmonton, Canada.

Kupiec, J. (1993). An algorithm for finding noun phrases correspondences in bilingual corpora. In *Proceedings of the 31st annual Meeting of the Association for Computational Linguistics*, pages 17–22, Columbus, Ohio, USA.

Lambert, P. and Banchs, R. (2005). Data inferred multi-word expressions for statistical machine translation. In *Proceedings of MT SUMMIT*.

Lambert, P. and Banchs, R. (2006). Grouping multi-word expressions according to part-of-speech in statistical machine translation. In *Proceedings of the Workshop on Multi-word Expressions in a multilingual context*.

Lefever, E., Macken, L., and 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 ACL (EACL 2009)*, pages 496–504, Athens, Greece. Association for Computational Linguistics.

Lopez, A. and Resnik, P. (2006). Word-based alignment, phrase based translation:what's the link? In *Proceedings of the association for machine translation in the Americas: visions for the future of machine translation*, pages 90–99.

Luka, N., Seretan, V., and Wehrli, E. (2006). Le problème de collocation en tal. In *Nouveaux cahiers de linguistiques Française*, pages 95–115.

Och, F. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of ACL*.

Okita, T., Guerra, M., Alfredo 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 COLING 2010*, pages 26–34, Beijing.

Papineni, k., Roukos, S., Ward, T., and Zhu, W. J. (2002). Bleu: a method for automatic evaluation of machine translation. In *40th Annual meeting of the Association for Computational Linguistics*.

Ren, Z., Lu, Y., Liu, Q., and Huang, Y. (2009). Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions : Identification, Interpretation, Disambiguation and Applications*, pages 47–57.

Sag, I., Baldwin, T., Francis Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions:a pain in the neck for nlp. In *CICLing 2002*, Mexico City, Mexico.

Sagot, B., Clément, L., De La Clergerie, É., Boullier, P., et al. (2005). Vers un méta-lexique pour le français: architecture, acquisition, utilisation. In *Actes de TALN*.

Salton, G., Wong, A., and Yang, C. (1975). A vector space model for automatic indexing. In *Communications of the ACM*, pages 61–620.

Seretan, V. and Wehrli, E. (2007). Collocation translation based on sentence alignment and parsing. In Benarmara, F., Hatout, N., Muller, P., and Ozdowska, S., editors, *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*.

Tufis, I. 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*, pages 183–195.

Vintar, S. and Fisier, D. (2008). Harvesting multi-word expressions from parallel corpora. In *Proceedings of LREC*, Marrakech, Morocco.

Wu, C. and Chang, S. J. (2004). Bilingual collocation extraction based on syntactic and statistical analyses. In *Computational Linguistics*, pages 1–20.