Improved Statistical Machine Translation Using MultiWord Expressions

Dhouha Bouamor
CEA-LIST, Vision and Content Engineering Laboratory
F-92265 Fontenay aux roses, France
dhouha.bouamor@cea.fr

Nasredine Semmar
CEA-LIST, Vision and Content Engineering Laboratory
F-92265 Fontenay aux roses, France
nasredine.semmar@cea.fr

Pierre Zweigenbaum
LIMSI-CNRS,
F-91403 Orsay, France
pz@limsi.fr

Abstract

Identifying and translating a MultiWord Expression (MWE) in a text represents an issue for numerous applications in Natural Language Processing (NLP) as MWEs appear in all text genres and pose significant problems for every kind of NLP tasks. In this paper, we describe a hybrid approach for extracting contiguous MWEs and their translations in a French-English parallel corpus. We evaluate both the alignment and the translation quality. Next, we implement a method that integrates these units to Moses, the state of the art Machine Translation (MT) system. Conducted experiments show that MWEs improve translation performance.

1 Introduction

Statistical Machine Translation (SMT) initially focused on word to word translations (Brown et al., 1993). Various improvements of SMT systems quality used phrase-based translation (Koehn et al., 2003), defined simply as n-grams consistently translated in a parallel corpora. To compensate the lack of semantic information in phrase based approaches, we study bilingual MultiWord Expressions (MWEs) and integrate them in an existing phrase-based SMT system.

(Sag et al., 2002) define MWEs very roughly as *"idiosyncratic interpretations that cross word boundaries (or spaces)"*. Theses lexical units are numerous and constitute a significant portion of the lexicon of any natural language. (Jackendoff, 1997:156) estimates 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 computational systems, due to their flexible and heterogeneous nature. SMT does not model MWEs explicitly. In phrase based MT systems, these units are indirectly captured but they are not distinguished from any other n-gram.

In recent years, a number of techniques have been applied to the problem of MWEs extraction (Kupiec, 1993; Okita et al., 2010; Dagan and Church, 1994). Most of them based on identifiying these units within a corpus, with the goal of including them in bilingual lexicons (Smadja, 1993). Having such type of terms is useful for a variety of NLP application such as information retrieval (Vechtomova, 2005) , word sense disambiguation (Finlayson and Kulkarni, 2011) and others.

Some researches exploited MWEs in MT systems. (Tanaka and Baldwin, 2003) described an approach of noun-noun compound machine translation, but not significant comparison was presented. In (Lambert and Banchs, 2005), authors introduce a method in which a bilingual MWEs corpus 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 their further study, they reported even lower BLEU scores after grouping MWEs by part-of-speech on a large corpus (Lambert and Banchs, 2006). Recently, (Ren et al., 2009) implemented a method integrating a domain bilingual MWE to Moses. The method yielded an improvement of 0.61 BLEU score compared with the baseline system.

In this paper, we describe a hybrid approach combining linguistic and statistical information to extract and align MWEs from a French-English parallel corpus. Extracted MWEs are then integrated into Moses. The conducted experiments show that MWEs identification improve the translation quality. The remainder of this paper is organized as follows. In section 2, we describe the proposed method for identifying and extracting bilingual MWEs. Experiments and results are discussed in section 3. We conclude and present our future work, in section 4.

2 Bilingual MultiWord Expressions Extraction

2.1 Related Work

A number of techniques have already been applied to the problem of MWEs extraction. Starting from a sentence aligned parallel corpus, most works rely on statistical, linguistic or hybrid approaches. The work of (Kupiec, 1993) is considered as one of the early work concerned with this task. The author focused essentially on noun groups. These units are identified on the basis of their part-of-speech tag. Then, based on the Expectation Maximization (EM) algorithm, bilingual correspondences are identified. It obtained a precision rate of 90% referred to the 100 first correspondences. An extension of this method is proposed by (Okita et al., 2010). To detect MWEs, a bidirectional version of Kupiec (1993) is applied. Then, in order to add prior information, they replace the maximum likelihood estimate in the M-Step of the EM algorithm with the Maximum-A-Posteriori (MAP) estimate. (Dagan and Church, 1994) describe a semi-automatic tool, Termight, which extracts technical noun groups using a syntactic pattern filter. They use a word alignment program to align MWEs. For each source term, the tool identifies a candidate translation by selecting a sequence of target words whose first and last word are aligned with any of the words in the source term. The accuracy obtained for 192 English-German correspondences is about 40%.

Other recent related work attempt to extend the linguistic based methods used in identifying MWEs. They use additional association measures such as Mutual Information (Daille, 2001) and the Log Likelihood Ratio (Wu and Chang, 2004; Seretan and Wehrli, 2007) to capture the degree of cohesion between the constituents of a MWE. However, these measures present two main shortcomings. They are designed for bigrams and require a definition of a threshold above which an extracted phrase is considered as a MWE. Afterwards, some heuristics, are applied for the alignment task. (Tufis and Ion, 2007) and (Seretan and Wehrli, 2007) assume that MWEs keep in most cases the same morphosyntactic structure in the source and target language, which is not universal such as the English MWE *small island developing* which is aligned with the French *insulaire en développement*. The Champollion system of (Smadja et al., 1996) can produce translations of a source MWEs in the target language. It is based on a multi-word unit extraction system, Xtract, developed by (Smadja, 1993). They first extracted source MWEs. After that, for each source term, they extracted its translations in the target language by testing Dice-score. Champollion was tested on the Hansard corpus and an accuracy of 73% was reported, taking into account only MWEs appearing at least 10 times in the corpus.

2.2 MWEs Identification

In this section, we describe the MWEs extraction method from a French-English parallel corpus. The process of extraction involves full morphosyntactic analysis of source and target texts. For this, we used the CEA LIST Multilingual Analysis platform (LIMA) (Besançon et al., 2010). The linguistic analyzer produces a set of part-of-speech tagged normalized lemmas. It is needed to only permit specific strings for extraction and filter out undesirable ones such as *of the, is a*. Since most MWEs consist of noun, adjectives and sometimes prepositions, we adopted a linguistic filter that accepts n-gram units ($2 \leq n \leq 4$) matching the morphosyntactic configurations presented in Table 1.

English Pattern	French Pattern
Adj-Noun	Noun-Adj
Noun-Noun	Adj-Noun
Past_Participle -Noun	Noun-Past_Participle
Adj-Adj-Noun	Noun-Noun-Adj
Adj-Noun-Adj	Noun-Adj-Adj
Adj-Noun-Noun	Adj-Noun-Adj
Noun-Prep-Noun	Noun-Prep-Noun
Noun-Prep-Adj-Noun	Noun-Prep-Noun-Adj
Adj-Noun-Prep-Noun	Noun-Adj-Prep-Noun

Table 1: French and English MWE's morphosyntactic structure

To this list are added some prepositional idiomatic expressions (*in particular, in the light of, as regards...*) and proper noun (*Midle East, South Africa, El-Salvador...*) recognized by the morphosyntactic analyzer. Then, we scored them 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 alternative of having a MWE that appears nested in a high number of terms. We followed (Frantzie et al., 2000) by discarding all longer MWEs. An example of extracted MWEs is in Table 2.

The presented approach does not use additional correlations statistics such as Mutual Infomation or Log Likelohood Ratio since these measures require a definition of a threshold above which an extracted phrase is considered as a MWE or not. Our method consider that all extracted units are effective and valid and include all of them in the translation process. To our knowledge, none of other approaches can make this claim.

Freq	French MWEs
144	Parlement européen
25	Prestation de service
29	Industrie automobile allemand
36	Chemin de fer
65	En particulier
32	Source d'énergie renouvelable
11	Mise en place
Freq	**English MWEs**
19	Court of first instance
316	Member state
19	Point of view
65	In particular
29	Plenary meeting
32	Rural development
21	European public prosecutor

Table 2: A sample of extracted French and English MWEs

2.3 MWEs Alignment

We present a method in which we try to find for each MWE in a source language, a translation to which is adequate in the target one. We focus only on many-to-many correspondences and do not use any dictionary nor simple-word alignment tools. Our algorithm is quite simple and 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. Each MWE is represented by a binary vector of size n[1] indicating for each sentence of the corpus whether it occurs in that sentence or not. Then, translation pairs of MWEs are extracted by means of the following iterative process:

1. Find the most frequent MWE in the source sentence.
2. Extract all translation candidates from the target parallel sentence.
3. Compute a confidence value for the translation relation.
4. Consider that the target MWE that maximize the confidence value is the best translation.
5. Discard the translation pair from the process and go back to 1.

To compute the confidence value, we adopted the Jaccard Index, a frequently used measure in information retrieval. It is defined as

$$IJ = \frac{NS_i}{NS_s + NS_t - NS_i} \quad (1)$$

[1]n=number of the aligned sentences of parallel corpora

and based on the number NS_i 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 (NS_s and NS_t) decreased by NS_i.

2.4 Extraction Method Evaluation

To evaluate the alignment quality, we followed the evaluation framework defined in the shared task on word alignment organized as part of the HLT/NAACL 2003 Workshop on building and using parallel corpora (Mihalcea and Pedersen, 2003). Within this framework, participating teams were provided with data and asked to provide automatically derived word alignments for all the words in the test set, following a specific format. This framework is defined to evaluate simple-word alignment algorithms, but we adapted it to evaluate our MWEs alignment system. The alignment results are compared to a manually aligned reference corpus scored with respect to precision, recall and F-measure, where A is the alignment proposed by the system and G is a gold standard alignment. Because the manual construction of the alignment reference is a difficult and time-consuming task, we conducted a small-scale evaluation based on a small set of 100 French-English aligned sentences derived from the Europarl corpus.

$$P = \frac{|A \cap G|}{|A|} \quad (2)$$

$$R = \frac{|A \cap G|}{|G|} \quad (3)$$

$$F = \frac{2P * R}{P + R} \quad (4)$$

Our method yeilds a precision of 63,93% , a recall of 62,46% and an F-measure of 63.19%. We consider that obtained results are satisfactory and encouraging. In table 3 we give an example of MWEs aligned by our technique.

French → English MWEs
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
cut in forestation / coupe forestier

Table 3: Sample of aligned MWEs

From observing some couples of MWEs, we have identified a class of error caused by the choice of n-gram's size. Since our system does not capture one-to-many correspondences, some MWEs were not aligned correctly. For example, the French MWE *chemin de fer* corresponding normally to the simpe word *railway* was aligned here by the MWE *railway sector*.

3 Experiments

3.1 Application of MWEs

In the previous section, we described the approach we followed to extract translation pairs of MWEs, and evaluated it by comparing the list of extracted MWEs to a hand-created reference list. As it lacks a common benchmark data sets for evaluation in MWE extraction and alignment researches, we decided to study in what respect these units are useful to improve the performance of phrase based SMT systems. We present a method that integrates extracted MWEs into the baseline system's phrase table being considered as very important element according to the following two ways. In the first way, we simply add MWEs and keep translation probabilities proposed by the aligner. We call this method "Baseline+MWE". In the second one, "Baseline+NPMWE", we assign 1 to the two translation probabilities (in both directions) for simplicity.

3.2 Baseline

We use the factored translation model of the Moses[2] SMT system as our baseline system (Koehn, 2005). It is an extention 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 annotations at the word level. We present a model that 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 baseline system are: two translation probability features, two languages models, one generation model and word penalty. For the "Baseline+MWE" and "Baseline+NPMWE" methods, translation pairs of MWEs were extracted from the training corpus and added to the phrase table. Consequently, a new phrase table is obtained. During the translation process, the decoder would search for each phrase in input sentence, all candidates translations in both original phrases and new MWEs.

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

3.3 Data

Training and Test data (Table 4) come from the French-English Europarl Corpus (Koehn, 2005). It groups a set of parallel sentences extracted from the Proceedings of the European Parliament. In this work, we focus on sentences consisting of at most 50 words.

	French	English
Training sentences	9002	
Words	213489	206562
Test sentences	500	
Words	13816	12736

Table 4: Caracteristics of Training and Test data

Since we use the factored translation model, training data are annotated with lemmas. Next, word-alignment for all the sentences in the parallel training corpus is established. Here, we use the same methodology as in phrase-based models (symmetrized GIZA++ alignments). The word alignment methods operates on lemmas. We also specified two language models using the IRST Language Modeling Toolkit [3] to train two tri-gram models. Besides the regular language model based on surface forms, we have a second language model which is trained on lemmas.

3.4 Results and discussion

We test translation quality on the test set described in the previous section and calculate the BLEU score. We also consider only one reference for each test sentence. Obtained BLEU results are reported in Table 5. The first notable observation is that using bilingual MWEs improves translation in the two cases. The "Baseline+MWE" method achieves the most improvement of 0.24 BLEU score compared to the baseline system. This method performs slightly higher than the "Baseline+NPMWE" method which in turn comes with 0.23 BLEU score improvement.

Method	BLEU
Baseline	0,1758
Baseline+MWE	**0,1782**
Baseline+NPMWE	0.1781

Table 5: Translation results using extracted MWEs

In order to know in what respects our method improves performance of translations, we manually analyzed the test sentence presented in Table 6. The french MWE "chemins de fer" is not correctly aligned in baseline system. It was translated to the english phrase "way of the

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

Source Sentence	Ce n'est que ces dernières années que la plupart des **états membres** ont investi dans l'amélioration des **chemins de fer** et parfois également dans la navigation intérieure.
Reference	Only in the last few years have most **member states** invested in improving the **railways** and sometimes inland shipping too.
Baseline	They will be that this last year that most **member states** have invested in improving the **way to go to fer** and sometimes also in the navigation internal.
Baseline+MWE	They will be that this last year that most **member states** have invested in improving the **railways sector** and sometimes also in the internal navigation.

Table 6: Translation example

fer". We can notice that in this case,a word-to-word alignment strategy is performed. It provides the following alignments:

- "chemin"="way to go to"
- "de"= Not Translated
- "fer"=Not translated

Here, the French word "chemin" was translated into the English phrase "way to go to" and the word "fer" was not translated since there is no entry in the baseline system's phrase table to which we can associate it. While it is aligned to the target MWE "railways sector" in baseline+MWE. We can consider that this is a correctly translated phrase as much as it keeps the same meaning.

4 Conclusion and Future Work

We described a method for extracting and aligning MWEs in a parallel corpus. The alignment algorithm we proposed checks only on many to many correspondences and can address both frequent and infrequent MWEs in a text. To evaluate the alignment quality, we used a small test set of 100 parallel sentences and reported an F-Measure value of 63,19%.

We also proposed a method for using extracted bilingual MWEs in Statistical Machine Translation. This method incorporates extracted MWEs in a baseline system's phrase table. Conducted experiments show that including such type of units in the translation process improves translation quality and yeilds an improvement of 0.24 BLEU score compared to a baseline system.

Although our initial experiments are positive, we believe that they can be improved in a number of ways. We fisrt intend to extend the morphosyntactic patterns to handle other forms of MWEs, e.g. starting with a verb. We will also try to develop and evaluate other statistical based methods to align MWEs.

Moreover, in the presented work the use of MWEs is actually restricted to the decoding step. We will also attempt to include these units in the training step using a larger set a parallel sentences and the two sides of MWEs as independent monolingual units.

Acknowledgments

This research work is supported by the FINANCIAL-WATCH (QNRF NPRP: 08-583-1-101) project.

References

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*. Proceedings of LREC, Valetta, Malta.

Brown P., Della Pietra S., Della Pietra V. and Mercer R. (1993). *The mathematics of statistical machine translation: Parameter estimation.*. Computational linguistics.

Daille B. (2001). *Extraction de collocation à partir de textes*. Proceedings of TALN, Tours, France.

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

Finlayson M. and Kulkarni N. (2011). *Detecting Multi-Word Expressions Improves Word Sense Disambiguation*. Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World. Portland, Oregon, USA. p.20-24

Frantzie C., Ananiadou S., and Mima H. (2000). *Automatic recognition of multi-word terms: the C-Value/NC-Value method*. Int. J. on Digital Libraries 3(2): 115-130.

Jackendoff R. (1997). *The Architecture of the Language Faculty*. Cambridge (Mass.), MIT Press.

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

Koehn P. (2005). *Europarl: A parallel Corpus for Statistical Machine Translation*. Proceeding of MT-SUMMIT

Koehn P and Hoang H. (2005). *Factored Translation Model*. Proceeding of MT-SUMMIT

Koehn P., Och F. and Marcu D. (2003). *Statistical Phrase-Based Translation*. Proceeding of the Human Language Technology Conference of the North American Chapter of

the Association for Computational Linguistics. Edmonton, Canada. p 115-124.

Lambert P. and Banchs R. 2005. *Data Inferred Multi-word Expressions for Statistical Machine Translation*. Proceeding of MT SUMMIT.

Lambert P. and Banchs R. 2006. *Grouping Multi-word Expressions According to Part-Of-Speech in statistical Machine Translation.*. Proceeding of the Workshop on Multi-word Expressions in a multilingual context.

Mihalcea R. and Pedersen T. (2003). *An evaluation exercise for Word Alignment*. Proceedings of the Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond. Edmonton, Canada. p. 1-10.

Okita T., Guerra M. Alfredo, Graham Y., and Way A. (2010). *Multi-Word Expression Sensitive Word Alignment*. Proceedings of the 4th International Workshop on Cross Lingual Information Access at COLING 2010, Beijing, p. 26–34.

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, p 47–54.

Sag I , Baldwin T., Francis Bond F, Copestake A, and Flickinger D. (2002). *Multiword Expressions:A Pain in the Neck for NLP*. CICLing 2002 Mexico City, Mexico.

Salton G. , Wong A. , and Yang C. S. (1975). *A Vector Space Model for Automatic Indexing*. Communications of the ACM, vol.18 p. 613–620.

Seretan V. and Wehrli E. (2007). *Collocation translation based on sentence alignment and parsing*. Proceedings of TALN. Toulouse, France.

Smadja F. (1993). *Retrieving collocations from text:Xtract*. Computational Linguistics. vol.19 p.143-177.

Smadja F., McKeown K., and Hatzivassiloglou V. (1996). *Translating collocations for bilingual lexicons: A Statistical Approach*. Computational Linguistics. p.1-38.

Tanaka T and Baldwin T. (2003). *Noun-noun compound machine translation: A feasibility study on shallow processing*. In Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment.

Tufis I. and Ion R. (2007). *Parallel corpora, alignment technologies and further prospects in multilingual resources and technology infrastructure*. Proceedings of the 4th International Conference on Speech and Dialogue Systems, Iaşi, Romania, p.183-195.

Vechtomova O. (2005). *The role of multi-word units in interactive information retrieval*. In D.E. Losada and J.M. Fernández-Luna, editors, ECIR 2005, LNCS 3408, p 403–420. Springer-Verlag, Berlin. alia Kordoni, Carlos Ramisch, Aline Villavicencio

Wu C. and Chang S. Jason. (2004). *Bilingual Collocation Extraction Based on Syntactic and Statistical Analyses*. Computational Linguistics. p.1-20.