Integration of statistical collocation segmentations in a phrase-based statistical machine translation system

Marta R. Costa-jussà*, Vidas Daudaravicius† and Rafael E. Banchs*
*Barcelona Media Research Center
Av Diagonal, 177, 9th floor, 08018 Barcelona, Spain
{marta.ruiz,rafael.banchs}@barcelonamedia.org
† Faculty of Informatics, Vytautas Magnus University
Vileikos 8, Kaunas, Lithuania
vidas@donelaitis.vdu.lt

Abstract

This study evaluates the impact of integrating two different collocation segmentations methods in a standard phrase-based statistical machine translation approach. The collocation segmentation techniques are implemented simultaneously in the source and target side. Each resulting collocation segmentation is used to extract translation units. Experiments are reported in the English-to-Spanish Bible task and promising results (an improvement over 0.7 BLEU absolute) are achieved in translation quality.

1 Introduction

Machine Translation (MT) investigates the use of computer software to translate text or speech from one language to another. Statistical machine translation (SMT) has become one of the most popular MT approaches given the combination of several factors. Among them, it is relatively straightforward to build an SMT system given the freely available software and, additionally, the system construction does not require of any language experts.

Nowadays, one of the most popular SMT approaches is the phrase-based system (Koehn et al., 2003) which implements a maximum entropy approach based on a combination of feature functions. The Moses system (Koehn et al., 2007) is an implementation of this phrase-based machine translation approach. An input sentence is first split into sequences of words (so-called phrases), which are then mapped one-to-one to target phrases using a large phrase translation table.

Introducing chunking in the standard phrase-based SMT system is a relatively frequent study (Zhou et al., 2004; Wang et al., 2002; Ma et al., 2007). Chunking may be used either to improve reordering or to enhance the translation table. For example, authors in (Zhang et al., 2007) present a shallow chunking based on syntactic information and they use the chunks to reorder phrases. Other studies report the impact on the quality of word alignment and in translation after using various types of multi-word expressions which can be regarded as a type of chunks, see (Lambert and Banchs, 2006) or sub-sentential sequences (Macken et al., 2008; Groves and Way, 2005). Chunking is usually performed on a syntactic or semantic basis which forces to have a tool for parsing or similar. We propose to introduce the collocation segmentation developed by (Daudaravicius, 2009) which is language independent. This collocation segmentation was applied in keyword assigment task and a high classification improvement was achieved (Daudaravicius, 2010).

We use this collocation segmentation technique to enrich the phrase translation table. The phrase translation table is composed of phrase units which generally are extracted from a word aligned parallel corpus. Given this word alignment, an extraction of contiguous phrases is carried out (Zens et al., 2002), specifically all extracted phrases fulfill the following restrictions: all source (target) words within a phrase are aligned only to target (source) words within the same phrase.

This paper is organized as follows. First, we detail the different collocation segmentation techniques proposed. Secondly, we make a brief description of the phrase-based SMT system and how we introduce the collocation segmentation to improve the phrase-based SMT system. Then, we present experiments performed in an standard phrase-based system comparing the phrase extraction. Finally, we present the conclusions.

2 Collocation segmentation

The Dice score is used to measure the association strength of two words. This score is used, for instance, in the collocation compiler XTract (Smadja, 1993) and in the lexicon extraction system Champollion (Smadja and Hatzivassiloglou, 1996). Dice is defined as follows:

$$Dice(x;y) = \frac{2f(x,y)}{f(x)+f(y)}$$

where $f(x,y)$ is the frequency of co-occurrence of x and y, and $f(x)$ and $f(y)$ the frequencies of occurrence of x and y anywhere in the text. If x and y tend to occur in conjunction, their Dice score will be high. The text is seen as a changing curve of the word associativity values (see Figure 1 and Figure 2).

The collocation segmentation is the process of detecting the boundaries of collocation segments within a text. A collocation segment is a piece of a text between boundaries. The boundaries are set in two steps. First, we set the boundary between two words within a text where the Dice value is lower than a threshold. The threshold value is set manually and is kept at the Dice value of *exp(-8)* in our experiment *CS-1* (i.e. Collocation Segmentation type 1), and the Dice value of *exp(-4)* in our experiment *CS-2* (i.e. Collocation Segmentation type 2). This decision was based on the shape of the curve found in (Daudaravicius and Marcinkeviciene, 2004). The threshold for CS-1 is kept very low, and many weak word associations are considered. The threshold for CS-2 is high to keep together only strongly connected words. The higher threshold value makes shorter collocation segments. Shorter collocation segments are more confident collocations and we may expect better transaltion results. Nevertheless, the results of our study show that longer collocation segments are preferable. Second, we introduce an average minimum law (AML). The average minimum law is applied to the three adjacent Dice values (i.e., four words). The law is expressed as follows:

$$\frac{Dice(x_{i-2},x_{i-1})+Dice(x_i,x_{i+1})}{2} >$$

$$Dice(x_{i-1},x_i) \longrightarrow x_{i-1} boundary x_i$$

The boundary of a segment is set at the point, where the value of collocability is lower than the average of preceding and following values of collocability. The example of setting the boundaries for English sentence is presented in Figure 1, and it shows a sentence and Dice values between word pairs. Almost all values are higher than an arbitrary chosen level of the threshold. Most of the boundaries in the example sentence are made by the use of the average minimum law. This law identifies segment or collocation boundaries by the change of Dice value. This approach is new and different from other widely used statistical methods (Tjong-Kim-Sang and S., 2000). For instance, the general method used by Choueka (Choueka, 1988) is the following: for each length n, $(1 \leq n \leq 6)$, produce all the word sequences of length n and sort them by frequency; impose a threshold frequency 14. Xtract is designed to extract significant bigrams, and then expands 2-Grams to n-Grams (Smadja, 1993). Lin (Lin, 1998) extends the collocation extraction methods with syntactic dependency triples. Such collocation extraction methods are performed on a dictionary level. The result of this process is a dictionary of collocations. Our collocation segmentation is performed within a text and the result of this process is a segmented text (see Figure 3).

The segmented text could be used later to create a dictionary of collocations. Such dictionary accepts all collocation segments. The main difference from Choueka and Smadja methods is that our proposed method accepts all collocations and no significance tests for collocations are performed. The main advantage of this segmentation is the ability to perform collocation segmentation using plain corpora only, and no manually segmented corpora or other databases and language processing tools are required. Thus, this approach could be used successfully in many NLP tasks such as statistical machine translation, information extraction, information retrieval and etc.

The disadvantage of collocation segmentation is that the segments do not always conform to the correct grammatical and lexical phrases. E.g., in Figure 1 an appropriate segmenation of the consecutive set of words *on the seventh day* would give segments *on* and *the seventh day*. But the collocation segmentation takes *on the* and *seventh day* segmentation. This happens because we have no extra information about structure of grammatical

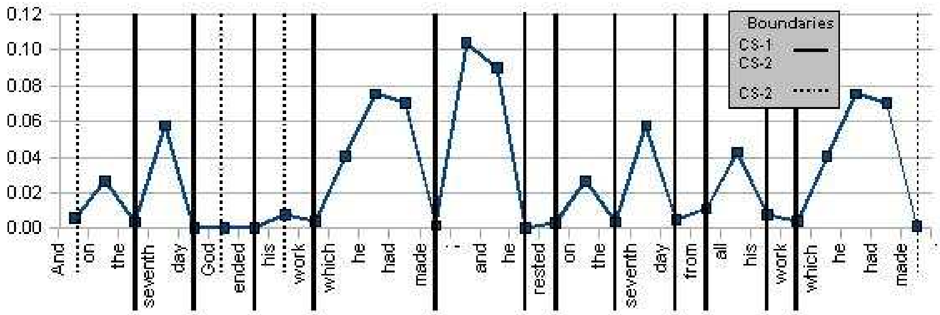

Figure 1: The segment boundaries of the English Sentence.

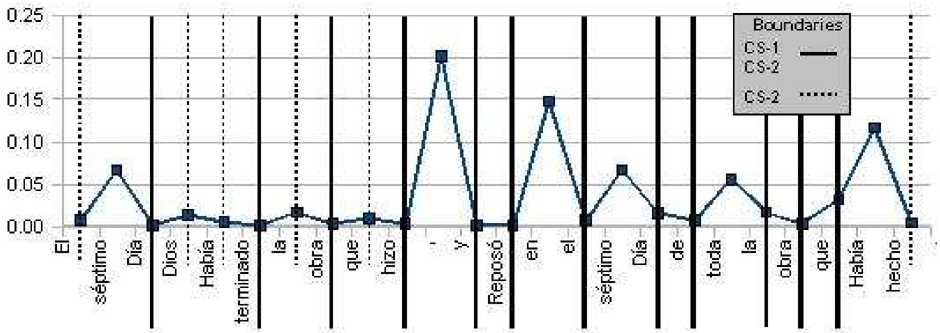

Figure 2: The segment boundaries of the Spanish Sentence.

phsases. On the other hand, it is important to notice that the collocation segmentation of the same translated text is similar for different languages, even if a word or phrase order is different (Daudaravicius, 2010). Therefore, even if collocation segments are not grammatically well formed, the collocation segments are more or less symetrical for different languages. The same sentence from Bible corpus is segmented and the result is shown in Figures 1 and 2. As future work, it is necessary to make a thorough evaluation of conformity of the proposed collocation segmentation method to phrase-based segmentation by using parsers.

3 Phrase-based SMT system

The basic idea of phrase-based translation is to segment the given source sentence into units (hereafter called phrases), then translate each phrase and finally compose the target sentence from these phrase translations.

Basically, a bilingual phrase is a pair of m source words and n target words. For extraction from a bilingual word aligned training corpus, two additional constraints are considered: words are consecutive, and, they are consistent with the word alignment matrix.

Given the collected phrase pairs, the phrase translation probability distribution is commonly estimated by relative frequency in both directions.

The translation model is combined together with the following six additional feature functions: the target language model, the word and the phrase bonus and the source-to-target and target-to-source lexicon model and the reordering model. These models are optimized in the decoder following the procedure described in *http://www.statmt.org/jhuws/*.

4 Integration of the collocation segmentation in the phrase-based SMT system

The collocation segmentation provides a new segmentation of the data. One straightforward approach is to use the collocation segments as words, and to build a new phrase-based SMT system from scratch. Therefore, phrases are composed from collocation segments. However, we have tested that this approach does not yield to better results. The reason for worse results could be the insufficient amount of data to build a transaltion table with reliable statistics. The collocation segmentaton increases the size of a dictionary more than 5 times (Daudaravicius, 2010), and we need a sufficient size corpus to get better results than base

In_the beginning **God_created** the_heaven **and_the** earth **.**
And_the earth **was** without_form **,_and** void **;_and** darkness_was **upon_the** face **of_the** deep **.** And_the **Spirit** of_God **moved** upon_the **face** of_the **waters** .
And_God said_, **Let** there_be **light** :_and **there_was** light **.**
And_God_saw the_light **,_that** it_was **good** :_and **God** divided **the_light** from_the **darkness** .
And_God called **the_light** Day **,_and** the_darkness **he_called** Night **.** And_the_evening **and_the** morning **were** the **first_day** .
And_God said_, **Let** there_be_a **firmament** in_the **midst** of_the **waters** ,_and **let_it** divide **the_waters** from_the **waters** .
And_God_made the_firmament **,_and** divided_the_waters **which_were** under_the **firmament** from_the **waters** which_were **above_the_firmament** :_and **it_was** so **.**
And_God called_the **firmament_Heaven** . **And_the_evening** and_the **morning** were **the** second_day **.**

Figure 3: The collocation segmentation of the begining of the Bible.

line. But the size of parallel corpora is limited by the number of texts we are able to gather. Therefore, we propose to integrate collocation segments into standard SMT. Instead of building a new SMT system from scrach, we enrich the base SMT with collocaton segments.

In this work, we integrate the collocation-segmentation as follows.

1. First, we build a baseline phrase-based system which is computed as reported in the section above.

2. Second, we build a collocation-based system which uses collocation segments as words. The main difference of this system is that phrases are composed of collocations instead of words.

3. Third, we convert the set of collocation-based phrases (which was computed in step 2) into a set of phrases composed by words. For example, given the collocation-based phrase *in_the_sight_of* ||| *delante*, it is converted into the phrase *in the sight of* ||| *delante*.

4. Fourth, we consider the union of the baseline phrase-based extracted phrases (computed in step 1) and the collocation-based extracted phrases (computed in step 2 and modified in step 3). That is, the set of standard phrases is combined with the set of modified collocation-phrases.

5. Finally, the phrase translation table is computed over the concatenated set of extracted phrases. This phrase table contains the standard phrase-based models which were named in section 3: relative frequencies, lexical probabilities and phrase bonus. Notice that some pairs of phrases can be generated in both extractions. Then this phrases will have a higher score when computing the relative frequencies. The IBM probabilities are computed at the level of words.

Hereinafter, this approach will be referred to as concatenate-based approach (*CONCAT*). Figure 4 shows an example of phrase extraction.

The goal of the integration of the collocations segmentation into the base SMT system is to introduce new phrases into translation table and smoothing of the relative frequencies of the translation phrases which appear in both segmentations. Additionally, the concatenation of two translation tables gives the possibility to highlight those translation phrases that are recognized in both translation tables. Therefore, this allows to 'vote' for the better translation phrases adding a new feature function which is '1' in case of appearing in both segmentations or '0' in the opposite case.

5 Experimental framework

The phrase-based system used in this paper is based on the well-known MOSES toolkit, which is nowadays considered as a state-of-the-art SMT system (Koehn et al., 2007). The training and weights tuning procedures are explained in details in the above-mentioned publication, as well as, on the MOSES web page: *http://www.statmt.org/moses/*.

5.1 Corpus statistics

Experiments were carried out on the English to Spanish Bible task, which have been proven to be a valid NLP resource (Chew et al., 2006). The main advantages of using this corpus are that it is the world's most translated book, with translations in over 2,100 languages (often, multiple translations per language) and easy availability, often in electronic form and in the public domain; it covers a variety of literary styles including narrative, poetry, and correspondence; great care is taken over the translations; it has a standard structure which

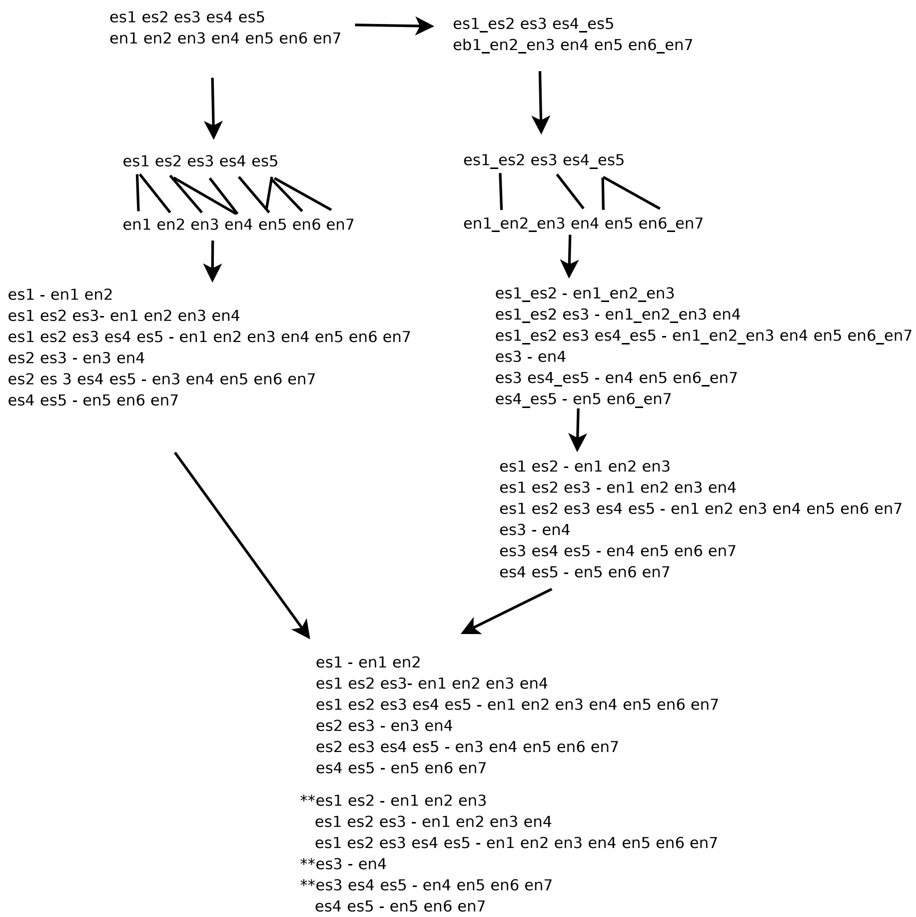

Figure 4: Example of the phrase extraction process in the *CONCAT* approach. New phrases added by the collocation-based system are marked with a $**$.

allows parallel alignment on a verse-by-verse basis; and, perhaps surprisingly, its vocabulary appears to have a high rate of coverage (as much as 85%) of modern-day language. The Bible is small compared to many corpora currently used in computational linguistics research, but still falls within the range of acceptability based on the fact that other corpora of similar size are used (see IWSLT International Evaluation Campaign [1]).

Table 1 shows the main statistics of the data used, namely the number of sentences, words and vocabulary, for each language.

5.2 Collocation Segment statistics

Here we analyse the collocation segment statistics. Table 2 shows the number of tokens and types of collocation segments. We see that the number of types of collocation segments is around 6 times higher than the number of types of words. The increase is different for Spanish and English. The *CS-1* segmentation increased the number of types for Spanish training set by 4 times, and for English by 6.5 times. Therefore, the dictionaries for Spanish and English become comparable in size. This allows to expect better alignment, and that is indeed in our experiments. The *CS-2* segmentation increased the number of types for Spanish train-

	Spanish	*English*
Training Sentences	28,887	28,887
Tokens	781,113	848,776
Types	28,178	13,126
Development Sentences	500	500
Tokens	13,312	14,562
Types	2,879	2,156
Test Sentences	500	500
Tokens	13,170	14,537
Types	2,862	2,095

Table 1: *Bible corpus: training, development and test data sets.*

[1] http://mastarpj.nict.go.jp/IWSLT2009/

	Spanish	*English*
Training Sentences	28,887	28,887
Tokens CS-1	407,505	456,608
Types CS-1	109,521	84,789
Tokens CS-2	524,916	549,585
Types CS-2	57,893	37,030

Table 2: *Tokens and types of collocation segments.*

ing set by 2 times, and for English by 2.8 times. The dictionaries are still comparably different in size. In section 4.5 we show that *CS-1* segmentation provides the best results. This result may indicate initial number of types before alignment is an important feature. The number of types should be comparable in order to achieve the best alignment, and the best translation results afterward. This may explain why *CS-1* segmentation contributes to obtain higher quality translations than *CS-2* segmentation, as will be shown in Section 4.5.

5.3 Experimental systems

We build four different systems: the phrase-based (*PB*), with two different phrase length limits, and the concatenate-based (*CONCAT*) SMT system, which has two versions: one for each type of segmentation presented above.

Phrase length is understood as the maximum number of words either in the source or the target part. In our experiments, the *CONCAT* systems catenated the baseline system which used phrases up to 10 words together with the units coming from the collocation segmentation which was limited to 10. This collocation segmentation limitation allowed for translation units of a maximum of 20 words. In order to make a fair comparison, we used two baseline systems, one with a maximum of 10 words (*PB-10*) and another of maximum of 20 words (*PB-20*) per translation unit.

5.4 Translation units analysis

This section analyses the translation units that were used in the test set (i.e. the highest scoring translation units found by the decoder).

Adding more phrases (in the *PB-20* system) without any selection leads to a phrase table of 7M translation units, whereas using our *CONCAT-1* proposal the phrase table contains 4.6M translation units and in the *CONCAT-2*, the phrase table contains 5.3M translation units. That means a 35% reduction of the total translation unit vocabulary.

Table 3 shows average and maximum length of the translation units used in the test set. The collocation segmentation influences the length of translation phrases. Neither the *CONCAT-1* nor *CONCAT-2* approach does not use longer phrases in average. In fact, the segmentation reduces the average length of the translation unit. This result may be surprising, because a segmentation which uses chunks instead of words may be expected to increase the average length of the translation units. In the next section, we will see that using longer phrases do not improve the translation. Notice that the literature showed that using longer phrases do not provide better translation (Koehn et al., 2003).

5.5 Automatic translation evaluation

The translation performance of the four experimental systems is evaluated and shown in Table 4.

In fact, an indirect composition of phrases with the help of the segmentation allows to get better results than a straightforward composition of translation phrases from single words. However, adding phrases using the standard algorithm can lead to slightly worse translations (Koehn et al., 2003).

The best translation results were achieved by integrating collocation segmentation 1, which uses longer collocation segments, into the SMT system. This result shows that shorter collocations, i.e. more confident collocations, do not improve results. This could be due to ability of the base SMT system to capture collocations in the similar way as the collocation segmentation 2 does. The collocation segmentation 1 introduces longer collocation that the base SMT system is not able to capture. Thus, longer collocations improves base SMT system better than shorter collocations.

The results show that the higher average of the length of translation phrases do not necessarly lead to better translations (see table 3). The improvement of translation quality (when using the collocation segmentation) may indicate that short phrases coming from the collocation segmentation have a better association between words and lead to a better translation. It is difficult to make a conclusion about the importance of the measure of the average length of the phrase in the translation table. Therefore, the average phrase length measure alone is not a reliable feature, and does not give important information and could cheat the conlusions. This is clearly seen in our results: the BLEU score of PB–10 and CONCAT-2 are very close,

	PB-10	*PB-20*	*CONCAT-1*	*CONCAT-2*
Source phrase average length	2.51	2.56	2.36	2.27
Source phrase maximum length	10	20	10	16
Target phrase average length	2.32	2.34	2.13	2.05
Target phrase maximum length	10	20	10	10

Table 3: *Translation unit length statistics used in the test set.*

but the average length of phrases are too different, and appear in the oposite sides of the CONCAT-1 value. Futher studies could show what features could be used to describe the quality of the translation dictionary.

Collocation segmentation is capable to introduce new translation units that are useful in the final translation system and to smooth the relative frequencies of those units which were already in the baseline translation table. The improvement is almost of +0.6 point BLEU in the test set. Further experiments could be dedicated to investigate the separate improvement due to (1) new translation units or (2) smoothing (in case they give independent gains). From now on, the comparison is made with the best baseline (*PB-10*) system and the best *CONCAT* (*CONCAT-1*) system, which obtained the best results in the automatic evaluation.

We found out that a certain number of sentences produced the same output with different segmentation. When comparing the best *CONCAT* with the best baseline (*PB-10*) systems' outputs, 165 sentences produced the same output (in most cases with different segmentation). The last row in table 4 shows BLEU when evaluating only the sentences which were different (Subset-Test, 335 sentences). In this case, the BLEU improvement reaches +0.75.

5.6 Translation analysis

We performed a manual analysis of the translation. We compared 100 output sentences from the baseline and the *CONCAT* system.

No significant advantages of the baseline system was tracked, whereas the collocation segmentation allows to improve translation quality in the following ways (only sentence subsegments are shown):

1. Not removal of words.

   Bas: llamó su nombre Noé :
   +CS: llamó su nombre Noé , diciendo:
   REF: llamó su nombre Noé , diciendo:

2. Better choice of prepositions.

   Bas: declarará por juramento
   +CS: declarará bajo juramento
   REF: declarará bajo juramento

3. Better choice of translation units.

   Bas: . ||| ;
   +CS: . ||| .
   REF: .

4. Better preservation of idiomacity.

   Bas: podrás comer pan
   +CS: comerás pan
   REF: comerás pan

5. Better selection of a phrase structure.

   Bas: cuando él conoce
   +CS: cuando él llegue a saberlo
   REF: cuando él llegue a saberlo

6 Conclusions and further research

This work explored the feasibility for improving a standard phrase-based statistical machine translation system by using a novel collocation segmentation method for translation unit extraction. Experiments were carried out with the English-to-Spanish Bible corpus task. A small but significant gain in translation BLEU was obtained when combining these units with the standard set of phrases.

Future research in this area is envisioned in the following main directions: to study how the collocations learned on the Bible corpus differ from those learned on more general corpora; to improve collocation segmentation quality in order to obtain more human-like translation unit segmentations; to explore the use of a specific feature function for helping the translation systems to select translation units from both categories (collocation segments and conventional phrases) according to their relative importance at each decoding step; and to evaluate the impact of new translation units vs. smoothing.

	PB-10	*PB-20*	*CONCAT-1*	*CONCAT-2*
Test	35.68	35.60	**36.28**	35.82
Subset-Test	33.65	–	**34.40**	–

Table 4: *Translation results in terms of BLEU.*

7 Acknowledgements

This work has been partially funded by the Spanish Department of Education and Science through the *Juan de la Cierva* fellowship program and the BUCEADOR project (TEC2009-14094-C04-01). The authors also wants to thank the Barcelona Media Innovation Centre for its support and permission to publish this research.

References

Chew, P. A, S. J Verzi, T. L Bauer, and J. T McClain. 2006. Evaluation of the bible as a resource for cross-language information retrieval. In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, pages 68–74.

Choueka, Y. 1988. Looking for needles in a haystack, or locating interesting collocational expressions in large textual databases. In *Proceedings of the RIAO Conference on User-Oriented Content-Based Text and Image Handling*, pages 21–24, Cambridge, MA.

Daudaravicius, V. and R Marcinkeviciene. 2004. Gravity counts for the boundaries of collocations. *International Journal of Corpus Linguistics*, 9(2):321–348.

Daudaravicius, V. 2009. Automatic identification of lexical units. *An international Journal of Computing and Informatics. Special Issue Computational Linguistics*.

Daudaravicius, Vidas. 2010. The influence of collocation segmentation and top 10 items to keyword assignment performance. In *11th International Conference on Intelligent Text Processing and Computational Linguistics, Springer Verlag, LNCS*, page 12, Iasi, Romania.

Groves, D. and A. Way. 2005. Hybrid data-driven models of machine translation. *Machine Translation*, 19(3):301323.

Koehn, P., F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the HLT-NAACL*, pages 48–54, Edmonton.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL*, pages 177–180, Prague, Czech Republic, June.

Lambert, P. and R. Banchs. 2006. Grouping multi-word expressions according to part-of-speech in statistical machine translation. In *Proceedings of the EACL*, pages 9–16, Trento.

Lin, D. 1998. Extracting collocations from text corpora. In *First Workshop on Computational Terminology*, Montreal.

Ma, Y., N. Stroppa, and A. Way. 2007. Alignment-guided chunking. In *Proc. of TMI 2007*, pages 114–121, Skvde, Sweden.

Macken, L., E. Lefever, and V. Hoste. 2008. Linguistically-based sub-sentential alignment for terminology extractionfrom a bilingual automotive corpus. In *Proceedings of COLING*, pages 529–536, Machester.

Smadja, F.and McKeown, K. R. and V. Hatzivassiloglou. 1996. Translation collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38.

Smadja, F. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

Tjong-Kim-Sang, E. and Buchholz S. 2000. Introduction to the conll-2000 shared task: Chunking. In *Proc. of CoNLL-2000 and LLL-2000*, pages 127–132, Lisbon, Portugal.

Wang, W., J. Huang, M. Zhou, and C. Huang. 2002. Structure alignment using bilingual chunks. In *Proc. of COLING 2002*, Taipei.

Zens, R., F.J. Och, and H. Ney. 2002. Phrase-based statistical machine translation. In Jarke, M., J. Koehler, and G. Lakemeyer, editors, *KI - 2002: Advances in artificial intelligence*, volume LNAI 2479, pages 18–32. Springer Verlag, September.

Zhang, Y., R. Zens, and H. Ney. 2007. Chunk-level reordering of source language sentences with automatically learned rules for statistical machine translation. In *Proc. of the Human Language Technology Conf. (HLT-NAACL'06):Proc. of the Workshop on Syntax and Structure in Statistical Translation (SSST)*, pages 1–8, Rochester, April.

Zhou, Y., C. Zong, and X. Bo. 2004. Bilingual chunk alignment in statistical machine translation. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 2, pages 1401–1406, Hague.