Using collocation segmentation to extract translation units in a phrase-based statistical machine translation system

Implementación de una segmentación estadística complementaria para extraer unidades de traducción en un sistema de traducción estadístico basado en frases

Marta R. Costa-jussà*, Vidas Daudaravicius† and Rafael E. Banchs*
*Barcelona Media Innovation 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

Resumen: Este artículo evalúa un nuevo método de segmentación en un sistema de traducción automática estadístico basado en frases. La técnica de segmentación se implementa tanto en la parte fuente como en la parte destino y se usa para extraer unidades de traducción. Los resultados mejoran el sistema de referencia en la tarea español-inglés del EuroParl.
Palabras clave: Traducción automática, segmentación

Abstract: This report evaluates the impact of using a novel collocation segmentation method for phrase extraction in the standard phrase-based statistical machine translation approach. The collocation segmentation technique is implemented simultaneously in the source and target side. The resulting collocation segmentation is used to extract translation units. Experiments are reported in the Spanish-to-English EuroParl task and promising results are achieved in translation quality.
Keywords: Machine translation, collocation segmentation

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 approach [Zhou et al.2004, Wang et al.2002, Ma et al.2007]. Several works use chunks for reordering purposes. 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 Banchs2006] or sub-sentential sequences [Macken et al.2008].

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 [Daudaravicius2009] which can be applied to any language.

Our idea is to introduce this collocation segmentation technique to further improve

the phrase translation table. The phrase translation table is composed of phrase units which generally are extracted from a word aligned parallel corpus.

This paper is organized as follows. First, we detail the collocation segmentation technique. 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

A text segment is a single word or a sequence of words. The Dice word associativity score is used to calculate the associativity of words and to produce a discrete signal of a text. This score is used, for instance, in the collocation compiler XTract [Smadja1993] and in the lexicon extraction system Champollion [Smadja et al.1996]. Dice is defined as follows [Smadja1993]:

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

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 Dice score is not sensitive to the corpus size and the level of collocability does not change while the corpus size is changing. The logarithm of Dice is used in order to discern small numbers [Daudaravicius and Marcinkeviciene2004].
The text is seen as a changing curve of the word associativity values.

A collocation segment is a piece of a text between boundaries and the segmentation is done by detecing the boundaries of collocation segments in a text. First, the boundary in a text is the point where the associativity score is lower than an arbitrarily chosen level of collocability. The associativity value above the level of collocability conjoin two words. Human experts have to set the level of collocability manually. We set the level of collocability at the Dice minus 8 in our experiment. This decision was based on the shape of the curve found in [Daudaravicius and Marcinkeviciene2004].

Second, we use an additional definition of the boundary, which is called as an average minimum law. The average minimum law is applied to the three adjacent collocability points. The law is expressed as follows [Daudaravicius2010]:

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

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

The boundary of a segment is set at the point, where the value of collocability is lower the average of preceding and following values of collocability. The example of setting the boundaries in English and Spanish sentence is presented in Figure 1 and 2, respectively.

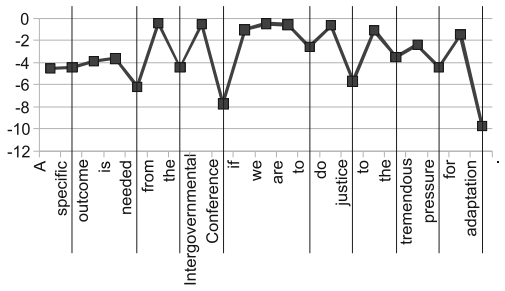

Figure 1: The segment boundaries of the English Sentence.

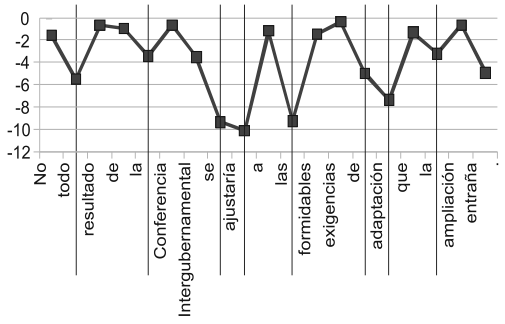

Figure 2: The segment boundaries of the Spanish Sentence.

The examples show a sentence and the logarithm of Dice values between word pairs. Almost all values are higher than an arbitrary chosen level of collocability. The boundaries in the example sentence is made by the use of the average minimum law. This law identifies segment or collocation boundaries by the change of collocability value [Tjong Kim Sang and S.2000]. The

main advantage of this segmentation is the ability to learn collocation segmentation using plain corpora and no manually segmented corpora or other databases and language processing tools are required. On the other hand, the disadvantage is that the segments do not always conform to correct grammatical phrases such as noun, verb or other phrases. Surprisingly, the collocation segments are similar for different languages even if word or phrase order is different, and can be easily aligned. An example of one segmented sentence translated into 21 official EU languages could be found in [Daudaravicius2010].

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. Given a 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 and words are consecutive.

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 feature models: the target language model, the word and the phrase bonus and the source-to-target and target-to-source lexical weights and the reordering model [Koehn et al.2003]. These models are optimized in the decoder following the minimum error rate procedure [Och2003].

The collocation segmentation provides a new segmentation of the data. As follows, we propose two techniques to integrate this collocation segmentation in an SMT system.

3.1 Collocation-based SMT system

One straightforward approach is to use the new segmentation to build from scratch a new phrase-based SMT system. This approach uses collocation segments instead of words. Therefore, phrases are sequences of collocation segments instead of words.

Hereinafter, this approach will be referred to as collocation-based approach (*CB*).

3.2 Integration of the collocation segmentation in a phrase-based SMT system

Another approach is to combine the phrases from the standard phrase-based approach together with the phrases from the collocation-based approach.

1. We build a baseline phrase-based system which is computed as reported in the section above.
2. We build a collocation-based system which instead of using words, uses collocations. The main difference of this system is that phrases are composed of collocations instead of words.
3. 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. 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 list of standard phrases is concatenated with the list of modified collocation-phrases.
5. Finally, the phrase translation table is computed over the concatenated set of extracted phrases. Notice that some pairs of phrases can be generated in both extractions. Then these phrases will have a higher score when computing the relative frequencies. Regarding the lexical weights, they are computed at the level of words.

Hereinafter, this approach will be referred to as concatenate-based approach (*CONCAT*).

4 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

EuroParl	*Spanish*	*English*
Training Sentences	727.1 k	727.1 k
Words	15.7 M	15.2 M
Vocabulary	108.7 k	72.3 k
Development Sentences	2000	2000
Words	60.6k	58.6k
Vocabulary	8.2k	6.5k
Test Sentences	2000	2000
Words	60.3k	57.9k
Vocabulary	8.3k	6.5k

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

EuroParl	*Spanish*	*English*
Training Sentences	727.1 k	727.1 k
Collocation Segments	8.4M	8.1M
Vocabulary	975.8 k	863.1 k

Table 2: *Running collocation segments and vocabulary.*

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/*. Note that we limit the length of the phrase (maximum number of words or segments in the source or in the target part) to 7 in all cases. The language model was built using the SRILM toolkit [Stolcke2002] using 5-grams and kneser-ney smoothing.

4.1 Corpus statistics

Experiments were carried out on the Spanish and English task of the WMT06 evaluation[1] (EuroParl corpus). It is a relatively large corpus. Table 1 shows the main statistics of the data used, namely the number of sentences, words and vocabulary, for each language.

4.2 Collocation Segment and phrase statistics

Here we analyse the collocation segment and phrase statistics. First, Table 2 shows the number of running collocation segments and vocabulary. We see that the vocabulary of collocation segments is around 10 times higher than the vocabulary of words.

Secondly, Figure 3 shows the quantity of phrases given the maximum number of words in the source or in the target side (which is considered the length of the phrase). We observe that the number of phrases of one word length is much larger in the standard set than in the segmentation set. However, within the segment set, we have a number of longer phrases than seven words. This happens because the limit of the phrase length is set to 7 segments in the segment-based set, and a segment may contain more than one word. In this case, we see that we have translation units which are longer. The quality of these longer translation units will determine the improvement in translation quality when using the concatenated system.

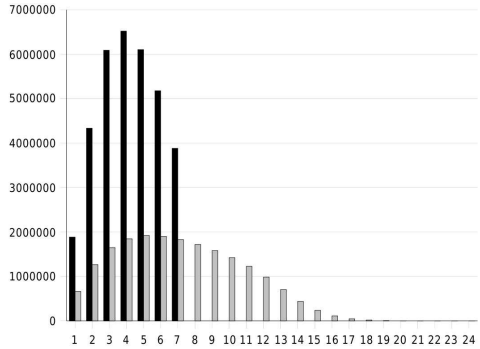

Figure 3: *Distribution of phrases according to the number of words in the source side for both, the phrase-based, PB, (in dark grey) and the collocation-based, CB, (in light grey) sets.*

4.3 Translation results

Finally, we build the three systems: the phrase-based (PB), the collocation-based (CB) and the concatenate-based (CONCAT) SMT systems. The translation performance is evaluated and shown in Table 3. Results show that the best performing system is the concatenate-based SMT system which uses both standard phrases and collocation-phrases.

[Koehn et al.2003] states that limiting the length to a maximum of only three words per phrase achieves top performance and learning longer phrases does not yield much improvement, and occasionally leads to worse results. Our approach provides an indirect composition of phrases with the help of the segmentation and this allows to get better results than a straightforward composition of translation phrases from single words. Our approach is not comparable to just composing longer phrases from single words. The fact of just increasing the length of phrases from single words would make the translation table increase a lot and would make the

[1] www.statmt.org/wmt06/shared-task/

EuroParl	*PB*	*CB*	*CONCAT*
Development	31.16	22.73	**32.32**
Test	30.85	21.74	**31.24**

Table 3: *Translation results in terms of BLEU.*

translation inefficient.

The segmentation allows to improve translation quality in following ways: the segmentation (1) introduces new translation phrases, and (2) smooths the relative frequencies. Collocation segmentation is capable to introduce new translation units that are useful in the final translation system. The improvement is over 1 point BLEU in the development set and almost of 0.4 point BLEU in the test set. The conclusion is that caring of strongly monolingual connected words can reduce the alignment noise, and improve translation dictionary quality.

5 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 Spanish-to-English EuroParl corpus task. Although the use of statistical collocation segmented translation units alone strongly deteriorates the system performance, 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 two main directions: first, to improve collocation segmentation quality in order to obtain more human-like translation unit segmentations; and, second, to explore the use of specific feature functions to select translation units from either collocation segments or conventional phrases according to their relative importance.

Acknowledgements

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

References

[Daudaravicius and Marcinkeviciene2004] V. Daudaravicius and R Marcinkeviciene. 2004. Gravity counts for the boundaries of collocations. *International Journal of Corpus Linguistics*, 9(2):321–348.

[Daudaravicius2009] V. Daudaravicius. 2009. Automatic identification of lexical units. *An international Journal of Computing and Informatics. Special Issue Computational Linguistics.*

[Daudaravicius2010] V. Daudaravicius. 2010. The influence of collocation segmentation and top 10 items to keyword assignment performance. In *CICLing*, pages 648–660.

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

[Koehn et al.2007] P. Koehn, 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 and Banchs2006] P. Lambert 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.

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

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

[Och2003] F.J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the ACL*, pages 160–167, Sapporo, July.

[Smadja et al.1996] F. Smadja, K. R. McKeown, and V. Hatzivassiloglou. 1996. Translation collocations for bilingual lexicons: A

statistical approach. *Computational Linguistics*, 22(1):1–38.

[Smadja1993] F. Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

[Stolcke2002] A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the ICSLP*, pages 901–904, Denver, USA, September.

[Tjong Kim Sang and S.2000] E. F. Tjong Kim Sang 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 et al.2002] W. Wang, J. Huang, M. Zhou, and C. Huang. 2002. Structure alignment using bilingual chunks. In *Proc. of COLING 2002*, Taipei.

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

[Zhang et al.2007] Y. Zhang, 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 et al.2004] Y. Zhou, 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.