Using collocation segmentation to augment the phrase table

Carlos A. Henríquez Q.*, **Marta R. Costa-jussà**†, **Vidas Daudaravicius**‡

Rafael E. Banchs†, **José B. Mariño***

*TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain
{carlos.henriquez,jose.marino}@upc.edu
†Barcelona Media Innovation Center, Barcelona, Spain
{marta.ruiz,rafael.banchs}@barcelonamedia.org
‡Faculty of Informatics, Vytautas Magnus University, Kaunas, Lithuania
vidas@donelaitis.vdu.lt

Abstract

This paper describes the 2010 phrase-based statistical machine translation system developed at the TALP Research Center of the UPC[1] in cooperation with BMIC[2] and VMU[3]. In phrase-based SMT, the phrase table is the main tool in translation. It is created extracting phrases from an aligned parallel corpus and then computing translation model scores with them. Performing a collocation segmentation over the source and target corpus before the alignment causes that different and larger phrases are extracted from the same original documents. We performed this segmentation and used the union of this phrase set with the phrase set extracted from the non-segmented corpus to compute the phrase table. We present the configurations considered and also report results obtained with internal and official test sets.

1 Introduction

The TALP Research Center of the UPC[1] in cooperation with BMIC[2] and VMU[3] participated in the Spanish-to-English WMT task. Our primary submission was a phrase-based SMT system enhanced with POS tags and our contrastive submission was an *augmented* phrase-based system using collocation segmentation (Costa-jussà et al., 2010), which mainly is a way of introducing new phrases in the translation table. This paper presents the description of both systems together with the results that we obtained in the evaluation task and is organized as follows: first, Section 2 and 3 present a brief description of a phrase-based SMT, followed by a general explanation of collocation segmentation. Section 4 presents the experimental framework, corpus used and a description of the different systems built for the translation task; the section ends showing the results we obtained over the official test set. Finally, section 5 presents the conclusions obtained from the experiments.

2 Phrase-based SMT

This approach to SMT performs the translation splitting the source sentence in segments and assigning to each segment a bilingual phrase from a phrase-table. Bilingual phrases are translation units that contain source words and target words, e.g. $<$ *unidad de traducción* $|$ *translation unit* $>$, and have different scores associated to them. These bilingual phrases are then sorted in order to maximize a linear combination of feature functions. Such strategy is known as the log-linear model (Och and Ney, 2003) and it is formally defined as:

$$\hat{e} = \arg\max_{e} \left[\sum_{m=1}^{M} \lambda_m h_m (e, f) \right] \tag{1}$$

where h_m are different feature functions with weights λ_m. The two main feature functions are the translation model (TM) and the target language model (LM). Additional models include POS target language models, lexical weights, word penalty and reordering models among others.

3 Collocation segmentation

Collocation segmentation is the process of detecting boundaries between collocation segments within a text (Daudaravicius and Marcinkeviciene, 2004). A collocation segment is a piece of text between boundaries. The boundaries are established in two steps using two different measures: the Dice score and a Average Minimum Law (AML).

The Dice score is used to measure the association strength between two words. It has been used before in the collocation compiler XTract (Smadja, 1993) and in the lexicon extraction system Champollion (Smadja et al., 1996). It is defined as follows:

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

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. It gives high scores when x and y occur in conjunction. The first step then establishes a boundary between

[1] Universitat Politècnica de Catalunya

[2] Barcelona Media Innovation Center

[3] Vytautas Magnus University

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 98–102,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

two adjacent words when the Dice score is lower than a threshold $t = exp(-8)$. Such a threshold was established following the results obtained in (Costa-jussà et al., 2010), where an integration of this technique and a SMT system was performed over the Bible corpus.

The second step of the procedure uses the AML. It defines a boundary between words x_{i-1} and x_i when:

$$\frac{Dice\,(x_{i-2}; x_{i-1}) + Dice\,(x_i; x_{i+1})}{2} > Dice\,(x_{i-1}; x_i) \quad (3)$$

That is, the boundary is set when the Dice value between words x_i and x_{i-1} is lower than the average of preceding and following values.

4 Experimental Framework

All systems were built using Moses (Koehn et al., 2007), a state-of-the-art software for phrase-based SMT. For preprocessing Spanish, we used Freeling (Atserias et al., 2006), an open source library of natural language analyzers. For English, we used TnT (Brants, 2000) and Moses' tokenizer. The language models were built using SRILM (Stolcke, 2002).

4.1 Corpus

This year, the translation task provided four different sources to collect corpora for the Spanish-English pair. Bilingual corpora included version 5 of the Europarl Corpus (Koehn, 2005), the News Commentary corpus and the United Nations corpus. Additional English corpora was available from the News corpus. The organizers also allowed the use of the English Gigaword Third and Fourth Edition, released by the LDC. As for development and internal test, the test sets from 2008 and 2009 translation tasks were available.

For our experiments, we selected as training data the union of the Europarl and the News Commentary. Development was performed with a section of the 2008 test set and the 2009 test set was selected as internal test. We deleted all empty lines, removed pairs that were longer than 40 words, either in Spanish or English; and also removed pairs whose ratio between number of words were bigger than 3.

As a preprocess, all corpora were lower-cased and tokenized. The Spanish corpus was tokenized and POS tags were extracted using Freeling, which split clitics from verbs and also separated words like *"del"* into "de el". In order to build a POS target language model, we also obtained POS tags from the English corpus using the TnT tagger. Statistics of the selected corpus can be seen in Table 1.

Corpora	Spanish	English
Training sent	$1,180,623$	$1,180,623$
Running words	$26,454,280$	$25,291,370$
Vocabulary	$118,073$	$89,248$
Development sent	$1,729$	$1,729$
Running words	$37,092$	$34,774$
Vocabulary	$7,025$	$6,199$
Internal test sent	$2,525$	$2,525$
Running words	$69,565$	$65,595$
Vocabulary	$10,539$	$8,907$
Official test sent	$2,489$	-
Running words	$66,714$	-
Vocabulary	$10,725$	-

Table 1: Statistics for the training, development and test sets.

	Internal test	Official test
Adjectives	137	72
Common nouns	369	188
Proper nouns	408	2,106
Verbs	213	128
Others	119	168
Total	1246	2662

Table 2: Unknown words found in internal and official test sets

It is important to notice that neither the United Nations nor the Gigaword corpus were used for bilingual training. Nevertheless, the English part from the United Nations and the monolingual News corpus were used to build the language model of our systems.

4.1.1 Unknown words

We analyzed the content from the internal and official test and realized that they both contained many words that were not seen in the training data. Table 2 shows the number of unknown words found in both sets, classified according to their POS.

In average, we may expect an unknown word every two sentences in the internal test and more than one per sentence in the official test set. It can also be seen that most of those unknown words are proper nouns, representing 32% and 79% of the unknown sets, respectively. Common nouns were the second most frequent type of unknown words, followed by verbs and adjectives.

4.2 Systems

We submitted two different systems for the translation task. First a baseline using the training data mentioned before; and then an *augmented* system, where the baseline-extracted phrase list was extended with additional phrases coming from a segmented version of the training corpus.

We also considered an additional system built with two different decoding path, a standard path from words to words and POS and an alternative path from stems to words and POS in the target side. At the end, we did not submit this system to the translation task because it did not provide better results than the previous two in our internal test.

The set of feature functions used include: source-to-target and target-to-source relative frequencies, source-to-target and target-to-source lexical weights, word and phrase penalties, a target language model, a POS target language model, and a lexicalized reordering model (Tillman, 2004).

4.2.1 Considering stems as an alternate decoding path.

Using Moses' framework for factored translation models we defined a system with two decoding paths: one decoding path using words and the other decoding path using stems in the source language and words in the target language. Both decoding paths only had a single translation step. The possibility of using multiple alternative decoding path was developed by Birch et. al. (2007).

This system tried to solve the problem with the unknown words. Because Spanish is morphologically richer than English, this alternative decoding path allowed the decoder translate words that were not seen in the training data and shared the same root with other known words.

4.2.2 Expanding the phrase table using collocation segmentation.

In order to build the augmented phrase table with the technique mentioned in section 3, we segmented each language of the bilingual corpus independently and then, using the collocation segments as words, we aligned the corpus and extracted the phrases from it. Once the phrases were extracted, the segments of each phrase were split again in words to have standard phrases. Finally, we use the union of this phrases and the phrases extracted from the baseline system to compute the final phrase table. A diagram of the whole procedure can be seen in figure 1.

The objective of this integration is to add new phrases in the translation table and to enhance the relative frequency of the phrases that were extracted from both methods.

4.2.3 Language model interpolation.

Because SMT systems are trained with a bilingual corpus, they ended highly tied to the domain the corpus belong to. Therefore, when the documents we want to translate belong to a different domain, additional domain adaptation techniques are recommended to build the system. Those techniques usually employ additional corpora that correspond to the domain we want to translate from.

	internal test
baseline	24.25
baseline+stem	23.45
augmented	23.9

Table 3: Internal test results.

	test	$\text{test}_{cased-detok}$
baseline	26.1	25.1
augmented	26.1	25.1

Table 4: Results from translation task

The test set for this translation task comes from the news domain, but most of our bilingual corpora belonged to a political domain, the Europarl. Therefore we use the additional monolingual corpus to adapt the language model to the news domain.

The strategy used followed the experiment performed last year in (R. Fonollosa et al., 2009). We used SRILM during the whole process. All language models were order five and used modified Kneser-Ney discount and interpolation. First, we build three different language models according to their domain: Europarl, United Nations and news; then, we obtained the perplexity of each language model over the News Commentary development corpus; next, we used `compute-best-mix` to obtain weights for each language model that diminish the global perplexity. Finally, the models were combined using those weights.

In our experiments all systems used the resulting language model, therefore the difference obtained in our results were cause only by the translation model.

4.3 Results

We present results from the three systems developed this year. First, the *baseline*, which included all the features mentioned in section 4.2; then, the system with an alternative decoding path, called *baseline+stem*; and finally the *augmented* system, which integrated collocation segmentation to the baseline. Internal test results can be seen in table 3. Automatic scores provided by the WMT 2010 organizers for the official test can be found in table 4. All BLEU scores are case-insensitive and tokenized except for the official test set which also contains case-sensitive and non-tokenized score.

We obtained a BLEU score of 26.1 and 25.1 for our case-insensitive and sensitive outputs, respectively. The highest score was obtained by University of Cambridge, with 30.5 and 29.1 BLEU points.

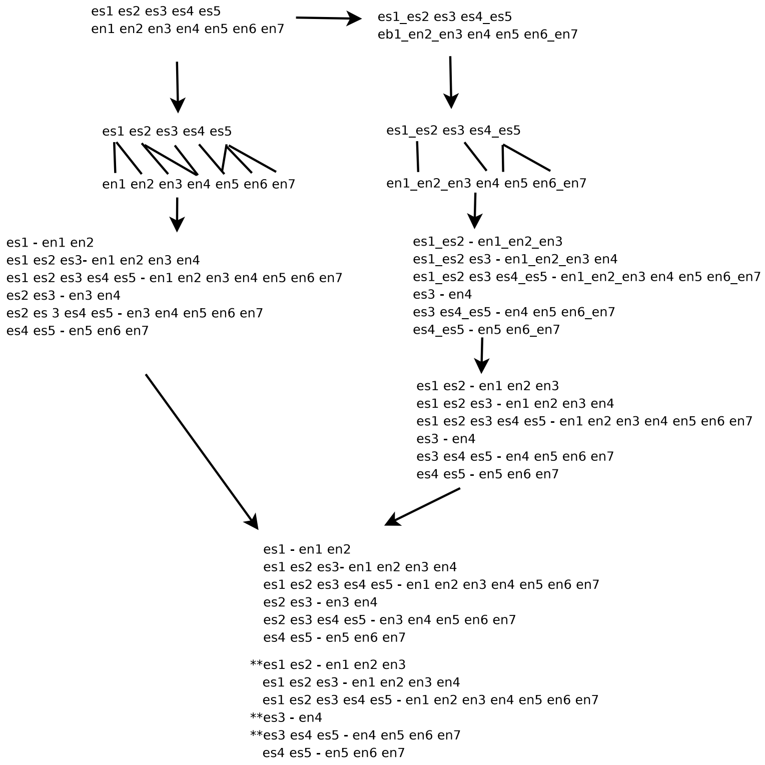
es1 es2 es3 es4 es5
en1 en2 en3 en4 en5 en6 en7
es1_es2 es3 es4_es5
eb1_en2_en3 en4 en5 en6_en7
es1 es2 es3 es4 es5
en1 en2 en3 en4 en5 en6 en7
es1_es2 es3 es4_es5
en1_en2_en3 en4 en5 en6_en7
es1 - en1 en2
es1 es2 es3- en1 en2 en3 en4
es1 es2 es3 es4 es5 - en1 en2 en3 en4 en5 en6 en7
es2 es3 - en3 en4
es2 es 3 es4 es5 - en3 en4 en5 en6 en7
es4 es5 - en5 en6 en7
es1_es2 - en1_en2_en3
es1_es2 es3 - en1_en2_en3 en4
es1_es2 es3 es4_es5 - en1_en2_en3 en4 en5 en6_en7
es3 - en4
es3 es4_es5 - en4 en5 en6_en7
es4_es5 - en5 en6_en7
es1 es2 - en1 en2 en3
es1 es2 es3 - en1 en2 en3 en4
es1 es2 es3 es4 es5 - en1 en2 en3 en4 en5 en6 en7
es3 - en4
es3 es4 es5 - en4 en5 en6 en7
es4 es5 - en5 en6 en7
es1 - en1 en2
es1 es2 es3- en1 en2 en3 en4
es1 es2 es3 es4 es5 - en1 en2 en3 en4 en5 en6 en7
es2 es3 - en3 en4
es2 es3 es4 es5 - en3 en4 en5 en6 en7
es4 es5 - en5 en6 en7
**es1 es2 - en1 en2 en3
es1 es2 es3 - en1 en2 en3 en4
es1 es2 es3 es4 es5 - en1 en2 en3 en4 en5 en6 en7
**es3 - en4
**es3 es4 es5 - en4 en5 en6 en7
es4 es5 - en5 en6 en7

Figure 1: Example of the expansion of the phrase table using collocation segmentation. New phrases added by the collocation-based system are marked with a $**$.

4.3.1 Comparing systems

Once we obtained the translation outputs from the baseline and the *augmented* system, we performed a manual comparison of them. Even though we did not find any significant advantages of the *augmented* system over the baseline, the collocation segmentation strategy chose a better morphological structures in some cases as can be seen in Table 5 (only sentence sub-segments are shown):

5 Conclusion

We presented two different submissions for the Spanish-English language pair. The language model for both system was built interpolating two big out-of-domain language models and one smaller in-domain language model. The first system was a baseline with POS target language model; and the second one an *augmented* system, that integrates the baseline with collocation segmentation. Results over the official test set showed no difference in BLEU between these two, even though internal results showed that the baseline obtained a better score.

We also considered adding an additional decoding path from stems to words in the baseline but internal tests showed that it did not improve translation quality either. The high number of unknown words found in Spanish suggested us that considering in parallel the simple form of stems could help us achieve better results. Nevertheless, a deeper study of the unknown set showed us that most of those words were proper nouns, which do not have inflection and therefore cannot benefited from stems.

Finally, despite that internal test did not showed an improvement with the *augmented* system, we submitted it as a secondary run looking for the effect these phrases could have over human evaluation.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 247762, from the Spanish Ministry of Science and Innovation through the Buceador project (TEC2009-14094-C04-01) and 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

Jordi Atserias, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluís Padró, and Muntsa Padró. 2006. FreeLing 1.3: Syntactic and semantic services in an open-source NLP

Original: sabiendo que **está recibiendo** el premio Baseline: knowing that **it receive** the prize Augmented: knowing that **he is receiving** the prize
Original: muchos de mis amigos **prefieren no separarla**. Baseline: many of my friends **prefer not to separate them**. Augmented: many of my friends **prefer not to separate it**.
Original: Los estadounidenses **contarán** con un teléfono móvil Baseline: The Americans **have** a mobile phone Augmented: The Americans **will have** a mobile phone
Original: es plenamente consciente del camino más largo **que debe emprender** Baseline: is fully aware of the longest journey **must undertake** Augmented: is fully aware of the longest journey **that need to be taken**

Table 5: Comparison between baseline and augmented outputs

library. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006), ELRA*, Genoa, Italy, May.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2007. Ccg supertags in factored statistical machine translation. In *StatMT '07: Proceedings of the Second Workshop on Statistical Machine Translation*, pages 9–16, Morristown, NJ, USA. Association for Computational Linguistics.

Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000)*, Seattle, WA.

Marta R. Costa-jussà, Vidas Daudaravicius, and Rafael E. Banchs. 2010. Integration of statistical collocation segmentations in a phrase-based statistical machine translation system. In *14th Annual Conference of the European Association for Machine Translation.*

Vidas Daudaravicius and Ruta Marcinkeviciene. 2004. Gravity counts for the boundaries of collocations. *International Journal of Corpus Linguistics*, 9:321–348(28).

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL '07: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180, Morristown, NJ, USA. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit.*

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29:19–51.

José A. R. Fonollosa, Maxim Khalilov, Marta R. Costa-jussá, José B. Mariño, Carlos A. Henríquez Q., Adolfo Hernández H., and Rafael E. Banchs. 2009. The TALP-UPC phrase-based translation system for EACL-WMT 2009. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 85–89, Athens, Greece, March. Association for Computational Linguistics.

Frank A. Smadja, Kathleen McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Comput. Linguist.*, 19(1):143–177.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. pages 901–904.

Christoph Tillman. 2004. A Unigram Orientation Model for Statistical Machine Translation. In *HLT-NAACL.*