UPM system for the translation task

Verónica López-Ludeña
Grupo de Tecnología del Habla
Universidad Politécnica de Madrid
veronicalopez@die.upm.es

Rubén San-Segundo
Grupo de Tecnología del Habla
Universidad Politécnica de Madrid
lapiz@die.upm.es

Abstract

This paper describes the UPM system for translation task at the EMNLP 2011 workshop on statistical machine translation (http://www.statmt.org/wmt11/), and it has been used for both directions: Spanish-English and English-Spanish. This system is based on Moses with two new modules for pre and post processing the sentences. The main contribution is the method proposed (based on the similarity with the source language test set) for selecting the sentences for training the models and adjusting the weights. With system, we have obtained a 23.2 BLEU for Spanish-English and 21.7 BLEU for English-Spanish.

1 Introduction

The Speech Technology Group of the Universidad Politécnica de Madrid has participated in the sixth workshop on statistical machine translation in the Spanish-English and English-Spanish translation task.

Our submission is based on the state-of-the-art SMT toolkit Moses (Koehn, 2010) adding a pre-processing and a post-processing module. The main contribution is a corpus selection method for training the translation models based on the similarity of each source corpus sentence with the language model of the source language test set.

There are several related works on filtering the training corpus by using a similarity measure based on the alignment score or based on sentences length (Khadivi and Ney, 2005; Sanchis-Trilles et al, 2010). However, these techniques are focused on removing noisy data, i.e., their idea is to eliminate possible errors in the databases.

The difference between these techniques and the method that we propose is that we do not search "bad" pairs of sentences, but we search those sentences in source training corpus that are more similar with the language model generated with the source test sentences and we select them for training.

Other interesting technique of corpus selection is based on transductive learning (Ueffing, 2007). In this work, authors use of transductive semi-supervised methods for the effective use of monolingual data from the source language in order to improve translation quality.

The method proposed in this paper is also applied to the validation corpus. There are other works related to select development set (Hui, 2010) that they combine different development sets in order to find the more similar one with test set.

2 Overall description of the system

The translation system used is based on Moses, the software released to support the translation task (http://www.statmt.org/wmt11/) at the EMNLP 2011 workshop on statistical machine translation.

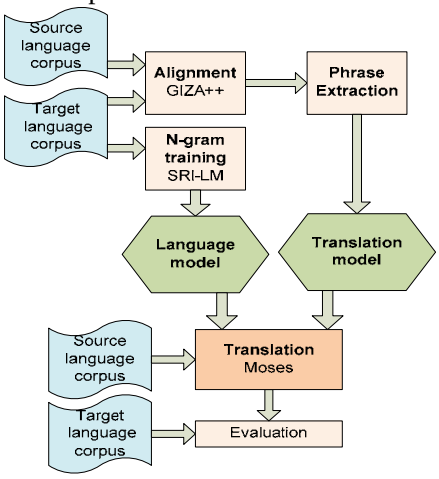

Figure 1: Moses translation system

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 420–425,
Edinburgh, Scotland, UK, July 30–31, 2011.

The phrase model has been trained following these steps (Figure 1):

- Word alignment computation. GIZA++ (Och and Ney, 2003) is a statistical machine translation toolkit that is used to calculate the alignments between Spanish and English words in both direction (Spanish-English and English-Spanish). To generate the translation model, the parameter "alignment" was fixed to "grow-diag-final" (default value), and the parameter "reordering" was fixed to "msd-bidirectional-fe" as the best option, based on experiments on the development set.
- Phrase extraction (Koehn et al 2003). All phrase pairs that are consistent with the word alignment (grow-diag-final alignment in our case) are collected. To extract the phrases, the parameter "max-phrase-length" was fixed to "7" (default value), based on experiments on the development set.
- Phrase scoring. In this step, the translation probabilities are computed for all phrase pairs. Both translation probabilities are calculated: forward and backward.

The Moses decoder is used for the translation process (Koehn, 2010). This program is a beam search decoder for phrase-based statistical machine translation models. In order to obtain a 3-gram language model, the SRI language modeling toolkit has been used (Stolcke, 2002).

In addition, a pre-processing module was developed for adapting the format of the corpus before training (pre-processing of training, development and test corpora). And a post-processing for ordering punctuations, recasing, etc. is also applied to Moses output.

3 Corpora used in these experiments

For the system development, we have only used the free corpora distributed in the EMNLP 2011 translation task.

In particular, we have considered the union of the Europarl corpus, the United Nations Organization (UNO) Corpus, the News Commentary Corpus and the test sets of 2000, 2006, 2007 and 2008.

For developing the system, we have developed and evaluated the system considering the union of 2009 and 2010 test sets.

All these files can be free downloaded from http://www.statmt.org/wmt11/.

A pre-processing of these databases is necessary for adapting the original format to our system.

We have not used the complete union of all corpora, but a corpus selection by filtering the union of the training set and also filtering the union of the development set. This selection will be explained in section 5.

The main characteristics of the corpus are shown in Table 1: the previous corpora and the filtered corpora.

		Original sentences	**Filtered sentences**
Training (Translation Model (TM) /Language Model (LM))	**Europarl Training Corpus**	1,650,152	150,000 (TM) 3,000,000 (LM)
	UNO Corpus	6,222,450	
	News commentary	98,598	
	Previous test sets	15,150	
Development	**news-test2009**	2,525	1,000
	news-test2010	2,489	
Test	**news-test2011**	3,003	3,003

Table 1: Main characteristics of the corpus

4 Preparing the corpora

In order to use the corpus described in section 3 with the mentioned translation systems, it is necessary a pre-processing. This pre-processing, for training files, consists of:

- UTF-8 to Windows format conversion, because our software adapted to Windows had several problems with the UTF-8 format: it does not know accent marks, ñ letter, etc.
- Deletion of blank lines and sentences that are comments (for instance: "<CHAPTER ID=1>")
- Deletion of special characters (.,;:¿?¡!-/\, etc.), except those that are next to numbers (for instance: "1.4", "2,000", "1/3"). We decided to remove these special characters to avoid including them in the translation model. During translation, these characters will be considered as phrase limits.

- Words were kept in their natural case, but the first letter of each sentence was lowercased, because first words of sentences are used to be lowercased as their most common form.
- Contracted words were separated for training each word separately. For instance, “it’s” becomes “it is”. For the ambiguous cases, like “he’s” that can be “he is” or “he has”, we have not done any further processing: we have considered the most frequent situation. For the case of Saxon genitive, when proper names are used (instead of pronouns), “’s” is a Saxon genitive most of the times. But, when using a pronoun, it is a contracted word.

For development and test sets, the same actions were carried out, but now, special characters were not deleted, but separated in tokens, i.e., a blank space was introduced between special characters and adjacent words. For instance, “*la bolsa de Praga , al principio del martes comercial , reaccionó inmediatamente a la caída del lunes cuando descendió aproximadamente a un 6 % .*”

So, special characters are considered as independent tokens in translation. The main idea was to force the system to consider special characters as phrase limits during the translation process.

5 Selecting the training corpus

Scattering of training data is a problem when integrating training material from different sources for developing a statistical system. In this case, we want to use a big training corpus joining all available corpora obtaining about 8 millions sentences.

But an excessive amount of data can produce an important scattering that the statistical model cannot learn properly.

The technique proposed by the Speech Technology Group at UPM in the translation task (Spanish-English and English-Spanish) consists of a filtering of the training data in order to obtain better results, without having memory problems.

The first step is to compute a language model of the source language considering sentences to translate (sentences from the 2011 source test set).

Secondly, the system computes the similarity of each source sentence in the training to the language model obtained in the first step. This similarity is computed with the following formula:

$$sim = \frac{1}{n}\sum_{i=0}^{n}\log(P_n) \qquad (1)$$

For example, if one sentence is “A B C D” (where each letter is a word of the sentence):

$$sim = \frac{1}{4}(P_A + P_{AB} + P_{ABC} + P_{BCD}) \quad (2)$$

Each probability is extracted from the language model calculated in the first step. This similarity is the negative of the source sentence perplexity given the language model.

With all the similarities, the mean and the standard deviation values are computed and used to define a threshold. For example, calculating the similarity of all sentences in our train corpus (about 8,000,000 of sentences) a similarity histogram is obtained (Figure 2).

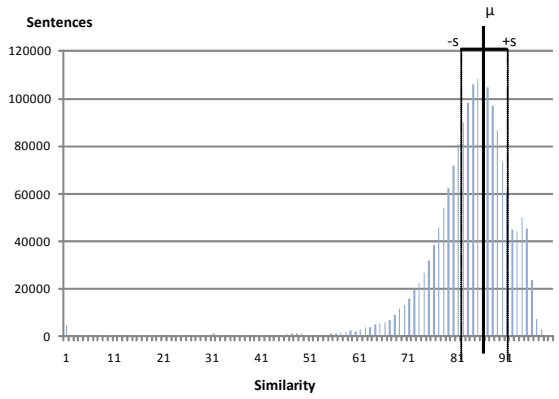

Figure 2: Similarity histogram of Spanish-English system

This histogram indicates the number of sentences inside each interval. There are 100 different intervals: the minimum similarity is mapped into 0 and the maximum one into 100.

Finally, source training sentences with a similarity lower than the threshold are eliminated from the training set (the corresponding target sentences are also removed).

The whole process is shown in Figure 3. This process takes 20 hours approximately for filtering

more than 8 million sentences in an Intel core 2 quad computer.

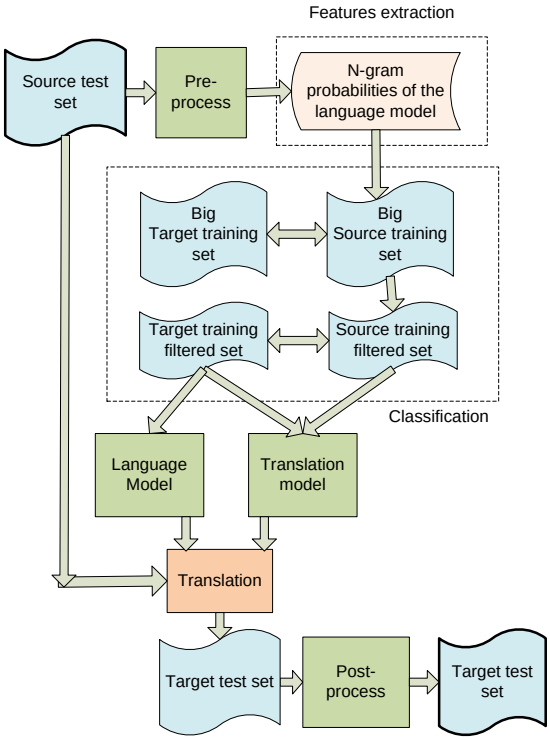

Figure 3: Diagram of complete process

Figure 4 shows the results of the experiments in Spanish-English system selecting the training corpus with different similarity thresholds. These results were obtained before filtering the development corpus, with the same filtered training corpus for translation and language models and before post-processing.

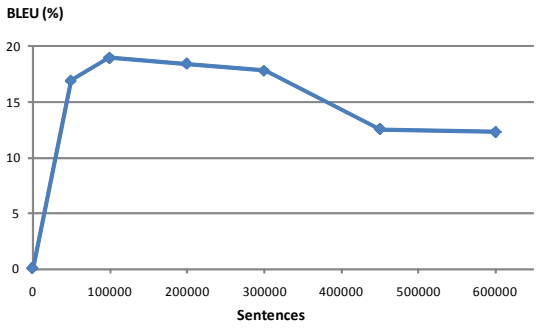

Figure 4: Translation results of baseline Spanish-English system with different number of training sentences

As can be observed, with more than 400,000 sentences there is a 12% BLEU (with an asymptotic tendency), but there is an important improvement filtering up to 100,000 (there is already not scattering). But results start to fall off when there are insufficient sentences (problem of sparseness of data with less than 100,000 sentences).

6 Post processing

After performing the statistical translation, we have incorporated a post-processing module with the following functions:

- To check the date format, detecting possible order errors and correcting them.
- To check the format of the numbers, numerical and ordinal ones: 1º into 1st and so on.
- Detokenization and ordering the punctuations marks when there are several ones consecutively (i.e. '".' or ').'), trying to follow, always, the same order.
- To put the first letter of the sentences in capital letters.
- To use a backup dictionary for translating isolated words. This aspect has improved 2% (BLEU) but it has also introduced some errors. For example in the case of English-Spanish, there was a checking process for translating English words into Spanish. But there were several English words that also are Spanish words. For example, "un" is an article in Spanish but in English means "United Nations" (Naciones Unidas) so some "un" were translated as "Naciones Unidas" by error.

7 Selecting the development corpus

The development corpus is used to adapt the different weights used in the translation process for combining the different sources of information. Weight computation is a sensible task. In order to better adapt these weights, the development corpus is also filtered considering the same strategy commented in section 5.

Our solution consists of using two different corpora (2009 and 2010 test sets) and "choosing" the best sentences to use in development task with

the same filtering technique explained in section 5. Finally, we select the 1,000 sentences with the greater similarity respect to the source language model of the test set.

Other action carried out in final experiments is using different corpora for training translation and language models. In order to generate the language model it is better to use a big corpus; so, we use 3,000,000 sentences that it is the biggest model that we can generate without memory problems.

But in order to generate the translation model, the final one is trained with 150,000 sentences.

The final results are shown in Table 2.

Spanish-English	**BLEU**	**BLEU cased**
Baseline	12.57	12.15
Best result	**23.20**	**21.90**
English-Spanish	**BLEU**	**BLEU cased**
Baseline	10.73	10.30
Best result	**21.70**	**20.90**

Table 2: Final results of the translation system

With this work, we have demonstrated that filtering the corpus for training the translation module, can improve the translation results. But there are still important problems that must be addressed like the high number of out of vocabulary words (OOVs) (more than 40% of the test corpus vocabulary) that they have to be improved in the selecting method.

About the selection, it is important to comment that this method more likely filters long sentences out: the average number of words in the selected corpus is 14 while in the whole training set and in the test set is higher than 25.

Other interesting aspect to comment is that in the selected training corpus, more than 70% of the sentences come from the Europarl or the News Commentary corpus, being the UNO corpus the biggest one.

Anyway, although the improvement is interesting, the system can not compete with other well-known translation systems until we incorporate additional modules for reordering or n-best post processing.

8 Conclusions

This paper has presented and described the UPM statistical machine translation system for Spanish-English and English-Spanish. This system is based on Moses with pre-processing and post-processing modules. The main contribution has been the proposed method for selecting the sentences used for training and developing the system. This selection is based on the similarity with the source language test set. The results have been 23.2 BLEU for Spanish into English and 21.7 for English into Spanish.

9 Future work

One of the main problems we have observed in the selection proposed method has been the high number of OOVs during translation. This problem has been addressed by incorporating a backup vocabulary in the post-processing module. This solution has solved some cases but it has not able to deal with order problems. Because of this, in the near future, we will try to improve the corpus selection method for reducing the number of OOVs.

Acknowledgments

The authors would like to thank discussions and suggestions from the colleagues at GTH-UPM. This work has been supported by Plan Avanza Exp Nº: TSI-020100-2010-489), INAPRA (MEC ref: DPI2010-21247-C02-02), and SD-TEAM (MEC ref: TIN2008-06856-C05-03) projects and FEDER program.

References

Hui, C., Zhao, H., Song, Y., Lu, B., 2010 “An Empirical Study on Development Set Selection Strategy for Machine Translation Learning” on Fifth Workshop on Statistical Machine Translation.

Koehn P., F.J. Och D. Marcu. 2003. “Statistical Phrase-based translation”. Human Language Technology Conference 2003 (HLT-NAACL 2003), Edmonton, Canada, pp. 127-133, May 2003.

Koehn, Philipp. 2010. “Statistical Machine Translation”. Cambridge University Press.

Khadivi, S., Ney, H., 2005. “Automatic filtering of bilingual corpora for statistical machine translation.” In Natural Language Processing and Information Systems, 10th Int. Conf. on Applications of Natural Language to Information Systems, volume 3513 of Lecture Notes in Computer Science, pages 263–274, Alicante, Spain, June. Springer.

Och J., Ney. H., 2003. “A systematic comparison of various alignment models”. Computational Linguistics, Vol. 29, No. 1 pp. 19-51, 2003.

Sanchis-Trilles, G., Andrés-Ferrer, J., Gascó, G., González-Rubio, J., Martínez-Gómez, P., Rocha, M., Sánchez, J., Casacuberta, F., 2010. “UPV-PRHLT English–Spanish System for WMT10”. On ACL Fifth Workshop on Statistical Machine Translation.

Stolcke A., 2002. “SRILM – An Extensible Language Modelling Toolkit”. Proc. Intl. Conf. on Spoken Language Processing, vol. 2, pp. 901-904, Denver.

Ueffing, N., Haffari, G., Sarkar, A., 2007. “Transductive learning for statistical machine translation”. On ACL Second Workshop on Statistical Machine Translation.