UPM system for WMT 2012

Verónica López-Ludeña, Rubén San-Segundo and Juan M. Montero
GTH-IEL-ETSI Telecomunicación
Universidad Politécnica de Madrid
{veronicalopez, lapiz, juancho}@die.upm.es

Abstract

This paper describes the UPM system for the Spanish-English translation task at the NAACL 2012 workshop on statistical machine translation. This system is based on Moses. We have used all available free corpora, cleaning and deleting some repetitions. In this paper, we also propose a technique for selecting the sentences for tuning the system. This technique is based on the similarity with the sentences to translate. With our approach, we improve the BLEU score from 28.37% to 28.57%. And as a result of the WMT12 challenge we have obtained a 31.80% BLEU with the 2012 test set. Finally, we explain different experiments that we have carried out after the competition.

1 Introduction

The Speech Technology Group at the Technical University of Madrid has participated in the seventh workshop on statistical machine translation in the Spanish-English translation task.

Our submission is based on the state-of-the-art SMT toolkit Moses (Koehn et al., 2007). Firstly, we have proved different corpora for training the system: cleaning the whole corpus and deleting some repetitions in order to have a better performance of the translation model.

There are several related works on filtering the training corpus by removing noisy data that use a similarity measure based on the alignment score or based on sentences length (Khadivi and Ney, 2005).

In this paper, we also propose a technique for selecting the most appropriate sentences for tuning the system, based on the similarity with the Spanish sentences to translate. This technique is an update of the technique proposed by our group in the last WMT11 challenge (López-Ludeña and San-Segundo, 2011). There are other works related to select the development set (Hui et al., 2010) that combine different development sets in order to find the more similar one with test set.

There are also works related to select sentences, but for training instead of tuning, based on the similarity with the source test sentences. Some of them are based on transductive learning: semi-supervised methods for the effective use of monolingual data from the source language in order to improve translation quality (Ueffing, 2007); methods using instance selection with feature decay algorithms (Bicici and Yuret, 2011); or using TF-IDF algorithm (Lü et al., 2007). There are also works based on selecting training material with active learning: using language model adaptation (Shinozaki et al., 2011); or perplexity-based methods (Mandal et al., 2008).

In this work, we have used the proposed selection method only for tuning.

The rest of the paper is organized as follows. Next section overviews the system. Section 3 describes the used corpora. Section 4 explains the experiments carried out before the competition. Section 5 describes the sentences selection technique for tuning. Section 6 summarizes the results: before the WMT12 challenge, the corresponding to the competition and the last experiments. Finally, section 7 shows the conclusions.

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/wmt12/) at the NAACL 2012 workshop on statistical machine translation.

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 338–344,
Montréal, Canada, June 7-8, 2012.

The Moses decoder is used for the translation process (Koehn et al., 2007). This program is a beam search decoder for phrase-based statistical machine translation models.

We have used GIZA++ (Och and Ney, 2003) for the word alignment computation. In order 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.

In order to extract phrases (Koehn et al 2003), the considered alignment was grow-diag-final. And the parameter "max-phrase-length" was fixed to "7" (default value), based on experiments on the development set.

Finally, we have built a 5-gram language model, using the IRSTLM language modeling toolkit (Federico and Cettolo, 2007).

Additionally, we have used the following tools for pre-processing the training corpus: tokenizer.perl, lowercase.perl, clean-corpus-n.perl. And the following ones for recasing, detokenizer and normalizing punctuation in the translation output: train-recaser.perl, recase.perl, detokenizer.perl and normalize-punctuation.perl.

In addition, we have used Freeling (Padró et al., 2010) in some experiments, an open source library of natural language analyzers, but we did not improve our experiments by using Freeling. We used this tool in order to extract factors for Spanish words in order to train factored translation models.

3 Corpora used in these experiments

For the system development, only the free corpora distributed in the NAACL 2012 translation task has been used, so any researcher can validate these experiments easily.

In order to train the translation model, we used the union of the Europarl corpus, the United Nations Organization (UNO) corpus and the News Commentary corpus.

A 5-gram language model was built joining the following monolingual corpora: Europarl, News commentary, United Nations and News Crawl. We have not used the Gigaword corpus.

In order to tune the model weights, the 2010 and 2011 test set were used for development. We did not use the complete set, but a sentences selection in order to improve the tuning process. This selection will be explained in section 5.

The main characteristics of the corpora are shown in Table 1. All the parallel corpora has been cleaned with clean-corpus-n.perl, lowercased with lowercase.perl and tokenized with tokenizer.perl.

All these tools can be also free downloaded from http://www.statmt.org/wmt12/.

We observed that the parallel corpora, specially the UNO corpus, have many repeated sentences. We noted that these repetitions can cause a bad training. So, after cleaning the parallel corpora with the clean-corpus-n.perl tool, we eliminated all repetitions that appear more than 3 times in the parallel corpus.

		Original sentences
Translation Model (TM)	Europarl (EU)	1,965,734
	UNO	11,196,913
	News commentary (NC)	157,302
	Total	13,319,949
	Total clean	9,530,335
	Total without repetitions	**4,907,778**
Language Model (LM)	Europarl	2,218,201
	UNO	11,196,913
	News commentary (NC)	212,517
	News Crawl (NCR)	51,827,710
	Total	**65,455,341**
Tuning	news-test2010	2,489
	news-test2011	3,003
	Total	5,492
	Total selected	**4,500**
Test	news-test2012	3,003

Table 1: Size of the corpora used in our experiments

4 Previous experiments

Several experiments were carried out by using different number of sentences, as it is shown in Table 2.

In these experiments, we used the 2010 test set for tuning (news-test2010) and the 2011 test set for test (news-test2011). And a 5-gram language model was built with the IRSTLM tool. For evaluating the performance of the translation system, the BLEU (BiLingual Evaluation Understudy) metric

has been computed using the NIST tool (mteval.pl) (Papipeni et al., 2002).

Firstly, we checked the contribution of UNO corpus in the final result. As it is shown in Table 2, the results improve when we add the UNO corpus, although this difference is small compared to the increasing of number of sentences: with 1,643,597 sentences we have a 28.24% BLEU and if we add around other 8 million sentences more, the BLEU score only increase 0.13 points (28.37%).

Training	Deleting repetitions	Number of sentences	BLEU (%)
EU+NC	NO	1,643,597	28.24
EU+NC+ UNO	NO	9,530,335	28.37
EU+NC+ UNO	YES (> 1)	2,112,968	28.12
EU+NC +UNO	**YES (> 3)**	**4,907,778**	**28.47**
EU+NC+ UNO	YES (> 5)	6,270,441	28.28

Table 2: Previous experiments using news-test2010 for tuning and news-test2011 as test set

We observed that UNO corpus have a lot of repeated sentences. So, we decided to remove repetitions in the whole corpus. With this action, we aimed to keep the UNO sentences that let us to improve the BLEU score and, on the other hand, to delete the sentences that do not contribute in any way, reducing the training time.

We did some experiments deleting repetitions: allowing 5 repetitions, 3 repetitions and, finally, 1 repetition (no repetitions). Table 2 shows how the results improve deleting more than 3 repetitions. So, finally, we improved the BLEU score from 23.24% without UNO corpus to 28.37% adding the UNO and to 28.47% deleting all sentences repeated more than 3 times.

5 Selecting the development corpus

When the system is trained, different model weights must be tuned corresponding to the main four features of the system: translation model, language model, reordering model and word penalty. Initially, these weights are equal, but it is necessary to optimize their values in order to get a better performance. Development corpus is used to adapt the different weights used in the translation process for combining the different sources of information. The weight selection is performed by using the minimum error rate training (MERT) for log-linear model parameter estimation (Och, 2003).

It is not demonstrated that the weights with better performance on the development set provide better results on the unseen test set. Because of this, this paper proposes a sentence selection technique that allows selecting the sentences of the development set that have more similarity with the sentences to translate (source test set): if the weights are tuned with sentences more similar to the sentence in the test set, the tuned weights will allow obtaining better translation results.

We have considered two alternatives for computing the similarity between a sentence and the test set. As it will be shown, with these methods the results improve.

The first alternative consists of the similarity method proposed in (López-Ludeña and San-Segundo, 2011), that computed a 3-gram language model considering the source language sentences from the test set. After that, the system computes the similarity of each source sentence in the validation corpus considering the language model obtained in the first step and, finally, a threshold is defined for selecting a subset with the higher similarity.

The second method that we propose now is a modification of the first one. With the formula of the first method, it was observed that, in some cases, the unigram probabilities had a relevant significance in the similarity, compared to 2-gram or 3-grams. The system was selecting sentences that have more unigrams that coincide with the source test sentences. However, these unigrams sometimes were not part of "good" bigrams or trigrams. Moreover, it was detected that the previous strategy was selecting short sentences, leaving the long ones out.

Considering the previous aspects, a second method was proposed and evaluated, trying to correct these effects. The proposal was to remove the unigram effect by normalizing the similarity measure with the unigram probabilities of the word sequence. So, the similarity measure is computed now using the following equation:

$$sim = \frac{1}{n}\sum_{i=1}^{n}\log(P_n) - \frac{1}{n}\sum_{i=1}^{n}\log(P_{unig,n})$$

Where Pn is the probability of the word 'n' in the sentence considering the language model trained with the source language sentences of the test set.

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

$$sim\_norm = \frac{1}{4}(\log(P_A) + \log(P_{AB}) + \log(P_{ABC}) + \log(P_{BCD})) - \frac{1}{4}(\log(P_A) + \log(P_B) + \log(P_C) + \log(P_D))$$

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 organized in a sorted list, it is possible to define a threshold selecting a subset with the higher similarity. For example, calculating the similarity of all sentences in our development corpus (around 2,500 sentences) a similarity histogram is obtained (Figure 1).

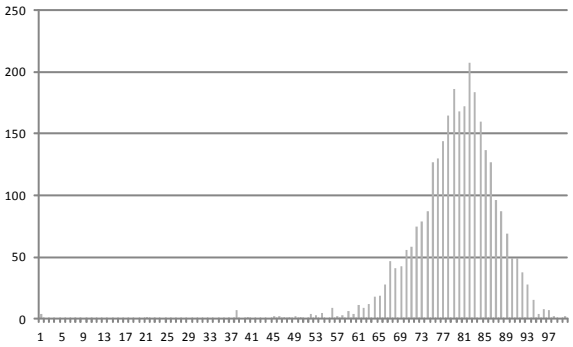

Figure 1: Similarity histogram of the source development sentences respect to the language model trained with the source language sentence of the test set

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. As it is shown, the similarity distribution is very similar to a Gaussian distribution.

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

All the experiments have been carried out in the Spanish into English translation system, using the corpora described in section 3 to generate the translation and language models.

In order to evaluate the system, the test set of the EMNLP 2011 workshop on statistical machine translation (news-test2011) was considered.

In order to adapt the different weights used in the translation process, the test set of the ACL 2010 workshop on statistical machine translation (news-test2010) has been used for weight tuning. The previous selection strategies allow filtering this validation set, selecting the most similar sentences to the test set.

Figure 2 and Table 3 show the different results with each number of selected sentences.

Sentences selected for development	BLEU results (%)	
	Normalized similarity	Similarity (López-Ludeña and San-Segundo, 2011)
500	28.01	28.36
1,000	28.11	28.47
1,500	**28.57**	**28.51**
2,000	28.57	28.36
2,489 (Baseline)	28.47	28.47
ORACLE	28.91	28.91

Table 3: Results with different number of development sentences

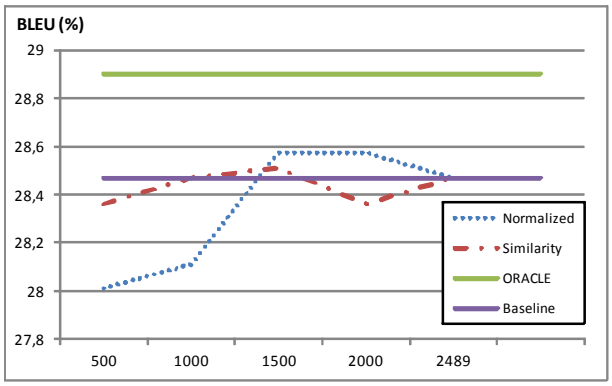

Figure 2: Results with different number of development sentences

Figure 2 shows that the BLEU score improves when the number of sentences of the development corpus increases from 0 to around 1,500 sentences with both methods. However, with more than 1,500 sentences (selected with the first similarity computation method) and more than 2,000 (select-

ed with the normalized similarity method), the BLEU score starts to decrease. This decrement reveals that there is a subset of sentences that are quite different from the test sentences and they are not appropriate for tuning the model weights.

The best obtained result has been 28.57% BLEU with 1,500 sentences of the development corpus, selected with the normalized similarity method. The improvement reached is 30% of the possible improvement (considering the ORACLE experiment). This result is better than using the complete development corpus (28.47% BLEU).

When comparing both alternatives to compute the similarity between a sentence (from the validation set) and a set of sentences (source sentences from the test set), we can see that the normalized similarity method allows a higher improvement. The main reason is that the similarity method selects sentences including information about similar unigrams, but sometimes, these unigrams are not part of "good" bigrams or trigrams. Moreover, this strategy selects short sentences, leaving the long ones out. When using the normalized similarity method, these two problems are reduced.

6 Results

	Test set	BLEU (%)	BLEU cased (%)	TER (%)
Baseline	news-test2011	28.37	25.76	59.9
Best result	news-test2011	28.57	25.98	59.8
WMT12 result	**news-test2012**	**31.80**	**28.90**	**57.9**

Table 4: Final results of the translation system

Table 4 shows the results with the 2011 test set: we have a 28.37% BLEU as baseline using the whole corpora and finally we obtain a 28.57% BLEU with the deletion of repetitions and the sentences selection for tuning.

With this configuration, we have obtained a 31.8% BLEU with the 2012 test set as a result of the competition of this year.

6.1 Other experiments

We have carried out other experiments with the 2012 test set: factored models, Minimum Bayes Risk Decoding (MBR) and other sets for tuning. However, they did not finish before the competition deadline.

- **Factored models using Freeling**

Firstly, we have trained factored models in Spanish with Moses (Koehn and Hoang, 2007). We have only factored the source language (Spanish) and, in order to obtain the factors for each Spanish word, we have used Freeling (http://nlp.lsi.upc.edu/freeling/).

When running the Freeling analyzer with a Spanish sentence and the output option "tagged", we obtain, for each word, an associated lemma, a coded tag with morphological and syntactic information, and a probability. For instance, with the sentence "*la inflación europea se deslizó en los alimentos*", we obtain:

word	*lemma*	*tag*	*probability*
la	**el**	DA0FS0	0.972
inflación	**inflación**	NCFS000	1.000
europea	**europeo**	AQ0FS0	0.900
se	**se**	P00CN000	0.465
deslizó	**deslizar**	VMIS3S0	1.000
en	**en**	SPS00	1.000
los	**el**	DA0MP0	0.976
alimentos	**alimento**	NCMP000	1.000

Table 5: Freeling analyzer output

We take advantage of the lemma (second column) associated to each word and we use it as factor. So, the previous sentence is factorized as "*la*|*el inflación*|*inflación europea*|*europeo se*|*se deslizó*|*deslizar en*|*en los*|*el alimentos*|*alimento*"

This way, two models are generated in the translation process. For the GIZA++ alignment we used the second factor (lemma) instead of the word.

Results show that there is not improvement by using Freeling. BLEU score is a bit lower (30.95% in contrast to the 31.80% obtained without Freeling). However, we want to continue doing experiments with Freeling with other different GIZA++ alignment options different to the default value "grow-diag-final".

On the other hand, we want to prove different sets for tuning. When using factored models, there are more weights to be adjusted and it is possible that 4,500 sentences are insufficient.

- **MBR**

The use of Minumum Bayes Risk (MBR) (Kumar and Byrne, 2004) consists of, instead of selecting the translation with the highest probability, minimum Bayes risk decoding selects the translation that is most similar to the highest scoring translations. The idea is to choose hypotheses that minimize Bayes Risk as oppose to those that maximize posterior probability.

If we set up this option for decoding, the results improve from 31.80% to 31.99%.

- **Tuning with a 2008-2011 test set sentences selection**

We have also changed the set for tuning, including the 2008 and 2009 test set in addition to the 2009 and 2010 sets. With the four sets we have around 10,000 sentences. For tuning, we have selected 8,000 of these sentences with the normalized similarity method explained in section 5.

Table 6 shows that the results are worse. However, we have established the threshold based on previous experiments with the 2010 and 2011 sets. Now, we should test different threshold with the four sets in order to determine the best one.

	BLEU (%)	**BLEU cased (%)**	**TER (%)**
WMT result	31.80	28.90	53.5
Freeling	30.95	28.03	54.9
MBR	**31.99**	**29.06**	**53.4**
Tuning sets (2008-2011)	31.55	28.62	53.8

Table 6: Results of the experiments after competition

7 Conclusions

This paper has described the UPM statistical machine translation system for the Spanish-English translation task at the WMT12. This system is based on Moses. We have checked that deleting repetitions of the corpus, we can improve lightly the results: we increase the BLEU score from 28.37% with the whole corpora to 28.47% allowing only 3 repetitions of each sentence. Although this improvement is not significant (we have a confidence interval of ±0.35), we can say that we obtain a similar result by reducing very much the training time.

We have also proposed a method for selecting the sentences used for tuning the system. This selection is based on the normalized similarity with the source language test set. With this technique we improve the BLEU score from 28.47% to 28.57%. Although this result is not significant, we can appreciate an improving tendency by selecting the training sentences.

As a result of WMT12 challenge, we have obtained a 31.8% BLEU in Spanish-English translation with the 2012 test set. Our system takes around 40 hours for training, 16 hours for tuning (with 5 minutes for the sentences selection) and 3 hours to translate and to recase the test sentences in an 3.33 GHz Intel PC with 24 cores.

Finally, we have presented other additional experiments after the competition. We can improve a bit more the results to 32% BLEU by using the MBR decoding option.

Acknowledgments

The work leading to these results has received funding from the European Union under grant agreement n° 287678. It has also been supported by TIMPANO (TIN2011-28169-C05-03), ITALIHA (CAM-UPM), INAPRA (MICINN, DPI2010-21247-C02-02), and MA2VICMR (Comunidad Autónoma de Madrid, S2009/TIC-1542), Plan Avanza Consignos Exp Nº: TSI-020100-2010-489 and the European FEDER fund projects.

References

E. Bicici, D. Yuret, 2011. *Instance Selection for Machine Translation using Feature Decay Algorithms*. In Proceedings of the 6th Workshop on Statistical Machine Translation, pages 272–283.

M. Federico, M. Cettolo, 2007 *Efficient Handling of N-gram Language Models for Statistical Machine Translation*. Proceedings of the Second Workshop on Statistical Machine Translation, pages 88–95.

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

S. Khadivi, H. Ney, 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.

P. Koehn and H. Hoang, 2007 *Factored Translation Models*, Conference on Empirical Methods in Natural Language Processing (EMNLP), Prague, Czech Republic.

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, E. Herbst, 2007. *Moses: Open Source Toolkit for Statistical Machine Translation*, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic.

P. Koehn, 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.

S. Kumar and W. J. Byrne. 2004. *Minimum bayes-risk decoding for statistical machine translation*. In HLT-NAACL, pages 169–176.

V. López-Ludeña and R. San-Segundo. 2011. *UPM system for the translation task*. In Proceedings of the Sixth Workshop on Statistical Machine Translation.

Y. Lü, J. Huang, Q. Liu. 2007. *Improving statistical machine translation performance by training data selection and optimization*. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 343–350.

A. Mandal, D. Vergyri, W. Wang, J. Zheng, A. Stolcke, G. Tur, D. Hakkani-Tur and N.F. Ayan. 2008. *Efficient data selection for machine translation*. In Spoken Language Technology Workshop. SLT 2008. IEEE, pages 261 –264.

F. J. Och, 2003. *Minimum error rate training in statistical machine translation*. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pages 160–167, Sapporo, Japan, July. Association for Computational Linguistics.

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

L. Padró, M. Collado, S. Reese, M. Lloberes, I. Castellón, 2010. *FreeLing 2.1: Five Years of Open-Source Language Processing Tools* Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010), ELRA La Valletta, Malta. May.

K. Papineni, S. Roukos, T. Ward, W.J. Zhu. 2002. *BLEU: a method for automatic evaluation of machine translation*. 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, PA, pp. 311-318.

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