recibido 24-12-11 revisado 10-03-12 aceptado 18-03-12

Sentence selection for improving the tuning process of a statistical machine translation system

Selección de frases para la mejora del proceso de ajuste de un sistema de traducción estadística

Verónica López-Ludeña, Rubén San-Segundo, Juan M. Montero, Jaime Lorenzo
Grupo de Tecnología del Habla
Universidad Politécnica de Madrid
veronicalopez@die.upm.es

Resumen: Este artículo describe una estrategia de selección de frases para hacer el ajuste de un sistema de traducción estadístico basado en el decodificador Moses que traduce del español al inglés. En este trabajo proponemos dos posibilidades para realizar esta selección de las frases del corpus de validación que más se parecen a las frases que queremos traducir (frases de test en lengua origen). Con esta selección podemos obtener unos mejores pesos de los modelos para emplearlos después en el proceso de traducción y, por tanto, mejorar los resultados. Concretamente, con el método de selección basado en la medida de similitud propuesta en este artículo, mejoramos la medida BLEU del 27,17% con el corpus de validación completo al 27,27% seleccionando las frases para el ajuste. Estos resultados se acercan a los del experimento ORACLE: se utilizan las mismas frases de test para hacer el ajuste de los pesos. En este caso, el BLEU obtenido es de 27,51%.

Palabras clave: Traducción estadística, selección de corpus, traducción basada en subfrases, traducción español-inglés, ajuste de pesos.

Abstract: This paper describes a sentence selection strategy for tuning a statistical machine translation system based on Moses that translates Spanish into English. This work proposes two techniques that allow selecting the more similar source sentences of the development corpus to the sentences to translate (source test sentences). With this selection, better model weights are obtained to be used later in the translation process and therefore, to obtain better translation results. In particular, with the similarity selection method proposed in this paper, experiments report a BLEU improvement from 27.17%, with the complete development set, to 27.27% BLEU, selecting the sentences for tuning. This result is closer to the result obtained for the ORACLE experiment: BLEU of 27.51%. The ORACLE experiment consists of using the same test set for tuning the system weights.

Keywords: Statistical Machine Translation, corpus selection, phrase-based translation, Spanish into English translation, weight tuning.

1 Introduction

This paper presents a sentence selection strategy for tuning a Spanish into English machine translation system based on the state-of-the-art Statistical Machine Translation toolkit Moses (Koehn, 2010).

Statistical translation systems usually are trained with all available corpora keeping out a number of sentences (development corpus) for tuning the different model weights that are used in the translation process. However, it is not demonstrated that the final weight values tuned with this development corpus would be the best for the sentences to translate (test set).

This paper proposes two techniques that allow selecting the more similar source sentences of the development corpus to the sentences to translate using only the source test sentences. With this selection, it is possible to obtain better model weights to use later in the translation process and, therefore, to get better translation results.

ISSN 1135-5948

The rest of the paper is organized as following. Section 2 describes a summary of the state of the art on sentence selection. Section 3 describes the phrase-based translation system used in this work. In section 4, the corpora used in the development of the system are described. Section 5 explained the two methods for selecting the development corpus and the results of the experiments are described and discussed in section 6. Finally, in section 7, several conclusions are extracted from the results of this work.

2 *State of the art*

There are several related works on filtering the available training corpora. On one hand, there are several works focused on selecting training sentences in order to clean the database and remove noisy data (Khadivi and Ney, 2005; Sanchis-Trilles et al, 2010). On the other hand, there are also works focused on selecting the most appropriate training sentences given the source test sentences (more similar to the sentences to translate) in order to better train the system. 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).

But there are also other works related to select the development sentences (Hui, 2010) that combine different development sets in order to find the more similar ones with the test set.

The methods proposed in this paper are focused on selecting the development data for tuning the weights that are used when combining translation and language models into the decoding process.

3 *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).

The phrase model has been trained following these steps:

- 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. 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).
- 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 4-gram language model, the SRI language modeling toolkit has been used (Stolcke, 2002).

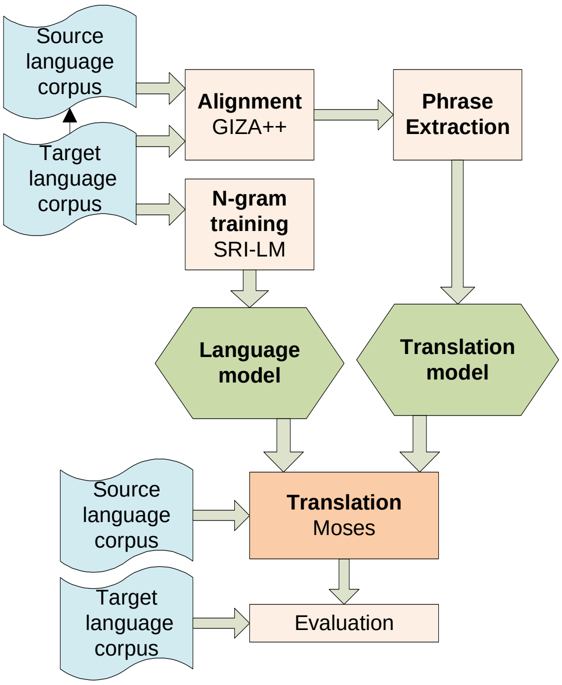

Figure 1: Moses translation system

4 *Corpora used in the experiments*

For the system development, only the free corpora distributed in the EMNLP 2011 translation task has been used, so any researcher can validate these experiments easily.

In particular, this work has considered the union of the Europarl corpus, the United Nations Organization (UNO) corpus and the News Commentary corpus to train the translation and the target language (English) model.

In order to tune the model weights, the 2010 test set was used for development. Indeed, the work presented in this paper is to select sentences from this set 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 these files can be free downloaded from http://www.statmt.org/wmt11/.

All the parallel corpora has been cleaned with clean-corpus-n.perl, lowercased with lowercase.perl and tokenized with tokenized.perl.

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

Task	Corpus	Sentences
Training translation and language models	Europarl	1,786,594
	UNO	10,662,993
	News commentary	132,571
	TOTAL	12,582,158
Tuning	news-test2010	2,489
Test	news-test2011	3,003

Table 1: Corpora used in all the experiments presented in this work.

5 *Sentence selection for tuning*

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. The 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.

Next section describes two alternatives proposed in this paper for computing the similarity between a sentence and the test set. As it will be shown in the experiments section, with these methods the results will improve.

5.1 Similarity

In the first proposal, the similarity is computed in several steps. The first step is to compute a 3-gram language model of the source language considering the source language sentences of the test set.

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

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

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):

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

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 2).

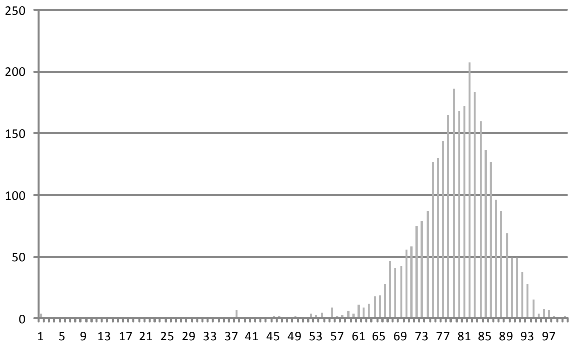

Figure 2: Similarity histogram of the source development sentences respect to the language model trained with the source language sentence of the test set

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).

5.2 Normalized similarity

With the formula of the previous 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 is 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 this equation:

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

Considering the same example described in the previous section, with the sentence "A B C D", the normalized similarity would be:

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

6 *Experiments*

All the experiments have been carried out in the Spanish into English translation system, using the corpora described in section 4 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 3 and Table 2 show the different results with each number of selected sentences. 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).

Sentences selected for development	**BLEU results (%)**	
	Similarity	**Normalized similarity**
500	27.05	26.71
1,000	27.17	26.83
1,500	27.21	**27.27**
2,000	27.07	27.27
2,489 (Baseline)	27.17	27.17
ORACLE	**27.51**	**27.51**

Table 2: Results with different number of development sentences

It is also shown the ORACLE and baseline experiments. In ORACLE experiment, the translation weights have been tuned using the same test set. In this situation, the obtained BLEU was 27.51%.

The baseline system consists of using all the sentences included in the validation set (without discarding any sentence). In the baseline case, the BLEU was 27.17%.

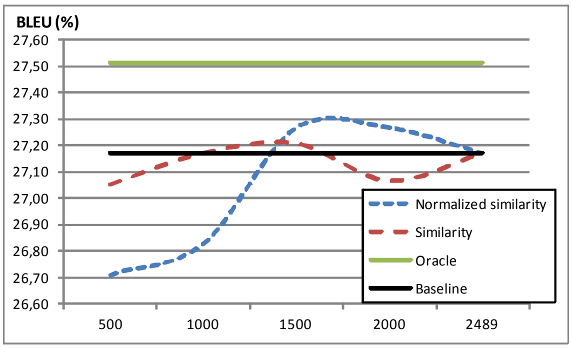

Figure 3: Results with different number of development sentences

Figure 4 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 (selected 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 adjusting the model weights.

The best obtained result has been 27.27% BLEU with 2,000 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 (27.17% 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.

7 *Conclusions*

This paper has described a sentence selection strategy for tuning a statistical machine translation system based on Moses that translates Spanish into English.

The proposed strategy consists of selecting the sentences of the development set that have more similarity with the sentences to translate (source test set). When using more similar sentences for tuning the weights using in the translation process, the tuned weights will allow obtaining better translation results.

In this work two alternatives for computing this similarity have been presented and evaluated. The first one consists of computing the negative of the perplexity of a given sentence compared to a language model trained with the source sentences of the test set. The second alternative is very similar by subtracting the probability of the sequence of unigrams (1-gram). The second alternative considers only how the similarity increases when considering 2-gram and 3-gram probabilities: removing the 1-gram effect as a normalization process.

In the experiments carried out in this work, the system performance in BLEU has increased from 27.17%, with the complete development set, to 27.27%.

Comparing both methods for computing the similarity, the normalized one obtains better results because this method is based on more reliable N-grams generating a better similarity measurement.

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 some domestic projects: TIMPANO (TIN2011-28169-C05-03), ITALIHA (CAM-UPM), INAPRA (MICINN, DPI2010-21247-C02-02), SD-TEAM (MEC, TIN2008-06856-C05-03) and MA2VICMR (Comunidad Autónoma de Madrid, S2009/TIC-1542) projects.

Authors also thank all the other members of the Speech Technology Group for the continuous and fruitful discussion on these topics.

References

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

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*.

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.

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

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.

Lü, Y., Huang, J., Liu, Q. 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.

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

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

Och, F. 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.

Papineni K., 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. 2002.

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*.