UPV-PRHLT English–Spanish system for WMT10

Germán Sanchis-Trilles and **Jesús Andrés-Ferrer** and **Guillem Gascó**
Jesús González-Rubio and **Pascual Martínez-Gómez** and **Martha-Alicia Rocha**
Joan-Andreu Sánchez and **Francisco Casacuberta**
Instituto Tecnológico de Informática
Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
{gsanchis|jandres|fcn}@dsic.upv.es
{ggasco|jegonzalez|pmartinez}@dsic.upv.es
{mrocha|jandreu}@dsic.upv.es

Abstract

In this paper, the system submitted by the PRHLT group for the Fifth Workshop on Statistical Machine Translation of ACL2010 is presented. On this evaluation campaign, we have worked on the English–Spanish language pair, putting special emphasis on two problems derived from the large amount of data available. The first one, how to optimize the use of the monolingual data within the language model, and the second one, how to make good use of all the bilingual data provided without making use of unnecessary computational resources.

1 Introduction

For this year's translation shared task, the Pattern Recognition and Human Language Technologies (PRHLT) research group of the Universidad Politécnica de Valencia submitted runs for the English–Spanish translation task. In this paper, we report the configuration of such a system, together with preliminary experiments performed to establish the final setup.

As in 2009, the central focus of the Shared Task is on Domain Adaptation, where a system typically trained using out-of-domain data is adjusted to translate news commentaries.

For the preliminary experiments, we used only a small amount of the largest available bilingual corpus, i.e. the United Nations corpus, by including into our system only those sentences which were considered similar.

Language model interpolation using a development set was explored in this work, together with a technique to cope with the problem of "out of vocabulary words".

Finally, a reordering constraint using walls and zones was used in order to improve the performance of the submitted system.

In the final evaluation, our system was ranked fifth, considering only primary runs.

2 Language Model interpolation

Nowadays, it is quite common to have very large amounts of monolingual data available from several different domains. Despite of this fact, in most of the cases we are only interested in translating from one specific domain, as is the case in this year's shared task, where the provided monolingual training data belonged to European parliamentary proceedings, news related domains, and the United Nations corpus, which consists of data crawled from the web.

Although the most obvious thing to do is to concatenate all the data available and train a single language model on the whole data, we also investigated a "smarter" use of such data, by training one language model for each of the available corpora.

3 Similar sentences selection

Currently, it is common to of huge bilingual corpora for SMT. For some common language pairs, corpora of millions of parallel sentences are available. In some of the cases big corpora are used as out-of-domain corpora. For example, in the case of the shared task, we try to translate a news text using a small in-domain bilingual news corpus (News Commentary) and two big out-of-domain corpora: Europarl and United Nations.

Europarl is a medium size corpus and can be completely incorporated to the training set. However, the use of the UN corpus requires a big computational effort. In order to alleviate this problem, we have chosen only those bilingual sentences from the United Nations that are similar to the in-domain corpus sentences. As a similarity measure, we have chosen the alignment score.

Alignment scores have already been used as a

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 172–176,
Uppsala, Sweden, 15-16 July 2010.

filter for noisy corpora (Khadivi and Ney, 2005). We trained an IBM model 4 using GIZA++ (Och and Ney, 2003) with the in-domain corpus and computed the alignment scores over the United Nations sentences. We assume that the alignment score is a good measure of similarity.

An important factor in the alignment score is the length of the sentences, so we clustered the bilingual sentences in groups with the same sum of source and target language sentence sizes. In each of the groups, the higher the alignment score is, the more similar the sentence is to the in-domain corpus sentences. Hence, we computed the average alignment score for each one of the clusters obtained for the corpus considered in-domain (i.e. the News-Commentary corpus). This being done, we assessed the similarity of a given sentence by computing the probability of such sentence with respect to the alignment model of the in-domain corpus, and established the following similarity levels:

- Level 1: Sentences with an alignment score equal or higher than the in-domain average.
- Level 2: Sentences with an alignment score equal or higher than the in-domain average, minus one standard deviation.
- Level 3: Sentences with an alignment score equal or higher than the in-domain average, minus two standard deviations.

Naturally, such similarity levels establish partitions of the out-of-domain corpus. Then, such partitions were included into the training set used for building the SMT system, and re-built the complete system from scratch.

4 Out of Vocabulary Recovery

As stated in the previous section, in order to avoid a big computational effort, we do not use the whole United Nations corpus to train the translation system. Out of vocabulary words are a common problem for machine translation systems. When translating the test set, there are test words that are not in the reduced training set (out of vocabulary words). Some of those out of vocabulary words are present in the sentences discarded from the United Nations Corpus. Thus, recovering the discarded sentences with out of vocabulary words is needed.

The out of vocabulary words recovery method is simple: the out of vocabulary words from the test, when taking into account the reduced training set, are obtained and then discarded sentences that contain at least one of them are retrieved. Then, those sentences are added to the reduced training set.

Finally, alignments with the resulting training set were computed and the usual training procedure for phrase-based systems was performed.

5 Walls and zones

In translation, as in other linguistics areas, punctuation marks are essential as they help to understand the intention of a message and organise the ideas to avoid ambiguity. They also indicate pauses, hierarchies and emphasis.

In our system, punctuation marks have been taken into account during decoding. Traditionally, in SMT punctuation marks are treated as words and this has undesirable effects (Koehn and Haddow, 2009). For example, commas have a high probability of occurrence and many possible translations are generated. Most of them are not consistent across languages. This introduces too much noise to the phrase tables.

(Koehn and Haddow, 2009) established a framework to specify reordering constraints with `walls` and `zones`, where commas and end of sentence are not mixed with various clauses. Gains between 0.1 and 0.2 of BLEU are reported. Specifying `zones` and `walls` with XML tags in input sentences allows us to identify structured fragments that the Moses decoder uses with the following restrictions:

1. If a <zone> tag is detected, then a block is identified and must be translated until a </zone> tag is found. The text between tags <zone> and </zone> is identified and translated as a block.
2. If the decoder detects a <wall/> tag, the text is divided into a prefix and suffix and Moses must translate all the words of the prefix before the suffix.
3. If both `zones` and `walls` are specified, then `local walls` are considered where the constraint 2 applies only to the area established by zones.

corpus	Language	$\|S\|$	$\|W\|$	$\|V\|$
Europarl v5	Spanish	1272K	28M	154K
	English		27M	106K
NC	Spanish	81K	1.8M	54K
	English		1.6M	39K

Table 1: Main figures of the Europarl v5 and News-Commentary (NC) corpora. K/M stands for thousands/millions. $|S|$ is the number of sentences, $|W|$ the number of running words, and $|V|$ the vocabulary size. Statistics are reported on the tokenised and lowercased corpora.

We used quotation marks, parentheses, brackets and dashes as zone delimiters. Quotation marks (when appearing once in the sentence), commas, colons, semicolons, exclamation and question marks and periods are used as wall delimiters.

The use of zone delimiters do not alter the performance. When using `walls`, a gain of 0.1 BLEU is obtained in our best model.

6 Experiments

6.1 Experimental setup

For building our SMT systems, the open-source SMT toolkit Moses (Koehn et al., 2007) was used in its standard setup. The decoder includes a log-linear model comprising a phrase-based translation model, a language model, a lexicalised distortion model and word and phrase penalties. The weights of the log-linear interpolation were optimised by means of MERT (Och, 2003). In addition, a 5-gram LM with Kneser-Ney (Kneser and Ney, 1995) smoothing and interpolation was built by means of the SRILM (Stolcke, 2002) toolkit.

For building our baseline system, the News-Commentary and Europarl v5 (Koehn, 2005) data were employed, with maximum sentence length set to 40 in the case of the data used to build the translation models, and without restriction in the case of the LM. Statistics of the bilingual data can be seen in Table 1.

In all the experiments reported, MERT was run on the 2008 test set, whereas the test set 2009 was considered as test set as such. In addition, all the experiments described below were performed in lowercase and tokenised conditions. For the final run, the detokenisation and recasing was performed according to the technique described in the Workshop baseline description.

corpus	$\|S\|$	$\|W\|$	$\|V\|$
Europarl	1822K	51M	172K
NC	108K	3M	68K
UN	6.2M	214M	411K
News	3.9M	107M	512K

Table 2: Main figures of the Spanish resources provided: Europarl v5, News-Commentary (NC), United Nations (UN) and News-shuffled (News).

6.2 Language Model interpolation

The final system submitted to the shared task included a linear interpolation of four language models, one for each of the monolingual resources available for Spanish (see Table 2). The results can be seen in Table 3. As a first experiment, only the in-domain corpus, i.e. the News-Commentary data (`NC data`) was used for building the LM. Then, all the available monolingual Spanish data was included into a single LM, by concatenating all the data together (`pooled`). Next, in `interpolated`, one LM for each one of the provided monolingual resources was trained, and then they were linearly interpolated so as to minimise the perplexity of the 2008 test set, and fed such interpolation to the SMT system. We found out that weights were distributed quite unevenly, since the News-shuffled LM received a weight of 0.67, whereas the other three corpora received a weight of 0.11 each. It must be noted that even the in-domain LM received a weight of 0.11 (less than the News-shuffled LM). The reason for this might be that, although the in-domain LM should be more appropriate and should receive a higher weight, the News-shuffled corpus is also news related (hence not really out-of-domain), but much larger. For this reason, the result of using only such LM (`News`) was also analysed. As expected, the translation quality dropped slightly. Nevertheless, since the differences are not statistically significant, we used the News-shuffled LM for internal development purposes, and the interpolated LM only whenever an improvement prooved to be useful.

6.3 Including UN data

We analysed the impact of the selection technique detailed in Section 3. In this case, the LM used was the interpolated LM described in the previous section. The result can be seen in Table 4. As it can be seen, translation quality as measured by

Table 3: Effect of considering different LMs

LM used	BLEU
NC data	21.86
pooled	23.53
interpolated	**24.97**
news	24.79

BLEU improves constantly as the number of sentences selected increases. However, further sentences were not included for computational reasons.

In the same table, we also report the effect of adding the UN sentences selected by our out-of-vocabulary technique described in Section 4. In this context, it should be noted that MERT was not rerun once such sentences had been selected, since such sentences are related with the test set, and not with the development set on which MERT is run.

Table 4: Effect of including selected sentences

system	BLEU
baseline	24.97
+ oovs	25.08
+ Level 1	24.98
+ Level 2	25.07
+ Level 3	25.13

6.4 Final system

Since the News-shuffled, UN and Europarl corpora are large corpora, a new LM interpolation was estimated by using a 6-gram LM on each one of these corpora, obtaining a gain of 0.17 BLEU points by doing so. Further increments in the n-gram order did not show further improvements.

In addition, preliminary experimentation revealed that the use of `walls`, as described in Section 5, also provided slight improvements, although using `zones` or combining both did not prove to improve further. Hence, only `walls` were included into the final system.

Lastly, the final system submitted to the Workshop was the result of combining all the techniques described above. Such combination yielded a final BLEU score of 25.31 on the 2009 test set, and 28.76 BLEU score on the 2010 test set, both in tokenised and lowercased conditions.

7 Conclusions and future work

In this paper, the SMT system presented by the UPV-PRHLT team for WMT 2010 has been described. Specifically, preliminary results about how to make use of larger data collections for translating more focused test sets have been presented.

In this context, there are still some things which need a deeper investigation, since the results presented here give only a small insight about the potential of the similar sentence selection technique described.

However, a deeper analysis is needed in order to assess the potential of such technique and other strategies should be implemented to explore new kids of reordering constraints.

Acknowledgments

This paper is based upon work supported by the EC (FEDER/FSE) and the Spanish MICINN under the MIPRCV "Consolider Ingenio 2010" program (CSD2007-00018),iTrans2 (TIN2009-14511) project, and the FPU scholarship AP2006-00691. This work was also supported by the Spanish MITyC under the erudito.com (TSI-020110-2009-439) project and by the Generalitat Valenciana under grant Prometeo/2009/014 and scholarships BFPI/2007/117 and ACIF/2010/226 and by the Mexican government under the PROMEP-DGEST program.

References

Shahram Khadivi and Hermann 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.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. *IEEE International Conference on Acoustics, Speech and Signal Processing*, II:181–184, May.

Philipp Koehn and Barry Haddow. 2009. Edinburgh's submission to all tracks of the WMT2009 shared task with reordering and speed improvements to Moses. In *The 4th EACL Workshop on Statistical Machine Translation*, ACL, pages 160–164, Athens, Greece, March. Springer.

P. Koehn et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of*

the ACL Demo and Poster Sessions, pages 177–180, Prague, Czech Republic.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

F.J. Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of ACL*, pages 160–167, Sapporo, Japan.

A. Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proc. of ICSLP'02*, pages 901–904, September.