Bilingual segmentation for phrasetable pruning in Statistical Machine Translation

Germán Sanchis-Trilles **Daniel Ortiz-Martínez** **Jesús González-Rubio**
Jorge González **Francisco Casacuberta**
Instituto Tecnológico de Informática
Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València
Valencia, Spain
{gsanchis,dortiz,jegonzalez,jgonzalez,fcn}@dsic.upv.es

Abstract

Statistical machine translation systems have greatly improved in the last years. However, this boost in performance usually comes at a high computational cost, yielding systems that are often not suitable for integration in hand-held or real-time devices. We describe a novel technique for reducing such cost by performing a Viterbi-style selection of the parameters of the translation model. We present results with finite state transducers and phrase-based models showing a 98% reduction of the number of parameters and a 15-fold increase in translation speed without any significant loss in translation quality.

1 Introduction

Nowadays, the key step of the process of statistical machine translation (SMT) involves inferring a large table of phrase pairs that are translations of each other from a large corpus of aligned sentences. The set of all phrase pairs, together with estimates of conditional probabilities and other useful features, is called *phrasetable*. Such phrases are applied during the decoding process, combining their target sides to form the final translation.

A variety of algorithms to extract phrase pairs has been proposed (Och and Ney, 2000; Marcu and Wong, 2002; Zens et al., 2002; Och and Ney, 2003; Vogel, 2005). Typically, these algorithms heuristically collect a highly redundant set of phrases from each training sentence pair generating phrasetables with a huge number of elements.

This bulk comes at a cost. Large phrasetables lead to large data structures that require more resources and more time to process. More importantly, effort spent in handling large tables could likely be more usefully employed in more features or more sophisticated search processes. Additionally, this is the main restriction for the widespread application of SMT techniques in small portable devices like cell phones, PDAs or hand-held game consoles; one can imagine many scenarios that could benefit from a lightweight translation device: tourism, medicine, military, etc.

In this paper, we show that is possible to prune phrasetables by removing those phrase pairs that have little influence on the final translation performance. Our approach consist in selecting only those phrase pairs extracted from the most probable segmentation of the training sentences.

The technique presented here has several advantages. It does not depend on the actual algorithm used to extract the phrase pairs, therefore can be applied to every imaginable method that assigns probabilities to phrase pairs. It provides a straightforward method for pruning the phrasetables, without the need of adjusting any additional parameter. It does not significantly affect translation quality, as measured by BLEU or TER scores, while very substantial savings in terms of computational requirements are reported.

The rest of the paper is organised as follows. Section 2 revised previously published techniques to prune the phrasetable. Section 3 introduces SMT and the different models used in the experimentation. Section 4 reviews the bilingual segmentation problem in order to present our technique to filter the phrasetable. Section 5 describes the experimentation carried out and presents the obtained results. The paper concludes with a summary and discussion of the results.

Mikel L. Forcada, Heidi Depraetere, Vincent Vandeghinste (eds.)
Proceedings of the 15th Conference of the European Association for Machine Translation, p. 257–264
Leuven, Belgium, May 2011

2 Related work

Most phrase-based decoders already include several built-in thresholds in order to prune the size of phrasetables estimated from training corpora (Ortiz et al., 2005; Koehn et al., 2007). They are usually related either to absolute scores of phrase pairs in the phrasetable or to relative scores between the phrase pairs sharing their source phrase.

Apart from phrasetable threshold pruning techniques, which are usually employed in SMT, different complementary methods in order to reduce even more the size of phrasetables have been explored within the last years. On the one hand, Johnson et al. (2007) propose to use significance testing in order to select only those phrase pairs which are the most co-occurring ones in the training corpus. On the other hand, Eck et al. (2007) considers usage statistics of phrase pairs, also based on either their scores or their ranks, in order to prune the ones below some minimal values.

Our work however does not perform an explicit statistical analysis of the phrases in phrasetables, but instead uses the concept of bilingual segmentation of each sentence pair to greatly reduce the number of parameters to be included in the final phrasetable. González et al. (2008) already proposed a segmentation-based technique using phrasetables which indirectly causes a reduction in their sizes. This technique was adopted by Sanchis-Trilles and Casacuberta (2008) in order to take advantage of the phrasetable pruning concept within a standard, phrasetable-based SMT system. Similarly, Wuebker et al. (2010) propose the use of a single bilingual segmentation in order to re-estimate translation probabilities by leaving-one-out. As a side effect, the amount of model parameters is also reduced. In our work however, the goal of reducing the size of phrasetables is directly targeted, thus achieving much larger reductions.

3 Statistical machine translation

Statistical Machine Translation (SMT) was defined by Brown et al. (1993) as follows: given a sentence $\mathbf{x}$ from a certain source language, a corresponding sentence $\hat{\mathbf{y}}$ in a given target language that maximises the posterior probability is to be found. State-of-the-art SMT systems model the translation distribution $p(\mathbf{y}|\mathbf{x})$ via the log-linear approach (Och and Ney, 2002):

$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\mathbf{y}} \Pr(\mathbf{y}|\mathbf{x}) \quad (1)$$

$$\approx \operatorname*{argmax}_{\mathbf{y}} \sum_{m=1}^{M} \lambda_m h_m(\mathbf{x}, \mathbf{y}) \quad (2)$$

where $h_m(\mathbf{x}, \mathbf{y})$ is a function representing an important feature for the translation of $\mathbf{x}$ into $\mathbf{y}$, M is the number of features (or models) and λ_m are the weights of the log-linear combination.

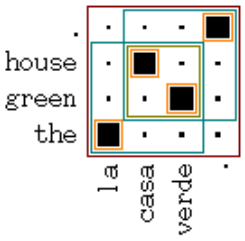
Current SMT systems are strongly based on the concept of *phrase*. A phrase is defined as a consecutive group of words of the source or the target sentences. In this work, we will conduct our experiments on two different machine translation models based on phrases: *phrase-based* (PB) models and *phrase-based stochastic finite state transducers* (PBSFSTs).

PB models (Tomas and Casacuberta, 2001; Och and Ney, 2002; Marcu and Wong, 2002; Zens et al., 2002), constitute the core of the current state-of-the-art in SMT. The basic idea of PB models is to segment the source sentence into phrases, then to translate each source phrase into a target phrase, and finally to reorder them in order to compose the final translation in the target language. The set of feature functions that compose the log-linear model used by state-of-the-art PB-SMT systems typically include an n-gram language model, phrase-based models estimated in both translation directions and some additional components such as word or phrase penalties. The word and phrase penalties allow the SMT system to limit the number of words or target phrases, respectively, that compose the translations of the source sentences.

PBSFSTs (González et al., 2008) are defined as a set of states, a set of labelled transitions between pairs of states (where labels are composed of a source phrase and a target phrase), and probabilistic distributions for the initial and the final states, and for the labelled transitions (Vidal et al., 2005). The inference of PBSFSTs is based on the use of monotonic bilingual segmentations of parallel training data and a language model of bilingual phrases (Casacuberta and Vidal, 2004). These models can also implement the log-linear approach as described for PB models, which the aforesaid PB bilingual language model is incorporated to as an additional feature.

4 Phrasetable pruning by bilingual segmentation

The problem of segmenting a bilingual sentence pair in such a manner that the resulting segmentation is the one that contains, without overlap, the



source phrase	target phrase
La	the
casa	house
verde	green
casa verde	green house
La casa verde	the green house
.	.
casa verde .	green house .
La casa verde .	the green house .

Figure 1: Consistent bilingual phrases (right) given a word alignment matrix (left).

best phrases that can be extracted from that pair is a difficult problem. First, because of the huge number of possible segmentations that are to be considered. Second, because a measure of optimality must be established. Consider the example:

Source: *La casa verde .*
Target: *The green house .*

When considering this example, one would probably state that a good segmentation for this bilingual pair is {{*La, The*}, {*casa verde , green house*}, {. , .}} . However, why is such a segmentation better than {{*La , The*},{*casa verde . , green house* .}}? As humans, we could argue with more or less convincing linguistic terms in favour of the first option, but that does not necessarily mean that such a segmentation is the most appropriate one for SMT. Furthermore, one could possibly think of several linguistically motivated segmentations for this small example.

In SMT, a variety of algorithms to extract phrase pairs have been proposed (Tomas and Casacuberta, 2001; Marcu and Wong, 2002; Och and Ney, 2003; Vogel, 2005). Typically, the bilingual phrases that compose phrasetables are extracted by using a heuristic algorithm (Zens et al., 2002). Such heuristic algorithm is driven by the following constraint: bilingual phrases must be *consistent* with their corresponding word alignment matrix. A phrase pair constitutes a consistent bilingual phrase if all aligned words in the source phrase are aligned with words of the target phrase and vice versa. Figure 1 exemplifies this phrase extraction process, together with the bilingual phrases extracted for a simple sentence. As shown, this process generates huge phrasetables with highly redundant phrase pairs.

The main purpose of this paper is to reduce the extremely high redundancy in the amount of phrase-pairs that current state-of-the-art SMT systems contain. With this purpose, we examine two different methods to obtain one single segmentation per sentence pair. These two methods rely on the concept of bilingual segmentation.

4.1 Bilingual segmentation

In SMT, the concept of bilingual segmentation can be easily derived from a phrase-based alignment, which can be stated formally as follows let $\mathbf{x}$ be a source sentence and $\mathbf{y}$ the corresponding target sentence in a bilingual corpus. A phrase-alignment between $\mathbf{x}$ and $\mathbf{y}$ is defined as a set $\mathcal{S}$ of ordered segment pairs included in $\mathcal{P}(\mathbf{x}) \times \mathcal{P}(\mathbf{y})$, where $\mathcal{P}(\mathbf{x})$ and $\mathcal{P}(\mathbf{y})$ are the set of all subsets of consecutive sequences of words, of $\mathbf{x}$ and $\mathbf{y}$, respectively. In addition, the ordered pairs contained in $\mathcal{S}$ have to include all the words of both the source and target sentences, without overlap. A phrase-based alignment $\tilde{A}(\mathbf{x}, \mathbf{y})$ of length K of a sentence pair $(\mathbf{x}, \mathbf{y})$ is defined as a specific one-to-one mapping $\tilde{\mathbf{a}}$ between $\mathcal{P}(\mathbf{y})$ and $\mathcal{P}(\mathbf{x})$. Then, the problem of finding the best PB-alignment $\tilde{A}_V(\mathbf{x}, \mathbf{y})$ (or Viterbi phrase-alignment) between $\mathbf{x}$ and $\mathbf{y}$ can be stated formally as

$$\tilde{A}_V(\mathbf{x}, \mathbf{y}) = \underset{\tilde{\mathbf{a}}}{\operatorname{argmax}}\, p(\tilde{\mathbf{a}}|\mathbf{x}, \mathbf{y}) \qquad (3)$$

One would suggest that we can perform a search process using a regular SMT system which filters its PT to obtain those translations of $\mathbf{x}$ that are compatible with $\mathbf{y}$. Unfortunately, such problem cannot be easily solved, since standard estimation tools such as Thot (Ortiz et al., 2005) and Moses (Koehn et al., 2007) do not guarantee complete coverage of sentence pairs seen in training due to the large number of heuristic decisions involved in the estimation process. This means that it is often the case that the SMT system is not able to produce the correct output sentence $\mathbf{y}$. This problem is exemplified in Figure 2. In this example, which has been extracted from a real training procedure, only three phrase pairs will be extracted, and the remaining words will not be included into the PT. It is shown that words such as `cannot`

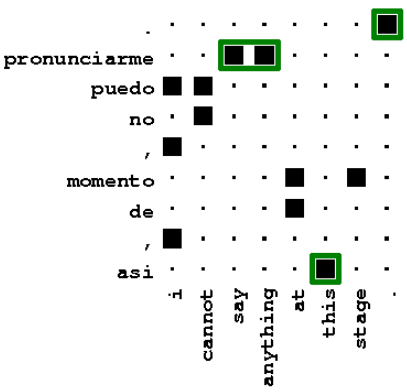


Figure 2: Example of word alignment that results in coverage problems. Maximum phrase length of 7 is assumed. Black squares represent word alignments, whereas extracted phrases are marked with a rectangle involving one or more squares.

present multiple alignments. In order to include target word `cannot` within a consistent alignment, one would need to include word `puedo` into the alignment, but including word `puedo` implies that word `I` is also included. Including `I` also forces the two commas to be included, together with whatever words appear between both. Continuing with this procedure leads to the necessity of including the whole sentence pair (except for the final dot) as a phrase before being able to include `cannot` into a consistent alignment. However, due to performance reasons, it is quite common to restrict the maximum length of the phrases to be extracted. If such maximum is set to e.g. 7, the complete sentence pair will not be included into the system, and `cannot` will remain unknown despite having been observed in training.

We propose two different solutions to this problem. The first one pursues the goal of obtaining *true* phrase-based alignments between $\mathbf{x}$ and $\mathbf{y}$, whereas the second one focuses on the primary goal of this work, i.e. reducing the amount of bilingual phrases derived from each sentence pair, leading to a *source-driven* bilingual segmentation.

4.2 True bilingual segmentation

As described in the previous section, coverage problems inherent to state-of-the-art SMT systems imply that it is often impossible to obtain the Viterbi segmentation of a given sentence pair. For this reason, a possible way of overcoming such coverage problems is proposed in (Ortiz-Martínez et al., 2008). In their work, the main idea is to consider every source phrase of $\mathbf{x}$ as a possible translation of every target phrase of $\mathbf{y}$. For this purpose, a general mechanism to assign probabilities to phrase pairs is needed, regardless if they are contained in the phrasetable or not.

Such mechanism can be implemented by means of the application of smoothing techniques over the phrasetable. As shown in (Foster et al., 2006), well-known language model smoothing techniques can be imported into the PB translation framework, and these can also be applied to obtain phrase-level alignments. According to (Ortiz-Martínez et al., 2008), the best smoothing techniques combine a maximum likelihood phrase-based model statistical estimator with a lexical distribution by means of linear interpolation or backing-off. The lexical distribution uses an IBM 1 alignment model (Brown et al., 1993) that allows to decompose phrase-to-phrase translation probabilities into word-to-word translation probabilities. In our experiments, we have combined a phrase-based statistical estimator with a lexical distribution by means of linear interpolation. In addition, (Ortiz-Martínez et al., 2008) also proposes the use of a log-linear model to control different aspects of the segmentation, such as the number of phrases in which the sentences are divided, the length of the source and the target phrases, the re-orderings and so on. In this work we have also adopted this strategy. Hence Equation 3 can be rewritten as:

$$\begin{aligned} \tilde{A}_V(\mathbf{x},\mathbf{y}) &= \operatorname*{argmax}_{\tilde{\mathbf{a}}} p(\tilde{\mathbf{a}}|\mathbf{x},\mathbf{y}) \\ &= \operatorname*{argmax}_{\tilde{\mathbf{a}}} \frac{p(\tilde{\mathbf{a}},\mathbf{y}|\mathbf{x})}{p(\mathbf{x}|\mathbf{y})} \\ &= \operatorname*{argmax}_{\tilde{\mathbf{a}}} p(\tilde{\mathbf{a}},\mathbf{y}|\mathbf{x}) \end{aligned} \quad (4)$$

Although it might seem that Equation 4 matches exactly the decoding problem in SMT, this is not so, since the maximisation takes place only over phrase-alignments, and is subject to the constraint that $\mathbf{y}$ is the actual reference sentence given.

Once the scoring function for phrase pairs has been defined, a search algorithm to find the bilingual segmentations is required. For this purpose, a search strategy based on the well-known *stack-decoding* algorithm (Jelinek, 1969) can be used.

The bilingual segmentation procedure that has been described above allows us to compute one true segmentation for each sentence pair. Once the segmentations for every sentence pair have been computed, it is possible to build a phrasetable by

only taking into account those segments that are contained in the set of true segmentations.

4.3 Source-driven bilingual segmentation

As it has been explained in Section 4.1, computing $\tilde{A}_V(\mathbf{x}, \mathbf{y})$ according to a given phrasetable is not an easy task. Specifically, the phrase alignments cannot often be generated due to coverage problems of the phrase-based alignment model. In the previous section it has been shown how to compute a true phrase-alignment between two given sentences. However, such method must bear with the constraint of having the output sentence fixed. Although such restriction seems logical at training time, it should not be underestimated that this will not be the case in translation time, and such restriction may introduce a non-intended bias. The bilingual segmentation technique described in Section 4.2 allows to overcome coverage problems by combining smoothing techniques with an appropriate search algorithm. This is done at the cost of modifying the scoring function used during the search process due to the application of smoothing techniques, and also by introducing new segment pairs. As said in Section 3, phrase-extraction is typically done by a heuristic algorithm, which has proved to provide appropriate bilingual segments, and altering such segments may not be a good idea.

Since our goal is to discard unnecessary segment pairs contained in the phrasetable, we propose an alternative bilingual segmentation technique that obtains *source-driven* bilingual segmentations, by relaxing the restriction considered in Equation 4, leading to

$$\tilde{A}_V(\mathbf{x}) \approx \underset{\tilde{\mathbf{a}},\mathbf{y}}{\operatorname{argmax}}\, p(\tilde{\mathbf{a}}, \mathbf{y}|\mathbf{x}) \qquad (5)$$

where the output sentence $\mathbf{y}$ is allowed to be different from the true reference, and the segmentation has been induced by taking into account only the input sentence. By using $\tilde{A}_V(\mathbf{x})$ instead of $\tilde{A}_V(\mathbf{x}, \mathbf{y})$, we ensure that only segments present in the current phrasetable are used, and no new segments are introduced.

The maximisation described in Equation 5 is exactly the same problem as the one of finding the best translation of a source sentence within a phrase-based system. Hence, for computing $\tilde{\mathbf{a}}$ we simply translate each source training sentence and include into the phrasetable those phrase pairs that compose the output hypothesis. We are aware that translating the source sentence will not necessarily produce the target sentence in the training pair, but on the other hand no artificial bilingual segments will be introduced into the phrasetable. In addition, as shown in Section 5, experiments show that this approach might be good enough to prune the PT without a significant loss in translation quality.

5 Experimental Setup

Both true and source-driven segmentations were conducted by means of a yet unpublished extension of the Thot (Ortiz et al., 2005) toolkit, which features a log-linear model and includes a state-of-the-art decoder and a phrase-based aligner, used here to obtain true alignments. Although such toolkit does not include lexical-based probabilities or a lexical-based distortion model, Sanchis-Trilles and Casacuberta (2008) show that the relationship between the baseline system and the reduced system via source-driven segmentation also holds for the Moses toolkit. The weights of the log-linear model were optimised by means of MERT (Och, 2003). This log-linear model includes direct and inverse phrase-based translation models, a language model and word and phrase penalties.

Once the source-driven or true segmentation is obtained, the new phrase pairs were used to build new phrasetables and new PBSFSTs. The probabilities assigned to the extracted segment pairs are obtained by normalising for the whole set of parameters resulting from the segmentation process.

Although PBSFSTs have the potential to use a log-linear combination of features to estimate $\Pr(\mathbf{y}|\mathbf{x})$, they were only used here to model the joint probability distribution $\Pr(\mathbf{x}, \mathbf{y})$, allowing us to determine the baseline associated to the segmentation method employed.

5.1 System evaluation and corpora

In this work, we measure the translation quality by means of BLEU and TER scores. BLEU measures the precision of n-grams (Papineni et al., 2001), whereas TER (Snover et al., 2006) is an error metric that computes the minimum number of edits required to modify the system hypotheses so that they match the references. In addition to this, we will also report the number of parameters that are used by the translation system and the *speedup* of the proposed system with respect to a conventional system. We define the speedup by means of the formula $S_p = T_b/T_r$, where T_b is the time taken by the baseline system and T_r is the time taken by

	Subset features	De	En	Es	En
Training	Sentences	751k		731k	
	Run. words	15.3M	16.1M	15.7M	15.2M
	Mean length	20.3	21.4	21.5	20.8
	Vocabulary	195k	66k	103k	64k
Dev.	Sentences	2000		2000	
	Run. words	55k	59k	61k	59k
	Mean length	27.6	29.3	30.3	29.3
	OoV words	432	125	208	127
Test	Sentences	3064		3064	
	Run. words	82k	85k	92k	85k
	Mean length	26.9	27.8	29.9	27.8
	OoV words	1020	488	470	502

Table 1: Main figures of the Europarl corpus. *OoV* stands for Out of Vocabulary, k for thousands of elements, and M for millions of elements.

the system with reduced PT.

We conducted our experiments on the Europarl corpus (Koehn, 2005), with the partition established in the Workshop on SMT of NAACL 2006 (Koehn and Monz, 2006). The Europarl corpus (Koehn, 2005) is built from the proceedings of the European Parliament published on the web, and was acquired in eleven different languages. We will only focus on the German–English (De–En) and Spanish–English (Es–En) tasks, since experiments with other language pairs yielded similar results. The corpus is divided into four separate sets: one for training, one for development, one for test and another test set which was the one used in the workshop for the final evaluation and included a *surprise* out-of-domain subset. We performed experiments on both test sets, yielding similar results for both of them. Because of this, and to avoid an overwhelming number of results, we only report those results obtained with the final evaluation test set, being these more interesting because of the out-of-domain data involved. The figures of the corpus are shown in Table 1.

6 Results

In the tables shown in this section, sizes are given in number of entries in the PT or number of transitions of PBSFSTs. Speed is reported in words per second (w/s), and S_p stands for *speedup*, as described in Section 5.1.

Confidence intervals at a confidence level of 95% were computed, following the bootstrap technique described by Koehn (2004). These turned to be, in every case and for BLEU and TER, around 0.65 points, and are omitted for the sake of clarity.

6.1 Phrase-based models

We carried out translation experiments using both source-driven bilingual segmentation and true bilingual segmentation. Results for both proposals and baseline system are displayed in Table 2.

In the case of the source-driven segmentation, translation quality is not significantly affected by the reduction of the size of the phrasetable we propose. On the one hand BLEU scores, are slightly lower than those of the baseline system, although confidence tests show that these differences are not statistically significant. On the other hand, TER scores seem to remain completely unaltered, even though a very slight variation can be observed

As for the number of parameters of the models used, it can be seen that such number is reduced in two orders of magnitude, i.e. the number of parameters remaining in the phrase table after applying our pruning technique is only around 2% the original number of parameters. Moreover, translation speed is increased by a factor between 9 and 16, all this without a significant loss in translation quality.

In the case of true segmentation, and as opposed to source-driven segmentation, translation quality does drop significantly (although not consistently) with respect to the baseline, ranging from 0.5 to 4.4 BLEU points and from 0.2 to 5.1 TER points.

6.2 Phrase-based SFSTs

Since our PBSFST estimation framework is based on the use of monotonic bilingual segmentations, there is no chance for the above-mentioned baseline setup to be applied given that it relies on multiple overlapping segmentations for each bilingual sentence pair. However, both segmentation techniques proposed here could actually be employed. As Section 6.1 has shown that source-driven segmentation method performs best, only these experiments were carried out then for PBSFSTs. The corresponding results are presented in Table 3.

Although baseline PB models are able to provide better translation quality, it must be stressed that, as described in Section 5, PBSFSTs were used to take into account only one feature model whereas PB models were a combination of five. Therefore, the differences between PBSFSTs and PB models may be welcome as an interesting

	Baseline				Source-driven					True				
Pair	BLEU	TER	size	w/s	BLEU	TER	size	w/s	S_p	BLEU	TER	size	w/s	S_p
Es–En	28.2	56.0	5.0	93	27.5	56.2	0.05	1500	16	23.8	60.8	0.07	380	4
En–Es	27.6	56.6	5.1	76	27.2	56.6	0.12	700	9	24.7	60.1	0.16	250	3
De–En	21.6	64.8	4.2	100	21.1	64.8	0.06	1500	15	17.5	69.9	0.22	280	3
En–De	15.2	70.9	5.5	46	15.1	70.2	0.14	400	9	14.7	71.1	0.31	170	4

Table 2: Translation quality, number of model parameters, number of translated words per second and speedup (S_p) obtained when using a PB translation system for both source-driven and true segmentation techniques. Monotonic search was considered. PB model size is given in millions of phrase-pairs.

	Source-driven				
Pair	BLEU	TER	size	w/s	S_p
Es–En	25.8	58.2	0.12	91730	986
En–Es	25.3	59.0	0.23	28411	374
De–En	18.8	68.3	0.12	41249	412
En–De	13.0	74.1	0.28	14205	309

Table 3: Translation quality, number of model parameters and number of translated words per second for the source-driven segmentation technique when using a PBSFST translation system. Size of PBSFSTs given in millions of single-word edges.

trade-off to achieve acceptable quality performance with a further increase in translation speed. It must be remarked that PBSFSTs are able to translate any of the test sets in just a few seconds (vs. tens of minutes taken by baseline PB models).

7 Discussion and conclusions

In this paper, we have presented a technique to reduce the size of the phrasetables used in state-of-the-art SMT systems. Our approach consist on selecting the phase pairs given by the most probable segmentation of the training sentences. We propose two different segmentation techniques. Both segmentation techniques allow to obtain substantial reductions in the size of the phrasetables as well as in the time cost of the translation process. Particularly, source-driven segmentation leads to important improvements in terms of decoding speed without a significant loss in translation quality. We think that the reductions in spatial and time costs of the proposed techniques can significantly help to implement state-of-the-art translation models into hand-held devices.

It is worth noting that, unexpectedly, in the experiments we carried out, the true bilingual segmentation technique obtained worse results than the source-driven segmentation technique.

One key difference between the two proposed techniques consists in the degree of similarity of the pruned phrasetables obtained by the techniques with respect to the original phrasetable. Although the true bilingual segmentation allows to obtain a complete segmentation of the source and target sentences, this comes at the cost of introducing smoothing techniques. Hence, the resulting segmentations contain phrase pairs that are not present in the original phrasetable. In the experiments we carried out, the pruned phrasetables generated by the true bilingual segmentation contained a relatively high number of phrase pairs that were not present in the original phrasetables, ranging from 10% to 50% depending on the language pair. In contrast, the source-driven bilingual segmentation, since it merely consists in translating the source sentence, always generates a pruned phrasetable that is a true subset of the original phrasetable. This suggests that the true segmentation technique not only prunes the original phrasetable, but also has an important role in the estimation of new model parameters, which could be the reason for the degradation of the translation quality. Nevertheless, a further analysis of the impact of the smoothing techniques used by true bilingual segmentation is required to better understand why this technique is not performing as expected.

Acknowledgements

This paper is based upon work supported by the EC (FEDER/FSE) and the Spanish MICINN under projects MIPRCV "Consolider Ingenio 2010" (CSD2007-00018) and iTrans2 (TIN2009-14511). Also supported by the Spanish MITyC under the erudito.com (TSI-020110-2009-439) project, by the Generalitat Valenciana under grant Prometeo/2009/014, and by the UPV under grant 20091027.

The authors would also like to thank the anonymous reviewers for their constructive and detailed comments.

References

Brown, P.F., S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. 1993. The mathematics of machine translation. In *Computational Linguistics*, volume 19, pages 263–311, June.

Casacuberta, F. and E. Vidal. 2004. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30:205–225, June.

Eck, M., S. Vogel, and A. Waibel. 2007. Translation model pruning via usage statistics for statistical machine translation. In *Proc. of the North American Chapter of the Association for Computational Linguistics*, pages 21–24.

Foster, G., R. Kuhn, and H. Johnson. 2006. Phrasetable smoothing for statistical machine translation. In *Proc. of Empirical Methods in Natural Language Processing*, pages 53–61.

González, J., G. Sanchis, and F. Casacuberta. 2008. Learning finite state transducers using bilingual phrases. In *Proc. of Computational linguistics and intelligent text processing*, pages 411–422.

Jelinek, F. 1969. Fast sequential decoding algorithm using a stack. *IBM Journal of Research Development*, 13:675–685, November.

Johnson, J.H., J. Martin, G. Foster, and R. Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proc. of Empirical methods in natural language processing*, pages 967–975.

Koehn, P. and C. Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proc. of Workshop on Statistical Machine Translation*, pages 102–121, June.

Koehn, P., 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, and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proc. of Association for Computational Linguistics*, pages 177–180.

Koehn, P. 2004. Statistical significance tests for machine translation evaluation. In *Proc. of Empirical methods in natural language processing*, pages 388–395.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proc. of the Machine Translation Summit*, pages 79–86.

Marcu, D. and W. Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proc. of Empirical methods in natural language processing*, pages 133–139.

Och, F.J. and H. Ney. 2000. Improved statistical alignment models. In *Proc. of Association for Computational Linguistics*, pages 440–447.

Och, F.J. and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of Association for Computational Linguistics*, pages 295–302.

Och, F.J. and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29:19–51, March.

Och, F.J. 2003. Minimum error rate training for statistical machine translation. In *Proc. of Association for Computational Linguistics*, pages 160–167, July.

Ortiz, D., I. García-Varea, and F. Casacuberta. 2005. Thot: a toolkit to train phrase-based statistical translation models. In *Proc. of the Machine Translation Summit*, pages 141–148.

Ortiz-Martínez, D., I. García-Varea, and F. Casacuberta. 2008. Phrase-level alignment generation using a smoothed loglinear phrase-based statistical alignment model. In *Proc. of European Association for Machine Translation*, pages 158–167.

Papineni, K., S. Roukos, T. Ward, and W. Jing-Zhu. 2001. Bleu: A method for automatic evaluation of machine translation. In *IBM Research Report RC22176 (W0109-022)*.

Sanchis-Trilles, G. and F. Casacuberta. 2008. Increasing translation speed in phrase-based models via suboptimal segmentation. In *Proc. of Workshop on Pattern Recognition in Information Systems*, pages 135–143.

Snover, M, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proc. of Association for Machine Translation in the Americas*, pages 223–231.

Tomas, J. and F. Casacuberta. 2001. Monotone statistical translation using word groups. In *Proc. of the Machine Translation Summit*, pages 357–361.

Vidal, E., F. Thollard, F. Casacuberta C. de la Higuera, and R. Carrasco. 2005. Probabilistic finite-state machines - part II (in Section IV-A). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1025–1039.

Vogel, S. 2005. PESA: Phrase Pair Extraction as Sentence Splitting. In *Proc. of the Machine Translation Summit*, pages 251–258.

Wuebker, J., A. Mauser, and H. Ney. 2010. Training phrase translation models with leaving-one-out. In *Proc. of Association for Computational Linguistics*, pages 475–484.

Zens, Richard, Franz Josef Och, and Hermann Ney. 2002. Phrase-based statistical machine translation. In *Proc. of Advances in Artificial Intelligence*, pages 18–32.