Increasing Translation Speed in Phrase-based Models via Suboptimal Segmentation

Germán Sanchis-Trilles and Francisco Casacuberta

Instituto Tecnológico de Informática
Camino de Vera s/n, 46022 Valencia, Spain

Abstract. Phrase-Based Models constitute nowadays the core of the state of the art in the statistical pattern recognition approach to machine translation. Being able to introduce context information into the translation model, they usually produce translations whose quality is often difficult to improve. However, these models have usually an important drawback: the translation speed they are able to deliver is mostly not sufficient for real-time tasks, and translating a single sentence can sometimes take some minutes. In this paper, we describe a novel technique for reducing significantly the size of the translation table, by performing a Viterbi-style selection of the phrases that constitute the final phrase-table. Even in cases where the pruned phrase table contains only 6% of the segments of the original one, translation quality is not worsened. Furthermore, translation quality remains the same in the worst case, achieving an increase of 0.3 BLEU in the best case.

1 Introduction

The grounds of modern Statistical Machine Translation (SMT), a pattern recognition approach to Machine Translation, were established in [1], where the problem of machine translation was defined as following: given a sentence $\mathbf{x}$ from a certain source language, an adequate sentence $\hat{\mathbf{y}}$ that maximises the posterior probability is to be found. Such a statement can be specified with the following formula:

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}}\, Pr(\mathbf{y}|\mathbf{x}) \tag{1}$$

Applying the Bayes theorem on this definition, one can easily reach the next formula

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}}\, \frac{Pr(\mathbf{y}) \cdot Pr(\mathbf{x}|\mathbf{y})}{Pr(\mathbf{x})} \tag{2}$$

and, since we are maximising over t, the denominator can be neglected, arriving to

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}}\, Pr(\mathbf{y}) \cdot Pr(\mathbf{x}|\mathbf{y}) \tag{3}$$

where $Pr(\mathbf{y}|\mathbf{x})$ has been decomposed into two different probabilities: the *statistical language model* of the target language $Pr(\mathbf{y})$ and *the (inverse) translation model* $Pr(\mathbf{x}|\mathbf{y})$.

Sanchis-Trilles G. and Casacuberta F. (2008).
Increasing Translation Speed in Phrase-based Models via Suboptimal Segmentation.
In *Proceedings of the 8th International Workshop on Pattern Recognition in Information Systems*, pages 135-143

Although it might seem odd to model the probability of the source sentence given the target sentence, this decomposition has a very intuitive interpretation: the translation model $Pr(\mathbf{x}|\mathbf{y})$ will capture the word or phrase relations between both input and output language, whereas the language model $Pr(\mathbf{y})$ will ensure that the output sentence is a well-formed sentence belonging to the target language.

In the last years, SMT systems have evolved to become the present state of the art, specially since the up-rise of Phrase Based (PB) models. Introducing information about context, PB models have widely outperformed word based models [2, 3]. However, an important drawback of the systems which implement the former models is the enormous size the phase tables need, which has as consequence the high requirements such models need, in terms of space but also time. In this paper, we propose a novel technique for reducing the amount of segment pairs needed for translating a given test set.

Related work was performed by [4]. In this work, the authors present a method for reducing the phrase table by performing significance testing. Our work, however, does not perform a statistical analysis of the phrases in the phrase table, but instead uses the concept of optimal segmentation of each sentence pair to reduce significantly the amount of segments to be included in the final phrase table. In addition, we also perform a speed analysis of the different systems built, both before and after the reduction.

The rest of the paper is structured as follows: In Section 2 we will briefly review the main ideas of Phrase Based models. In Section 3 we propose the algorithm which has been used for pruning the phrase table. Section 4 presents the experiments we performed showing that BLEU and WER scores are not affected by the pruning. In Section 5 we analyse the results and give some insight on why this pruning can be performed. Lastly, we conclude on Section 6.

2 Phrase-based Models

Phrase based (PB) [5–8] models have succeeded to achieve predominance in the state of the art in SMT. One would only need to take a look at the most recent international competitions [2, 3] to realise that PB models have succeeded to achieve predominance in the state of the art in SMT. Under this framework, *phrases* (i.e. word sequences) are extracted automatically from a word-aligned bilingual corpus. Because of their nature, PB models make use of context information, which has led them to outperform single-word SMT models.

Common assumptions under PB models are that only sequences of contiguous words are considered, that the number of source phrase (or segment) is equal to the number of target segments, and that a given source segment is aligned with exactly one target segment. Hence, when learning a PB model, the purpose is to compute a *phrase translation table*, where each input phrase is assigned to one or more output phrases with a given probability.

In the last years, a wide variety of techniques to produce PB models have been researched and implemented [9]. Firstly, a direct learning of the probabilities of each segment was proposed [5, 6]. At the same time, heuristics for extracting all possible segmentations coherent with a word-aligned corpus [7], where the alignments were learnt by means of the GIZA++ toolkit [10], were also proposed. Other approaches

have been suggested, exploring more linguistically motivated techniques [11, 12]. In this paper, we report experiments using the heuristic, (word) alignment-based phrase extraction algorithm.

However, these models have an important drawback, which must be tackled with whenever being applied to real time tasks: PB models tend to produce huge phrase tables, which entail slow translation speeds. In this paper, we propose a Viterbi style reduction of the phrase table, as it is done in the Viterbi re-estimation of Hidden Markov Models, achieving size reductions of over 90% and multiplying translation speed, measured as words per second, by almost a factor of 10.

3 Phrase Table Reduction via Suboptimal 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 best phrases that can be extracted from that pair is a difficult problem. In the first place, because all possible segmentations must be considered, and this number is a combinatorial number. In the second place, because a measure of *"optimality"* must be established. Consider the following example:

Source: *The table is red .*
Target: *La mesa es roja .*

At the sight of this example, one would probably state that {{*The table , La mesa*}, {*is red, es roja*}, {. , .}} is a good segmentation for this bilingual pair. However, why is such a segmentation better than {{*The , La*},{*table is , mesa es*},{*red . , roja .*}}? 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, and, moreover, one could easily think of several *linguistically appropriate* segmentations of this small example. To overcome this problem, PB SMT systems are forced to extract a large number of possible overlapping segmentations, and hope that one of them will be useful. Obviously, such an aggressive approach is bound to be computationally costly, and decoding time greatly suffers because of this issue.

When considering all possible segmentations of a bilingual sentence pair and assuming a "bag of words" model for the target sentence, the probability $Pr(\mathbf{x}|\mathbf{y})$ in Equation 3 can be modelled as:

$$P(\mathbf{x}|\mathbf{y}) = \sum_{K} \sum_{\mu} \sum_{\gamma} \prod_{k=1}^{K} p(x_{\gamma_{k-1}+1}^{\gamma_k} | y_{\mu_{k-1}+1}^{\mu_k}) \tag{4}$$

where K is the number of bilingual segments into which each bilingual pair is divided, μ is the set of possible segmentations of the source sentence $\mathbf{x}$ and γ the set of possible segmentations of the target sentence $\mathbf{y}$. In this formula we have assumed monotonic translation, in which no word (or segment) reordering is performed for the sake of simplicity.

Our approach for solving the problem of the overwhelming amount of possible segmentations, and the consequent increase of the phrase table, is based on the concept of

Viterbi re-estimation [13]. Following this idea, we can approximate $P(\mathbf{x}|\mathbf{y})$ by changing the summations by maximisations:

$$P(\mathbf{x}|\mathbf{y}) \approx \hat{P}(\mathbf{x}|\mathbf{y}) = \max_{K} \max_{\mu} \max_{\gamma} \prod_{k=1}^{K} p(x_{\gamma_{k-1}+1}^{\gamma_k} | y_{\mu_{k-1}+1}^{\mu_k}) \tag{5}$$

Given that the phrase table establishes the probability of an input segment given a certain output segment, we can use the scores within the phrase table to compute $\hat{P}(\mathbf{x}|\mathbf{y})$, and then build a phrase table by only taking into account those segments used to compute the optimal segmentation of each bilingual sentence in the training corpus.

However, computing $\hat{P}(\mathbf{x}|\mathbf{y})$ according to a given phrase table is not an easy task: if we establish a certain maximum length for the segments contained in the phrase table, it is common that, due to non-monotonic alignments, certain words of a sentence will not be contained in the segments extracted. Observing all possible segments without constraining the maximum length is not a solution either, since the number of entries in the phrase table would grow too much. This implies that the phrase table has coverage problems even on the training set.

However, our intention is to discard unnecessary segment pairs contained in the phrase table. To this purpose, a *suboptimal* bilingual segmentation, in which we *translate* the source sentence, may be enough. We are aware, nevertheless, that translating the input sentence will not necessarily produce the output sentence in the training pair, but our experiments show that this might be good enough to prune the phrase table without a significant loss in translation quality.

4 Experiments

We conducted our experiments on the Europarl corpus [14], with the partition established in the Workshop on Statistical Machine Translation of the NAACL 2006 [15].

The Europarl corpus [14] is built from the proceedings of the European Parliament, which are published on the web, and was acquired in 11 different languages. However, in this work we will only focus on the German–English, Spanish–English and French–English tasks, since these were the language pairs selected for the cited workshop. 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. This test set will be referred to as "Test", whereas the test set provided for evaluation purposes outside the final evaluation will be referred to as "Devtest". It must be noted that the Test set included a surprise out-of-domain subset, and hence the translation quality on this set will be significantly lower. The characteristics of the corpus can be seen in Table 1. It might seem surprising that the average sentence length in the training set is significantly lower than in the rest of the subsets. This is due to the fact that, for the competition, the training corpus pruned to contain only those sentences with a maximum length of 40, whereas this restriction was not imposed on the other subsets. The translation systems were tuned using the development set with the MERT [16] optimisation procedure, where the measure to be optimised was BLEU [17].

We performed experiments on both test sets, yielding similar results for both of them. Because of this, and in order not to provide an overwhelming amount of results,

Table 1. Characteristics of the Europarl corpus.

		German	English	Spanish	English	French	English
Training	Sentences	751088		730740		688031	
	Running words	15.3M	16.1M	15.7M	15.2M	15.6M	13.8M
	Average length	20.3	21.4	21.5	20.8	22.7	20.1
	Vocabulary size	195291	65889	102886	64123	80349	61627
Development	Sentences	2000		2000		2000	
	Running words	55147	58655	60628	58655	67295	58655
	Average length	27.6	29.3	30.3	29.3	33.6	29.3
	Out of vocabulary	432	125	208	127	144	138
Devtest	Sentences	2000	2000	2000	2000	2000	
	Running words	54260	57951	60332	57951	66200	57951
	Average length	27.1	29.0	30.2	29.0	33.1	29.3
	Out of vocabulary	377	127	207	125	139	133
Test	Sentences	3064		3064		3064	
	Running words	82477	85232	91730	85232	100952	85232
	Average length	26.9	27.8	29.9	27.8	32.9	27.8
	Out of vocabulary	1020	488	470	502	536	519

we only report the results obtained with the Test set, being this result more interesting because of the out-of-domain data it contains.

4.1 Suboptimal Segmentation Filtering

As a baseline system, we used the same system as the one used in the workshop. To filter the phrase table as described in the previous section, we translated the whole training subcorpus using the baseline model, and kept only those entries of the phrase table which were used while doing this. Since the baseline system uses lexicalised reordering [18], we also filtered the reordering table according to the segments used. The result of this setup can be seen in Table 2.

In this table, the sizes are given in number of entries in the phrase table and the speed is given in words per second. *fsize* is the size of the phrase table after filtering out all segments which will not be needed for translating the current test set, which is usual when dealing with big phrase tables. In this context, it must be noted that the translation speed detailed in Table 2 was measured in all cases when translating using the filtered phrase table, since loading the complete phrase table into memory without any filtering is unfeasible with the baseline model. Moreover, the speed does not take into account the time the system needs to load the model files (i.e. phrase table and lexicalised reordering table), which is reduced in a factor of ten due to the difference in model size. S_p is the *speedup*, which is given by 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 the filtered system. The values appearing as "size red." in the table represent the *fsize* reduction in percentage with respect to the original *fsize*. Hence, this column displays the effective reduction of data loaded into the decoder when translating.

Translation quality, as measured with BLEU [17] is not affected by the reduction of the size of the phrase table we proposing. Moreover, we can see that, in the worst case,

Table 2. Performance comparison between the baseline system and our suboptimal-segmentation-reduced approach. Lexicalised reordering is considered. *Speed* is measured in number of translated source words per second, and *fsize* is the size of the phrase table when filtered for the test set.

	baseline					reduced						
pair	WER	BLEU	size	fsize	speed	WER	BLEU	size	fsize	speed	size red.	S_p
Es–En	57.8	30.6	19M	1.6M	5.3	57.5	30.9	1.9M	0.15M	13.1	91%	2.5
En–Es	57.5	30.3	19M	1.8M	5.7	57.4	30.6	1.7M	0.16M	11.3	92%	2.0
De–En	68.1	23.7	12M	1.1M	6.6	68.2	23.9	1.8M	0.18M	11.4	84%	1.7
En–De	72.5	16.4	13M	1.7M	4.3	72.4	16.5	1.9M	0.23M	9.0	86%	2.1
Fr–En	60.2	28.3	15M	1.6M	5.6	60.1	28.3	1.5M	0.12M	17.7	92%	3.2
En–Fr	60.5	30.5	16M	1.7M	4.5	60.1	30.9	1.6M	0.15M	9.5	91%	2.1

we get exactly the same score than with the baseline system, and in the best case we are improving BLEU by 0.35 points. As measured with WER, which is an adaptation of the edit distance used in Speech Recognition, the translation quality is slightly worsened in some cases (with a maximum of 0.1 points), and in some cases improved. The behaviour difference between BLEU and WER can be explained because of the measure being optimised in MERT, which was BLEU.

Although the differences named in the previous paragraph are not significant, it is important to stress that we are improving translation speed by a factor of 3.2 in the best case and 1.7 in the worst case, without a significant loss of translation quality even in cases where out-of-domain sentences were translated.

4.2 Increasing Translation Speed Further

Although the speeds achieved in the previous subsection are already competitive, they may not be enough for real time applications: translating an average sentence of 25 words may take more than two seconds, and this might not be enough for the user who is waiting for the translation.

A common resource for increasing translation speed is to consider only monotonic translation. Under this decoding strategy, a given bilingual segment must occupy the same position in both input and output sentences. For example, if the source part of a certain bilingual segment is placed at the start of the source sentence, it cannot be placed at the end of the target sentence (or anywhere else but at the start). Although it is true that some translation quality is lost by doing so, the difference is relatively small the language pairs considered in our work. Our phrase table reduction technique can also be applied to monotonic translation. The results for this setup are shown in Table 3, yielding, again, no significant worsening (or improvement) of the translation scores, but achieving speedups ranging from 3.2 to 9.5, depending mainly on the language pair chosen and when compared to the non-reduced monotonic search.

In this case, it must be emphasised that the *fsize* of the baseline is the same as in the case of the lexicalised reordering search, since the reordering has no effect on the number of phrases extracted. This is not so, however, with our suboptimal segmentation, since the monotonicity constraint is also imposed when obtaining the segments that will

Table 3. Performance comparison between the baseline system and our suboptimal-segmentation-reduced approach. Monotonic search is considered. *Speed* is measured in number of translated source words per second, and *fsize* is the size of the phrase table when filtered for the test set.

	baseline				reduced				
pair	WER	BLEU	fsize	speed	WER	BLEU	fsize	speed	S_p
Es–En	58.8	29.6	1.6M	17.6	58.4	29.7	0.13M	91.5	5.2
En–Es	58.5	29.2	1.8M	19.1	58.6	29.2	0.08M	125.0	6.5
De–En	68.9	22.6	1.1M	20.6	69.0	22.5	0.14M	107.0	5.2
En–De	73.1	16.0	1.7M	23.5	72.6	16.2	0.20M	80.0	3.4
Fr–En	60.3	27.6	1.6M	15.8	60.9	27.4	0.11M	147.0	9.3
En–Fr	61.7	29.4	1.7M	19.0	61.5	29.4	0.16M	74.7	3.9

Table 4. Performance as measured by BLEU and WER for the re-normalised system. Both monotonic and non-monotonic search are considered.

	baseline				re-normalised			
	monotonic		reordering		monotonic		reordering	
pair	WER	BLEU	WER	BLEU	WER	BLEU	WER	BLEU
Es–En	58.8	29.6	57.8	30.6	59.0	29.1	57.8	30.5
En–Es	58.5	29.2	57.5	30.3	58.8	29.0	57.6	30.4
De–En	68.9	22.6	68.1	23.7	69.1	22.5	68.3	23.8
En–De	73.1	16.0	72.5	16.4	72.7	16.3	72.7	16.4
Fr–En	60.3	27.6	60.2	28.3	61.0	27.2	60.2	28.1
En–Fr	61.7	29.4	60.5	30.5	61.8	29.3	60.4	30.9

be part of the final phrase table, which implies that fewer (but shorter) segments will be kept.

5 Analysis and Side Notes

A question which could be asked at this point is whether we can truly obtain the same translation quality by just taking into account the suboptimal segmentation, or rather what we are doing is simply a filtering, but we actually would need the probabilities contained within the complete phrase table. In order to clarify this, we re-normalised the phrase table, assigning to each segment the score obtained by only taking into account those phrase pairs contained within the reduced phrase table. In Table 4 we can see the results of performing such a renormalisation.

As can be seen in the table, the performance is not significantly affected by the renormalisation. In our opinion, this clearly reveals that computing the phrase translation probabilities by only taking into account the segments used to translate the training set obtains a similar result than taking into account all possible segmentations that are consistent with the word alignments, as is common in regular SMT systems. A possible interpretation is that those segments which were selected to stay in the final, filtered table are those which account for the biggest part of the probability mass.

Table 5. BLEU and WER scores for the Training set, with both monotonic and non-monotonic search.

pair	monotonic		reordering	
	WER	BLEU	WER	BLEU
Es-En	44.9	48.2	43.2	50.6
En-Es	47.1	46.3	44.8	49.4
De-En	53.9	41.6	51.8	43.6
En-De	55.6	37.9	55.6	37.9
Fr-En	46.7	45.9	46.9	46.0
En-Fr	51.5	44.4	46.4	49.8

Lastly, and since we already had translated the training set, we found interesting to compute the BLEU and WER scores over the training data. These scores, which can be seen in Table 5, constitute an upper bound of the score that could be achieved in the test set. However, these results are not as good as could be expected, which hints towards a relatively weak (but even though state-of-the-art) performance of the translation models and (or) decoding algorithm.

6 Conclusions and Future Work

In this work we have presented a straight-forward method for reducing the size of the phrase table by a factor of ten, and increasing translation speed up to nine times. By doing so, the translation quality as measured by WER and BLEU remains unaffected, for both in-domain and out-of-domain data. Given that translation speed is a serious issue in systems implementing phrase-based models, the approach presented in this paper provides an efficient solution for the problem.

As future work, we are planning on researching ways to obtain the optimal segmentation of the sentences in the training corpus, without going through the drawback of having to translate the corpus. This includes both segmenting the sentences according to a phrase table, and without having the phrase table as a starting point.

Acknowledgements

This work has been partially supported by the Spanish MEC under scholarship AP2005-4023 and under grant CONSOLIDER Ingenio-2010 CSD2007-00018, and by the EC (FEDER) and the Spanish MEC under grant TIN2006-15694-CO2-01.

References

1. Brown, P.F., Pietra, S.A.D., Pietra, V.J.D., Mercer, R.L.: The mathematics of machine translation. In: Computational Linguistics. Volume 19. (1993) 263–311
2. Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., Schroeder, J.: (meta-) evaluation of machine translation. In: Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic, Association for Computational Linguistics (2007) 136–158

3. Fordyce, C.S.: Overview of the IWSLT 2007 evaluation campagin. In: International Workshop on Spoken Language Translation, Trento, Italy (2007)
4. Johnson, J., Martin, J., Foster, G., Kuhn, R.: Improving translation quality by discarding most of the phrasetable. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Czech Republic (2007)
5. Tomas, J., Casacuberta, F.: Monotone statistical translation using word groups. In: Proceedings of the Machine Translation Summit VIII, Santiago de Compostela, Spain (2001) 357–361
6. Marcu, D., Wong, W.: Joint probability model for statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP02), Pennsylvania, Philadelphia, USA (2002)
7. Zens, R., Och, F., Ney, H.: Phrase-based statistical machine translation. In: Advances in artificial intelligence. 25. Annual German Conference on AI. Lecture Notes in Computer Science. Volume 2479. (2002) 18–32
8. Zens, R., Ney, H.: Improvements in phrase-based statistical machine translation. In: Proceedings of the Human Language Technology Conference (HLT-NAACL), Boston, USA (2004) 257–264
9. Koehn, P., Och, F., Marcu, D.: Statistical phrase-based translation. In: Proceedings of the 2003 Conf. of the NAACL on Human Language Technology. Volume 1., Edmonton, Canada (2003) 48–54
10. Och, F., Ney, H.: A systematic comparison of various statistical alignment models. In: Computational Linguistics. Volume 29. (2003) 19–51
11. Sánchez, J., Benedí, J.: Stochastic inversion transduction grammars for obtaining word phrases for phrase-based statistical machine translation. In: Proceedings of the Workshop on SMT, New York City (2006) 130–133
12. Watanabe, T., Sumita, E., Okuno, H.: Chunk-based statistical translation. In: Proceedings of the 41st. Annual Meeting of the ACL, Sapporo, Japan (2003)
13. Viterbi, A.: Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. **13** (1967) 260 – 269
14. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: MT Summit. (2005)
15. Koehn, P., Monz, C., eds.: Proceedings on the Workshop on Statistical Machine Translation. Association for Computational Linguistics, New York City (2006)
16. Och, F.: Minimum error rate training for statistical machine translation. In: ACL 2003: Proc. of the 41st Annual Meeting of the ACL, Sapporo, Japan (2003)
17. Papineni, Kishore, A., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY (2001)
18. Koehn, P., Axelrod, A., Mayne, A.B., Callison-Burch, C., Osborne, M., Talbot, D.: Edinburgh system description for the 2005 IWSLT speech translation evaluation. In: International Workshop on Spoken Language Translation, Pittsburgh, Pennsylvania, USA (2005)