Analysis of statistical and morphological classes to generate weighted reordering hypotheses on a Statistical Machine Translation system

Marta R. Costa-jussà and **José A. R. Fonollosa**
Department of Signal Theory and Communications
TALP Research Center (UPC)
Barcelona 08034, Spain
(mruiz,adrian)@gps.tsc.upc.edu

Abstract

One main challenge of statistical machine translation (SMT) is dealing with word order. The main idea of the statistical machine reordering (SMR) approach is to use the powerful techniques of SMT systems to generate a weighted reordering graph for SMT systems. This technique supplies reordering constraints to an SMT system, using statistical criteria.

In this paper, we experiment with different graph pruning which guarantees the translation quality improvement due to reordering at a very low increase of computational cost.

The SMR approach is capable of generalizing reorderings, which have been learned during training, by using word classes instead of words themselves. We experiment with statistical and morphological classes in order to choose those which capture the most probable reorderings.

Satisfactory results are reported in the WMT07 Es/En task. Our system outperforms in terms of BLEU the WMT07 Official baseline system.

1 Introduction

Nowadays, statistical machine translation is mainly based on phrases (Koehn et al., 2003). In parallel to this phrase-based approach, the use of bilingual n-grams gives comparable results, as shown by Crego et al. (2005). Two basic issues differentiate the n-gram-based system from the phrase-based: training data is monotonically segmented into bilingual units; and, the model considers n-gram probabilities rather than relative frequencies. The n-gram-based system follows a maximum entropy approach, in which a log-linear combination of multiple models is implemented (Mariño et al., 2006), as an alternative to the source-channel approach.

Introducing reordering capabilities is important in both systems. Recently, new reordering strategies have been proposed such as the reordering of each source sentence to match the word order in the corresponding target sentence, see Kanthak et al. (2005) and Mariño et al. (2006). These approaches are applied in the training set and they lack of reordering generalization.

Applied both in the training and decoding step, Collins et al. (2005) describe a method for introducing syntactic information for reordering in SMT. This approach is applied as a pre-processing step.

Differently, Crego et al. (2006) presents a reordering approach based on reordering patterns which is coupled with decoding. The reordering patterns are learned directly from word alignment and all reorderings have the same probability.

In our previous work (Costa-jussà and Fonollosa, 2006) we presented the SMR approach which is based on using the powerful SMT techniques to generate a reordered source input for an SMT system both in training and decoding steps. One step further, (Costa-jussà et al., 2007) shows how the SMR system can generate a weighted reordering graph, allowing the SMT system to make the final reordering decision.

In this paper, the SMR approach is used to train the SMT system and to generate a weighted reordering graph for the decoding step. The SMR system uses word classes instead of words themselves and we analyze both statistical and morphological classes. Moreover, we present experiments regarding the reordering graph efficiency: we analyze different graph pruning and we show the very low increase in computational cost (compared to a monotonic translation). Finally, we compare the performance our system in terms of BLEU with the WMT07 baseline system.

This paper is organized as follows. The first two sections explain the SMT and the SMR baseline systems, respectively. Section 4 reports the study of statistical and

Proceedings of the Second Workshop on Statistical Machine Translation, pages 171–176,
Prague, June 2007.

morphological classes. Section 5 describes the experimental framework and discusses the results. Finally, Section 6 presents the conclusions and some further work.

2 Ngram-based SMT System

This section briefly describes the Ngram-based SMT (for further details see (Mariño et al., 2006)). The Ngram-based SMT system uses a translation model based on bilingual n-grams. It is actually a language model of bilingual units, referred to as tuples, which approximates the joint probability between source and target languages by using bilingual n-grams. Tuples are extracted from any word alignment according to the following constraints:

1. a monotonic segmentation of each bilingual sentence pairs is produced,

2. no word inside the tuple is aligned to words outside the tuple, and

3. no smaller tuples can be extracted without violating the previous constraints.

As a result of these constraints, only one segmentation is possible for a given sentence pair.

In addition to the bilingual n-gram translation model, the baseline system implements a log-linear combination of feature functions, which are described as follows:

- A **target language model**. This feature consists of a 4-gram model of words, which is trained from the target side of the bilingual corpus.

- A **class target language model**. This feature consists of a 5-gram model of words classes, which is trained from the target side of the bilingual corpus using the statistical classes from (Och, 1999).

- A **word bonus function**. This feature introduces a bonus based on the number of target words contained in the partial-translation hypothesis. It is used to compensate for the system's preference for short output sentences.

- A **source-to-target lexicon model**. This feature, which is based on the lexical parameters of the IBM Model 1 (Brown et al., 1993), provides a complementary probability for each tuple in the translation table. These lexicon parameters are obtained from the source-to-target alignments.

- A **target-to-source lexicon model**. Similarly to the previous feature, this feature is based on the lexical parameters of the IBM Model 1 but, in this case, these parameters are obtained from target-to-source alignments.

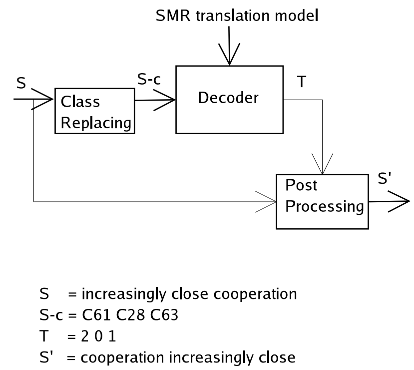

Figure 1: *SMR block diagram.*

3 SMR Baseline System

As mentioned in the introduction, SMR and SMT are based on the same principles.

3.1 Concept

The aim of SMR consists in using an SMT system to deal with reordering problems. Therefore, the SMR system can be seen as an SMT system which translates from an original source language (*S*) to a reordered source language (*S'*), given a target language (*T*).

3.2 Description

Figure 1 shows the SMR block diagram and an example of the input and output of each block inside the SMR system. The input is the initial source sentence (*S*) and the output is the reordered source sentence (*S'*). There are three blocks inside SMR: (1) the class replacing block; (2) the decoder, which requires an Ngram model containing the reordering information; and, (3) the post-processing block which either reorders the source sentence given the indexes of the decoder output 1-best (training step) or transforms the decoder output graph to an input graph for the SMT system (decoding step).

The decoder in Figure 1 requires a *translation* model which is an Ngram model. Given a training parallel corpus this model has been built following the next steps:

1. Select source and target word classes.

2. Align parallel training sentences at the word level in both translation directions. Compute the union of the two alignments to obtain a symmetrized many-to-many word alignment.

3. Use the IBM1 Model to obtain a many-to-one word alignment from the many-to-many word alignment.

4. Extract translation units from the computed many-to-one alignment. Replace source words by their

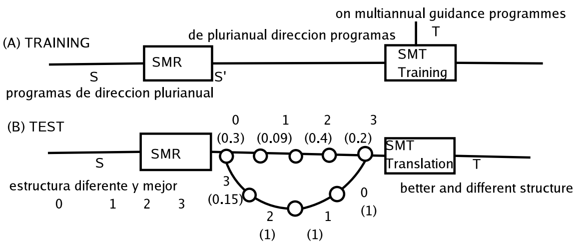

Figure 2: *SMR approach in the (A) training step (B) in the test step (the weight of each arch is in brackets).*

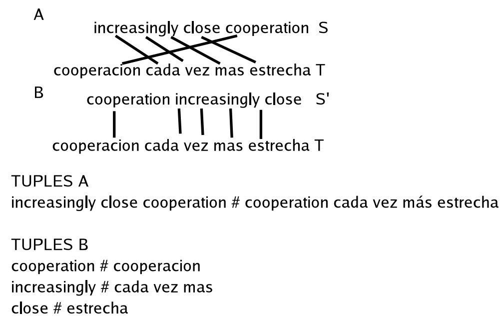

Figure 3: *Alignment and tuple extraction (A) original training source corpus (B) reordered training source corpus.*

classes and target words by the index of the linked source word. An example of a translation unit here is: *C61 C28 C63#2 0 1*, where # divides source (word classes) and target (positions).

5. Compute the sequence of the above units and learn the language model

For further information about the SMR training procedure see (Costa-jussà and Fonollosa, 2006).

3.3 Improving SMT training

Figure 2 (A) shows the corresponding block diagram for the training corpus: first, the given training corpus *S* is translated into the reordered training source corpus *S'* with the SMR system. Then, this reordered training source corpus *S'* and the given training target corpus *T* are used to build the SMT system

The main difference here is that the training is computed with the *S'2T* task instead of the *S2T* given task. Figure 3 (A) shows an example of the word alignment computed on the given training parallel corpus *S2T*. Figure 3 (B) shows the same links but with the reordered source training corpus *S'*. Although the quality in alignment is the same, the tuples that can be extracted change (notice that tuple extraction is monotonic). We now are able to extract smaller tuples which reduce the translation vocabulary sparseness. These new tuples are used to build the SMT system.

3.4 Generation of multiple weighted reordering hypotheses

The SMR system, having its own search, can generate either an output 1-best or an output graph. In decoding, the SMR technique generates an output graph which is used as an input graph by the SMT system. Figure 2 (B) shows the corresponding block diagram in decoding: the SMR output graph is given as an input graph to the SMT system. Hereinafter, this either SMR output graph or SMT input graph will be referred to as (weighted) reordering graph. The monotonic search in the SMT system is extended with reorderings following this reordering graph. This reordering graph has multiple paths and each path has its own weight. This weight is added as a feature function in the log-linear model.

4 Morphological vs Statistical Classes

Previous SMR studies (Costa-jussà and Fonollosa, 2006) (Costa-jussà et al., 2007) considered only statistical classes. On the one hand, these statistical classes performed fairly well and had the advantage of being suitable for any language. On the other hand, it should be taken into account the fact of training them in the training set allows for unknown words in the development or in the test set. Additionally, they do not have any reordering information because they are trained on a monolingual set.

The first problem, unknown words which appear in the development or in the test set, may be solved by using a disambiguation technique. Unknown words can be assigned to one class by taking into account their own context. The second problem, incorporating information about order, might be solved by training classes in the reordered training source corpus. In other words, we monotonized the training corpus with the alignment information (i.e. reorder the source corpus in the way that matches the target corpus under the alignment links criterion). After that, we train the statistical classes, hereinafter, called statistical reordered classes.

In some pair of languages, as for example English/Spanish, the reordering that may be performed is related to word's morphology (i.e. *TAGS*). Some *TAGS* rules (with some lexical exceptions) can be extracted as in (Popovic and Ney, 2006) where they were applied with reordering purposes as a preprocessing step. Another approach that has related *TAGS* and reordering was presented in (Crego and Mariño, 2006) where instead of rules, they learned reordering patterns based on *TAGS* as named in this paper's introduction. Hence, the SMR tech-

		Spanish	English
Train	Sentences	1,3M	
	Words	37,9M	35,5M
	Vocabulary	138,9k	133k
Dev	Sentences	2 000	2 000
	Words	60.5k	58.7k
	Vocabulary	8.1k	6.5k
Test	Sentences	2 000	2 000
	Words	60,2k	58k
	Vocabulary	8,2k	6,5k

Table 1: *Corpus Statistics.*

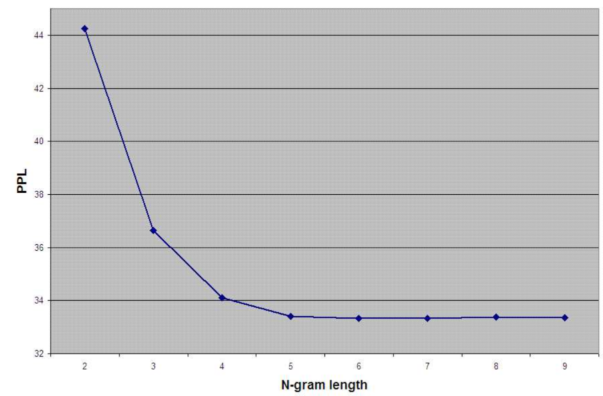

Figure 5: *Perplexity over the manually aligned test set given the SMR Ngram length.*

nique may take advantage of the morphological information. Notice that an advantage is that there is a *TAG* for each word, hence there are not unknown words.

5 Evaluation Framework

5.1 Corpus Statistics

Experiments were carried out using the data in the second evaluation campaign of the WMT07 [1].

This corpus consists in the official version of the speeches held in the European Parliament Plenary Sessions (EPPS), as available on the web page of the European Parliament. Additionally, there was available a smaller corpus (News-Commentary). Our training corpus was the catenation of both. Table 1 shows the corpus statistics.

5.2 Tools and preprocessing

The system was built similarly to (Costa-jussà et al., 2007). The SMT baseline system uses the Ngram-based approach, which has been explained in Section 2. Tools used are defined as follows: word alignments were computed using GIZA++ [2]; language model was estimated using SRILM [3]; decoding was carried out with MARIE[4]; an n-best re-ranking strategy is implemented which is used for optimization purposes just as proposed in *http://www.statmt.org/jhuws/* using the simplex method (Nelder and Mead, 1965) and BLEU as a loss function.

The SMT system we use a 4gram translation language model, a 5gram target language model and a 5gram class target language model.

Spanish data have been processed so that the pronouns which are attached to verbs are split up. Additionally, several article and prepositions words are separated (i.e. *del* goes into *de el*). This preprocessing was performed using Freeling software (Atserias et al., 2006). Training and evaluation were both true-case.

5.3 Classes and Ngram length Study for the SMR-Graph generation

This section evaluates several types of classes and n-gram lengths in the SMR model in order to choose the SMR configuration which provides the best results in translation in terms of quality. To accomplish this evaluation, we have designed the following experiment. Given 500 manually aligned parallel sentences of the EPPS corpora (Lambert et al., 2006), we order the source test in the way that better matches the target set. This ordered source set is considered our reference as it is based on manual alignments. On the other hand, the 500 sentences set is translated using the SMR configurations to be tested. Finally, the *Word Error Rate (WER)* is used as quality measure.

Figure 4 shows the WER behavior given different types of classes. As statistical classes (*cl50,cl100,cl200*) we used the Och monolingual classes (Och, 1999), which can be performed using 'mkcls' (a tool available with GIZA). Also we used the statistical reordered classes (*cl100mono*) which were explained in Section 4. Both statistical and statistical reordered classes used the *disamb* tool of SRILM in order to classify unknown words. As morphological classes we used the *TAGS* provided by Freeling. Clearly, statistical classes perform better than *TAGS* and best results can be achieved with 100 and 200 classes and an n-gram length of 5.

For the sake of completeness, we have evaluated the perplexity of the SMR Ngram model over the aligned test set above and choosing 200 classes. Figure 5 is coherent with the WER results above and it shows that perplexity is not reduced for an n-gram length greater than 5.

[1] http://www.statmt.org/wmt07/
[2] http://www.fjoch.com/GIZA++.html
[3] http://www.speech.sri.com/projects/srilm/
[4] http://gps-tsc.upc.es/veu/soft/soft/marie/

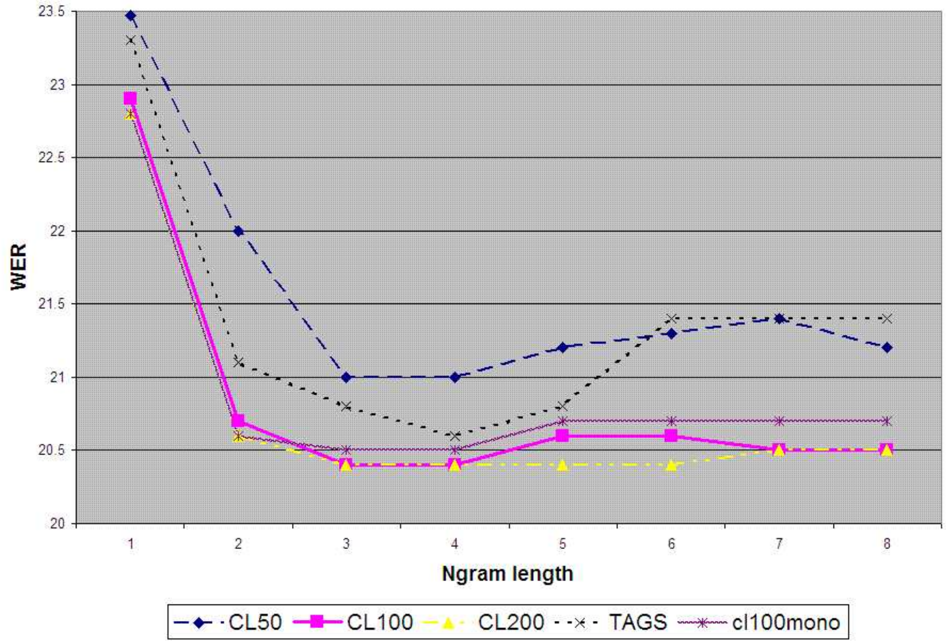

Figure 4: *WER over the reference given various sets of classes and Ngram lengths.*

5.4 Graph pruning

The more complex is the reordering graph, the less efficient is the decoding. That is why, in this section, we experiment with several ways of graph pruning. Additionally, for each pruning we see the influence of considering the graph weights (i.e. reordering feature importance).

Given that the reordering graph is the output of a beam search decoder, we can consider pruning the reordering graph by limiting the SMR beam, i.e. limiting the size of hypothesis stacks.

Given a reordering graph, another option is to prune states and arches only used in paths *s* times worse than the best path.

Table 2 gives the results of the proposed pruning. Note that computational time is given in terms of the monotonic translation time (and it is the same for both directions). It is shown that graph pruning guarantees the efficiency of the system and even increases the translation's quality. Similar results are obtained in terms of BLEU for both types of pruning. In this task and for both translation directions, it seems more appropriate to limit directly the beam search in the SMR step to 5.

As expected, the influence of the reordering feature, which takes into account the graph weights, tends to be more important as pruning decreases (i.e. when the graph has more paths).

Pruning	W_r	$BLEU_{En2Es}$	$BLEU_{Es2En}$	TIME
b5	yes	31.32	32.64	$2.4T_m$
b5	no	31.25	31.82	$2.5T_m$
b50	yes	30.95	32.28	$5.3T_m$
b50	no	30.90	27.44	$4.8T_m$
b50 s10	yes	31.19	32.20	$1.5T_m$
b50 s10	no	31.07	32.41	$1.4T_m$

Table 2: Performance in BLEU in the test set of different graph pruning (*b* stands for beam and *s* for states); the use of reordering feature function (W_r indicates its use); and the time increase related to T_m (monotonic translation time).

5.5 Results and discussion

Table 3 shows the performance of our Ngram-based system using the SMR technique. First row is the WMT07 baseline system which can be reproduced following the instructions in *http://www.statmt.org/wmt07/baseline.html*. This baseline system uses a non-monotonic search. Second row shows the results of the Ngram-based system presented in section 2 using the weighted reordering graph trained with the best configuration found in the above section (200 statistical classes and an Ngram of length 5).

System	$BLEU_{es2en}$	$BLEU_{en2es}$
WMT07 Of. Baseline	31.21	30.74
Ngram-based	32.64	31.32

Table 3: *BLEU Results.*

6 Conclusions and further work

The proposed SMR technique can be used both in training and test steps in a SMT system. Applying the SMR technique in the training step reduces the sparseness in the translation vocabulary. Applying SMR technique in the test step allows to generate a weighted reordering graph for SMT system.

The use of classes plays an important role in the SMR technique, and experiments have shown that statistical classes are better than morphological ones.

Moreover, we have experimented with different graph pruning showing that best translation results can be achieved at a very low increase of computational cost when comparing to the monotonic translation computational cost.

Finally, we have shown that our translation system using the SMR technique outperforms the WMT07 Official baseline system (which uses a non-monotonic search) in terms of BLEU.

As further work, we want to introduce the SMR technique in a state-of-the-art phrase-based system.

7 Acknowledgments

This work has been funded by the European Union under the TC-STAR project (IST- 2002-FP6-506738) and the Spanish Government under grant TEC2006-13964-C03 (AVIVAVOZ project).

References

J. Atserias, B. Casas, E. Comelles, M. González, L. Padró, and M. Padró. 2006. Freeling 1.3: Syntactic and semantic services in an open-source nlp library. In *5th Int. Conf. on Language Resource and Evaluation (LREC)*, pages 184–187.

P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–311.

M. Collins, P. Koehn, and I. Kucerová. 2005. Clause restructuring for statistical machine translation. In *43st Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 531 – 540, Michigan.

M.R. Costa-jussà and J.A.R. Fonollosa. 2006. Statistical machine reordering. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 71–77, Sydney.

M. R. Costa-jussà, P. Lambert, J.M. Crego, M. Khalilov, J.A.R. Fonollosa, J.B. Mariño, and R. Banchs. 2007. Ngram-based statistical machine translation enhanced with multiple weighted reordering hypotheses. In *ACL: Workshop of Statistical Machine Translation (WMT07)*, Prague.

J.M. Crego and J.B. Mariño. 2006. Reordering experiments for n-gram-based smt. *Ist IEEE/ACL International Workshop on Spoken Language Technology (SLT'06)*, pages 242–245.

J. M. Crego, M. R. Costa-jussà, J. Mariño, and J. A. Fonollosa. 2005. Ngram-based versus phrase-based statistical machine translation. In *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'05*, pages 177–184, Pittsburgh, October.

S. Kanthak, D. Vilar, E. Matusov, R. Zens, and H. Ney. 2005. Novel reordering approaches in phrase-based statistical machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pages 167–174, Ann Arbor, MI, June.

P. Koehn, F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proc. of the Human Language Technology Conference, HLT-NAACL'2003*, pages 48 – 54, Edmonton, Canada, May.

P. Lambert, A. de Gispert, R. Banchs, and J. Mariño. 2006. Guidelines for word alignment and manual alignment. *Language Resources and Evaluation*, 39(4):267–285.

J.B. Mariño, R.E. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R. Fonollosa, and M.R. Costa-jussà. 2006. N-gram based machine translation. *Computational Linguistics*, 32(4):527–549.

J.A. Nelder and R. Mead. 1965. A simplex method for function minimization. *The Computer Journal*, 7:308–313.

F.J. Och. 1999. An efficient method for determining bilingual word classes. In *9th Conf. of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 71–76, June.

M. Popovic and H. Ney. 2006. Pos-based word reorderings for statistical machine translation. In *5th International Conference on Language Resources and Evaluation (LREC)*, pages 1278–1283, Genova, May.