Statistical Machine Reordering

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

Reordering is currently one of the most important problems in statistical machine translation systems. This paper presents a novel strategy for dealing with it: statistical machine reordering (SMR). It consists in using the powerful techniques developed for statistical machine translation (SMT) to translate the source language (*S*) into a reordered source language (*S'*), which allows for an improved translation into the target language (*T*). The SMT task changes from *S2T* to *S'2T* which leads to a monotonized word alignment and shorter translation units. In addition, the use of classes in SMR helps to infer new word reorderings. Experiments are reported in the EsEn WMT06 tasks and the ZhEn IWSLT05 task and show significant improvement in translation quality.

1 Introduction

During the last few years, SMT systems have evolved from the original word-based approach (Brown et al., 1993) to phrase-based translation systems (Koehn et al., 2003). In parallel to the phrase-based approach, the use of bilingual n-grams gives comparable results, as shown by Crego et al. (2005a). Two basic issues differentiate the n-gram-based system from the phrase-based: training data are monotonously segmented into bilingual units; and, the model considers n-gram probabilities rather than relative frequencies. This translation approach is described in detail by Mariño et al. (2005). The n-gram-based system follows a maximum entropy approach, in which a log-linear combination of multiple models is implemented (Och and Ney, 2002), as an alternative to the source-channel approach.

In both systems, introducing reordering capabilities is of crucial importance for certain language pairs. Recently, new reordering strategies have been proposed in the literature on SMT such as the reordering of each source sentence to match the word order in the corresponding target sentence, see Kanthak et al. (2005) and Crego et al. (2005b). Similarly, Matusov et al. (2006) describe a method for simultaneously aligning and monotonizing the training corpus. The main problems of these approaches are: (1) the fact that the proposed monotonization is based on the alignment and cannot be applied to the test sets, and (2) the lack of reordering generalization.

This paper presents a reordering approach called statistical machine reordering (SMR) which improves the reordering capabilities of SMT systems without incurring any of the problems mentioned above. SMR is a first-pass translation performed on the source corpus, which converts it into an intermediate representation, in which source-language words are presented in an order that more closely matches that of the target language. SMR and SMT are performed using the same modeling tools as n-gram-based systems but using different statistical log-linear models.

In order to be able to infer new reorderings we use word classes instead of words themselves as the input to the SMR system. In fact, the use of classes to help in the reordering is a key difference between our approach and standard SMT systems.

This paper is organized as follows: Section 2 outlines the baseline system. Section 3 describes the reordering strategy in detail. Section 4 presents and discusses the results, and Section 5 presents our conclusions and suggestions for further work.

Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pages 70–76,
Sydney, July 2006.

2 N-gram-based SMT System

This section briefly describes the n-gram-based SMT which 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 (de Gispert and Mariño, 2002).

Bilingual units (tuples) are extracted from any word alignment according to the following constraints:

1. a monotonous 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.

Figure 1 presents a simple example which illustrates the tuple extraction process.

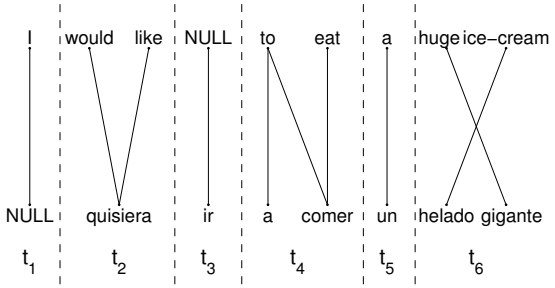

Figure 1: *Example of tuple extraction from an aligned bilingual sentence pair.*

Two important issues regarding this translation model must be considered. First, it often occurs that large number of single-word translation probabilities are left out of the model. This happens for all words that are always embedded in tuples containing two or more words. Consider for example the word "ice-cream" in Figure 1. As seen from the Figure, "ice-cream" is embedded into tuple t_6. If a similar situation is encountered for all occurrences of "ice-cream" in the training corpus, then no translation probability for an independent occurrence of this word will exist.

To overcome this problem, the tuple 4-gram model is enhanced by incorporating 1-gram translation probabilities for all the embedded words detected during the tuple extraction step. These 1-gram translation probabilities are computed from the intersection of both, the source-to-target and the target-to-source alignments.

The second issue has to do with the fact that some words linked to NULL end up producing tuples with NULL source sides. Consider for example the tuple t_3 in Figure 1. Since no NULL is actually expected to occur in translation inputs, this type of tuple is not allowed. Any target word that is linked to NULL is attached either to the word that precedes or the word that follows it. To determine this, we use the $IBM1$ probabilities, see Crego et al. (2005a).

In addition to the bilingual n-gram translation model, the baseline system implements a log-linear combination of four 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 **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.

All these models are combined in the decoder. Additionally, the decoder allows for a non-monotonous search with the following distorsion model.

- A word distance-based **distorsion model**.

$$P(t_1^K) = exp(-\sum_{k=1}^{K} d_k)$$

where d_k is the distance between the first word of the k^{th} tuple (unit), and the last word+1 of the $(k-1)^{th}$ tuple. Distance are measured in words referring to the units source side.

To reduce the computational cost we place limits on the search using two parameters: the distortion limit (the maximum distance measured in words that a tuple is allowed to be reordered, m) and the reordering limit (the maximum number of reordering jumps in a sentence, j). This feature is independent of the reordering approach presented in this paper, so they can be used simultaneously.

In order to combine the models in the decoder suitably, an optimization tool is needed to compute log-linear weights for each model.

3 Statistical Machine Reordering

As mentioned in the introduction, SMR and SMT are based on the same principles. Here, we give a detailed description of the SMR reordering approach proposed.

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*). Then, the translation tasks changes from *S2T* to *S'2T*. The main difference between the two tasks is that the latter allows for: (1) monotonized word alignment, and (2) higher quality monotonized translation.

3.2 Description

Figure 2 shows the SMR block diagram. The input is the initial source sentence (*S*) and the output is the reordered source sentence (*S'*). There three blocks inside SMR: (1) class replacing ; (2) the decoder, which requires the translation model; and, (3) the block which reorders the original sentence using the indexes given by the decoder. The following example specifies the input and output of each block inside the SMR.

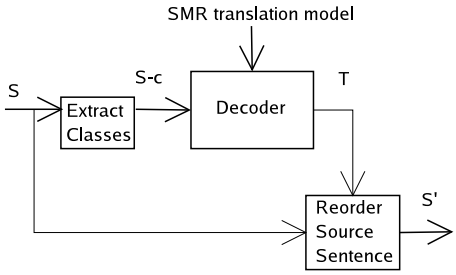

Figure 2: *SMR block diagram.*

1. Source sentence (*S*):

   El compromiso sólo podría mejorar

2. Source sentence classes (*S-c*):

   C38 C43 C49 C42 C22

3. Decoder output (translation, T):

   C38#0 | C43 C49 C42#1 2 0 | C22#0

   where | indicates the segmentation into translation units and # divides the source and target. The source part is composed of word classes and the target part is composed of the new positions of the source word classes, starting at 0.

4. SMR output (*S'*). The reordering information inside each translation unit of the decoder output (T) is applied to the original source sentence (S):

   El sólo podría compromiso mejorar

3.3 Training

For the reordering translation, we used an n-gram-based SMT system (and considered only the translation model). Figure 3 shows the block diagram of the training process of the SMR translation model, which is a bilingual n-gram-based model. The training process uses the training source and target corpora and consists of the following steps:

1. Determine 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. Extract reordering tuples, see Figure 4.

   (a) From union word alignment, extract bilingual *S2T* tuples (i.e. source and target fragments) while maintaining the

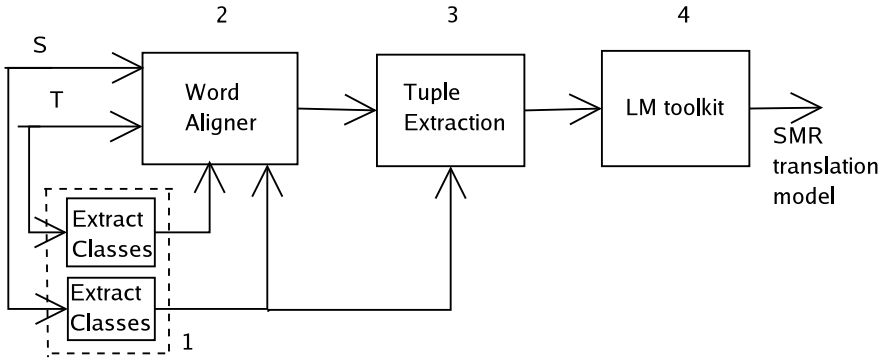

Figure 3: *Block diagram of the training process of the SMR translation model.*

(a) bilingual S2T tuple

only possible compromise # compromiso solo podria # 0-1 1-1 1-2 2-0
(target) (source) (word alignment)
(wrd_src-wrd_trg)

(b) many-to-many word alignment → many-to-one word alignment
P_ibm (only, solo)>P_ibm(possible, solo)

only possible compromise # compromiso solo podria # 0-1 1-2 2-0

(c) bilingual S2S' tuple

compromiso solo podria # 1 2 0
(source) (new order)

(e) classes substitution

C43 C49 C42 # 1 2 0

Figure 4: *Example of the extraction of reordering tuples (step 3).*

alignment inside the tuple. As an example of a bilingual *S2T* tuple consider: *only possible compromise # compromiso sólo podría # 0-1 1-1 1-2 2-0*, as shown in Figure 4, where the different fields are separated by # and correspond to: (1) the target fragment; (2) the source fragment; and (3) the word alignment (in this case, the fields that respectively correspond to a target and source word are separated by $-$).

(b) Modify the many-to-many word alignment from each tuple to many-to-one. If one source word is aligned to two or more target words, the most probable link given IBM Model 1 is chosen, while the other are omitted (i.e. the number of source words is the same before and after the reordering translation). In the above example, the tuple would be changed to: *only possible compromise # compromiso sólo podría # 0-1 1-2 2-0*, as $P_{ibm1}(\text{only, sólo})$ is higher than $P_{ibm1}(\text{possible, sólo})$.

(c) From bilingual *S2T* tuples (with many-to-one inside alignment), extract bilingual *S2S'* tuples (i.e. the source fragment and its reordering). As in the example: *compromiso sólo podría # 1 2 0*, where the first field is the source fragment, and the second is the reordering of these source words.

(d) Eliminate tuples whose source fragment consists of the NULL word.

(e) Replace the words of each tuple source fragment with the classes determined in Step 1.

4. Compute the bilingual language model of the bilingual *S2S'* tuple sequence composed of the source fragment (in classes) and its reorder.

Once the translation model is built, the original source corpus *S* is translated into the reordered source corpus *S'* with the SMR system, see Figure 2. The reordered training source corpus and the original training target corpus are used to train the SMT system (as explained in Section 2). Finally, with this system, the reordered test source corpus is translated.

4 Evaluation Framework

In this section, we present experiments carried out using the EsEn WMT06 and the ZhEn IWSLT05 parallel corpus. We detail the tools which have been used and the corpus statistics.

EuroParl	*Spanish*	*English*
Training Sentences	727.1 k	727.1 k
Words	15.7 M	15.2 M
Vocabulary	108.7 k	72.3 k
Development Sentences	500	500
Words	15.2 k	14.8 k
Vocabulary	3.6 k	3 k
Test Sentences	3064	3064
Words	91.9 k	85.2 k
Vocabulary	11.1 k	9.1 k

Table 1: *Spanish to English task. EuroParl corpus: training, development and test data sets.*

BTEC	*Chinese*	*English*
Training Sentences	20 k	20 k
Words	176.2 k	182.3 k
Vocabulary	8.7 k	7.3 k
Development Sentences	506	506
Words	3.5 k	3.3 k
Vocabulary	870	799
Test Sentences	506	506
Words	4 k	3 k
Vocabulary	916	818

Table 2: *Chinese to English task. BTEC corpus: training, development and test data sets. Development and test data sets have 16 references.*

4.1 Tools

- The word alignments were computed using the GIZA++ tool (Och, 2003).
- The word classes were determined using 'mkcls', a freely-available tool with GIZA++.
- The language model was estimated using the SRILM toolkit (Stolcke, 2002).
- We used MARIE as a decoder (Crego et al., 2005b).
- The optimization tool used for computing log-linear weights (see Section 2) is based on the simplex method (Nelder and Mead, 1965).

4.2 Corpus Statistics

Experiments were carried out on the Spanish and English task of the WMT06 evaluation[1] (EuroParl Corpus) and on the Chinese to English task of the IWSLT05 evaluation[2] (BTEC Corpus). The former is a large corpus, whereas the latter is a small corpus translation task. Table 1 and 2 show the main statistics of the data used, namely the number of sentences, words, vocabulary, and mean sentence lengths for each language.

4.3 Units

In this section different statistics units of both approaches (*S2T* and *S'2T*) are shown (using the ZhEn task). All the experiments in this section were carried out using 100 classes in the SMR step.

[1] www.statmt.org/wmt06/shared-task/

[2] www.slt.atr.jp/IWSLT2005

Table 3 shows the vocabulary of bilingual n-grams and embedded words in the translation model. Once the reordering translation has been computed, alignment becomes more monotonic. It is commonly known that non-monotonicity poses difficulties for word alignments. Therefore, when the alignment becomes more monotonic, we expect an improvement in the alignment, and, therefore in the translation. Here, we can observe a significant enlargement of the number of translation units, which leads to a growth of the translation vocabulary. We also observe a decrease in the number of embedded words (around 20%). From Section 2, we know that the probability of embedded words is estimated independently of the translation model. Reducing embedded words allows for a better estimation of the translation model.

Figure 5 shows the histogram of the tuple size in the two approaches. We observe that the number of tuples is similar over length 5. However, there are a greater number of shorter units in the case of SMR+NB (shorter units lead to a reduction in data sparseness).

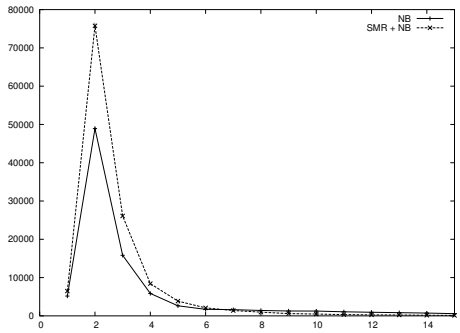

Figure 5: *Comparison of the histogram of the tuple size in the two approaches (NB and SMR+NB).*

System	1gr	2gr	3gr	4gr	Embedded
NB	34487	57597	3536	1918	5735
SMR + NB	35638	70947	5894	3412	4632

Table 3: *Vocabulary of n-grams and embedded words in the translation model.*

System	Total	Vocabulary
NB	4460	959
SMR + NB	4628	1052

Table 4: *Tuples used to translate the test set (total number and vocabulary).*

Table 4 shows the tuples used to translate the test set (total number and vocabulary). Note that the number of tuples and vocabulary used to translate the test set is significantly greater after the reordering translation.

4.4 Results

Here, we introduce the experiments that were carried out in order to evaluate the influence of the SMR approach in both tasks EsEn and ZhEn. The log-linear translation model was optimized with the simplex algorithm by maximizing over the BLEU score. The evaluation was carried out using references and translation in lowercase and, in the ZhEn task, without punctuation marks.

We studied the influence of the proposed SMR approach on the n-gram-based SMT system described using a monotonous search (NBm or monotonous baseline configuration) in the two tasks and a non-monotonous search (NBnm or non-monotonous baseline configuration) in the ZhEn task. In allowing for reordering in the SMT decoder, the distortion limit (m) and reordering limit (j) (see Section 2) were empirically set to 5 and 3, as they showed a good trade-off between quality and efficiency. Both systems include the four features explained in Section 2: the language model, the word bonus, and the source-to-target and target-to-source lexicon models.

Tables 5 and 6 show the results in the test set. The former corresponds to the influence of the SMR system on the EsEn task (NBm), whereas the latter corresponds to the influence of the SMR system on the ZhEn task (NBm and NBnm).

4.5 Discussion

Both BLEU and NIST coherently increase after the inclusion of the SMR step when 100 classes are used. The improvement in translation quality can be explained as follows:

- SMR takes advantage of the use of classes and correctly captures word reorderings that are missed in the standard SMT system. In addition, the use of classes allows new reorderings to be inferred.
- The new task *S'2T* becomes more monotonous. Therefore, the translation units tend to be shorter and SMT systems perform better.

The gain obtained in the SMR+NBnm case indicates that the reordering provided by SMR system and the non-monotonous search are complementary. It means that the output of the SMR could still be further monotonized. Note that the ZhEn task has complex word reorderings.

These preliminary results also show that SMR itself provides further improvements to those provided by the non-monotonous search.

5 Conclusions and Further Research

In this paper we have mainly dealt with the reordering problem for an n-gram-based SMT system. However, our approach could be used similarly for a phrase-based system. We have addressed the reordering problem as a translation from the source sentence to a monotonized source sentence. The proposed SMR system is applied before a standard SMT system. The SMR and SMT systems are based on the same principles and share the same type of decoder.

In extracting bilingual units, the change of order performed in the source sentence has allowed the modeling of the translation units to be improved (shorter units mean a reduction in data sparseness). Also, note that the SMR approach allows the coherence between the change of order in the training and test source corpora to be maintained.

System	Classes	BLEU	NIST	WER	PER
NBm	-	27.69	7.31	61.6	45.34
SMR + NBm	-	28.60	7.53	59.89	43.53
SMR + NBm	100	30.89	7.75	55.77	42.85

Table 5: *Results in the test set of the EsEn task using a monotonous search.*

System	Classes	BLEU	NIST	WER	PER
NBm	-	42.42	8.3	42.87	33.44
NBnm	-	43.58	8.9	43.89	34.05
SMR + NBm	100	43.75	8.49	42.45	33.85
SMR + NBnm	100	45.97	9.0	40.92	32.32

Table 6: *Results in the test set of the ZhEn task using a monotonous and a non-monotonous search.*

Performing reordering as a preprocessing step and independently from the SMT system allows for a more efficient final system implementation and a quicker translation. Additionally, using word classes helps to infer unseen reorderings. These preliminary results show consistent and significant improvements in translation quality.

As further research, we would like to add extra features to the SMR system, and study new types of classes for the reordering task.

6 Acknowledgments

This work has been partially funded by the European Union under the integrated project TC-STAR - Technology and Corpora for Speech to Speech Translation - (IST-2002-FP6-506738, http://www.tc-star.org) and the Spanish government under a FPU grant.

References

E. Matusov A. Mauser and H. Ney. 2006. Training a statistical machine translation system without giza++. *5th Int. Conf. on Language Resources and Evaluation, LREC'06*, May.

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

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

J.M. Crego, J. Mariño, and A. de Gispert. 2005b. An Ngram-based statistical machine translation decoder. *Proc. of the 9th Int. Conf. on Spoken Language Processing, ICSLP'05*.

A. de Gispert and J. Mariño. 2002. Using X-grams for speech-to-speech translation. *Proc. of the 7th Int. Conf. on Spoken Language Processing, ICSLP'02*, September.

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

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

J.B. Mariño, R.E. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R. Fonollosa, and M. Ruiz. 2005. Bilingual n-gram statistical machine translation. In *Proc. of the MT Summit X*, pages 275–82, Pukhet (Thailand), May.

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

F.J. Och and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. *40th Annual Meeting of the Association for Computational Linguistics*, pages 295–302, July.

F.J. Och. 2003. Giza++ software. http://www-i6.informatik.rwth-aachen.de/~och/ software/giza++.html.

A. Stolcke. 2002. Srilm - an extensible language modeling toolkit. *Proc. of the 7th Int. Conf. on Spoken Language Processing, ICSLP'02*, September.