# Computing multiple weighted reordering hypotheses for a statistical machine translation phrase-based system

**Marta R. Costa-jussà and José A. R. Fonollosa**
Universitat Politècnica de Catalunya (UPC)
TALP Research Center
Campus Nord 08034 Barcelona
{mruiz,adrian}@gps.tsc.upc.edu

## Abstract

Reordering is one source of error in statistical machine translation (SMT). This paper extends the study of the statistical machine reordering (SMR) approach, which uses the powerful techniques of the SMT systems to solve reordering problems. Here, the novelties yield in: (1) using the SMR approach in a SMT phrase-based system, (2) adding a feature function in the SMR step, and (3) analyzing the reordering hypotheses at several stages. Coherent improvements are reported in the TC-STAR task (Es/En) at a relatively low computational cost.

## 1 Introduction

Statistical machine translation (SMT) has evolved from the initial word-based translation models to more advanced models that take the context surrounding the words into account, i.e. the so-called phrase-based system (Koehn et al., 2003). The phrase-based model is usually the main feature in a log-linear framework, reminiscent of the maximum entropy modeling approach.

One of the best known reordering approach is permitting arbitrary word-reorderings. However, the exact decoding problem was shown to be NP-hard (Knight, 1999). To solve this problem, several approaches have defined different kinds of constraints as for example heuristic (Berger et al., 1996) (Crego et al., 2005) or linguistic (Wu, 1996). Other approaches try to reorder the source language in a way that better matches the target language (Popovic and Ney, 2006) (Collins et al., 2005).

A natural evolution of the source reordering strategies consists in using a word graph, containing the $N$-best reordering decisions, instead of the single-best used in the above strategies. The reordering problem is equally approached by alleviating the difficulty of needing highly accurate reordering decisions in preprocessing. The final decision is delayed, to be subsequently in the global search, where all the information is then available. Inspired by (Knight and Al-Onaizan, 1998), they permute the source sentence to provide a source input graph that extends the search graph. In (Kanthak et al., 2005), they train the system using a monotonized source corpora and they translate the test set allowing source reorderings which are limited by constraints such as IBM or ITG. Similarly in (Crego and Mariño, 2007; Zhang et al., 2007), reordering is addressed through a source input graph. In this case, the reordering hypotheses are defined from a set of linguistically motivated rules (either using *Part of Speech*; chunks; or parse trees).

Previous work (Costa-jussà and Fonollosa, 2006) presents the SMR approach which is based on using the powerful SMT techniques to generate a reordered source input for an Ngram-based SMT system both in training and decoding steps. One step further, (R. Costa-jussà and R. Fonollosa, 2007) shows how the SMR system generates a weighted reordering graph, allowing the SMT decoder to make the reordering decision.

In this paper, we use the above mentioned weighted reordering graph in a standard state-of-the-art phrase-based system. Moreover, we introduce an additional feature function in the SMR sys-

tem which is a class language model. Therefore, the SMR graph provides two feature functions to the log-linear SMT framework. We report experiments in the *European Parliament Plenary Sessions* (EPPS) task (Spanish/English), showing improvements in BLEU, NIST and METEOR. Finally, we analyze the reordering hypotheses (i.e. how many hypotheses are proposed for the SMR system and which ones are chosen for the SMT system).

This paper is organized as follows. The next section describes the baseline system. Section 3 reports the SMR approach. Section 4 describes the evaluation framework, discusses the results and analyzes the reordering hypotheses at different stages. Finally, Section 5 presents the conclusions.

## 2 Phrase-based System

The basic idea of the phrase-based translation is to segment the given source sentence into units (here called phrases), then translate each phrase and finally compose the target sentence from these phrase translations.

In order to train these phrase-based models, an alignment between the source and target training sentences is found by using the standard IBM models in both directions (source-to-target and target-to-source) and combining the two obtained alignments. Given this alignment an extraction of contiguous phrases is carried out, specifically we extract all phrases that fulfill the following restrictions: all source (target) words within the phrase are aligned only to target (source) words within the phrase.

The probability of these phrases is normally estimated by relative frequencies, normally in both directions, which are then combined in a log-linear way.

### 2.1 Feature Functions

The probability of the phrases is combined in a log-linear way with several additional feature functions: a target 4-gram language model, a forward and a backward lexicon model, a word bonus, a phrase bonus and a POS target language model.

## 3 Weighted Reordering Hypotheses

As mentioned in the introduction, the weighted reordering hypotheses are generated using an statistical approach, which we call the SMR technique.

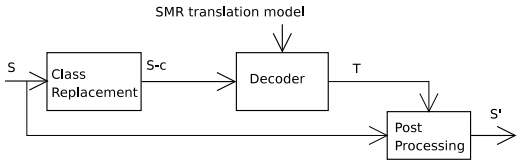


Figure 1: *SMR block diagram.*

S: ***we offer a better and different structure***
CLASS REPLACING: ***62 150 130 36 88 185 178***
DECODING: ***62# | 150#0 | 130#0 | 36 88 185 178#4 1 2 3***
POST PROCES.: ***we offer a structure better and different***

Figure 2: *Example of a source sentence reordering performed by the SMR module. The decoding output is shown in units: '|' marks the unit boundaries and '#' marks the two components of the bilingual units. The reference target sentence is: Nosotros ofrecemos una estructura mejor y differente.*

### 3.1 Concept

The 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

The SMR module (see Figure 1) is in charge of translating the source sentence(*S*) into a reordered source sentence (*S'*). Figure 1 shows the block diagram, and each block works as follows:

1. Class replacement. Use the correspondence of word to word class to substitute each source word by its word class.

2. Decoding. A monotonic decoding using the SMR Model allows to assign reordering tuples to the input sequence.

3. Post Processing. The decoder output is post-processed to build the reordered sentence.

An example of the input and output of each step is shown in Figure 2.

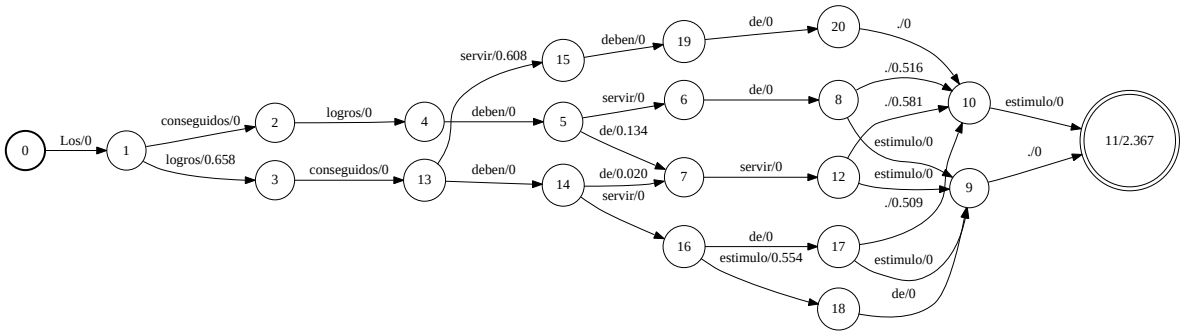


Figure 3: *SMR output graph. The source sentence is:* Los logros conseguidos deben servir de estímulo. *The target sentence could be:* The achieved goals should be an encouragement.

### 3.3 SMR Model

The training process for the SMR Model requires 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 both 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 alignment inside the tuple. As an example of a bilingual *S2T* tuple consider: *better and different structure # estructura mejor y diferente # 1-1 1-2 2-1 3-4 4-1*, as shown in Figure 4, where the fields are separated by # and correspond to: (1) the source fragment; (2) the target fragment; and (3) the word alignment (in this case, the fields that correspond to a target and source word, respectively, are separated by -).

   (b) Change the many-to-many word alignment 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 kept). Following our example, the tuple would be changed to: *better and different structure # estructura mejor y diferente # 1-2 2-3 3-4 4-1*, as $P_{ibm1}(\text{better, mejor})$ is higher than $P_{ibm1}(\text{better, estructura})$.

   (c) From the bilingual *S2T* tuples (with many-to-one inside alignment), extract bilingual *S2S'* tuples (i.e. the source fragment and its reordering). Example: *better and different structure # 4 1 2 3*, 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 NULL.

   (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 reordering.

For further details, see (Costa-jussà, 2008).

### 3.4 Additional Reordered Source Language Model

We propose to add a feature function in the SMR system which is the reordered source language model. This language model is trained on the reordered source corpus (in word classes). Hence, Figure 1 changes to Figure 5.

The reordered source corpus has been obtained by using the word alignment links (i.e. *also I welcome* would be the reordered source corpus of the source corpus *I also welcome* given the alignment in Figure 6 (A)). Additionally, words themselves have

(A) BILINGUAL S2T TUPLE:
**better and different structure # estructura mejor y diferente # 1-1 1-2 2-3 3-4 4-1**
(B) MANY-TO-MANY WORD ALIGNMENT $\longrightarrow$ MANY-TO-ONE WORD ALIGNMENT:
$P_ibm1(better, mejor) > P_ibm1(better, estructura)$
**better and different structure # estructura mejor y diferente # 1-2 2-3 3-4 4-1**
(C) BILINGUAL S2S' TUPLE:
**better and different structure # 4 1 2 3**
(D) CLASS SUBSTITUTION:
**C36 C88 C185 C176 # 4 1 2 3**

Figure 4: *Example of the extraction of reordering tuples (Step 3). # divides the fields: source, target and word alignment, which includes the source and final position separated by -.*

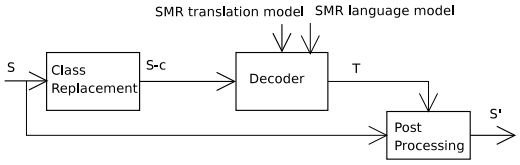


Figure 5: *SMR block diagram adding a target language model to the SMR decoder.*

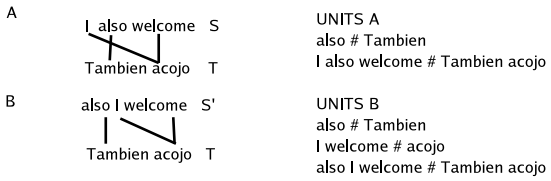


Figure 6: *Unit extraction (A) original training source corpus (B) reordered training source corpus.*

been substituted by statistical word classes, which were trained on the given source corpus.

### 3.5 Coupling SMR and SMT

The SMR module can output either a single sentence or a word graph (see Figure 7). The former is a reordered sentence like the one shown in the fourth row of Figure 2. This gives a unique reordering option and this leads to a deterministic reordering. The latter contains several possible reorderings coded in a graph (see an example in Figure 3).

In the training step, we propose coupling the SMR and SMT systems with a single best translation. In the test step, we propose using a graph to couple both systems (R. Costa-jussà and R. Fonollosa, 2007).

**Coupling SMR and SMT in the training step.** Figure 7 (A) shows the corresponding block diagram for the training corpus: first, the given source corpus *S* is translated into the reordered source corpus *S'* with the SMR system.

The main difference between corpus *S'* and *S* is that the former matches the target language model order much better than the latter. The reordered training source corpus and the original training target corpus are used to train the SMT system. Using the SMT baseline system (S2T task) or using the SMR plus the SMT system (S'2T task) generates a different set of translation units. Figure 6 shows an example.

**Coupling SMR and SMT in the test step.** The SMR technique can generate an output graph that can be used as an input graph for the SMT system. Figure 7 (B) shows the corresponding block diagram in decoding: the SMR outputs graph is given as an input graph to the SMT system. The graph without including probabilities in the arches is referred to as reordering graph, in short SMR-Graph. The monotonic search in the SMT system is extended with a reordering graph and the feature functions in the SMT system (as the target or the POS target language model) can provide new reordering information. The final reordering decision is taken in the SMT decoding. Note that the SMR technique takes advantage of the generalizing information added by word classes.

One step further, if we consider each reordering hypothesis owns their probability (as shown in Figure 7 (B)), the graph will be referred to as weighted

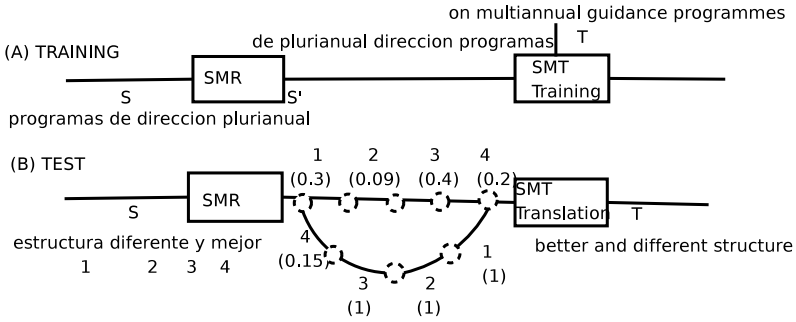


Figure 7: *SMR and SMT coupling.*

reordering graph, in short SMR-Graph$_R$. Each arch has a probability associated. Notice that these probabilities have been computed taking advantage of the smoothing techniques of a language model, the use of Ngram context and the use of statistical classes. In addition, the reordering graph has one additional feature function (SMR-Graph$_T$) given by the reordered source language model. Finally, both probabilities extend the SMT log-linear combination of feature functions.

## 4 Evaluation Framework

### 4.1 Data

| | | Spanish | English |
|---|---|---|---|
| Train | Sentences | 1.35M | |
| | Words | 39M | 37M |
| | Vocabulary | 147K | 109K |
| Dev | Sentences | 699 | 1122 |
| | Words | 21K | 26K |
| Test | Sentences | 1 117 | 894 |
| | Words + Punct. | 29K | 29K |
| | Words | 26K | 26K |
| | OOV Words | 72 | 150 |

Table 1: *Statistics of the EPPS Corpora (official training set of the 3rd TC-STAR Evaluation and official test set of the 2nd TC-STAR Evaluation).*

The corpus consists of the official version of the speeches held in the *European Parliament Plenary Sessions* (EPPS), as available on the web page of the European Parliament. The task is the so-called Final Text Edition (FTE) in the Es/En language pair. Table 1 shows some statistics of the corpus. The training set was used in the *TC-STAR*[1] official 3rd Evaluation and the test set was used in the official 2nd Evaluation. The development set used to tune the system consists of a subset (first half sentences) of the official development set. This allows reducing the optimization time without affecting the translation quality on the test set.

### 4.2 System Configuration Details

*Word Alignment.* The word alignment is automatically computed by using GIZA++ [2] in both directions, which are symmetrized by using the union operation. Instead of aligning words themselves, stems are used for aligning. Afterwards case sensitive words are recovered.

*Word Classes.* 200 statistical classes, which were built using 'mkcls', are the SMR vocabulary.

*Spanish Morphology Reduction.* We implement a morphology reduction of the Spanish language as a preprocessing step. As a consequence, training data sparseness due to Spanish morphology is reduced improving the performance of the overall translation system. In particular the pronouns attached to the verb are separated and contractions as *del* or *al* are splitted into *de el* or *a el*.

*Pruning parameters.* The current version of the SMR system uses a 5-gram language model and a beam pruning of 5 (best results experimented in (R. Costa-jussà and R. Fonollosa, 2007)). The phrase-based SMT system uses a beam pruning of 50.

[1] http://www.tc-star.org/

[2] http://www.fjoch.com/GIZA++.html

*Optimization.* We implement an n-best re-ranking strategy which is used for optimization purposes [3]. This procedure allows for a faster and more efficient adjustment of model weights by means of a double-loop optimization, which provides significant reduction of the number of translations that should be carried out. The current optimization procedure uses the Simplex algorithm. BLEU score is used as the loss function.

*Case sensitive evaluation.* Standard automatic measures are used for evaluation: BLEU, NIST and METEOR.

Notice that no reordering model is added to the baseline system. The first idea was to use the standard distance-based reordering model in our baseline system but it has a high computational cost and, in this Spanish-English EPPS task, this model is proven not to significantly improve the translation quality (Crego and Mariño, 2007) of a monotonic baseline system.

### 4.3 Translation Results

| System | BLEU | NIST | METEOR | W/S |
|---|---|---|---|---|
| PB | 51.48 | 10.54 | 67.82 | 4.76 |
| +SMR-Graph | 52.64 | 10.65 | 68.39 | 1.70 |
| +SMR-Graph$_R$ | 53.54 | 10.67 | 68.63 | 1.68 |
| +SMR-Graph$_T$ | 53.07 | 10.68 | 68.49 | 1.60 |
| +SMR-Graph$_{R+T}$ | **53.70** | **10.74** | **68.65** | 1.56 |

Table 2: *Translation results and words per second of: SMT system and for SMR+SMT system using none, one (R or T) or two (R+T) weights in the reordering graph.*

Table 2 presents the automatic scores obtained for the 2006 test data set comparing the phrase-based SMT system and the SMR plus the SMT system using different reordering feature functions configurations.

The SMR approach allows for an improvement in all measures, specially, using all reordering feature functions. Moreover, we point out the relatively low increase of the computational cost in time.

### 4.4 Analysis of the Reordering Hypotheses

The SMR system proposes several reordering hypotheses to the SMT system. Here we analyze these hypotheses. Figure 8 shows the number of hypotheses proposed for the SMR system in average given the test sentence size (measured in words).

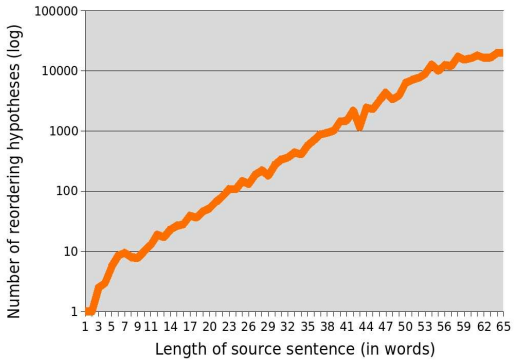


Figure 8: *SMR reordering hypotheses in logarithmic scale and in average given the sentence size in the test set*

Using the reordering graph allows the SMT decoding to choose the final order. Figure 9 shows the position of the final reordering hypothesis inside the graph. Actually, the graph shows the percentage the 1 best ($\approx 51.5\%$), the 2-3 best and the 4-9 best hypotheses of the SMR-Graph$_{R+T}$ are chosen and, in any other case, the percentage one of the 10 best or higher is chosen ($\approx 12\%$). Therefore, we could try to further prune the reordering graph in order to reduce even more the computational cost of the reordering model.

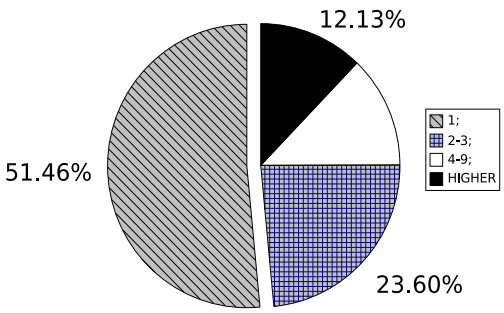


Figure 9: *Given the test sentences (Es), this figure shows the final reordering hypothesis position (in percentage) inside the SMR-Graph$_{R+T}$.*

[3] as proposed in *http://www.statmt.org/jhuws/*

Table 3 shows the most frequent reorderings which have been performed in the test set. Regarding the test set, there are more than 2,000 reorderings and a vocabulary of 36 reorderings. There is no reordering limit and up to a reordering of 9 words has been performed, see the example in Figure 10.

*S: lograr en la región una paz justa y duradera*
*S': justa y duradera una paz lograr en la región*
*T: fair and lasting peace in the region*
*Ref: a fair and lasting peace in the region*

Figure 10: *Example of long reordering: source (S), reordered source (S'), translation (T) and reference (R)*

| Reordering | Num Appearence |
|---|---|
| 2 1 | 1020 |
| 2 3 1 | 171 |
| 3 2 1 | 36 |
| 2 3 4 1 | 25 |
| 3 4 1 2 | 16 |
| 2 4 3 1 | 11 |
| 4 1 2 3 | 8 |
| 3 4 5 1 2 | 4 |
| 2 3 4 5 6 1 | 3 |

Table 3: *Most frequent reorderings performed in the test set.*

## 5 Conclusions

This paper extends previous SMR studies. Particularly it shows the SMR can successfully be applied to a phrase-based system and an additional feature function in the SMR system provides a slight improvement in the SMT translation. Moreover, the reordering hypotheses analysis shows that long reorderings are performed in the Es/En task. The SMT system chooses in almost the 88% of cases a reordering included in the 10-best SMR hypotheses. Therefore, this could be used in the future to further prune the reordering graph.

## 6 Acknowledgments

This work has been funded by the Spanish Government under an FPU grant and TEC2006-13964-C03 (AVIVAVOZ project).

## References

A. Berger, S. Della Pietra, and V. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–72.

M. Collins, P. Koehn, and I. Kucerová. 2005. Clause restructuring for statistical machine translation. In *Proc. of the 43th Annual Meeting of the ACL*, pages 531 – 540, Michigan.

M.R. Costa-jussà and J.A.R. Fonollosa. 2006. Statistical machine reordering. In *Proc. of the Conf. EMNLP*, pages 71–77, Sydney, July.

M. R. Costa-jussà. 2008. *New Reordering and Modeling Approaches for Statistical Machine Translation.* Ph.D. thesis, Universitat Politècnica de Catalunya (UPC), Barcelona, July.

J.M. Crego and J.B. Mariño. 2007. Improving SMT by coupling reordering and decoding. *Machine Translation*, 20(3):199–215.

J.M. Crego, J. Mariño, and A. de Gispert. 2005. An Ngram-based statistical machine translation decoder. In *Proc. of the 9th Int. Conf. ICSLP*, pages 3185–3188, Lisboa, April.

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

K. Knight and Y. Al-Onaizan. 1998. Translation with finite-state devices. In *Proc. of the 4th Conf. AMTA*, pages 421–437, Langhorne, December.

K. Knight. 1999. Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25(4), December.

P. Koehn, F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proc. of the Human Language Technology Conf., HLT-NAACL'03*, pages 48–54, Edmonton, Canada, May.

M. Popovic and H. Ney. 2006. Pos-based word reorderings for statistical machine translation. In *5th Int. Conf. LREC*, pages 1278–1283, Genoa, May.

M. R. Costa-jussà and J. A. R. Fonollosa. 2007. Analysis of atatistical and morphological classes to generate weighted reordering hypotheses on a statistical machine translation system. In *Proc. of the Second ACL Workshop on Statistical Machine Translation (WMT)*, pages 171–176, Prague, June.

D. Wu. 1996. A polynomial-time algorithm for statistical machine translation. In *Proc. of the 34th Annual Meeting of the ACL*, Santa Cruz, June.

Y. Zhang, R. Zens, and H. Ney. 2007. Chunk-level reordering of source language sentences with automatically learned rules for statistical machine translation. In *Proc. of the Workshop on Syntax and Structure in Statistical Translation (SSST)*, pages 1–8, Rochester, April.