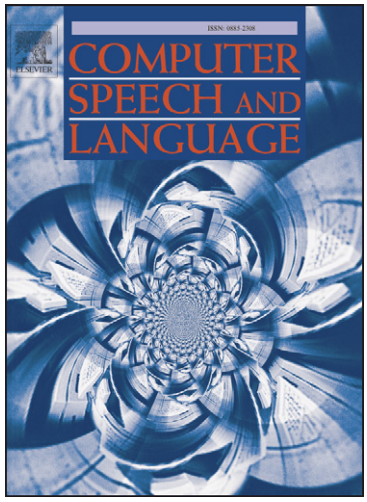
# An Ngram-based Reordering Model

Marta R. Costa-Jussà, José A.R. Fonollosa




# An Ngram-based Reordering Model

Marta R. Costa-jussà, José A.R. Fonollosa

PII: S0885-2308(08)00057-0
DOI: 10.1016/j.csl.2008.12.002
Reference: YCSLA 394

To appear in: *Computer Speech and Language*

Received Date: 5 April 2008
Revised Date: 19 September 2008
Accepted Date: 12 December 2008

Please cite this article as: Costa-jussà, M.R., Fonollosa, J.A.R., An Ngram-based Reordering Model, *Computer Speech and Language* (2008), doi: 10.1016/j.csl.2008.12.002

Elsevier Editorial System(tm) for Computer Speech and Language
Manuscript Draft

Manuscript Number: CSL08-25R1

Title: An Ngram-based Reordering Model

Article Type: Research Article

Section/Category:

Keywords: Statistical Machine translation, Language modeling, Word Reordering

Corresponding Author: Ms Marta Ruiz Costa-jussà,

Corresponding Author's Institution: UPC-TALP Research Center

First Author: Marta Ruiz Costa-jussà

Order of Authors: Marta Ruiz Costa-jussà; José A. R. Fonollosa

Manuscript Region of Origin:

Abstract: This paper describes in detail a novel approach to the reordering challenge
in statistical machine translation (SMT).
This Ngram-based Reordering (NbR) approach uses
the powerful techniques of SMT systems
to generate a weighted reordering graph.
Thus, statistical criteria
reordering constraints are supplied to an SMT system,
and this allows an extension to the SMT decoding search.

The NbR approach is capable of generalizing reorderings
that have been learned during training, through the use of word classes
instead of words themselves.

Improvement in translation performance
is demonstrated with the EPPS task (Spanish and German to English)
and the BTEC task (Arabic to English).

# An Ngram-based Reordering Model

Marta R. Costa-jussà and José A. R. Fonollosa

*Universitat Politècnica de Catalunya*
*TALP Research Center*
*Department of Signal Theory and Communications, Campus Nord, Barcelona 08034, Spain*
{*mruiz,adrian*}*@gps.tsc.upc.edu*

*Key words:* Statistical Machine translation, Language modeling, Word Reordering

## 1 Introduction

Statistical machine translation (SMT) constitutes a research sub-area of machine translation (MT) that has recently gained much popularity. In fact, this technology has experienced real growth motivated by the development of computer resources needed to implement translation algorithms based on statistical methods (1; 2).

SMT is based on the principle that every target sentence $e$ is a possible translation of the source sentence $f$. The problem is formulated as the search for the target sentence with the highest likelihood target sentence among all target sentences. Present SMT systems have evolved from their original predecessors. However, they are distinct in two ways: first, word-based translation models have been replaced by phrase-based translation models (23; 11) which are directly estimated from aligned bilingual corpora by considering relative frequencies, and second, the noisy channel approach has been expanded to a more general maximum entropy approach in which a log-linear combination of multiple feature functions is implemented (15).

Although, significant quality improvements have been produced in SMT, many difficulties, such as word reordering or word correspondence, have not yet been overcome. This paper focuses on the introduction of reordering capabilities. Incorporating these capabilities into the search process generates a high computational cost. Nevertheless, reordering plays an important role in some language pairs, as shown by the high number of works on reordering. Some extended theories may view statistical translation as a concatenation of two sub-tasks: predicting the collection of words in a translation and deciding the order of the predicted words.

Reordering between two languages is a widely studied challenge in MT. Reordering

*Preprint submitted to Elsevier* *Received: date / Accepted: date*

may be solved in the target or source language or even at the bilingual unit level. The main difference is that the movement distance and/or reordering constraints are applied on the foreign side, on the source side or on the bilingual units, respectively. Several overviews of recent reordering approaches can be found in (9; 24).

We briefly describe some previously published reordering approaches related to the Ngram-based Reordering (NbR) algorithm proposed in this paper. These approaches attempt to reorder the source language in a way that better matches the target language. The reordering rules and/or constraints are defined in the source language:

- Deterministic reordering rules (16; 5) The source corpus is reordered following a set of rules. These rules have been automatically learned using lexical and/or morphological information, i.e. *Part of Speech* (POS). The decoder search is monotonic.
- Clause restructuring (22; 3; 21). These methods, which are applied both in training and decoding steps, use syntactic information to reorder source words in SMT as a preprocessing step. This source reordering is complemented with a local reordering in search.
- Input reordering graph (9; 13). The word alignment is then used as a function of source words to reorder the source corpus. Inspired by (10), they permute the source sentence to provide a source input graph which extends the search graph. The reordering hypotheses of the source input graph are limited by several constraints, as IBM or ITG. Similarly in (6; 24), the reordering search problem 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* or chunks).
- Syntax structure (17) and others. This is carried out using standard phrases extended with syntax information from the source side, through suing dependency trees.

This paper describes in detail a novel approach to solve reordering problems in the SMT framework. The NbR approach uses the powerful SMT techniques to convert the source corpora into an intermediate representation, in which source-language words are presented in an order that more closely matches that of the target language. Reorderings hypotheses are learned from the aligned parallel corpus and are successfully smoothed by taking advantage of the extensively investigated area of language modeling.

A natural harmonization of the NbR and SMT system is by generating multiple intermediate representations, i.e. reordering hypotheses, which extend the SMT decoder search. The SMT translation is nearly as efficient as a monotonic translation because the input reordering graph can be highly pruned without affecting the translation quality. Each reordering hypothesis provided by the NbR system has a score which is used to extend the SMT log-linear framework with a reordering feature

function. Moreover, an important characteristic of the proposed approach is the use of word classes for reordering generalization. The NbR approach offers some versatility because, depending on the pair of languages, the reordering hypothesis may be better captured by using a particular type of word class, i.e. statistical, morphological or syntactical.

This paper is organized as follows: Section 2 briefly reviews the particulars of the Ngram-based SMT system used in this work. Section 3 details the NbR approach, while Section 4 reports the experiments conducted to asses the accuracy and efficiency of the NbR approach. Finally, Section 5 concludes and outlines some further work.

## 2 Ngram SMT Baseline System

This section briefly describes the Ngram-based SMT baseline system which uses a translation model based on bilingual n-grams. It is actually a language model composed of bilingual units, referred to as tuples, which approximates the joint probability between source and target languages by using bilingual n-grams. This Ngram-based SMT approach is described in detail in (12).

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

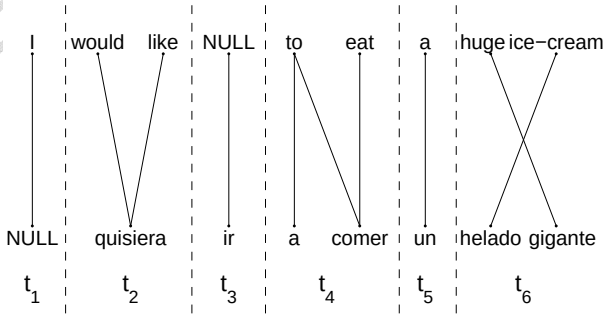



Fig. 1. *Example of tuple extraction from an aligned bilingual sentence pair.*

The first important observation from Figure 1, is related to the possible occurrence of tuples containing unaligned elements in its target side. This is the case of tuple

3. This kind of tuple should be handled in an alternate way in order to allow the system to be able to provide appropriate translations for such unaligned elements. The problem of how to handle this kind of situation is discussed in detail in (12). In short, 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 IBM-1 probabilities. More specifically, the IBM-1 lexical parameters (2) are used for computing the translation probabilities of two possible new tuples: the one resulting when the null-aligned-word is attached to the previous word, and the one resulting when it is attached to the following one. Then, the attachment direction is selected according to the tuple with the highest translation probability.

The second observation from Figure 1, is that it often occurs that a 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, this word 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.

Another important observation from Figure 1 is that each tuple length is implicitly defined by the word-links in the alignment. In contrast to phrase extraction procedures, for which a maximum phrase length should be defined in order to avoid a vocabulary explosion, tuple extraction procedures do not have any control over tuple lengths. Because of this characteristic, the tuple approach will strongly benefit from the structural similarity between the languages under consideration. Then, for close language pairs, tuples are expected to successfully handle those short reordering patterns that are included in the tuple structure, as in the case of "huge ice-cream : helado gigante" presented in Figure 1. On the other hand, in the case of distant pairs of languages, for which a large number of long tuples are expected to occur, this baseline approach will more easily fail to provide a good translation model due to tuple sparseness.

### *2.1 Features functions*

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 5-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 com-

pensate for the system's preference for short output sentences.

- A **source-to-target lexicon model**. This feature, which is based on the IBM-1 lexical parameters, 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**. Similar to the previous feature, this feature is based on the IBM-1 lexical parameters; in this case, these parameters are obtained from target-to-source alignments.

All the above models are combined in an in-house beam search decoder, MARIE [1]. It implements a beam-search strategy based on dynamic programming (7).

## 3 The Ngram-based Reordering Approach

This section describes the Ngram-based Reordering approach. The aim of the NbR technique is to use an SMT system to deal with reordering problems. Therefore, the NbR technique can be seen as an SMT system that translates from an original source language (*S*) into a reordered source language (*S'*), given a target language (*T*).

The NbR approach uses a bilingual n-gram language model (hereafter, NbR Model) to translate from *S* to *S'*. This NbR Model is learned similarly to the bilingual n-gram translation language model. Here, the bilingual units contain reordering information. The last sentence in Figure 2 is a NbR bilingual unit and, therefore, is a candidate reordering hypothesis. Note that a bilingual unit of length 1 does not generate any reordering change. Therefore, the length of the reordering that can be captured depends directly on the length of the NbR bilingual units. Theoretically, the NbR approach manages to deal with local as well as long reorderings because there is no limit on the NbR bilingual units size. However, in practice, the probability of a match between input word classes and long reordering bilingual units is very low.

The NbR approach is used as a preprocessor for both training and test sentences, transforming the source sentences to be much closer to the target language. When reordering training sentences, NbR outputs only one reordered sentence. Whereas, when reordering test sentences, it additionally outputs several reordering hypotheses encoded in a graph.

[1] http://gps-tsc.upc.es/veu/soft/soft/marie/

(A) BILINGUAL S2T TUPLE:

**better and different structure # estructura mejor y differente # 1-1 1-2 2-3 3-4 4-1**

(B) MANY-TO-MANY WORD ALIGNMENT ⟶ MANY-TO-ONE WORD ALIGNMENT:

**better and different structure # estructura mejor y differente # 1-2 2-3 3-4 4-1**

(C) BILINGUAL S2S' TUPLE:

**better and different structure # 4 1 2 3**

(D) CLASS REPLACING:

**C36 C88 C185 C176 # 4 1 2 3**

Fig. 2. *Example of the extraction of NbR bilingual units. In (a) and (b) # divides the fields: source, target and word alignment, which includes the source and final position separated by -. In (c) and (d) # divides the source and positions of the reordered source.*

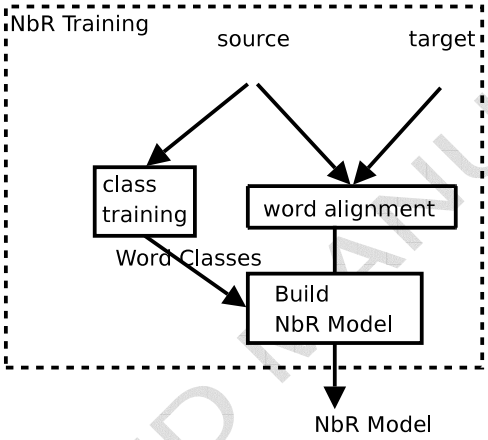



Fig. 3. *Block diagram of the NbR training.*

### *3.1 NbR training*

Figure 3 shows the block diagram of the NbR training, which mainly consists in training the word classes and the NbR Model.

Given the training source and target corpora (parallel at the sentence level), the NbR training is developed as follows:

(1) Use the source corpus to train statistical word classes (14).
(2) Align parallel training sentences at the word level.
(3) Extract reordering tuples, see Figure 2.
   (a) From word alignment and following the criteria in section 2, extract bilingual *S2T* units while keeping the word alignment information. Figure 2 (A) shows an example.
   (b) Modify the many-to-many word alignment to many-to-one. If one source

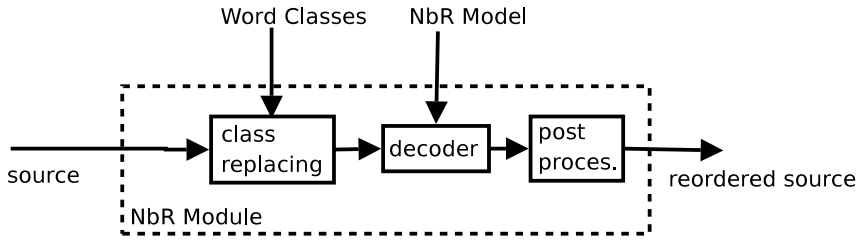



Fig. 4. *Block diagram of the NbR module.*

word is aligned to two or more target words, the most probable link given IBM-1 lexical probabilities ($P_{ibm1}$) is chosen, while the others are omitted. If $P_{ibm1}$(better, mejor) is higher than $P_{ibm1}$(better, estructura), then Figure 2 (A) leads to Figure 2 (B).

(c) From these *S2T* units, extract *S2S'* units which consist of a source fragment and its reordering. See Figure 2 (C).

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

(e) Replace the words of each source fragment with the classes determined in Step 1 (Figure 2 (D)).

(4) Compute the NbR Model, given the *S2S'* sequence.

### *3.2 NbR module*

Given the word classes and the NbR Model, we build the NbR module. Figure 5 shows the block diagram of the NbR module. The input is the initial source sentence (*S*) and the output is the reordered source sentence (*S'*). The NbR module is performed in three steps:

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

(2) Decoding. A monotonic decoding using the NbR Model is used 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 5.

### *3.3 NbR used before the SMT system*

**Training step.** The source corpus is processed by the NbR module to a reordered source corpus. The SMT system uses as training this reordered source corpus instead of the source corpus. As word alignment tends to be a computationally ex-

S: ***Sólo una sociedad plural , democrática y segura puede garantizar el ejercicio pleno***

**pleno de las libertades**

CLASS REPLACING: ***4 166 197 140 53 71 67 112 159 115 155 134 39 43 127***

DECODING: ***4#1 | 166#1 | 197 140 53 71 67 112# 2 3 4 5 6 1 | 119#1 | 159#1 | 115#1 | 155#1***

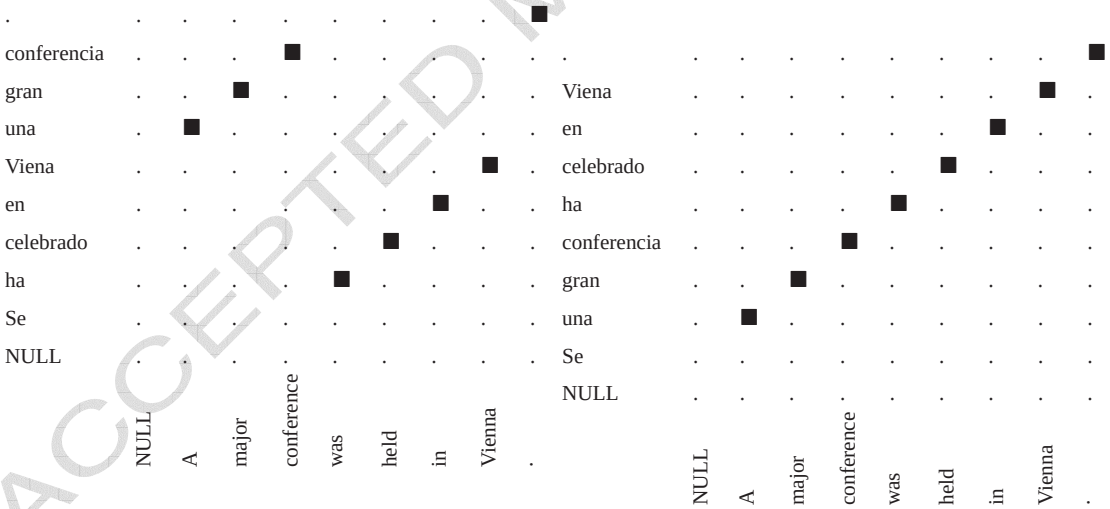
**| 134#1 | 39#1 | 43#1 | 127#1 |**

POST PROCESSING (S'): ***Sólo una plural , democrática y segura sociedad puede garantizar el***

**ejercicio pleno de las libertades**

Fig. 5. *Example of a source sentence reordering performed by the NbR module.The decoding output is shown in units: | separates the units and # separates the source and target part of the units. Note that the reordered sentence follows the order of the reference target sentence:* *Only a plural , democratic and secure society can guarantee the full exercise of freedoms.*

pensive task, the word alignment links are not recomputed. However, the alignment matrix may change. As an example see the difference from Table 1, which shows the *S2T* (left) and *S'2T* (right) word alignment and bilingual units. The SMT system (except for the word alignment) is trained on the S'2T task. Although the links from word alignments in Table 1 (left and right) remain the same, the extracted units change. In general, because of the unique segmentation, this modification in the word alignment matrix has benefits in the tuple extraction. The main advantage is the reduction of the unit vocabulary sparseness.




| | | |
|---|---|---|
| TUPLES | Se#NULL<br>ha celebrado en Viena una gran conferencia#<br>#A major conference was held in Viena<br>.#. | TUPLES | Se # NULL<br>**ha#was \|celebrado#held**<br>**en#in \| Viena#Vienna \| una#A**<br>**gran#major \| conferencia#conference** .#. |

Table 1
*S2T (left) and S'2T (right) word alignment and bilingual units*

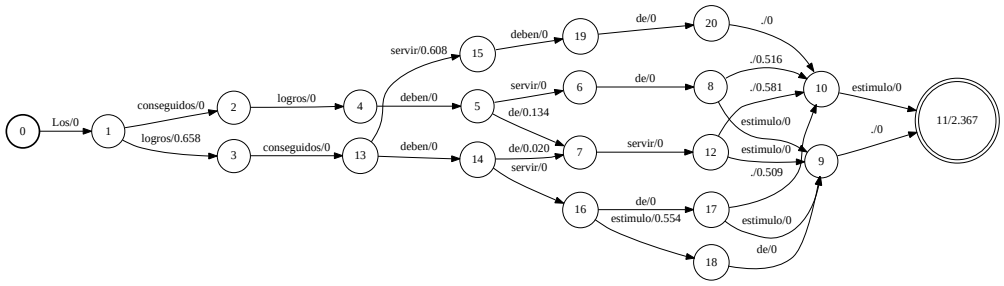



Fig. 6. *Weighted graph. The source sentence is:* Los logros conseguidos deben servir de estímulo. *The target sentence could be:* The achieved goals should be an encouragement.

**Test step.** The source corpus is processed by the NbR module and, afterwards, its output is given as input of the SMT system. The main difference at this stage is that this output/input may be a single-best or a graph:

**Single-best (NbR-1best).** The best output of the NbR system is the input of the SMT system. Here, the SMT decoder performs a monotonic search. Therefore, this approach does not increase the computational cost.

**Weighted Graph (NbR-WGraph).** The NbR technique generates an output graph that is introduced as an input graph for the SMT system. See a graph example in Figure 6. The weights of the reordering graph are the probabilities given by the NbR Model.

The SMT system in this case uses a non-monotonic decoding search. Therefore, all the SMT feature functions contribute to the search of the final reordering.

## 4 Evaluation Framework

### *4.1 Data*

Experiments are reported using three tasks: from Spanish, German and Arabic to English.

Table 2 shows some corpus statistics of the Spanish-to-English task. The corpus is based on the official version of the speeches held in the European Parliament Plenary Sessions (EPPS), as available on the web page of the European Parliament. Training, development and test set were used in the TC-STAR [2] official evaluations.

Additionally, experiments are shown on the German-to-English task with the data provided by the ACL 2007 Second Workshop on Statistical Machine Translation

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

| | | Spanish | English |
|---|---|---|---|
| Train | Sentences | 1.35M | |
| | Words | 39M | 37M |
| | Vocabulary | 147k | 109k |
| Dev | Sentences | 430 | |
| | Words | 15.7k | 16.2k |
| | Vocabulary | 3.2k | 2.7k |
| Test | Sentences | 892 | |
| | Words | 29.1k | 29.6k |
| | Vocabulary | 4.8k | 3.8k |
| | OOV | 94 | - |

Table 2
*Corpus Statistics for the Spanish to English task (EPPS). Official TC-STAR Evaluation.*

Evaluation (WMT) [3] (see Figure 3). One of the challenges of the evaluation was domain adaptation and we provide results with the out-of-domain task (News Commentary corpus).

Finally, experiments are also reported on the Arabic-to-English task with the *Basic Traveling Expression Corpus* (BTEC) provided by the evaluations of the International Workshop on Spoken Language Translation (IWSLT). This corpus consists of typical sentences from phrase books for tourists in several languages (20). This task provides a very limited amount of resources (see Figure 4) in comparison to the above tasks. We report results on the official test set of the IWSLT'07 evaluation, with six reference translations.

### *4.2 System Configuration Details*

**Word Alignment.** The word alignment was automatically computed by using GIZA++ [4] in both directions, which were symmetrized by using the union operation. Instead of aligning words themselves, stems were used for aligning. Afterwards, case sensitive words were recovered.

**Spanish Morphology.** A morphology reduction of the Spanish language was performed as a preprocessing step. As a consequence, training data sparseness due to

[3] http://www.statmt.org/wmt07/
[4] http://www-i6.informatik.rwth-aachen.de/õch/software/GIZA++.html

| | | German | English |
|---|---|---|---|
| Train EPPS | Sentences | 1.3M | |
| | Words | 34.7M | 36.7M |
| | Vocabulary | 321.7k | 121.1k |
| Train News-Commentary | Sentences | 59.7k | |
| | Words | 1,5M | 1,0M |
| | Vocabulary | 80,0k | 36,7k |
| Dev | Sentences | 1 057 | |
| | Words | 26.1k | 25.7k |
| | Vocabulary | 6.4k | 5.0k |
| Test | Sentences | 2 007 | |
| | Words | 50,9k | 49,7k |
| | Vocabulary | 10,2k | 7,5k |
| | OOV | 692 | - |

Table 3
*Corpus Statistics for the German to English task (EPPS and News-Commentary). Official WMT07 Evaluation.*

Spanish morphology was reduced improving the performance of the overall translation system. The pronouns attached to the verb were separated and contractions as *del* or *al* were split into *de el* or *a el*.

**Arabic Morphology.** We used the Buckwalter Arabic Morphological Analyzer [5] to obtain possible word analysis, and disambiguate them using the Morphological Analyzer and Disambiguation for Arabic (MADA) tool (8), kindly provided by the University of Columbia.

**Word classes.** We consider some conclusions from previous works regarding the use of classes. The use of word classes in NbR was empirically justified in (4). Moreover, several experiments were performed in (18) comparing statistical versus morphological classes in the EPPS Es2En task. Statistical classes, which were built with 'mkcls' [6], outperformed morphological classes. Therefore, for the EPPS Es2En task, word classes that were extracted without taking into account any linguistic information seemed to perform better than the other way round. This conclusion was assumed to hold for the De2En and the Ar2En tasks. Notice that several

[5] Version 2.0. Linguistic Data Consortium Catalog: LDC2004L02
[6] A free tool available with GIZA++

| | | Arabic | English |
|---|---|---|---|
| Train | Sentences | 24,4k | |
| | Words | 189k | 170k |
| | Vocabulary | 10,9k | 6,9k |
| Dev | Sentences | 489 | |
| | Words | 5,9k | 6,4k |
| | Vocabulary | 1,2k | 1,1k |
| Test | Sentences | 489 | |
| | Words | 3,2k | 3,5k |
| | Vocabulary | 976 | 920 |
| | OOV | 115 | - |

Table 4
*Corpus Statistics for the Arabic to English task (BTEC). Official IWSLT07 Evaluation.*

related approaches from Section 1 may work using statistical classes, although no comparison is presented in any of them.

**NbR parameters.** The bilingual unit extraction did not have any limit over unit lengths. The NbR Model was a 5gram (4gram in the BTEC task) back-off language model with Kneser-Ney smoothing and was built with the SRILM toolkit (19).

**SMT parameters.** Again, the tuple extraction did not have any limit over tuple lengths. The Ngram translation model was a 4gram back-off language model with Kneser-Ney smoothing. Pruning was performed by keeping the $N$ most frequent tuples with common source sides ($N = 20$). The target language model was a 5gram (4gram in the BTEC task) back-off language model with Kneser-Ney smoothing.

**Optimization.** An n-best re-ranking strategy was implemented for optimization purposes [7] . The optimization search used the Simplex algorithm with the BLEU score as the objective function.

**Case sensitive evaluation.** Translation results were evaluated in terms of BLEU [8] , NIST, mPER and mWER.

[7] as proposed in *http://www.statmt.org/jhuws/*
[8] http://www.nist.gov/speech/tests/mt/resources/scoring.htm

| | NB | NbR+NB | Relative increment |
|---|---|---|---|
| **Es2En** 1-word tuples | 439.6k | 507.5k | 15.4% |
| >1-word tuples | 1.8M | 1.3M | -36% |
| **De2En** 1-word tuples | 618.1k | 711k | 15% |
| >1-word tuples | 2.1M | 1.7M | -25% |
| **Ar2En** 1-word tuples | 9.5k | 10.3k | 8% |
| >1-word tuples | 20.4k | 18.8k | -8% |

Table 5
*Variation in the size of the translation vocabulary (1-word and longer than 1-word tuples).*

### *4.3 Translation Units Analysis*

This section shows an analysis of the translation units of the baseline SMT system (NB) and the baseline SMT system enhanced with the NbR system (NbR+NB).

As a consequence of reordering the source training, there are fewer crossings in word alignments. In an Ngram-based system the non-monotonicity poses difficulties for units extraction. The tuple length is defined as the number of words in the source side. There are a greater number of shorter units in the case of the NbR+NB system (shorter units lead to a reduction in data sparseness). In Spanish to English, the most common reordering is the swapping of two words. Hence, the most important reduction is seen in tuples of length two.

NB: hablar en esta Asamblea de manera provechosa e interesante#useful and interesting discussions in this House

NbR+NB: de#NULL | manera#NULL | provechosa#useful | e#and | interesante#interesting | hablar#discussions | en#in | esta#this | Asamblea#House

NB: aus der Presse und dem Fernsehen wissen#aware from the press and television

NbR+NB: wissen#aware | aus#from | der#the | Presse#press | und#and | dem#NULL | Fernsehen#television

NB: hl#is | ywjd#there | any grfp b+ sEr >rxS#cheaper room

NbR+NB: hl#is | ywjd#there any | b+#NULL sEr >rxS#cheaper | grfp#room

Fig. 7. *Examples of tuples extracted from a training sentence pair in the baseline (NB) and in the enhanced (NbR+NB) system. Symbol | separates units and symbol # separates source and target inside a unit. Arabic is written in Buckwalter.*

Figure 7 shows an example of how SMT units are modified when using the NbR approach as preprocessing in SMT training. Clearly, the SMT units are reduced in length. As a consequence, there is a reduction in SMT vocabulary shown in Table 5.

### *4.4 Translation Results*

Table 6 presents the BLEU, NIST, mPER and mWER scores obtained for the EPPS data set comparing the NB and NbR+NB approaches.

The NbR approach improves all measures, especially when using the graph approach. The gain in BLEU goes from 2 points absolute to more than 4 (BLEU in %). Additionally, Table 6 provides the number of words and the time of the translation. There is a moderate increase of the computational cost in the non-monotonic search: time increases around 50% in the longer tasks and a little bit more in the smallest task.

| System | BLEU | NIST | PER | WER | Words | Time |
|---|---|---|---|---|---|---|
| Es2En | | | | | | |
| NB | 52.57 | 10.64 | 26.63 | 36.97 | 29.2k | 50.0’ |
| NbR-1best + NB | 52.95 | 10.62 | 26.84 | 36.96 | 29.4k | 52.8’ |
| NbR-WGraph + NB | **54.51** | **10.81** | 26.24 | **35.67** | 29.0k | 78.7’ |
| De2En | | | | | | |
| NB | 21.30 | 6.69 | 47.90 | 65.61 | 45.0k | 85.9’ |
| NbR-1best + NB | 21.61 | 6.82 | 46.80 | 64.79 | 44.6k | 76.9’ |
| NbR-WGraph + NB | **23.30** | **7.02** | **46.10** | **62.98** | 43.9k | 125’ |
| Ar2En | | | | | | |
| NB | 45.00 | 7.65 | 34.92 | 39.15 | 3.6k | 3.1’ |
| NbR-1best+NB | 46.45 | 7.90 | 32.32 | 36.73 | 3.6k | 2.1’ |
| NbR-WGraph+NB | **49.35** | **8.13** | **30.70** | **34.23** | 3.5k | 6.5’ |

Table 6
*Translation results and computational time for several couplings of the NbR and SMT.*

Finally, Figure 8 shows typical examples of translated sentences, where the NB baseline system is compare to the NbR+NB system. Es2En language pair usually requires local reorderings like noun plus adjective that swaps from one language to the other. German has traditionally been considered problematic because of the position of the verb, which is in second position in a main clause and at the end in a subordinate clause. De2En examples report reorderings involving up to five words which handle this change in verb position. Finally, in general the Ar2En task tends to present local reorderings. Additionally, given the tourist domain of the Ar2En BTEC task, test sentences are not very long. Examples show local reorderings either in question or affirmative sentences.

| |
|---|
| NB: not only through a compromise economic immediately<br>NbR+NB: not only through an **immediate economic commitment**<br>REF: not only through an immediate economic commitment |
| NB: The European Union must be a political element essential for the fight against terrorism<br>NbR+NB: The UE should be an **essential political element** in the fight against terrorism<br>REF: The UE must be an essential political element to fight against terrorism |
| NB: The Group of the European Peoples has asked (...)<br>NbR+NB: The **European Peoples Group has requested (...)**<br>REF:The European Popular Group asked (...) |
| NB: Iraq needs several years a new constitution to write<br>NbR+NB: Iraq needs **several years to write a new constitution,**<br>REF: Iraq needs several years to write a new constitution, |
| NB: EU membership has result in a state decisive measures must accept<br>NbR+NB: the EU membership result, **a state must accept radical measures,**<br>REF: EU membership entails having to accept incisive measures. |
| NB: (..) that the death penalty threatened murderers go further would her arrest to escape,<br>NbR+NB: (..) **that the death penalty threatened murderers would go even further,**<br>REF: (..) that capital punishment may make a murderer fight harder |
| NB: Broke one of them room.<br>NbR+NB: Someone **broke them our room.**<br>REF: Someone broke into our room. |
| NB: Can you discount a little it?<br>NbR+NB: **Can you discount it a little?**<br>REF: Can't you lower the price? |
| NB: I'm sorry but not this what I think.<br>NbR+NB: **I'm sorry but this is not what I think.**<br>REF: I'm sorry but this is not what I have in mind. |

Fig. 8. *Translation examples from the* NB *and* NBR+NB *systems: Es2En, De2En and Ar2En (from top to bottom).*

### *4.5 Reordering Comparison*

In order to provide a comparison to the NbR approach, this section shows results of an Ngram-based system with a standard distance-based reordering model. Typically, a distance-based reordering model is used during the search to penalize longest reorderings, only allowed when well supported by the rest of models. Here, the implemented distance-based reordering model will be referred to as *m5j3* which corresponds to a search allowing for a fully reordered search constrained to a five words window limit and a maximum of three reorderings per sentence. This configuration introduces a distance-based reordering model in the log-linear combination corresponding to the next equation:

$$p_{db}(t_K) = exp(\sum_{k=1}^{K} d_k)$$

where $d_k$ is the distance between the first word of the $k$th tuple ($t_k$), and the last word +1 of the $k$1th tuple (distances are measured in words referring to the units source side).

Here, we compare the distance-based reordering model with the NbR reordering approach in the Ar2En task. Automatic measures in Table 7 show that NbR+NB outperforms the distance-based reordering in the Ar2En tasks. A manual analysis of the translations show that reorderings were better solved when using the NbR approach. Moreover, the computational cost of the $m5j3$ search is clearly higher (almost 10 times) than the cost of the NbR search, despite of being both algorithms of the same complexity. The $m5j3$ search graph contains about three times more partial hypotheses (thus archs) than the corresponding NbR search graph.

| System | BLEU | NIST | PER | WER | Words | Time |
|---|---|---|---|---|---|---|
| NB + m5j3 | 46.28 | 7.75 | 30.96 | 35.01 | 3.3k | 59.1’ |
| NbR-WGraph+NB | **49.35** | **8.13** | **30.70** | **34.23** | 3.5k | 6.5’ |

Table 7
*Distance-based reordering vs best configuration of the NbR reordering approach (Arabic to English task).*

We do not report results with the Es2En and De2En tasks, because the distance-based reordering is shown not to improve the Ngram-based baseline system in the same EPPS tasks (6).

## 5 Conclusions

The Ngram-based reordering approach addresses the reordering challenge in SMT by using the same powerful statistical translation techniques to generate source reordering hypotheses. The use of an *n*-gram language modeling permits to further learn ordering context.

NbR allows for a reduction of the vocabulary sparseness of the Ngram-based SMT system during the training phase.

The fact of using classes to train the reordering hypothesis (instead of words themselves) allows to generalize in the test phase. Therefore, the NbR technique is able to generate reordering hypotheses of sequences of words which were not seen during training. Additionally, the NbR technique provides a smoothed context-based weight to each reordering hypothesis by taking advantage of the highly developed language model techniques.

Although introducing reordering abilities increases the system computational cost, experiments show that using the NbR technique guides the final translation decoding in an efficient manner.

Reordering with the NbR technique highly outperforms our monotonic baseline system and a non-monotonic baseline system with a standard distance-based reordering.

## References

[1] P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, P. Roossin, A statistical approach to machine translation, Computational Linguistics 16 (2) (1990) 79–85.

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

[3] M. Collins, P. Koehn, I. Kucerová, Clause restructuring for statistical machine translation, in: Proc. of the 43th Annual Meeting of the Association for Computational Linguistics, Michigan, 2005.

[4] M. Costa-jussà, J. Fonollosa, Statistical machine reordering, in: Proc. of the Conf. on Empirical Methods in Natural Language Processing, EMNLP'06, Sydney, 2006.

[5] M. Costa-jussà, J. Fonollosa, Using reordering in statistical machine translation based on alignment block classification, internal Report:http://gps-tsc.upc.es/veu/personal/mruiz/ (2006).

[6] J. Crego, J. Mariño, Improving SMT by coupling reordering and decoding, Machine Translation 20 (3) (2007) 199–215.

[7] J. Crego, J. Mariño, A. de Gispert, An Ngram-based statistical machine translation decoder, in: Proc. of the 9th Int. Conf. on Spoken Language Processing, ICSLP'05, Lisboa, 2005.

[8] N. Habash, O. Rambow, Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop, in: Proc. of the 43th Annual Meeting of the Association for Computational Linguistics, Ann Arbor, MI, 2005.

[9] S. Kanthak, D. Vilar, E. Matusov, R. Zens, H. Ney, Novel reordering approaches in phrase-based statistical machine translation, in: Annual Meeting of the Association for Computational Linguistics: Proc. of the ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond (WMT), Ann Arbor, MI, 2005.

[10] K. Knight, Y. Al-Onaizan, Translation with finite-state devices, in: Proc. of the 4th Conf. of the Association for Machine Translation in the Americas, AMTA'02, Langhorne, 1998.

[11] P. Koehn, F. Och, D. Marcu, Statistical phrase-based translation, in: Proc. of the Human Language Technology Conf., HLT-NAACL'03, Edmonton, Canada, 2003.

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

[13] A. Mauser, E. Matusov, H. Ney, Training a statistical machine translation system without giza++, in: 5th Int. Conf. on Language Resources and Evaluation, LREC'06, Genova, 2006.

[14] F. Och, An efficient method for determining bilingual word classes, in: Proc. of the 9th Conf. of the European Chapter of the Association for Computational Linguistics, 1999.

[15] F. Och, H. Ney, Discriminative training and maximum entropy models for statistical machine translation, in: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, USA, 2002.

[16] M. Popovic, H. Ney, Pos-based word reorderings for statistical machine translation., in: 5th Int. Conf. on Language Resources and Evaluation, LREC'06, Genoa, 2006.

[17] C. Quirk, A. Menezes, C. Cherry, Dependency treelet translation: Syntactically informed phrasal SMT, in: Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Ann Arbor, Michigan, 2005.

[18] M. R. Costa-jussà, J. A. R. Fonollosa, Analysis of atatistical and morphological classes to generate weighted reordering hypotheses on a statistical machine translation system, in: Annual Meeting of the Association for Computational Linguistics: Proc. of the Second Workshop on Statistical Machine Translation (WMT), Prague, 2007.

[19] A. Stolcke, SRILM - an extensible language modeling toolkit, in: Proc. of

the 7th Int. Conf. on Spoken Language Processing, ICSLP'02, Denver, USA, 2002.

[20] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, S. Tamamoto, Toward a broad coverage bilingual corpus for speech translation of travel conversations in the real world, in: 3th Int. Conf. on Language Resources and Evaluation, LREC'02, Las Palmas, 2002.

[21] C. Wang, M. Collins, P. Koehn, Chinese syntactic reordering for statistical machine translation, in: Proc. of the 2007 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, 2007.

[22] F. Xia, M. McCord, Improving a statistical mt system with automatically learned rewrite patterns, in: Proc. of the 20th International Conference on Computational Linguistics, Morristown, 2004.

[23] R. Zens, F. Och, H. Ney, Phrase-based statistical machine translation, in: M. Jarke, J. Koehler, G. Lakemeyer (eds.), KI - 2002: Advances in artificial intelligence, vol. LNAI 2479, Springer Verlag, 2002, pp. 18–32.

[24] Y. Zhang, R. Zens, H. Ney, Chunk-level reordering of source language sentences with automatically learned rules for statistical machine translation, in: Proc. of the Human Language Technology Conf. (HLT-NAACL'06):Proc. of the Workshop on Syntax and Structure in Statistical Translation (SSST), Rochester, 2007.