N-BEST REORDERING IN STATISTICAL MACHINE TRANSLATION

Germán Sanchis[*], *Francisco Casacuberta*[*]

[*]Departamento de Sistemas Informaticos y Computación
Universidad Politécnica de Valencia
Camino de Vera, s/n. 46071 Valencia, Spain
{gsanchis,fcn}@dsic.upv.es

ABSTRACT

As statistical machine translation (SMT) systems strive to improve the translation quality they are able to deliver, the word reordering problem is being unveiled as a major problem that must be addressed, whenever these systems are to be improved. While most works published focus their results in corpora involving English, Chinese and Arabic, such a translation problem can also be found within Spain itself: its origin being unknown, Basque presents a very peculiar word order, which is very different to most other european languages, and specially very different to Spanish word order. Because of this fact, SMT systems not including some sort of word reordering yield unsatisfactory results, involving serious training problems, when confronted with the Basque-Spanish task. Although some efforts have been made towards including word or phrase reordering in the decoding algorithms, these approaches usually imply a computational overhead that obliges the designers of such algorithms to assume sub-optimal restrictions, which often lead to a significant dimish in the translation quality. In this work, we present a reordering method based on the extraction and exploitation of monotized corpora, which prove to be specially useful for the language pairs presenting severe word reorderings. Our system has been tested on the Basque Tourist task, where very promissing results have been obtained.

1. INTRODUCTION

SMT systems have proved in the last years to be an efficient way of building machine translation systems which, with little need of human supervision, are able to deliver a translation quality very similar, if not better, than most commercial machine translation systems, which are usually characterized by having a very important development effort behind each pair of languages they intend to translate.

Statistically, the machine translation problem can be defined as, given a sentence s from a certain source language, the search for an adequate sentence $\hat{t}$ that maximises the posterior probability:

$$\hat{t} = \underset{t}{\text{argmax}}\, Pr(t|s)$$

However, this probability is decomposed, in the vast majority of cases, into two different probabilities, the first being the target statistical language model and the second one being the translation model. Hence, the previous equation is transformed by means of the Bayes' theorem to obtain the following equation:

$$\hat{t} = \underset{t}{\text{argmax}}\, Pr(t) \cdot Pr(s|t)$$

More intuitively, the translation model $Pr(s|t)$ will capture word relations between both languages, and is normally based on stochastic dictionaries and alignment models, whereas the language model $Pr(t)$ will award a higher probability to well-formed sentences from the target language.

Monotonous SMT systems are nowadays the most used in research for Machine Translation. The reason is simple, but conclusive: non-monotonous models are computationally too expensive, needing restrictions to be applied for them to work. Being necessary, these restrictions are quite often applied in training and search, leading to a substantial reduction of the model's performance.

Because of this reason, most SMT systems that have been developped in the last years, including the vast majority of the state of the art systems, either implement a Phrase-Based model [1, 2] or Weighted Finite State Transducers for Machine Translation [3, 4]. These systems are inherently monotonous, and are usually estimated using word-level aligned corpora, hence often incurring in ordering related errors. Although they try to allow some sort of reordering, such reorderings are quite often very limited, and are not able to account for the more wild reorderings that take place in languages from very different origins. This does not only imply that the output sentence will present a grammatically incorrect word order, but, moreover, the parameters of the systems trained can be estimated incorrectly, by assuming that a given input phrase is the translation of a certain output phrase, leading this assumption in many cases to incorrect translations.

The problem of word reordering has been tackled with already since the origin of what is now a days known as machine translation: Berger et al. [5] already introduced in their alignment models what they called distortion models, in an effort towards including in their SMT system a solution for the reordering problem. However, these distortion models are usually implemented within the decoding algorithms and imply serious computational problems, leading ultimately to restrictions being applied to the set of possible permutations of the output sentence. Hence, the search performed turns sub-optimal, and an important loss in the representational power of the distortion models takes place.

The most nave - but effective - approach to the reordering problem would be to allow for arbitrary word reorderings, and choose the one that obtains the highest score in some reordering scoring model. However, when allowing all possible word permutations the search has been proved to be NP-hard [6].

In our work we present a novel approach to solve the reordering problem, based on the work of Zens, Matusov and Kanthak [7, 8, 9], who introduce the idea of monotonizing a corpus, i.e. using the alignments produced by the IBM models to reorder the input sentence s and produce a new bilingual pair, composed by the reordered input sentence s' and the output sentence t, whose translation is monotonous. Once this is achieved, any random monotonous translation model may be trained without the problems derived from word reorderings. In this paper, the monotonized corpus will be used to train a Phrase Based model.

However, a crucial problem has to be addressed when trying to learn models from monotonized corpora: in search time, the output sentence is not available, hence the need for a model that will be able to learn how to reorder efficiently a given input sentence. Our approach copes this problem by learning a very simple reordering model and generating an n-best list of reordering hypothesis. After this set of hypothesis has been translated, only the best scoring sentence with respect to the translation model is kept as final output sentence.

The next section briefly describes some of the latest efforts made towards solving this problem. Then, in section 3, we will explain our approach, based on an n-best list of best reordering hypothesis. In section 4 we will describe the experiments performed with our systems and the results obtained. Finally, in section 5 we will present the conclusions that can be elucidated from the experiments described, as well as the work that is currently in progress.

2. BRIEF REVIEW OF EXISTING APPROACHES

When facing the reordering problem for Machine Translation, three main possibilities exist: output sentence reordering, input sentence reordering and the reordering of both, the latter being as of yet unexplored.

- Let s be a source sentence, t a target sentence.
- Let C be a cost matrix, such that c_{ij} = cost of aligning s_j to t_i.
- Let $\{s^r\}$ be the set of all possible permutations of s.

```
1. compute alignment $A_D(j) = \operatorname{argmin}_i c_{ij}$
2. $s' = \{s^r | \forall j : A_D(j) \leq A_D(j+1)\}$
3. recompute (reorder) $C$, obtaining $C'$.
4. set $A'_I(i) = \operatorname{argmin}_j c'_{ij}$.
5. Optional: Compute the minimum-cost monotonic path through the cost matrix $C'$.
```

Figure 1. Algorithm for obtaining a monotonic alignment by reordering the source sentence.

2.1. Output sentence reordering

Two main approaches have been tested in this direction, the first one being originally developed by J.M.Vilar et al. [10], and more recently taken up again at the TALP Research Center [11]. The idea behind this approach is to monotonize the most probable non-monotonous alignment patterns and add a mark in order to be able to remember the original word order. This being done, a new output language has been defined and a new language and translation model can be trained, being the translation process now monotonous.

Another approach more recently tested by [12] involves learning weighted finite state transducers that account for local reorderings of two or three positions, allowing each word to jump a maximum of one or two positions. They applied these models for phrase reordering, but the training of the models did not yield statistically significant results with respect to the introduction of the models with fixed probabilities.

2.2. Input sentence reordering and corpora monotonization

The main idea behind this approach developed at the RWTH-Aachen [7, 8, 9], is to avoid the non-monotonous translation problem by reordering the input sentence in such a way that the translation model will not need to account for possible word reorderings. To achieve this, alignment models are used, in order to establish which word order should be the appropriate for the translation to be monotonous, and then the input sentence is reordered in such a manner that the alignment is monotonous. The algorithm performing such a reordering is described in Figure 1.

However, this approach has an obvious problem, since the output sentence is not available in search time and the sentence pair cannot be monotonized. The nave solution, test on all possible permutations of the input sentence, has already been discussed earlier, being NP-hard [6], as $J!$ possible permutations can be obtained from a sentence of length J. Hence, the search space must be restricted.

On their paper, Kanthak et al. [9] describe four constraints.

- IBM constraints, described in Berger et al. [5]. At each moment, the first k yet uncovered word positions can be chosen to be the next word, being an uncovered word position one such position that has not yet been chosen.
- Inverse IBM constraints. In this case, any word can be chosen, except if $k-1$ words on positions $j' > j$ have already been chosen, in which case j must be the next word. This is specially useful for language pairs in which we need to translate words at the end of the sentence, and then translate the rest nearly monotonically.
- Local constraints. Here, the next word must be contained within the first k positions, starting from the first yet uncovered position, and counting both covered and not covered positions. This constraint is useful for language pairs in which words are only shifted a few positions.
- ITG constraints. Being inspired by the bilingual bracketing of inverse transduction grammars (ITG) [13], these constraints have experimentally proven to be useful in Machine Translation. In this approach, the sentence is parsed into a binary tree. In the nodes of such a tree, we can choose either to invert the original order, or leave it as is. Because of its nature, ITG constraints forbid combinations such as (3, 1, 4, 2) or (2, 4, 1, 3).

3. THE REORDERING MODEL AND N-BEST REORDERINGS

However, and although the constraints described above did yield interesting results, the search space containing the permutations still remains huge, paying a very expensive computational price for the relatively small benefit obtained.

In addition, the reordering constraints described in the work by Kanthak et al. [9] disregards the crucial information that can be extrated from the monotonized corpora, for it is within this type of corpus that lies the information needed to train a reordering model which, ideally,would be capable of reordering the input sentences in such a manner that the translation following would be monotonous. Although such a model is bound to introduce error into the overall system, but the benefits obtained may well be worth it.

The main idea behind our approach is that monotonized alignments define a new source ”*language*”, and hence a reordered language model can be trained with the reordered input sentences s'. Such a reordered language will have the same vocabulary as the source language, but the word ordering of the target language, and hence a reordering model can be trained from the monotonized corpus. Moreover, the reordering model will most likely not need to depend on the output sentence, whenever the word-by-word translation is accurate enough.

Hence, the reordering problem can be defined as follows:

$$s' = \underset{s^r}{\operatorname{argmax}} Pr(s^r) \cdot Pr(s|s^r)$$

where $Pr(s^r)$ is the reordered language model, and $Pr(s|s^r)$ is the reordering model. Being this problem very similar to the translation problem but with a very constrained translation table, it seems only natural to use the same methods developed to solve the translation problem to face the reordering problem. Hence, in this paper we will be using an exponential model as reordering model, defined as:

$$Pr(s|s') \approx exp(-\sum_i d_i)$$

where d_i is the distance between the last reordered word position and the current candidate position.

However, and in order to reduce the error that will introduce a reordering model into the system, we found that it is very useful to compute an n-best list of reordering hypothesis and translate them all, selecting then as final output sentence the one which obtains the highest probability according to the models $Pr(t) \cdot Pr(s|t)$. Ultimately, what we are actually doing with this procedure is to contrain the search space of permutations of the source sentence as well, but taking into account the information that monotonized alignments entail, together with a much stronger restriction of the search space than previous approaches, reducing significantly the computational effort needed.

4. TRANSLATION EXPERIMENTS

4.1. Corpus characteristics

Our system has been tested on the Basque-Spanish translation task, a tough problem with no trivial solution in which reordering plays a crucial role.

The corpus chosen for this experiment is the *Tourist* corpus [14], which is an adaptation of a set of Spanish-German grammars generating bilingual sentence pairs [15] in such languages. Hence, the corpus is semi-synthetic. In this task, the sentences describe typical human dialogs in the reception desk of a hotel, and they have been mainly extracted from tourist guides. However, because of its design, there is some asymmetry between both languages,

Table 1. Characteristics of the Tourist corpus.

		Spanish	Basque
Training	Sentences	38940	
	Different pairs	20318	
	Words	368314	290868
	Vocabulary	722	884
	Average length	9.5	7.5
Test	Sentences	1000	
	Test independent	434	
	Words	9507	7453
	Average length	9.5	7.5

and a concept being expressed in several manners in the source language will always be translated in the same manner in the target language. Because of this, the target language is meant to be simpler than the source language. Since the input language during the design of the corpus was Spanish, the vocabulary size of Basque should be smaller. Actually, however, the vocabulary size of Basque is bigger than that of Spanish, and this is due to the agglomerative nature of the Basque language. The corpus has been divided into two separate subsets, a bigger one for training and a smaller one for test. The characteristics of this corpus can be seen in Table 1.

4.2. System evaluation

The SMT system developed has been automatically evaluated by measuring the following rates:

WER *(Word Error Rate)*: The WER criterion is similar to the edit distance for Speech Recognition, computing the minimum number of editions (substitution, insertion and deletion operations) needed to convert the translated sentence into the sentence considered ground truth. This measure is because of its nature a pessimistic one. For example, input sentences may allow many different translation, but the WER criterion will penalize all those translations which hold a difference with respect to the translation considered ground truth.

PER *(position-independent WER)*: The PER criterion is similar to WER, but word order is ignored. This criterion accounts for the fact that a system may produce an acceptable translation differing only in word order. This sentence may even be grammatically correct, but will be nevertheless penalized when using the WER criterion.

BLEU *(BiLingual Evaluation Understudy) score*: This score measures the precision of unigrams, bigrams, trigrams, and 4-grams with respect to a whole set of reference translations, with a penalty for too-short sentences [16]. It must be noted that BLEU measures accuracy, not error rate, which means that the higher the BLEU score, the better.

Table 2. Results for Spanish to Basque translation.

	Baseline translation	Reordered translation
WER	19.5%	10.9%
BLEU	81.0%	87.1%
PER	6.2%	4.9%

4.3. Experimental setup and translation results

We used the reordering technique described above to obtain an n-best reordering hypothesis list and translate them, keeping the best scoring one. In this case, n was set to 5.

First, the bilingual pairs were aligned using IBM model 4 by means of the GIZA++ toolkit [17]. After this, the alignments were monotonized in the fashion described in [9] and a new alignment was recalculated, determining the new monotonous alignment between the reordered source sentence and the target, and a reordered source sentence language model was built. Phrase extraction was performed by using the Thot toolkit [18].

For the next step, the reordering model, we used the reordering model built in the toolkit Pharaoh. This was done by including in the translation table only the words contained in the vocabulary of the desired source language, and allowing the toolkit to reorder the words by taking into account the language model and the phrase-reordering model it implements, which is an exponential model. Since in this case, the phrases are just words, what results is an effective implementation of an exponential word-reordering model, just as we wanted.

Once the 5 best reordering hypothesis had been calculated, we translated them all by using the toolkit Pharaoh once again, and kept just the best scoring translation, where the score is determined as the product of the (inverse) translation model and the language model.

The results of this setup can be seen in Table 2. As a baseline, we took the results of translating the same test set, but without the reordering pipeline, i.e. just using GIZA++ for aligning, Thot for phrase extraction and Pharaoh for translating.

In these results it can be seen that the reordering pipeline established does have significant benefits on the overall quality of the translation, almost achieving even a relative improvement of 50% of the Word Error Rate. Furthermode, it is interesting to point out that even in the case of the PER criterion the results obtained are better. At first sight, this might seem odd, since the PER criterion does not take into account possible ordering errors within the sentence, which is a main problem that every reordering technique tries to solve. However, we found that this improvement is due to the fact that reordering the source sentence allows for better phrases to be extracted.

5. CONCLUSIONS AND FUTURE WORK

A reordering technique has been implemented, taking profit of the information that monotonized corpora provide. By doing so, better quality phrases can be extracted and the overall performance of the system improves significantly in the case of a pair of languages which present heavy reordering complications.

This technique has been applied to translate a semi-synthetic corpus which deals with the task of Spanish-Basque translation, and the results obtained prove to be statistically signficant and show to be very promissing. Moreover, the technique we propose in this paper is learnt authomatically, without any need of linguistic annotation or manually specified syntatic reordering rules, which means that out technique can be applied to any language pair without need for any additional development effort.

Both reordered corpora and reordering techniques seem to have a very important potential for the case of very different language pairs, which are the most difficult translation tasks.

As future work, we are planning on obtaining results with other non-synthetic, richer and more complex corpora, as may be other Spanish-Basque corpora or corpora involving language pairs such as Arabic, Chinese or Japanese. In addition, we are planning on developping more specific reordering models, which will be more suitable for this task than the exponential model described here.

Acknowledgements

This work has been partially supported by Spanish project TIC 2003-08681-C02-02 and by the MEC scholarship AP2005-4023.

6. BIBLIOGRAPHY

[1] P. Koehn, F.J. Och, y D. Marcu, "Statistical phrase-based translation," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Edmonton, Canada, 2003, vol. 1, pp. 48–54.

[2] F.J. Och y H. Ney, "The alignment template approach to statistical machine translation," *Computational Linguistics*, vol. 30, no. 1, pp. 417–449, 2004.

[3] F. Casacuberta y E. Vidal, "Machine translation with inferred stochastic finite-state transducers," *Computational Linguistics*, vol. 30, no. 2, pp. 205–225, 2004.

[4] S. Kumar y W. Byrne, "A weighted finite state transducer implementation of the alignment template model for statistical machine translation," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Edmonton, Canada, 2003, vol. 1, pp. 63–70.

[5] A.L. Berger, P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, J.R. Gillet, A.S. Kehler, y R.L. Mercer, "Language translation apparatus and method of using context-based translation models," in *United States Patent 5510981*, 1996.

[6] K. Knight, "Decoding complexity in word-replacement translation models," *Computational Linguistics*, vol. 25, no. 4, pp. 607–615, 1999.

[7] R. Zens, H. Ney, T. Watanabe, y E. Sumita, "Reordering constraints for phrase-based statistical machine translation," in *COLING '04: The 20th Int. Conf. on Computational Linguistics*, Geneva, Switzerland, 2004, pp. 205–211.

[8] E. Matusov, S. Kanthak, y H. Ney, "Efficient statistical machine translation with constrained reordering," in *In Proceedings of EAMT 2005 (10th Annual Conference of the European Association for Machine Translation)*, Budapest, Hungary, 2005, pp. 181–188.

[9] S. Kanthak, D. Vilar, E. Matusov, R. Zens, y H. Ney, "Novel reordering approaches in phrase-based statistical machine translation," in *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, Ann Arbor, Michigan, 2005, pp. 167–174.

[10] J.M. Vilar, E. Vidal, y J.C. Amengual, "Learning extended finite-state models for language translation," in *Proc. of Extended Finite State Models Workshop (of ECAI'96)*, Budapest, August 1996, pp. 92–96.

[11] A. de Gispert y J.B. Mario, "Experiments in word-reordering and morphological preprocessing for transducer-based statistical machine translation," in *Proc. of 2003 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'03)*, USA, 2003.

[12] S. Kumar y W. Byrne, "Local phrase reordering models for statistical machine translation," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Vancouver, Canada, 2005, pp. 161–168.

[13] D. Wu, "Stochastic iversion transduction grammars and bilingual parsing of parallel corpora," *Computational Linguistics*, vol. 23, no. 3, pp. 377–403.

[14] A. P´erez, F. Casacuberta, M.I. Torres, y V. Guijarrubia, "Finite-state transducers based on k-tss grammars for speech translation," in *Proceedings of Finite-State Methods and Natural Language Processing (FSMNLP 2005)*, Helsinki, Finland, September 2005, pp. 270–272.

[15] E. Vidal, "Finite-state speech-to-speech translation," in *Int. Conf. on Acoustics Speech and Signal Processing (ICASSP-97), proc., Vol.1*, Munich, Germany, 1997, pp. 111–114.

[16] Papineni, A. Kishore, S. Roukos, T. Ward, y W. Jing Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Technical Report RC22176 (W0109-022)*, IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY, 2001.

[17] F. J. Och y H. Ney, "Improved statistical alignment models," Hongkong, China, October 2000, pp. 440–447.

[18] D. Ortiz, I. Garca-Varea, y F. Casacuberta, "Thot: a toolkit to train phrase-based statistical translation models," in *Tenth Machine Translation Summit*, pp. 141–148. Asia-Pacific Association for Machine Translation, Phuket, Thailand, September 2005.