Reordering via N-Best Lists for Spanish-Basque Translation

Germán Sanchis, Francisco Casacuberta
Instituto Tecnológico de Informática
Universidad Politcnica de Valencia
Camino de Vera, s/n. 46022 Valencia, Spain
{gsanchis,fcn}@dsic.upv.es

Abstract

In Statistical Machine Translation (SMT), one of the main problems they are confronted with is the problem stemming from the different word order that different languages imply. Most works addressing this issue centre their effort in pairs of languages involving Arabic, Japanese or Chinese because of their utmost different origin with respect to western languages. However, Basque is also a language with an extremely different word order with respect to most other European languages, linguists being unable to determine its origins with certainty. Hence, SMT systems which do not tackle the reordering problem in any way are mostly unable to yield satisfactory results. In this work, a novel source sentence reordering technique is presented, based on monotonized alignments and n-best lists, endorsed by very promissing results obtained from a Basque-Spanish translation task.

1 Introduction

SMT systems have proved in the last years to be an important alternative to rule-based machine translation systems, being even able of outperforming commercial machine translation systems in the tasks they have been trained on. Moreover, the development effort behind a rule-based machine translation system and an SMT system is dramatically different, the latter being able to adapt to new language pairs with little or no human effort, whenever suitable corpora are available.

The grounds of modern SMT were established in (Brown et al., 1993), where the problem of machine translation was defined as following: given a sentence s from a certain source language, an adequate sentence $\hat{t}$ that maximises the posterior probability is to be found. Such a statement can be specified with the following formula:

$$\hat{t} = \operatorname*{argmax}_t Pr(t|s)$$

Applying the Bayes theorem on this definition, one can easily reach the next formula

$$\hat{t} = \operatorname*{argmax}_t \frac{Pr(t) \cdot Pr(s|t)}{Pr(s)}$$

and, since we are maximising over t, the denominator can be neglected, arriving to

$$\hat{t} = \operatorname*{argmax}_t Pr(t) \cdot Pr(s|t)$$

where $Pr(t|s)$ has been decomposed into two different probabilities: the *statistical language model* of the target language $Pr(t)$ and *the (inverse) translation model* $Pr(s|t)$.

Although it might seem odd to model the probability of the source sentence given the target sentence, this decomposition has a very intuitive interpretation: the translation model $Pr(s|t)$ will capture the word relations between both input and output language, whereas the language model $Pr(t)$ will ensure that the output sentence is a well-formed sentence belonging to the target language.

In the last years, SMT systems have evolved to become the present state of the art, two of the most representative techniques being the phrase based models (Koehn et al., 2003; Och and Ney, 2004) and the Weighted Finite State Transducers for Machine Translation (Casacuberta and Vidal, 2004; Kumar and Byrne, 2003). Both of these frameworks typically rely on word-aligned corpora, which often lead them to incur in word ordering related errors. Although there have been different efforts aiming towards enabling them to deal with non-monotonicity, the algorithms developed often only account for very limited reorderings, being unable to tackle with the more complex reorderings that e.g. some Asian languages introduce with respect to european languages. Because of this, not only will monotone systems present incorrectly ordered translations, but, in addition, the parameters of such models will be incorrectly estimated, whenever a certain input phrase is erroneously assumed to be the translation of a certain output phrase in training time.

Although no efficient solution has still been found, this problem is well known already since the origin of what is known as statistical machine translation: (Berger et al., 1996) 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.

On the other hand, dealing with arbitrary word reordering and choosing the one which best scores given a translation model has been shown not to be a viable solution, since when allowing all possible word permutations the search is NP-hard (Knight, 1999).

In the present work we develop a new approach to the problem, based on the work of Zens, Matusov and Kanthak (Zens et al., 2004; Matusov et al., 2005; Kanthak et al., 2005), who introduced the idea of monotonizing a corpus. A very preliminary result of our work was published in a Spanish workshop (Sanchis and Casacuberta, 2006). The key idea behind this concept is to use the IBM alignment models to efficiently reorder the input sentence s and produce a new bilingual, monotone pair, composed by the reordered input sentence s' and the output sentence t. Hence, once this new bilingual pair has been produced, the translation model to be applied will not have to tackle with the problems derived from different word reorderings, since this problem will not be present any more. Still, there is one more problem to be solved: in search time, only the input sentence is available, and hence the pair cannot be monotonized. To solve this, a very simple reordering model will be introduced, together with a *reordered* language model and n-best hypothesis generation. In this work, a phrase based model is trained using these monotone pairs.

In the following section, a brief overview of the latest efforts made towards solving the reordering problem will be pointed. In section 3, the approach presented in this work will be described, and in section 4 the experiments performed with this system will be shown. Finally, in section 5 the conclusions from this work will be elucidated, as well as the work that is still to be done.

2 Brief overview of existing approaches

Three main possibilities exist when trying to solve the reordering problem: input sentence reordering, output sentence reordering, or reordering both. The latter is, to the best of our knowledge, as yet unexplored.

Vilar et al. (1996), tried to partially solve the problem by monotonizing the most probable non-monotone alignment patterns and adding 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, making the translation process now monotone.

More recently, Kumar and Byrne (2005) learned weighted finite state transducers accounting for local reorderings of two or three positions. These models were applied to 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.

When dealing with input sentence reordering (Zens et al., 2004; Matusov et al., 2005; Kanthak et al., 2005), the main idea is to reorder 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 monotone, and then the input sentence is reordered in such a manner that the alignment is monotone.

However, this approach has an obvious problem, since the output sentence is not available in search time and the sentence pair cannot be made monotone.

The naïve solution, test on all possible permutations of the input sentence, has already been discussed earlier, being NP-hard (Knight, 1999), as $J!$ possible permutations can be obtained from a sentence of length J. Hence, the search space must be restricted, and such restrictions are bound to yield sub-optimal results. In their work, Kanthak et al. present four types of constraints: IBM, inverse IBM, local and ITG constraints.

Although the restrictions presented in their work (IBM, inverse IBM, local and ITG constraints) did yield interesting results, the search space still remained huge, and the computational price paid for a relatively small benefit was far too high.

- Let:
  - s a source sentence, and s_j its j-th word
  - t a target sentence, and t_i its i-th word
- Let C be a cost matrix
  $c_{ij} = \text{cost}(\text{align}(s_j, t_i))$
- Let $\{s^r\} = \{\text{all possible permutations of } s\}$.
  1. compute alignment $A_D(j) = \underset{i}{\text{argmin}}\, c_{ij}$
  2. $s' = \{s^r | \forall j : A_D(j) \leq A_D(j+1)\}$
  3. recompute (reorder) C, obtaining C'.
  4. set $A'_I(i) = \underset{j}{\text{argmin}}\, {c'}_{ij}$.
  5. Optional: Compute minimum-cost *monotonic* path through cost matrix C'.

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

3 The reordering model and N-Best reorderings

An important motivation behind the approach in this work is that the reordering constraints presented by Kanthak et al. (Kanthak et al., 2005) do not take into account extremely significant information that can be extracted from monotonized corpora: while reordering the input sentence in such a fashion that the alignment turns monotone, we are performing the reordering step needed further on when this action is needed to be taken on the input test set. Hence, what we would ideally want to do is learn a model using this information that will be capable of reordering a given, unseen, input sentence in the same way that the monotonization procedure would have done, in the hope that the benefits introduced will be greater than the error that an additional model will add into the translation procedure.

Once the alignments made monotonic according to the algorithm shown in Figure 1 (Kanthak et al., 2005), a new source "*language*" has been established, meaning that a reordered language model can be trained with the reordered input sentences s'. Such a language will have the words of the

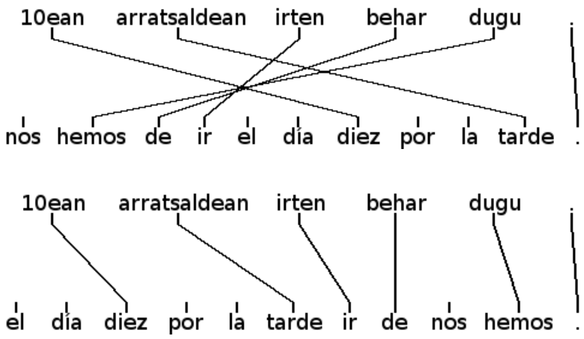

Figure 2: Alignment produced by GIZA (top) and alignment after the monotonization procedure (bottom). This is an example extracted from the Spanish→Basque corpus (i.e. Spanish is the source language). Although these sentences mean "We have to go day 10 in the evening.", the reordered spanish sentence would mean something like "Day ten in the evening go to we have.".

original source language, but the distinctive ordering of the target language. An example of this procedure is shown in Figure 2. Hence, a reordering model can be learnt from the monotonized corpus, which will most likely not depend on the output sentence, whenever the word-by-word translation is accurate enough.

Hence, the reordering problem can be defined as:

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

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.

		Spanish	Basque
Tranining	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

Table 1: Characteristics of the Tourist corpus.

However, and in order to reduce the error that will introduce a reordering model into the system, we found to be 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 $Pr(t) \cdot Pr(s^r|t)$. Ultimately, what we are actually doing with this procedure is to constrain the search space of permutations of the source sentence as well, but taking into account the information that monotonized alignments entail. In addition, this technique implies 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 a Basque-Spanish translation task, a tough machine translation problem in which reordering plays a crucial role.

The corpus chosen for this experiment is the *Tourist* corpus (Pérez et al., 2005), which is an adaptation of a set of Spanish-German grammars generating bilingual sentence pairs (Vidal, 1997) in such languages. Hence, the corpus is semi-synthetic. In this task, the sentences describe typical human dialogues in the reception desk of a hotel, being mainly extracted from tourist guides. However, because of its design, there is some asymmetry between both languages, 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. In spite of this fact, the vocabulary size of Basque is bigger than that of Spanish, and this is due to the agglutinative 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 computes the minimum number of editions (substitutions, insertions and deletions) needed to convert the translated sentence into the sentence considered ground truth. This measure is because of its nature a pessimistic one, when applied to Machine Translation.

PER *(position-independent WER)*: This criterion is similar to WER, but word order is ignored, accounting for the fact that an acceptable (and even grammatically correct) translation may be produced that differs only in word order.

BLEU *(Bilingual Evaluation Understudy) score*: This score measures the precision of unigrams, bigrams, trigrams, and 4-grams with respect to a set of reference translations, with a penalty for too short sentences (Papineni et al., 2001). BLEU is not an error rate, i.e. the higher the BLEU score, the better.

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.

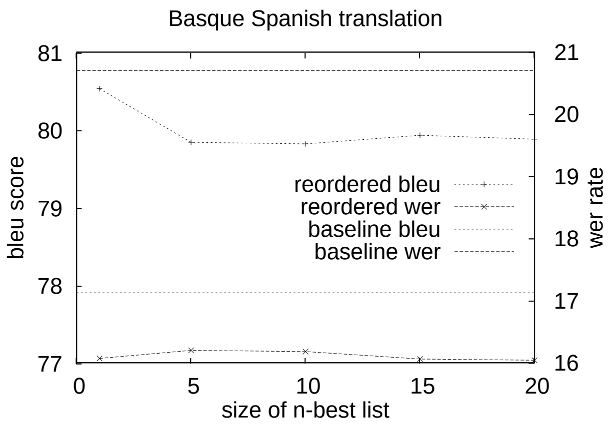

Figure 3: Evolution of translation quality when increasing n for Basque to Spanish.

	Baseline	Reordered, $n = 5$
WER	20.7%	16.2%
BLEU	77.9%	79.8%
PER	12.6%	11.0%

Table 2: Results for Basque to Spanish translation.

First, the bilingual pairs were aligned using IBM model 4 by means of the GIZA++ toolkit (Och and Ney, 2000). After this, the alignments were made monotone in the way described in Figure 1 and a new alignment was recalculated, determining the new monotone alignment between the reordered source sentence and the target, and a reordered source sentence language model was built. In addition, a phrase based model involving reordered source sentences and target sentences was learned by using the Thot toolkit (Ortiz et al., 2005).

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

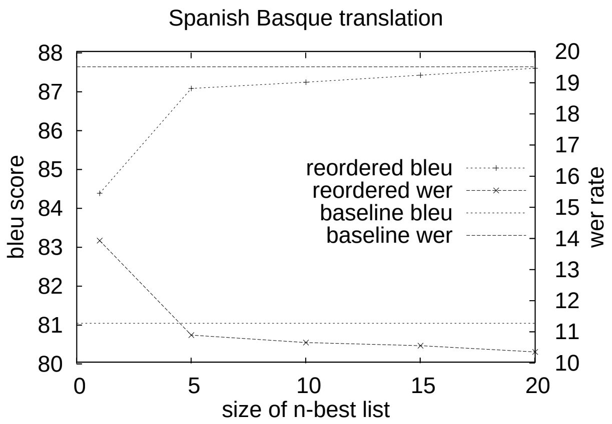

Figure 4: Evolution of translation quality when increasing n for Spanish to Basque.

	Baseline	Reordered, $n = 5$
WER	19.5%	10.9%
BLEU	81.0%	87.1%
PER	6.2%	4.9%

Table 3: Results for Spanish to Basque translation.

an exponential word-reordering model, just as we wanted.

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

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. The results of this setup can be seen in Table 3 and Table 2, with n-best list size set to 5. At this point, it must be noted that Pharaoh by itself also performs some reordering of the output sentence, but only on a per-phrase basis.

These results show that the reordering pipeline established does have significant benefits on the overall quality of the translation, almost achieving a relative improvement of 50% in WER. 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 word order errors within a sentence, which is the main problem reordering techniques try to solve. However, this improvement is explained because reordering the source sentence allows for better phrases to be extracted.

It is also interesting to point out that the translation quality when translating from Spanish to Basque is much higher than in the opposite sense. This is due to the corpus characteristics described in the previous section: Spanish being the input language of the corpus, it is only natural that the translation quality will worsen when reversing the meant translation direction. In addition, it can also be observed that the reordering pipeline has less beneficial effects when translating from Basque to Spanish.

Lastly, in Figure 4 and Figure 3, the result of increasing the size of the n-best reordering hypothesis list can be seen. In the case of Spanish-Basque translation, it can be seen how the translation quality still increases until size 20, where as in the case of Basque-Spanish the translation quality already reaches its maximum with the first 5 best hypothesis. However, it can also be seen that just using the best reordering hypothesis already yields better results than without introducing the reordering pipeline. Hence, these figures also show that the phrase extraction process obtains better quality phrases when the monotonization procedure has been implemented before the extraction takes place.

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 with 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 significant and show to be very promising, specially taking into account that Basque is an extremely complex language that poses many problems for state of the art systems.

Moreover, the technique we propose in this paper is learnt automatically, without any need of linguistic annotation or manually specified syntactic 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, as well as searching and developing integrated approaches trying to solve the reordering problem.

Acknowledgements

This work has been partially supported by the EC (FEDER) and the Spanish MEC under grant TIN2006-15694-CO2-01 and by the MEC scholarship AP2005-4023.

References

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

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of machine translation. In *Computational Linguistics*, volume 19, pages 263–311, June.

F. Casacuberta and E. Vidal. 2004. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30(2):205–225.

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

K. Knight. 1999. Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25(4):607–615.

P. Koehn, F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conf. of the NAACL on Human Language Technology*, volume 1, pages 48–54, Edmonton, Canada.

S. Kumar and W. Byrne. 2003. A weighted finite state transducer implementation of the alignment template model for statistical machine translation. In *Proceedings of the 2003 Conference of the NAACL on Human Language Technology*, volume 1, pages 63–70, Edmonton, Canada.

S. Kumar and W. Byrne. 2005. 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)*, pages 161–168, Vancouver, Canada.

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

F. J. Och and H. Ney. 2000. Improved statistical alignment models. In *ACL 2000*, pages 440–447, Hongkong, China, October.

F.J. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(1):417–449.

D. Ortiz, I. Garca-Varea, and F. Casacuberta. 2005. Thot: a toolkit to train phrase-based statistical translation models. In *Tenth Machine Translation Summit*, pages 141–148. Asia-Pacific Association for Machine Translation, Phuket, Thailand, September.

Papineni, A. Kishore, S. Roukos, T. Ward, and W. Jing Zhu. 2001. 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.

A. Pérez, F. Casacuberta, M.I. Torres, and V. Guijarrubia. 2005. Finite-state transducers based on k-tss grammars for speech translation. In *Proceedings of Finite-State Methods and Natural Language Processing (FSMNLP 2005)*, pages 270–272, Helsinki, Finland, September.

G. Sanchis and F. Casacuberta. 2006. N-best reordering in statistical machine translation. In *IV Jornadas en Tecnología del Habla*, pages 99–104, Zaragoza, Spain, November.

E. Vidal. 1997. Finite-state speech-to-speech translation. In *Proceedings of ICASSP-97*, volume 1, pages 111–114, Munich, Germany.

J.M. Vilar, E. Vidal, and J.C. Amengual. 1996. Learning extended finite-state models for language translation. In *Proc. of Extended Finite State Models Workshop (of ECAI'96)*, pages 92–96, Budapest, August.

R. Zens, H. Ney, T. Watanabe, and E. Sumita. 2004. Reordering constraints for phrase-based statistical machine translation. In *COLING '04: The 20th Int. Conf. on Computational Linguistics*, pages 205–211, Geneva, Switzerland.