Reordering on Spanish-Basque SMT

Arantza Díaz de Ilarraza, Gorka Labaka and Kepa Sarasola
Euskal Herriko Univertsitatea/Universidad del País Vasco
jipdisaa@ehu.es, gorka.labaka@ehu.es, jipsagak@ehu.es

Abstract

In this work we have deal with the reordering problem in Spanish-Basque statistical machine translation, comparing three different approaches and analyzing their strength and weakness. Tested approaches cover the more usual techniques: lexicalized reordering implemented on Moses, preprocessing based on hand defined rules over the syntactic analysis of the source and statistical translation.

According with the obtained results, the three reordering techniques improves the results of the baseline. We observe different behaviour at combining techniques. While the use of the Syntax-Based reordered corpus together with the lexicalized reordering get the best results, training the lexicalized reordering on the statistically reordered source does not improve the performance of the single methods.

1 Introduction

Basque language has many particularities which differences it from most European languages. Those differences makes the translation between Spanish (or English) and Basque an interesting challenge which involves both morphology and syntax features. Besides, Basque is low resourced which makes the development of a SMT system even more difficult.

Basque is an agglutinative language and many morpho-syntactic information which is expressed in separate words in most of the European languages is expressed using suffixes in Basque. In such a way, the information of prepositions or articles in Spanish, is expressed by means of suffixes which are added to the last word of the noun-phrase (similarly the information of conjunctions is attached at the end of the verbal phrase). Those morphological differences are discussed in a previous work (Díaz de Ilarraza et al., 2009) where we split Basque words in order to harmonise tokens in both languages (the results of those experiments are used in this work).

Furthermore, there are also syntactic differences which affect to the word order, that have a negative impact on the translation. As we said before, the agglutinative being of the Basque entails that the prepositions have to be translated into suffixes at the end of the phrase. Longer range differences, which have a worse impact on the translation, are also present. Modifiers of both verbs and noun-phrases are ordered differently in Basque and in Spanish. PP attached to noun-phrases are placed preceding the noun phrase instead of following it. The order of the constituents in Basque sentences is very flexible, nevertheless, in the most common order the verb is placed at the end of the sentence after the subject, the object and the rest of the verb modifiers. Figure 1 shows an example of a sentence's word alignment.

Those differences on the word order has an extremely negative impact on most of the steps of the Statistical Machine Translation, such as word alignment, phrase extraction and decoding. In this work, we have explored different approaches to deal with the reordering at SMT, and we have tried to determine the strength and the weakness of each approach.

The rest of the paper is structured as follows: In Section 2, we do a quick revision of the most relevant research on the area. Later, we describe the used reordering techniques (Section 3) and the SMT systems developed for this paper (Sections 4).

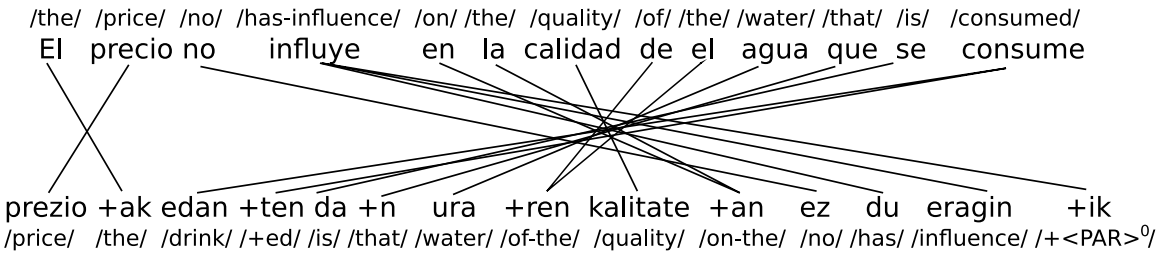

Figure 1: Example of word alignment. */the price does not affect the quality of the drinking water/*

We continue presenting and analyzing the results on Section 5. Finally, Section 6 presents conclusions and the future work.

2 Related work

Different researches have carried out trying to deal with word order differences at statistical machine translation. Among those publications which deal with order differences, the most commonly used approach is the preprocessing of the source in order to obtain a word-order which match with the word-order of the target language, allowing an almost monotonous translation.

Two main approaches are found on the bibliography; those where the reordering rules are hand-defined based on the linguistic analysis of the source, and those where the reordering is automatically inferred from the training corpus.

On (Collins et al., 2005), the authors get a significant improvement reordering German sentences based on the syntactic parsing. They define a small amount of rules to reorder verbal clauses in German, obtaining a English-like word order. In this way, they get an significant improvement both in BLEU and human judgments. Later, similar attempts are carried out for different languages. For example, Popovic and Ney (2006) proposed different reordering rules depending on the languages involved on the translation. They defined long-range reordering when translate into German and some local reordering for English-Spanish and German-Spanish language pairs. More recently, on (Ramanathan et al., 2008), authors combine Hindi language segmentation with some reordering applied on the syntactic analysis of the source to improve the quality of the English-Hindi SMT baseline system.

Many other research works try to learn the possible reordering automatically from the training corpus, instead of defining them manually. Some of those extract source reordering rules from the word alignment, based on different levels of linguistic analysis, from Part-of-Speech labelling (Chen et al., 2006) to shallow parsing (Zhang et al., 2007). Some other research works (Sanchis and Casacuberta, 2007; Costa-jussà and Fonollosa, 2006) consider the source reordering as a translation process, learning a SMT system to "translate" from the original source sentences to the reordered source sentences.

3 Reordering techniques

The main deal of this work is to analyse the impact of different reordering techniques on SMT. For this purpose, we have compared the results obtained by Spanish-Basque translation systems which implement the following reordering techniques.

3.1 Lexicalized reordering

The first method we have tried in this work is the lexicalized reordering[1] implemented in Moses. This method is the only one of the different methods we have tried which does not consist on the preprocessing of the source. In contrast, this method adds new features to the log-linear framework, in order to determine the order of the target phrases at decoding.

At extracting phrases from the training corpora the orientation of each occurrence is also extracted and the probability distribution is estimated in order to be added to the long-linear framework. Three different orientations are defined (See Figure 2):

[0] '+<*PAR*>' represents the Partitive Basque postposition suffix which appears on the direct object of negative sentences.

[1] http://www.statmt.org/moses/?n=Moses.AdvancedFeatures

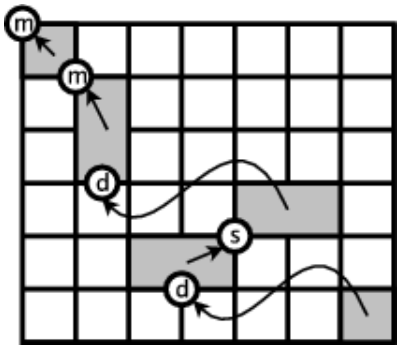

Figure 2: Possible orientation of phrases defined on the lexicalized reordering: monotone (m), swap (s), or discontinuous (d)

- monotone: a word alignment point to the top left exists.
- swap: an alignment point to the top right exists.
- discontinuous: no alignment points to the top left of top right.

Finally, at decoding, automatically inferred reordering models are used to score each hypothesis according the orientation of the used phrases.

3.2 Syntax-Based reordering

The second method presented here consists on the preprocessing of the Spanish sentences to adapt their word order to the order in Basque. This preprocessing is based on the dependency tree obtained with the morphological analyser Freeling (Carreras et al., 2004). We have defined ten rules to reorder the source sentence. Some of them imply local reordering (movements of single words inside the noun phrase) and others imply long-range reordering (movements of whole phrases along the sentence).

3.2.1 Local reordering

The main aim of the local reordering is to deal with the differences between both languages in the way that the phrases are constructed. As we have already explain, prepositions are translate into suffixes at the end of the noun-phrase. So we have defined reordering rules that use the POS tags and the chunk boundaries obtained with Freeling to move Spanish prepositions and articles to the end of the noun-phrase, since all those elements have to be translated as suffixes which appear at that position.

On the following example we can see an example of local reordering. In this example chunk boundaries are mark with '|', and elements which are moved (articles and prepositions) are in bold.

El precio | no | influye | en la calidad | de el agua | que | se consume

precio **El** | no | influye | calidad **la en** | agua **el de** | que | se consume

3.2.2 Long-range reordering

In order to deal with long-range reordering, we have defined rules which move whole phrases along the sentence based on its dependency tree. We have implemented rules which implies the following four movements (Figure 2 shows an example of the application of these rules):

(a) The verb is moved to the end of the clause, after all its modifiers.

(b) In negative sentences the particle 'no' is moved together with the verb to the end of the clause.

(c) Prepositional phrases and subordinated relative clauses which are attached to nouns are placed at the beginning of the whole noun phrase where they are included.

(d) Conjunctions (and relative pronouns) placed at the beginning of Spanish subordinated (or relative) clauses are moved to the end of the clause, after the subordinated verb.

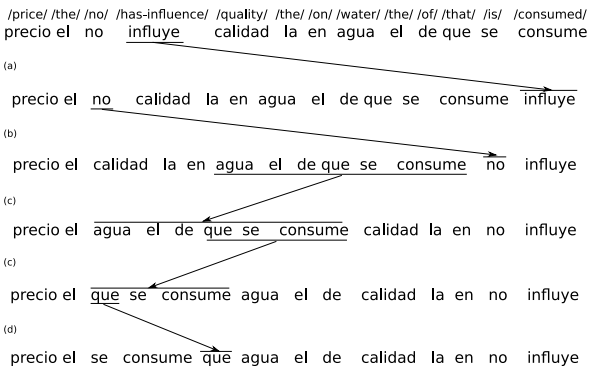

Figure 2: Example of long-range reordering.
/the price does not affect the quality of the drinking water/

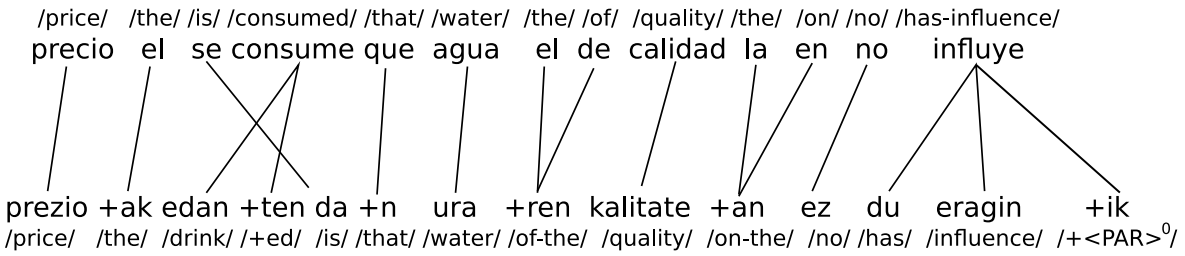

Figure 3: Example of word alignment after Syntax-Based reordering. */the price does not affect the quality of the drinking water/.*

3.3 Statistical Reordering

The Statistical Reordering considers the reordering preprocessing as the translation of the source sentences into a reordered source language, which allows a better translation into the target language.

Unlike the Syntax-Based reordering presented above, on Statistical Reordering all the information is extracted from the corpus and it is not necessary any linguistic parsing or hand-made rule.

The training process consists on the following steps; (1) align source and target training corpora in both directions and combine words alignments to obtain many-to-many word alignments, (2) Modify the many-to-many word alignments to many-to-one, (3) reorder source sentences in order to obtain a monotone alignment, (4) train a state-of-the-art SMT system to translate from original source sentences into reordered source. After Statistical Reordering, another SMT system is necessary to translate from reordered source to target.

4 Systems' overview

In order to measure the impact of the reordering techniques presented above, we built systems which uses those techniques (as well as baselines which uses distance-based reordering) and we compared their performance. The development of all those systems has been carried out using freely available tools:

- GIZA++ toolkit (Och and H. Ney, 2003) was used for training the word alignment.
- SRILM toolkit (Stolcke, 2002) was used for building the language model.
- Moses Decoder (Koehn et al., 2007) was used for translating the test sentences.

In order to deal with the agglutinative nature of the Basque, and according with our previous work (Díaz de Ilarraza et al., 2009), we have used segmented Basque text, where words are split into different tokens, to train all our systems. After translation a postprocessing has carried out which generates the final translation based on the segmented output of the decoder. After generation, a word based language model is incorporated using nbest lists reranking. Figure 4 shows the general design of the system used in this work.

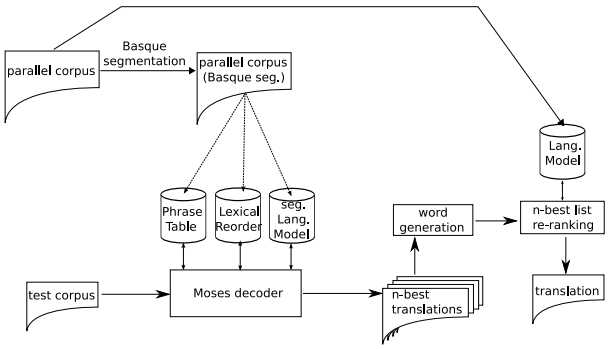

Figure 4: Design of the segmentation-based SMT system

Over the segmented target text, we have trained nine different systems combining three possible source text preprocessing (without reordering, Syntax-Based reordering and statistical reordering) and three reordering configurations at decoding (monotone, distance-based and lexicalized reordering).

Besides, we have also trained three more SMT systems, one for each reordering configurations at decoding, over the original (unsegmented) target text, in order to compare the results obtained using segmentation with a real state-of-the-art system.

		sentences	tokens	vocabulary	singletons
training	Spanish	58,202	1,284,089	46,636	19,256
	Basque (tokenized)		1,010,545	87,763	46,929
	Basque (segmented)		1,546,304	40,288	19,031
development	Spanish	1,456	32,740	7,074	4,351
	Basque (tokenized)		25,778	9,030	6,339
	Basque (segmented)		39,429	6,189	3,464
test	Spanish	1,446	31,002	6,838	4,281
	Basque (tokenized)		24,372	8,695	6,077
	Basque (segmented)		37,361	5,974	3,301

Table 1: Some statistics of the corpora.

All the systems use a log-linear combination (Och and Ney, 2002) of several common feature functions: phrase translation probabilities (in both directions), word-based translation probabilities (lexicon model, in both directions), a phrase length penalty, a word length penalty and a target language model. Both the language model used at decoding (based on the segmented text) and the language model which is incorporated after generation (based on the final words) are 5-gram models trained on the Basque portion of the bilingual corpus, using the SRI Language Modeling Toolkit, with modified Kneser-Ney smoothing.

We have used Minimum-Error-Rate Training (Och, 2003) within a log-linear framework for parameter optimization. The metric used to carry out this optimization is BLEU (Papineni et al., 2002).

5 Experimental results

5.1 Data and evaluation

In order to carry out this experiment we used the *Consumer Eroski* parallel corpus (Alcázar, 2005). This corpus is a collection of 1036 articles written in Spanish (January 1998 to May 2005, Consumer Eroski magazine, http://revista.consumer.es) along with their Basque, Catalan and Galician translations. It contains more than 1,200,000 Spanish words and more than 1,000,000 Basque words. This corpus was automatically aligned at sentence level[2] and it is available[3] for research. Consumer Eroski magazine is composed by the articles which compare the quality and prices of commercial products and brands.

We have divided this corpus in three sets, training set (60,000 sentences), development set (1,500 sentences) and test set (1,500 sentences), more detailed statistics are shown in Table 1.

In order to assess the quality of the translation obtained using the systems, we used four automatic evaluation metrics. We report two accuracy measures: BLEU (Papineni et al., 2002), and NIST (Doddington, 2002); and two error measures: Word Error Rate (WER) and Position independent word Error Rate (PER). In our test set, we have access to one Basque reference translation per sentence. Evaluation is performed in a case-insensitive manner.

5.2 Results

The evaluation results for the test corpus are reported on Table 2. First, we want to remark that the results are consistent with those obtained in our previous work (Díaz de Ilarraza et al., 2009), since systems using segmentation outperforms those which are trained over the unsegmented text. Furthermore, according to BLEU scores all single reordering methods outperforms the baseline ($10.37 < 11.03 < 11.13 < 11.27$), which is trained on the tokenized source corpus (without reordering) and uses distance-based reordering at decoding. The best results are obtained by the system which combines Syntax-Based reordering as preprocessing and the lexicalized reordering at decoding (11.51 BLEU score).

Considering those systems which uses single reordering methods, lexicalized reordering get the best results (11.27 BLEU), followed by the statistical reordering (11.13 BLEU). Finally, the Syntax-Based reordering (11.03 BLEU) get the smaller improve-

[2]Corpus was collected and aligned by Asier Alcázar from the University of Missouri-Columbia

[3]The Consumer corpus is accessible online via Universidade de Vigo (http://sli.uvigo.es/CLUVI/, public access) and Universidad de Deusto (http://www.deli.deusto.es, research intranet).

		BLEU	NIST	WER	PER
unreordered source unsegmented target	monotone	10.00	4.42	81.43	61.70
	distance	10.31	4.46	81.22	61.64
	lexicalized	10.82	4.55	80.10	61.11
unreordered source segmented target	monotone	10.01	4.40	80.59	61.79
	distance	10.37	4.54	79.47	60.59
	lexicalized	11.27	4.65	79.50	60.67
Statistical source-reordering segmented target	monotone	10.89	4.60	79.26	60.78
	distance	11.13	**4.69**	78.21	**59.66**
	lexicalized	11.12	4.66	78.69	60.19
Syntax-Based source-reordering segmented target	monotone	10.29	4.48	80.15	61.98
	distance	11.03	4.60	78.79	61.35
	lexicalized	**11.51**	**4.69**	**77.94**	60.45

Table 2: BLEU, NIST, WER and PER evaluation metrics.

ment over the baseline. In three cases, the improvement using sophisticated reordering methods is substantial.

The results obtained at combining the methods based on preprocessing (statistical reordering and Syntax-Based reordering) and the lexicalized reordering show different behaviour. While the use of the Syntax-Based reordered together with the lexicalized reordering get the best results, training the lexicalized reordering on the statistically reordered source does not improve the performance of the single methods.

6 Conclusions and Future work

Results obtained in this work allow us to compare different reordering methods on a specially demanding task as the Spanish-Basque translation. According with those results, the three reordering methods tested here (which could be considered as representative of the nowadays research) outperforms baseline, getting the best results with the lexicalized reordering implemented at decoding.

We have also tested different combination of methods, obtaining a significant improvement at using together the Syntax-Based and the lexicalized reordering. Each method takes advantage of different information and they are able to complement each other. For instance, order differences of noun and adjectives are not treat on Syntax-Based reordering and they are probably corrected by the lexicalized reordering.

On the other hand, the combination of the statistical reordering used at preprocessing and the lexicalized reordering at decoding gets worse results than the ones obtained by the single methods by their own. The performance dropping probably indicates that both methods use the same information about word alignment, so they could not achieved any improvement from the method combination.

As future work, we are planning to rerun experiments on a bigger training corpus and a different language pair (such as English-Basque) to confirm the results obtained in this work. Regarding the Syntax-Based reordering, we are planning to define more reordering rules, since the actual ones do not cover all order differences of both languages. Furthermore, we are considering a way to allow the decoder to chose among different reordering proposed by the Syntax-Based preprocessing (using a nbest list of reordering or a word-graph as input of the decoder).

Acknowledgments

This research was supported in part by the Spanish Ministry of Education and Science (OpenMT: Open Source Machine Translation using hybrid methods, TIN2006-15307-C03-01). Gorka Labaka is supported by a PhD grant from the Basque Government (grant code, BFI05.326).

Consumer corpus has been kindly supplied by Asier Alcázar from the University of Missouri-Columbia and by Eroski Fundazioa.

References

Asier Alcázar. 2005. Towards linguistically searchable text. In *Proceedings of BIDE 2005*, Bilbao.

X. Carreras, I. Chao, L. Padró, and M. Padró. 2004. Freeling: an Open-Source Suite of Language Analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*.

Boxing Chen, Mauro Cettolo, and Marcello Federico. 2006. Reordering Rules for Phrase-based Statistical Machine Translation. In *IWSLT 2006*, pages 182–189.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause Restructuring for Statistical Machine Translation. In *ACL*. The Association for Computer Linguistics.

Marta R. Costa-jussà and José A. R. Fonollosa. 2006. Statistical Machine Reordering. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 70–76, Sydney, Australia, July. Association for Computational Linguistics.

Arantza Díaz de Ilarraza, Gorka Labaka, and Kepa Sarasola. 2009. Relevance of different segmentation options in Spanish-Basque SMT. In *Proceedings of the EAMT 2009*, Barcelona. European Association for Machine Translation.

G. Doddington. 2002. Automatic evaluation of Machine Translation quality using n-gram cooccurrence statistics. In *Proceedings of HLT 2002*, San Diego, CA.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic, June.

F. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och and Hermann Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *ACL*, pages 295–302.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *ACL*, pages 160–167.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of 40th ACL*, Philadelphia, PA.

Maja Popović and Hermann Ney. 2006. POS-based Word Reorderings for Statistical Machine Translation. In *International Conference on Language Resources and Evaluation*, pages 1278–1283, Genoa, Italy, May.

Ananthakrishnan Ramanathan, Pushpak Bhattacharya, Jayprasad Hegde, Ritesh M.Shah, and Sasikumar M. 2008. Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation. In *IJCNLP 2008: Third International Joint Conference on Natural Language Processing*, Hyderabad, India.

G. Sanchis and F. Casacuberta. 2007. Reordering via N-Best Lists for Spanish-Basque Translation. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*, pages 191–198, Skövde, Sweden, September 7-9.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In *Proc. Intl. Conf. Spoken Language Processing*, Denver, Colorado, September.

Yuqi Zhang, Richard Zens, and Hermann Ney. 2007. Chunk-Level Reordering of Source Language Sentences with Automatically Learned Rules for Statistical Machine Translation. In *Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting*, Rochester, NY, April.