Relevance of Different Segmentation Options 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

Segmentation is widely used in adapting Statistical Machine Translation to highly inflected languages as Basque. The way this segmentation is carried out impacts on the quality of the translation. In order to look for the most adequate segmentation for a Spanish-Basque system, we have tried different segmentation options and analyzed their effects on the translation quality.

Although all segmentation options used in this work are based on the same morphological analysis, translation quality varies significantly depending on the segmentation criteria used. Most of the segmentation options outperform the baseline according to all metrics, except the one which splits words according the morpheme boundaries. From here we can conclude the importance of the development of the segmentation criteria in SMT.

1 Introduction

In this paper we present the work done for adapting a baseline SMT system to carry out the translation into a morphologically-rich agglutinative language such as Basque. In translation from Spanish to Basque, some Spanish words, such as prepositions or articles, correspond to Basque suffixes, and, in case of ellipsis, more than one of those suffixes can be added to the same word. In this way, based on the Basque lemma 'etxe' */house/* we can generate 'etxeko' */of the house/*, 'etxekoa' */the one of the house/*, 'etxekoarengana' */towards the one of the house/* and so on.

Besides, Basque is a low-density language and there are few corpora available comparing to other languages more widely used as Spanish, English, or Chinese. For instance, the parallel corpus available for this work is 1M word for Basque (1.2M words for Spanish), much smaller than the corpora usually used on public evaluation campaigns such as NIST.

In order to deal with the problems presented above, we have split up Basque words into the lemma and some tags which represent the morphological information expressed on the inflection. Dividing Basque words in this way, we expect to reduce the sparseness produced by the agglutinative being of Basque and the small amount of training data.

Anyway, there are several options to define Basque segmentation. For example, considering all the suffixes all together as a unique segment, considering each suffix as a different segment, or considering any other of their intermediate combinations. In order to define the most adequate segmentation for our Spanish-Basque system, we have tried some of those segmentation options and have measured their impact on the translation quality.

The remainder of this paper is organized as follows. In Section 2, we present a brief analysis of previous works adapting SMT to highly inflected languages. In Section 3, we describe the systems developed for this paper (the baseline and the morpheme based systems) and the different segmentation used by those systems. In Section 4, we evaluate the different systems, and report and discuss our experimental results. Section 5 concludes the paper and gives avenues for future work.

Proceedings of the 13th Annual Conference of the EAMT, pages 74–80,
Barcelona, May 2009

2 Related work

Many researchers have tried to use morphological information in improving machine translation quality. In (Koehn and Knight, 2003), the authors got improvements splitting compounds in German. Nießen and Ney (2004) achieved a similar level of alignment quality with a smaller corpora restructuring the source based on morpho-syntactic information when translating from German to English. More recently, on (Goldwater and McClosky, 2005) the authors achieved improvements in Czech-English MT optimizing a set of possible source transformations, incorporating morphology.

In general most experiments are focused on translating from morphologically rich languages into English. But last years some works have experimented on the opposite direction. For example, in (Ramanathan et al., 2008), the authors segmented Hindi in English-Hindi statistical machine translation separating suffixes and lemmas and, in combination with the reordering of the source words based on English syntactic analysis, they got a significant improvement both in automatic and human evaluation metrics. In a similar way Oflazer and El-Kahlout (2007) also segmented Turkish words when translate from English. The isolated use of segmentation does not get any improvement at translation, but combining segmentation with a word-level language model (incorporated by using n-best list re-scoring) and setting as unlimited the value of the *distortion limit* (in order to deal with the great order difference between both languages) they achieve a significant improvement over the baseline.

Segmentation is the most usual way to translate into highly inflected languages, but other approaches have been also tried. In (Bojar, 2007) factored translation have been used on English-Czech translation. Words of both languages are tagged with morphological information creating different factors which are translated independently and combined in a generation stage. Finally, in (Minkov et al., 2007) the authors have divided translation in two steps where they first use usual SMT system to translate from English to Russian lemmas and in a second step they decide the inflection of each lemma using bilingual information.

3 SMT systems

The main deal of this work is to measure the impact of different segmentation options on a Spanish-Basque SMT system. In order to measure this impact we have compared the quality of the baseline system which does not use segmentation at all, with systems that use different segmentation options. the development of 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.

3.1 Baseline

We have trained Moses on the tokenized corpus (without any segmentation) as baseline system. Moses and the scripts provided with it allow to easily train a state-of-the-art phrase-based SMT system. We have used a log-linear (Och and Ney, 2002) combination of several common feature functions: phrase translation probabilities (in both directions), word-based translation probabilities (lexicon model, in both directions), a phrase length penalty and a target language model.

The decoder also relies on a target language model. The language model is a simple 5-gram language model trained on the Basque portion of the training data, using the SRI Language Modeling Toolkit, with modified Kneser-Ney smoothing. Finally, we have also used a lexical reordering model (one of the advanced features provided by Moses[1]), trained using Moses scripts and *'msd-bidirectional-fe'* option. The general design of the baseline system is presented on Figure 1.

Moses also implements 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).

3.2 Morpheme-based statistical machine translation

Basque is an agglutinative language, so words may be made up several morphemes. Those morphemes are added as suffixes to the last word of

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

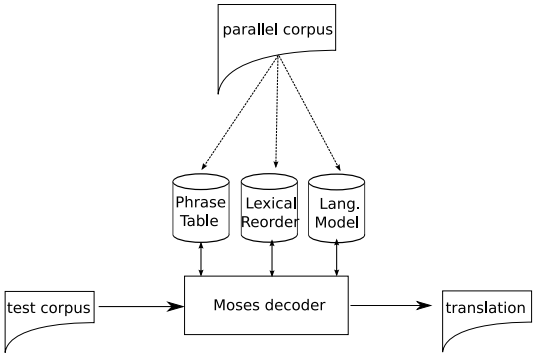

Figure 1: Basic design of a SMT system

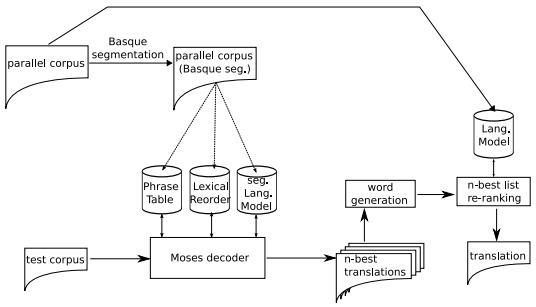

Figure 2: Design of the morpheme-based SMT system

noun phrases and verbal chains. Suffixes represent the morpho-syntactic information associated to the phrase, such as number, definiteness, grammar case and postposition.

As a consequence, many words only occur once in the training corpus, leading to serious sparseness problems when extracting statistics from the data. In order to overcome this problem, we segmented each word into a sequence of morphemes, and then we worked at this representation level. Working at the morpheme level we reduced the number of tokens that occur only once and, at the same time, we reduce the 1-to-n alignments. Although 1-to-n alignments are allowed in IBM model 4, training can be harmed when the parallel corpus contains many cases.

Adapting the baseline system to work at the morpheme level mainly consists on training Moses on the segmented text (same training options are used in baseline and morpheme-based systems). The system trained on these data will generate a sequence of morphemes as output and a generation post-process will be necessary in order to obtain the final Basque text. After generation, we have integrated a word-level language model using n-best list re-ranking. The general design of the morpheme-based system is presented on Figure 2.

3.2.1 Segmentation options for Basque

Segmentation of Basque words can be made in different ways and we want to measure the impact those segmentation options have on the translation quality. In order to measure this impact, we have tried different ways to segment Basque words and we have trained a different morpheme-based system on each segmentation.

The different segmentation options we have tried are all based on the analysis obtained by Eustagger (Aduriz and Díaz de Ilarraza, 2003), a tagger for Basque based on two-level morphology (Koskeniemmi, 1983) and statistical disambiguation. Based on those analysis we have divided each Basque word in different ways. From the most fine-grained segmentation, where each morpheme is represented as a token, to the most coarse-grained segmentation where all morphemes linked to the same lemma are put together in an unique token. Figure3 shows an analysis obtained by Eustagger the lemma and the morphological information added by the morphemes is represented marking the morphemes boundaries with a '+'.

Following we define the four segmentation options we are experimenting with.

Eustagger Segmentation: In our first approach we have strictly based on the lexicon of Eustagger, and we have created a separate token for each morpheme recognized by the analyzer. This lexicon has been created following a linguistic perspective and, although it has been proved very useful for the develop of several applications, it is probably not the most adequate for this work. As the lexicon is very fine-grained, some suffixes, which could be considered as a unique morpheme, are represented as a concatenation of several fine-grained morphemes in the Eustagger lexicon. Furthermore, some of those morphemes have not any effect on the word form, and they only adds some morphological features. Figure 3 shows segmentation of 'aukeratzerakoan' */at the election time/* word according to the segmentation produced by Eustagger.

One suffix per word: Taking into account that the Eustagger lexicon is too fine-grained and that it generates too many tokens at segmentation, our next approach consisted on putting together all suffixes linked to a lemma in one token. So, at splitting one Basque word we will generate at most

Analysis	aukeratu<adi><sin>+<adize>+<ala><gel>+<ine>				
Eustagger seg.	aukeratu<adi><sin>	+<adize>	+<ala>	+<gel>	+<ine>
Automatic seg.	aukeratu<adi><sin>	+<adize><ala>	+<gel>	+<ine>	
Hand defined seg.	aukeratu<adi><sin><adize>	+<ala><gel><ine>			
OneSuffix seg.	aukeratu<adi><sin>	+<adize><ala><gel><ine>			

Figure 3: Analysis obtained by Eustagger for 'aukeratzerakoan' */at the election time/* word. And the distinct segmentation inferred from it.

three tokens (prefixes, lemma and suffixes). We can see 'aukeratzerakoan' */at the election time/* word's segmentation on Figure 3.

Manual morpheme-grouping: After realizing the impact of the segmentation in translation, we tried to obtain an intermediate segmentation which optimizes the translation quality. Our first attempt consists on defining by hand which morphemes can be grouped together in one token and which ones can be considered a token by their own. In order to decide which morphemes to group, we have analyzed the alignment errors occurred at previous segmentation experiments, defining a small amount of rules to grouping morphemes. For instance, '+<adize>'[2] morpheme is usually wrongly aligned when it is considered as a token, so we have decided to join it to the lemma at segmentation. On Figure 3 we can see the segmentation corresponding to 'aukeratzerakoan' */at the election time/* word.

Automatic morpheme-grouping: Anyway, the morpheme-grouping defined by hand depends on the language pair and if we change it, we should redefine the grouping criteria, analyzing again the detected errors. So, in order to find a language independent way to define the most appropriate segmentation, we focus our research in establishing a statistical method to decide which morphemes have to be put into the same token. We observed that the morphemes which generates most of the errors are those which have not their own *meaning*, those that *need* another morpheme to complete their meaning. We thought on using the *mutual information* metric in order to measure statistical dependence between two morphemes. We will group those morphemes that are more dependent than a threshold. On this experiment we tried different thresholds and we obtained the best results when it is set to 0.5 (value that involve grouping most of the morphemes). In Figure 3 we can see 'aukeratzerakoan' */at the election time/* word segmented in this way.

[2]suffix for verb normalisation

3.2.2 Generating words from morphemes

When working at the morpheme level, the output of our SMT system is a sequence of morphemes. In order to produce the proper Basque text, we need to generate the words based on this sequence, so the output of the SMT system is post-processed to produce the final Basque translation.

To develop generation post-processing, we reuse the lexicon and two-level rules of our morphological tool Eustagger. The same generation engine is useful for all the segmentation options defined in section 3.2.1 since we have produced them based on the same analysis. However, we have to face two main problems:

- Unknown lemmas: some lemmas such as proper names are not in the Eustagger lexicon and could not be generated by it. To solve this problem and to be able to generate inflection of those words, the synthesis component has been enriched with default rules for unknown lemmas.

- Invalid sequences of morphemes: the output of the SMT system is not necessarily a well-formed sequence from a morphological point of view. For example, morphemes can be generated in a wrong order or they can be missed or misplaced (i.e. a nominal inflection can be assigned to a verb). In the current work, we did not try to correct these mistakes, and when the generation module can not generate a word it outputs the lemma without any inflection. A more refined treatment is left for future work.

3.3 Incorporation of word-level language model

When training our SMT system over the segmented test the language model used in decoding is a language model of morphemes (or groups of morphemes depending on the segmentation option). Real words are not available at decoding, but, after generation we can incorporate a second

		sentences	words	morph	word-vocabulary	morph-vocabulary
training	Spanish Basque	58,202	1,284,089 1,010,545	- 1,699,988	46,636 87,763	- 35,316
development	Spanish Basque	1,456	32,740 25,778	- 43,434	7,074 9,030	- 5,367
test	Spanish Basque	1,446	31,002 24,372	- 41,080	6,838 8,695	- 5,170

Table 1: Some statistics of the corpora.

language model based on words. The most appropriate way to incorporate the word-level language model is using n-best list as was done in (Oflazer and El-Kahlout, 2007). We ask Moses to produce a n-best list, and after generating the final translation based on Moses output, we estimate the new cost of each translation incorporating word-level language model. Once new cost is calculated the sentence with the lowest cost is selected as the final translation.

The weight for the word-level language model is optimized at Minimum Error Rate Training with the weights of the rest of the models. Minimum Error Rate Training procedure has been modified to post-process Moses output and to include word-level language model weight at optimization process.

4 Experimental results

4.1 Data and evaluation

In order to carry out this experiment we used the *Consumer Eroski* parallel corpus. 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[3] and it is available[4] 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 on Table 1.

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

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

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, 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.

4.2 Results

The evaluation results for the test corpus is reported in Table 2. These results show that the differences at segmentation have a significant impact at translation quality. Segmenting words according to the morphemes boundaries of the Eustagger lexicon does not involve any improvement. Compared to the baseline, which did not use any segmentation, the results obtained for the evaluation metrics are not consistent and varies depending on the metric. According to BLEU segmentation harms translation, but according the rest of the metrics the segmentation slightly improves translation, but this improvement is probably not statistically significant.

The rest of the segmentation options, which are based on the same analysis of Eustagger and contains the same morpheme sequences, consistently outperforms baseline according to all the metrics. Best results are obtained using the hand defined criteria (based on the alignment errors), but automatically defined segmentation criteria obtains similar results.

Due to the small differences on the results obtained for the evaluation metrics we have carried out a statistical significance test (Zhang et al., May 2004) over BLEU. According with this, the system using hand defined segmentation significantly outperforms both the system using OneSuffix segmentation and the system using segmentation based on mutual information. Difference between the system using OneSuffix segmentation and the system based on mutual information are

	BLEU	NIST	WER	PER
Baseline	10.78	4.52	80.46	61.34
MorphemeBased-Eustagger	10.52	4.55	79.18	61.03
MorphemeBased-OneSuffix	11.24	**4.74**	**78.07**	**59.35**
MorphemeBased-AutoGrouping	11.24	4.66	79.15	60.42
MorphemeBased-HandGrouping	**11.36**	4.69	78.92	60.23

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

Segmentation option	Running tokens	Vocabulary size	BLEU
No Segmentation	1,010,545	87,763	10.78
Hand Defined grouping	1,546,304	40,288	**11.36**
One Suffix per word	1,558,927	36,122	11.24
Statistical morph. grouping	1,580,551	35,549	11.24
Eustagger morph. boundaries	1,699,988	35,316	10.52

Table 3: Correlation between token amount on the train corpus and BLEU evaluation results

not statistically significant.

Finally, given the low scores obtained, we would like to make two additional remarks. First, it shows the difficulty of the task of translating into Basque, which is due to the strong syntactic differences with Spanish. Second, the evaluation based on words (or n-grams of words) always gives lower scores to agglutinative languages like Basque. Often one Basque word is equivalent to two or three Spanish or English words, so a 3-gram matching in Basque is harder to obtain having a highly negative effect on the automatic evaluation metrics.

4.3 Correlation between segmentation and BLEU

Analyzing the obtained results, we have realized that there are a correlation between the amount of tokens generated at segmentation and the results obtained at evaluation. Before segmentation, there are 1M words for Basque, which together with the 1.2M words for Spanish, make the word alignment more difficult (due to the 1-to-n alignment amount). Anyway, after segmenting the Basque words according with the morpheme boundaries of Eustagger, the Basque text contains 1.7M tokens (the same alignment problem is generated but in the opposite direction) see Table 3.

Intermediate segmentation options, where morphemes marked by Eustagger are grouped in different ways, get better results when the amount of the generated tokens is closer to the amount of tokens we have in Spanish part. We leave for future work to experiment ways to reduce the different number of tokens of both languages.

5 Conclusions and Future work

We have proved that the quality of the translation varies significantly when applying different options for word segmentation. Based on the same output of morphological analyzer, we have segmented words in different ways creating more fine or coarse grained segments (from one token per each morpheme to a unique token for all suffixes of a word). Surprisingly, the criteria based on considering each morpheme as a separate token obtains worse results than the system without segmentation. Other segmentation options outperforms the baseline, getting the best results with a hand defined intermediate grouping based on an alignment error analysis.

Anyway, the work done by hand is language dependent and could not be reused for a different pair of languages, so we also tried a statistical way to determine the morpheme grouping criteria which gets almost as accurate results as those obtained with the hand defined criterion. So we could use this statistical grouping criteria to adapt our system to a different language pair such as English-Basque.

As future work, we thought on trying a different measure to determine the statistical independence of the morphemes, as χ^2. Besides, as the dependence between morphemes is calculated on the monolingual text, a bigger monolingual corpus could be used (instead of using just the Basque side of the bilingual corpus) for this.

Taking into account the obtained correlation between the token amount and translation quality. We want to redefine the segmentation criteria to reduce the amount of tokens obtained. In such a way that the difference in the number of tokens of

both languages would be reduced.

Acknowledgement

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) and the Regional Branch of the Basque Government (AnHITZ 2006: Language Technologies for Multilingual Interaction in Intelligent Environments., IE06-185). 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

Aduriz, I. and A. Díaz de Ilarraza. 2003. Morphosyntactic disambiguation ands shallow parsing in Computational Processing of Basque. In *Inquiries into the lexicon-syntax relations in Basque. Bernarrd Oyharabal (Ed.)*, Bilbao.

Bojar, Ondrej. 2007. English-to-Czech Factored Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic.

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

Goldwater, S. and D. McClosky. 2005. Improving Statistical MT Through Morphological Analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Vancouver.

Koehn, P. and K. Knight. 2003. Empirical Methods for compound splitting. In *Proceedings of EACL 2003*, Budapest, Hungary.

Koehn, Philipp, 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.

Koskeniemmi, K. 1983. Two-level Model for Morphological Analysis. In *Proceedings of the Eigth International Joint Conference on Artificial Intelligence*, Karlsruhe, Germany.

Minkov, E., K. Toutanova, and H. Suzuki. 2007. Generating Complex Mophology for Machin Translation. In *Proceedings of 45th ACL*, Prague, Czech Republic.

Nießen, S. and H. Ney. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Comput. Linguist.*, 30(2):181–204.

Och, F. and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Och, Franz Josef and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *ACL*, pages 295–302.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *ACL*, pages 160–167.

Oflazer, Kemal and Ilknur Durgar El-Kahlout. 2007. Exploring Different Representation Units in English-to-Turkish Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic.

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

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

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

Zhang, Ying, Stephan Vogel, and Alex Waibel. May 2004. Interpreting Bleu/NIST scores: How much improvement do we need to have a better system? In *Proceedings of LREC 2004*, Lisbon, Portugal.