EUSMT: Incorporating Linguistic information to Statistical Machine Translation for a morphologically rich language. Its use in preliminary SMT-RBMT-EBMT hybridization

Gorka Labaka
Advisors: Arantza Díaz de Ilarraza and Kepa Sarasola
University of the Basque Country
Department of Computer Languages and Systems

Abstract

We have proposed and successfully tested new techniques to deal with the problems found in applying Statistical Machine Translation (SMT) to language pairs with great morphological and syntactical differences. These techniques are based on segmentation and reordering and we have evaluated them in the context of Spanish-Basque translation.

Dealing with morphology, we first proved that the quality of the translation varies significantly when applying different options for word segmentation. Then, we identified which is the best option to translate from Spanish to Basque and finally, we established a language independent statistical method to automatically decide which sequences of morphemes have to be included in the same token in order to optimize the translation.

Dealing with syntax, we experimented with three reordering methods implemented in decoding and in preprocessing. Preprocessing methods are based on manually defined rules for the syntactic analysis and on using a separate SMT to "translate" the original source language to a reordered source language which makes the translation easier. We obtained a significant improvement by using syntax-based and lexicalized reordering together.

In addition, we carried out two hybridization experiments, where the SMT system developed in this thesis was combined with previously developed Rule-Based (RBMT) and Example-Based (EBMT) Machine Translation systems. These are based on Statistical Post-Edition of RBMT output and on a Multi-Engine approach. We confirmed that both hybrid systems outperform the single systems.

Presentation

This document contains a summary of the dissertation that will be submitted to the *Department of Computer Languages and Systems* of the *University of the Basque Country* to obtain the degree of Doctor in Computer Science. The document is organized into four sections. Section 1 introduces the problems encountered in translating into Basque. Section 2 presents a discussion of the related work. In Section 3, we present the organization of the dissertation and provide a summary of contents for each chapter. In the last section, we describe the current state of our work. Finally the papers supporting the dissertation are attached to this document.

Contents

1 Introduction

The information society we live in is undoubtedly a multilingual one. Every day, hundreds of thousands of documents are being generated and translated in order to cover the linguistic diversity of the target population. For example, one of the largest translation services in the world (*The Directorate-General for Translation of the European Commission*) translated 1,805,689 pages in 2008. This figure has grown exponentially during the last few years (from 1.1 million pages in 1997 to 1.3 million pages in 2004 and 1.8M pages in 2008[1]). Even so, the high translation cost in terms of money and time is a bottleneck that prevents all information from being easily spread across languages.

In this context, machine translation is becoming more and more attractive. There are many automatic translation services which are freely offered on the World Wide Web and every day they are used to translate thousands of web pages, even though the translation performance is still far from being perfect.

Additionally, much research efforts has been focused on machine translation during the last 50 years. In the 1950s, The Georgetown experiment (1954) involved the fully automatic translation of over sixty Russian sentences into English. The experiment was a great success and ushered in an era of substantial funding for machine translation research. The authors claimed that within three to five years, machine translation issues would be fully resolved. Real progress was much slower, however, and following the ALPAC report (1966), which found that the ten-year-long research had failed to fulfil expectations, funding was greatly reduced.

Early machine translation systems carried out direct word-by-word translation. Later, the use of linguistic information and abstract levels of representation increased, giving rise to transfer-based and interlingua-based approaches (Hutchins, 1986). In the late 1980s, the huge increase in computational power and availability of written translated texts allowed the development of statistical (Brown *et al.*, 1988) and other corpus-based (Nagao, 1984) approaches. Over the years the statistical approach archived a prominent status at the expense of the linguistic approach, until it became the dominant paradigm with little research work on from older paradigms. Nowadays, on the other hand, most research efforts are focused on enhancing SMT by incorporating any kind of linguistic knowledge.

In its pure form, Statistical Machine Translation (Brown *et al.*, 1993)

[1] http://ec.europa.eu/dgs/translation/bookshelf/tools_and_workflow_en.pdf

systems do not make use of traditional linguistic data, and all the knowledge required is statistically extracted from bilingual (human translated) documents. The essence of this method is first to align word sequences (phrases) and individual words of the parallel texts and calculate the probabilities of each phrase in the source sentence being translated into a phrase with which is aligned in the other language. Finally, the translation process consists of finding the target language sentence which maximizes the translation probability according to the models extracted from the corpus.

The performance of SMT systems depends heavily on both the bilingual corpus used and the "distance" between the languages involved in the translation. Thus, the translation quality decreases when there are few corpora available or when the translation is carried out from/into a morphologically rich language. For example, in experiments carried out on the Europarl corpus the results obtained for different language pairs varies drastically, from a 40.2 BLEU score for Spanish-French translation to a 10.3 BLEU score for Dutch-Finnish translation (Koehn, 2005). Although the BLEU metric could not be used to compare systems trained in such different environments the great difference between the two scores certainly shows the difference in the complexity of the task.

The Basque language, a morphologically rich language, has many peculiarities which differentiate it from most European languages. Those differences make translation between Spanish (or English) and Basque an interesting challenge which involves both morphological and syntactical features.

The agglutinative nature of Basque means that much of the morpho-syntactic information which is expressed in separate words in most European languages is expressed using suffixes in Basque. Thus, the information expressed via prepositions or articles in Spanish is expressed by means of suffixes added to the last word of the noun-phrase (similarly the information expressed via conjunctions is attached to the end of the verb phrase). Thus, 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.

Furthermore, there are also syntactic differences related to the word order that have a negative impact on the translation. As we have already explained, the agglutinative nature of Basque means that prepositions, usually placed at the beginning of the phrase, 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. For example, prepositional phrases attached to noun-phrases are laced before the noun

phrase instead of after it. The order of the constituents in Basque sentences is very flexible, but, 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.

These differences have an extremely negative impact on most of the steps of Statistical Machine Translation, such as word alignment, phrase extraction and decoding. In addition, Basque is a low-density language and there are few corpora available compared to other more widely used languages, such as Spanish, English, or Chinese. Although the parallel corpus available for Spanish-Basque has increased from 1 million Basque words (1.2 million Spanish words) to 6 million Basque words (8 million Spanish words) during the development of this thesis, it is still far bellow the corpora available for other languages. In Europarl there are at least 30 million words for each language.

Nowadays, in order to overcome the limitations encountered by the different approaches to machine translation, most research effort is being focused on combining them. Thus, there are many attempts to try to include linguistic information (usually used in knowledge-based approaches) on the corpus-based systems. In the same way, other attempts are focused on improving the translation quality by combining different system's outputs (usually based on different translation techniques).

This thesis has been developed in the context presented above and our objectives have been to:

- Deal with the agglutinative nature of Basque by splitting words into smaller tokens, which allows a better statistical translation. By splitting words into morphemes and working at this level of representation we reduced the number of tokens that occur only once and, at the same time, we reduce the 1-to-n alignments. In any case, several criteria could be used to segment words and the way the segmentation is carried out impacts on the quality of the translation. In order to determine the most appropriate segmentation for a Spanish-Basque system, we tried different segmentation options and analyzed their effects on the translation quality.

- Implement different techniques to deal with word order differences in statistical machine translation. The tested techniques cover the most common techniques, and they are applied both at decoding (a lexicalized reordering model has been integrated) and preprocessing. At the same time, we tested two different methods of carrying out this

preprocessing, based on manually defined rules on the result of the syntactic analysis and using a separate SMT system to "translate" the original source language into a reordered source language which makes the translation easier.

- Improve MT results by combining the SMT system developed in this PhD thesis with the Rule-Based and Example-Based Machine Translation systems previously developed in our research group for the same language pair. For this purpose we carried out two different hybridization experiments. In the first experiment, we translated each sentence using the three systems we have available (SMT, RBMT and EBMT systems) and the most appropriate translation was chosen for each sentence. In the second experiment, we built a Statistical Post-Editing system in order to fix the errors made by the RBMT system. For this purpose, an SMT system was trained to post-edit the translation of the RBMT system; in other words, "translate" from the output of the RBMT system to the real target language.

- Carry out a human evaluation based on the HTER metric, allowing us to complete and contrast the results obtained by the automatic evaluation based on the BLEU metric.

- Measure the impact of the size and nature of the corpora on the different techniques developed during the thesis. In order to do this, we reran our experiments using corpora from different domains and of different sizes.

We achieved positive results for all the objectives set for this thesis, improving on the results obtained with the Spanish-Basque SMT baseline system. In this summary, we present the work performed and the results obtained.

2 Related work

2.1 Treatment of morphology in SMT

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

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

Segmentation is the most usual way to translate into highly inflected languages, but other approaches have been also tried. In (Bojar, 2007) factored translation has been used in 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 into two steps, where they first use the usual SMT system to translate from English to Russian lemmas and, then, in a second step, they determine the inflection of each lemma, making use of bilingual information.

2.2 Treatment of syntactic differences in SMT

Different studies have attempted to deal with word order differences in statistical machine translation. The most commonly used approach is the prepro-

cessing of the source sentence in order to obtain a word order that matches with the target language's word order, allowing an almost monotonous translation. The two main approaches are found in the bibliography; those where the reordering rules are manually defined based on the linguistic analysis of the source, and those where the reordering is automatically inferred from the training corpus.

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

Many other research works attempt to learn the possible reordering automatically from the training corpus, instead of defining them hand-manually. Some of them 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, training an SMT system to "translate" from the original source sentences to the reordered source sentences.

2.3 Combination of MT paradigms: Multi Engine hybridization

In (van Zaanen and Somers, 2005; Matusov *et al.*, 2006; Macherey and Och, 2007) there are a set of references about MEMT (Multi-Engine MT) including the first attempt by Frederking and Nirenburg (1994). All the papers on MEMT reach the same conclusion: combining the outputs results in a better translation. Most of the approaches generate a new consensus translation by combining different SMT systems using different language models and in some cases also combining with RBMT systems. Some of the approaches require confidence scores for each of the outputs. The improvement in trans-

lation quality is always lower than an 18% relative increase in the BLEU score.

Chen *et al.* (2007) report an 18% relative increase for in-domain evaluation, and 8% for out-domain, by incorporating phrases (extracted from alignments from one or more RBMT systems with the source texts) into the phrase table of the SMT system and using the Moses open-source decoder to find good combinations of phrases from the SMT training data with the phrases derived from RBMT.

Matusov *et al.* (2006) report a 15% relative increase in the BLEU score by using consensus translation computed by voting on a confusion network. Pairwise word alignments of the original translation hypotheses were estimated for an enhanced statistical alignment model in order to explicitly capture reordering.

Macherey and Och (2007) presented an empirical study on how different selections of translation outputs affect translation quality in system combination. Composite translations were computed using (i) a candidate selection method based on inter-system BLEU score matrices, (ii) a ROVER-like combination scheme, and (iii) a novel two-pass search algorithm which determines and re-orders bags of words that build the constituents of the final consensus hypothesis. All methods gave statistically significant relative improvements of up to 10% in the BLEU score. They combine large numbers of different research systems.

Mellebeek *et al.* (2006) report improvements of up to 9% in the BLEU score. Their experiment is based on the recursive decomposition of the input sentence into smaller chunks, and a selection procedure based on majority voting that finds the best translation hypothesis for each input chunk using a language model score and a confidence score assigned to each MT engine.

Huang and Papineni (2007) and Rosti *et al.* (2007) combine multiple MT system outputs at the word, phrase and sentence levels. They report improvements of up to 10% in the BLEU score.

2.4 Combination of MT paradigms: Automatic Post-Editing

In the experiments related by Simard *et al.* (2007a) and Isabelle *et al.* (2007) the Statistical Post-Editing (SPE) task is viewed as translation from the language of RBMT outputs into the language of their manually post-edited counterparts. So they don't use a parallel corpus created by human translation. Their RBMT system is SYSTRAN and their SMT system PORTAGE. Simard *et al.* (2007a) report a reduction in post-editing effort of up to a third compared to the output of the rule-based system (i.e. the input to the SPE),

and an improvement of as much as 5 BLEU points over the direct SMT approach. Isabelle *et al.* (2007) conclude that such an SPE system appears to be an excellent way of improving the output of a vanilla RBMT system, and constitutes a worthwhile alternative to the costly manual adaptation efforts for such systems. Thus, an SPE system using a corpus with no more than 100,000 words of post-edited translations is enough to outperform an expensive lexicon-enriched baseline RBMT system.

The same group recognizes (Simard *et al.*, 2007b) that this sort of training data is seldom available, and they conclude that the training data for the post-editing component does not need to be hand-manually post-edited translations, that but can even be generated from standard parallel corpora. Their new SPE system again outperforms both the RBMT and SMT systems. The experiments show that although post-editing is more effective when little training data is available, it also remains competitive with SMT translation even when larger amounts of data are available. Following a linguistic analysis they conclude that the main improvement is due to lexical selection.

In (Dugast *et al.*, 2007), the authors of SYSTRAN's RBMT system present a huge improvement in the BLEU score for an SPE system when compared to raw translation output. They achieved an improvement of around 10 BLEU points for German-English using the Europarl test set of WMT2007.

Ehara (2007) presents two experiments to compare RBMT and SPE systems. Two different corpora are issued: one is the reference translation (PAJ, Patent Abstracts of Japan); the other is a large-scale target language corpus. In the former case SPE wins and in the later case RBMT wins. Evaluation is performed using NIST scores and a new evaluation measure, NMG, which counts the number of words in the longest sequence matched between the test sentence and the target language reference corpus.

Finally, Elming (2006) works in the more general field known as Automatic Post-Processing (APE). The author uses transformation-based learning (TBL), a learning algorithm for extracting rules to correct MT output by means of a post-processing module. The algorithm learns from a parallel corpus of MT output and human-corrected versions of this output. The machine translations are provided by a commercial MT system, PaTrans, which is based on Eurotra. Elming reports a 4.6 point increase in the BLEU score.

3 Main Contributions

3.1 Basque morphology in SMT (chapter 4)

The agglutinative nature of Basque means that much of the morpho-syntactic information that is expressed in separate words in most European languages is expressed using suffixes in Basque. Thus, the information expressed using prepositions or articles in Spanish, is expressed by means of suffixes that are added to the last word of the noun-phrase. 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 we then worked at this level of representation. Working at the morpheme level we reduced the number of tokens that occur only once and, at the same time, we reduced 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 of training Moses on the segmented text (the same training options are used in the 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 integrated a word-level language model using n-best list re-ranking.

We proved that the quality of the translation varies significantly when applying different options for word segmentation. Based on the same output of the morphological analyzer, we segmented words in different ways, creating more fine- or coarse-grained segments (from one token per 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, the best results being obtained with a hand-manually defined intermediate grouping based on an error analysis of the word-alignments.

In any case, the work done by hand is language dependent and could not be reused for a different pair of languages. Thus, in order to find a language independent way to define the most appropriate segmentation, we focused our research on establishing a statistical method to decide which morphemes have to be put into the same token. We observed that the morphemes that generate most of the errors are those which do not have their own *meaning*: i.e. those that *need* another morpheme to complete their meaning. We used

the *mutual information* metric in order to measure statistical dependence between two morphemes and we grouped those morphemes that are more dependent than a set threshold. In these experiments we tried different thresholds and we obtained the best results when the threshold was set to 0.5 (a value that involves grouping most of the morphemes). This statistical criterion achieves results that are almost as accurate as those obtained with the hand-manually defined criterion. Thus, we could use this statistical grouping criterion to adapt our system to a different language pair such as English-Basque.

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

We have also observed an interesting correlation between the token number and translation quality. Thus, in order to obtain better translation results, we want to refine the hand-manually defined segmentation criteria to reduce the difference in the number of tokens of both languages.

We have published a paper (Díaz de Ilarraza *et al.*, 2009) based on the work explained on this section, which can be checked in order to obtain a more in-depth explanation.

3.2 Reordering in SMT (chapter 5)

In this chapter, we will deal with the reordering problem in Spanish-Basque statistical machine translation, comparing different approaches and analyzing their strengths and weaknesses. The tested approaches cover the more usual techniques, including techniques implemented for decoding and preprocessing. We tried two different methods of carrying out this preprocessing, based on hand-manually defined rules for the syntactic analysis and using a separate SMT to "translate" the original source language, to a reordered source language which makes the translation easier.

The results obtained in this experiment allow us to compare different reordering methods for an specially demanding task such as Spanish-Basque translation. According to the results, the three reordering methods tested here (which can be considered as representative of the present-day research) outperform the baseline, with the best results being achieved by the lexicalized reordering implemented at decoding.

We also tested different combinations of methods, obtaining a significant improvement by using the syntax-based and the lexicalized reordering

together. Each method takes advantage of different information and they are able to complement each other. For instance, order differences between nouns and adjectives are not treated in Syntax-Based reordering and they are probably corrected by the lexicalized reordering.

On the other hand, the combination of the statistical reordering used in preprocessing and the lexicalized reordering in decoding achieves worse results than those obtained by the single methods on their own. The drop in performance dropping probably indicates that both methods use the same information about word alignment, so that no improvement is achieved by combining the methods.

As future work, we are planning to rerun the experiments on a larger 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 current ones do not cover all the order differences between the two languages. Furthermore, we are considering a way to allow the decoder to choose from among different reordering options proposed by the Syntax-Based preprocessing (using a list of possible reorderings or a word-graph as input to the decoder).

We have published a paper (Díaz de Ilarraza *et al.*, 2009) based on the work explained in this section which can be checked in order to obtain a more in-depth explanation.

3.3 Multi-Engine hybridization (chapter 6)

In order to continue improving the results for translation and taking into account that there were already Rule-Based (Mayor, 2007) and Example-Based Machine Translation systems available for Spanish-Basque, we experimented with a simple mixing approach. This first hybridization attempt consists of selecting the best output in a multi-engine system (MEMT, Multi-engine MT). In our case, we combined the previously existing RBMT, EBMT and SMT systems developed in this thesis.

For this first attempt, we combined the three approaches in a very simple hierarchical way, processing each sentence with the three engines (RBMT, EBMT and SMT) and then trying to choose the best translation among them. First, we divided the text into sentences. Then we processed each sentence using each engine (parallel processing when possible). Finally, we selected one of the translations, dealing with the following:

- The precision of the EBMT approach is very high, but its coverage is

low.

- The SMT engine gives a confidence score.
- RBMT translations are more appropriate for human post-edition than those of the SMT engine, but SMT gets better scores when BLEU and NIST are used, with only one reference (Labaka *et al.*, 2007).

With these results for the single approaches we decided to apply the following combinatorial strategy:

1. If the EBMT engine covers the sentence, we chose its translation.
2. We chose the translation from the SMT engine if its confidence score was higher than a given threshold.
3. Otherwise, we chose the output from the RBMT engine.

A significant improvement was achieved in translation quality for BLEU in comparison with the improvements obtained by other systems. These improvements would be difficult to achieve with single engine systems. The RBMT contribution seems to be very small with automatic evaluation, but we expect that HTER evaluation will show better results.

In spite of trying the strategy for a domain, we belive that our translation system is a major advance in the field of language tools for Basque. The restriction of using a corpus in a domain is due to the absence of large, reliable Spanish-Basque corpora.

In the near future, we plan to carry out new experiments using a combination of the outputs based on a language model. We also plan to define confidence scores for the RBMT engine (including penalties when suspicious or very complex syntactic structures are present in the analysis). Furthermore, we are planning to detect other types of translation patterns, especially at the phrase or chunk level.

We have published a paper (Alegria *et al.*, 2008) based on the work explained in this section which can be checked in order to obtain a more in-depth explanation.

3.4 Statistical Post Edition (chapter 7)

Following the hybridization attempts we tried cascade hybridization, where different machine translation systems are used one after the other. In this

approach, we train a Statistical Machine Translation system which post-edits the output of the Rule-based Machine Translation system.

In order to carry out experiments with Statistical Post-Editing (SPE), we first translated Spanish sentences in the parallel corpus using our rule-based translator (Matxin) and, then, using these automatically translated sentences and their corresponding Basque sentences in the parallel corpus, we built a new parallel corpus to be used in training our statistical post-editor.

The statistical post-editor is the same corpus-based system as explained earlier. This system is based on freely available tools but enhanced in two main ways:

- In order to deal with the morphological richness of Basque, the system works at morpheme level, so that, in Section 3.1, a generation phase is necessary after the SPE is applied.
- Following the work carried out in collaboration with the DCU, the statistically extracted phrases are enriched with linguistically motivated chunk alignments.

We performed two experiments to verify the improvement obtained for other languages by using statistical post-editing. However, our experiments differ from other similar works because we use a morphological component in both RBMT and SMT translations, and because the size of the available corpora is small.

Our results are consistent with the huge improvements when using a SPE approach on a restricted domain presented by (Dugast *et al.*, 2007; Ehara, 2007; Simard *et al.*, 2007b). We obtained a 200% improvement in the BLEU score for a SPE system working with Matxin the RBMT system, when compared with raw translation output, and 40% when compared with the SMT system.

Our results also are consistent with a smaller improvement when using more general corpora, as presented by (Ehara, 2007; Simard *et al.*, 2007b).

We cannot work with manually post-edited corpora as did Simard *et al.* (2007a) and Isabelle *et al.* (2007), because there is no such large corpus for Basque, but we plan to compile one and compare the results obtained using a real post-edition corpus and the results presented here.

Finally, we also tried automatic extracting rules to correct MT output by means of a post-processing module (Elming, 2006), but the result obtained were not positive, probably because of the big difference between the MT

output and the reference we used (which is a general reference instead of a close detailed manual post-edition).

We have published a paper (Díaz de Ilarraza *et al.*, 2008) based on the work explained in this section, which can be checked in order to obtain a more in-depth explanation.

4 Current State and Work to Complete the Thesis

We consider that we have achieved the expected results for an efficient treatment of morphologically rich languages in SMT. In addition, we have improved the translation results with some attempts at hybridization, based on the combination of SMT with other MT paradigms, such as RBMT and EBMT, for which we are currently completing the writing up phase. Thus, we are preparing the dissertation based on the results obtained in (Alegria *et al.*, 2008; Díaz de Ilarraza *et al.*, 2008; Díaz de Ilarraza *et al.*, 2009; Díaz de Ilarraza *et al.*, 2009).

References

Alegria I., Casillas A., Díaz de Ilarraza A., Igartua J., Labaka G., Lersundi M., Mayor A. and Sarasola K. *Spanish-to-Basque MultiEngine Machine Translation for a Restricted Domain*. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas*. AMTA, Hawaii, Canada, 2008.

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

Brown P.F., Cocke J., Della Pietra S.A., Della Pietra V.J., Jelinek F., Mercer R.L. and Roossin P.S. *A Statistical Approach to Language Translation*. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING)*, 1988.

Brown P.F., Pietra V.J., Pietra S.A.D. and Mercer R.L. *The Mathematics of Statistical Machine Translation: Parameter Estimation*. Computational Linguistics, volume 19, pp. 263–311, 1993.

Chen B., Cettolo M. and Federico M. *Reordering Rules for Phrase-based Statistical Machine Translation*. In *IWSLT 2006*, pp. 182–189, 2006.

Chen Y., Eisele A., Federmann C., Hasler E., Jellinghaus M. and Theison S. *Multi-engine machine translation with an open-source decoder for statistical machine translation*. In *proceedings of the Second Workshop on Statistical Machine Translation*, pp. 193–196, 2007.

Collins M., Koehn P. and Kucerova I. *Clause Restructuring for Statistical Machine Translation*. In *ACL*. The Association for Computer Linguistics, 2005.

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

Díaz de Ilarraza A., Labaka G. and Sarasola K. *Statistical Post-Editing: A Valuable Method in Domain Adaptation of RBMT Systems*. In *Proceedings of the Mixing Approaches to Machine Translation workshop*. Donostia, Spain, 2008.

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

Díaz de Ilarraza A., Labaka G. and Sarasola K. *Reordering in Spanish-Basque SMT*. In *Proceedings of the MT Summit 2009*. International Association for Machine Translation, Ottawa, Canada, 2009.

Dugast L., Sennellart J. and koehn P. *Statistical post-editing in SYSTRAN's rule-based translation system*. In *ACL 2007: proceedings of the Second Workshop on Statistical Machine Translation*, pp. 220–223. Prague, Czech Republic, 2007.

Ehara T. *Rule based machine translation combined with statistical post editor for Japanese to English patent translation*. In *MT-Summit XI Workshop on patent translation*, pp. 13–18. Copenhagen, Denmark, 2007.

Elming J. *Transformation-based correction of rule-based MT*. In *11th Annual Conference of the European Association for Machine Translation*, pp. 219–226. Oslo, Norway, 2006.

Frederking R. and Nirenburg S. *Three heads are better than one*. In *Proceedings of the fourth ANLP*, 1994.

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

Huang F. and Papineni K. *Hierarchical system combination for machine translation*. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 277–286, 2007.

Hutchins W.J. *Machine translation: past, present, future*. John Wiley & Sons, Inc., New York, NY, USA, 1986.

Isabelle P., Goutte C. and Simard M. *Domain adaption of MT systems through automatic post-editing*. In *Proceedings of MT-Summit XI*, pp. 255–261. Copenhagen, Denmark, 2007.

Koehn P. *Europarl: A Parallel Corpus for Statistical Machine Translation*. In *MT Summit X*, 2005.

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

Labaka G., Stroppa N., Way A. and Sarasola K. *Comparing Rule-Based and Data-Driven Approaches to Spanish-to-Basque Machine Translation*. In *Proceedings of MT-Summit XI*, 2007.

Macherey W. and Och F.J. *An Empirical Study on Computing Consensus Translations from Multiple Machine Translation Systems*. In *Proceedings of the EMNLP and CONLL 2007*, 2007.

Matusov E., Ueffing N. and Ney H. *Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment*. In *Proceedings of EACL 2006*, 2006.

Mayor A. *Matxin: erregeletan oinarritutako itzulpen automatikoko sistema*. Ph.D. thesis, Euskal Herriko Unibertsitatea, 2007.

Mellebeek B., Owczarzak K., Genabith J.V. and Way A. *Multi-engine machine translation by recursive sentence decomposition*. In *Proceedings of the 7^{th} Conference of the Association for Machine Translation in the Americas, Visions for the Future of Machine Translation*, pp. 110–118, 2006.

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

Nagao M. *A framework of a mechanical translation between Japanese and English by analogy principle*. In *Proc. of the international NATO symposium on Artificial and human intelligence*, pp. 173–180. Elsevier North-Holland, Inc., New York, NY, USA, 1984.

Nießen S. and H. Ney. *Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information*. Comput. Linguist., volume 30(2), pp. 181–204, 2004.

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

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

Ramanathan A., Bhattacharya P., Hegde J., M.Shah R. and M S. *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, 2008.

Rosti A.V.I., Ayan N.F., Xiang B., Matsoukas S., Schwartz R. and Dorr B.J. *Combining outputs from multiple machine translation systems*. In *NAACL-HLT-2007 Human Language Technology: the conference of the North American Chapter of the Association for Computational Linguistics*, pp. 228–235, 2007.

Sanchis G. and Casacuberta F. *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)*, pp. 191–198. Skövde, Sweden, 2007.

Simard M., Goutte C. and Isabelle P. *Statistical Phrase-Based Post-Editing*. In C.L. Sidner, T. Schultz, M. Stone and C. Zhai, eds., *HLT-NAACL*, pp. 508–515. The Association for Computational Linguistics, 2007a.

Simard M., Ueffing N., Isabelle P. and Kuhn R. *Rule-based translation with statistical phrase-based post-editing*. In *ACL 2007: proceedings of the Second Workshop on Statistical Machine Translation*, pp. 203–206. Prague, Czech Republic, 2007b.

van Zaanen M. and Somers H. *DEMOCRAT: Deciding between Multiple Outputs Created by Automatic Translation*. In *MT Summit X*, 2005.

Zhang Y., Zens R. and Ney H. *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, 2007.

Publications

Following we include the papers supporting this PhD dissertation. They are organized according to our chapter distribution:

Chapter 4

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

Labaka G., Díaz de Ilarraza A. and Sarasola K. *Descripción de los sistemas presentados por IXA-EHU a la evaluación ALBAYCIN'08*. In *V Jornadas en tecnología del Habla*. Bilbao, Spain, 2008.

Agirre E., Díaz de Ilarraza A., Labaka G. and Sarasola K. *Uso de información morfológica en el alineamiento Español-Euskara*. In *XXII congreso de la SEPLN*. Zaragoza, Spain, 2006.

Labaka G., Stroppa N., Way A. and Sarasola K. *Comparing Rule-Based and Data-Driven Approaches to Spanish-to-Basque Machine Translation*. In *Proceedings of MT-Summit XI*, 2007.

Chapter 5

Díaz de Ilarraza A., Labaka G. and Sarasola K. *Reordering in Spanish-Basque SMT*. In *Proceedings of the MT Summit 2009*. International Association for Machine Translation, Ottawa, Canada, 2009.

Chapter 6

Alegria I., Casillas A., Díaz de Ilarraza A., Igartua J., Labaka G., Lersundi M., Mayor A. and Sarasola K. *Spanish-to-Basque MultiEngine Machine Translation for a Restricted Domain*. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas*. AMTA, Hawaii, Canada, 2008.

Díaz de Ilarraza A., Labaka G. and Sarasola K. *Mixing Approaches to MT for Basque: Selecting the best output from RBMT, EBMT and SMT*. In *Proceedings of the Mixing Approaches to Machine Translation workshop*. Donostia, Spain, 2008a.

Chapter 7

Díaz de Ilarraza A., Labaka G. and Sarasola K. *Statistical Post-Editing: A Valuable Method in Domain Adaptation of RBMT Systems*. In *Proceedings of the Mixing Approaches to Machine Translation workshop*. Donostia, Spain, 2008b.

Other publications

Alegria I., Arregi X., Artola X., Díaz de Ilarraza A., Labaka G., Lesundi M., Mayor A. and Sarasola K. *Strategies for sustainable MT for Basque: incremental design, reusability, standaritation and open-source*. In *Proceedings of the IJCNLP-08 Workshop on NLP for less Privileged Languages*. Hyderabab, India, 2008a.

Alegria I., Díaz de Ilarraza A., Labaka G., Lersundi M., Mayor A. and Sarasola K. *Transfer-based MT from Spanish into Basque: reusability, standardization and open source*. In *Springer Lecture Notes in Computer Science 4394*, pp. 374–384. Mexico City, Mexico, 2007.

Alegria I., Díaz de Ilarraza A., Labaka G., Lersundi M., Mayor A. and Sarasola K. *An FST grammar for verb chain transfer in a Spanish-Basque MT System*. In *Proceedings of Finite-State Methods and Natural Language Processing*, pp. 295–296. Helsinki, Finland, 2005.

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.

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

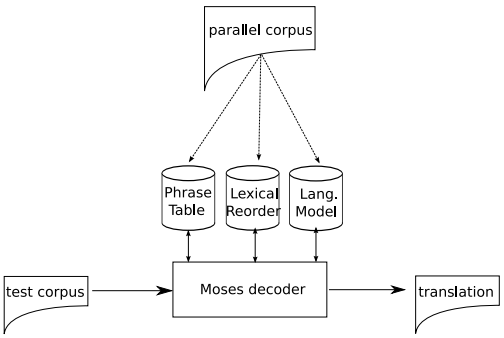


Figure 1: Basic design of a SMT system

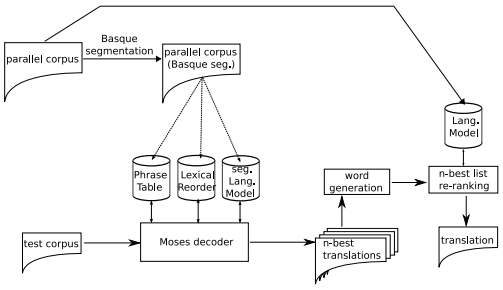


Figure 2: Design of the morpheme-based SMT system

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

DESCRIPCIÓN DE LOS SISTEMAS PRESENTADOS POR IXA-EHU A LA EVALUACIÓN ALBAYCIN'08

Gorka Labaka, Arantza Díaz de Ilarraza, Kepa Sarasola

Grupo IXA
Universidad del País Vasco
{jiblaing, jipdisaa, jipsagak}@ehu.es

RESUMEN

En este artículo describimos los sistemas presentados por el grupo IXA-EHU a la evaluación ALBAYCIN'08. Dada las características de los pares de lenguas a tratar y la naturaleza aglutinativa del euskara hemos procedido a la segmentación de las palabras en morfemas para, de este modo, facilitar el alineamiento. Además este proceso habilita la posibilidad de aprender pseudosintagmas (secuencias de palabras sin estructura sintáctica) que pueden estar compuestos además de por palabras, por morfemas considerados de manera independiente a la palabra a la que van unidos; por ejemplo, en el caso de 'etxe-ra' noa (voy a casa), el pseudosintagma '-ra noa' se puede alinear con 'voy a'.

Además de la segmentación hemos incorporado a la tabla de traducción pares de pseudosintagmas que han sido extraídos utilizando técnicas de traducción basada en ejemplos de MaTrEx [1]. Estos nuevos pseudosintagmas, a diferencia de los extraídos por las técnicas estadísticas, coinciden con sintagmas desde un punto de vista lingüístico. Al ampliar la tabla de traducción con estos nuevos pseudosintagmas, se amplia la cantidad de pseudosintagmas disponibles por el decodificador, además de favorecer aquellas traducciones sintácticamente correctas.

1. INTRODUCCIÓN

En este artículo describimos los sistemas presentados por el grupo IXA de la Universidad del País Vasco a la evaluación Albaycin'08.

El alto nivel de flexión del euskara, junto con el hecho de que no haya gran cantidad de corpus paralelo accesible, complica la tarea de traducción español-euskara convertiendola en un reto interesante.

Para hacer frente a su alto nivel de flexión hemos segmentado las palabras en euskara dividiéndolas en morfemas, de modo similar al realizado en otros trabajos para otros pares de lenguas de gran flexión, como en el caso del inglés-checo [2] y el inglés-turco [3].

El articulo está organizado del siguiente modo: en la sección 2 explicamos las distintas técnicas que usaremos en nuestros sistemas; la sección 3 está dedicada a mostrar los sistemas que hemos evaluado y que se basan en combinaciones de las técnicas previamente explicadas; en la sección 4 resumimos los resultados conseguidos por cada uno de los sistemas; finalmente comentamos las conclusiones extraídas de esos resultados (sección 5).

Este trabajo ha sido subvencionado por Gobierno Vasco, mediante la ayuda predoctoral concedida a Gorka Labaka (código BFI05.326)

2. TÉCNICAS UTILIZADAS

En esta sección explicamos las técnicas que hemos utilizado en la implementación de los sistemas que presentamos a la evaluación Albaycin'08.

2.1. Segmentación del texto en euskara

Dada la naturaleza aglutinante de la lengua y, continuando con el trabajo presentado en un publicación anterior [4], hemos llevado a cabo la segmentación del texto en euskera. De esta manera, tendremos en tokens independientes los morfemas que conforman una palabra.

Para llevar a cabo esta segmentación hemos analizado el texto en euskara utilizando EusTagger[5] y cada palabra se ha separado en como máximo tres tokenes: los prefijos, el lema y los sufijos. De este modo, para una palabra como 'etxekoa' (el de la casa) se crean dos tokenes: 'etxe' y '+koa'. En una primera aproximación, pensamos en crear un token por cada morfema pero, dadas las características de la salida de EusTagger que genera una segmentación muy fina (con muchos morfemas por palabra), decidimos unir todos los sufijos en un único token; los prefijos también fueron tratados de la misma manera. La razón principal para llevar a cabo esta segmentación es facilitar el alineamiento, ya que de este modo habrá menos alineamientos múltiples a la vez que se reduce la dispersión.

Gracias a este proceso de segmentación podemos aprender pseudosintagmas en los que toman parte sólo algunos de los morfemas de una palabra. De este modo se podría extraer el par de pseudosintagmas 'voy a' '-ra noa', donde la preposición 'a' se traduce con el sufijo '-ra' cuando acompaña al verbo 'ir' independientemente del lema al que esté unido. Sin la segmentación no sería posible esta clase de generalización teniendo que extraer pseudosintagmas distintos para cada ejemplo.

El hecho de usar el texto segmentado para entrenar el traductor estadístico, conlleva que necesitemos generar el texto final en euskara basándonos en la salida del traductor, ya que esta estará segmentada al igual que el corpus utilizado en el entrenamiento. Para generar el texto final hemos utilizado el módulo de generación del traductor basado en reglas matxin[6] (que utiliza en el mismo léxico que el analizador).

A la hora de generar el texto final hay que tener en cuenta que el traductor estadístico puede producir combinaciones de morfemas que no correctas pudiéndole asignar a un nombre la flexión correspondiente a un verbo o incluso llegando a asignarle algún tipo de flexión a tokenes que no se pueden flexionar como los signos de puntuación. En este caso y, como primera aproximación, eliminamos la flexión dejando únicamente el lema.

Finalmente, para poder incorporar un modelo de lenguaje basado en palabras (el decodificador usará uno basado en el texto segmentado), en vez de obtener sólo la mejor traducción que el decodificador es capaz de encontrar, obtenemos una lista de las n traducciones más probables y, tras la generación, reordenamos la lista de traducciones incorporando el modelo de lenguaje basado en palabras como si fuera un modelo más.

2.2. Hibridación SMT-EBMT: sistema MaTrEx

En colaboración con Nacional Centre for Language Technology de la Dublin City University hemos adaptado su sistema MaTrEx[1] para utilizarlo con el euskara. Este sistema consiste en enriquecer la tabla de traducción con pares de pseudosintagmas extraídos usando técnicas de la traducción automática basada en ejemplos.

Para extraer los nuevos pseudosintagmas, se analizan sintácticamente ambas partes del corpus paralelo y se marcan los sintagmas (hemos usado Freeling [7] para procesar el español y Eustagger para el euskara). En un segundo paso y basándose en los alineamientos palabra por palabra se alinean estos sintagmas y se incorporan a tabla de traducción.

3. SISTEMAS PRESENTADOS

Para crear nuestros sistemas hemos utilizado las siguientes herramientas:

- Alineador de palabras GIZA++ [8].
- Modelo de lenguaje SRILM [9]
- Moses SMT Toolkit [10]

Mediante estas herramientas de uso libre y los corpora habilitados por la organización (en la tabla 1 se muestra algunos datos de los corpora) hemos creado un sistema *baseline*, usando los *scripts* y los parámetros que Moses trae por defecto. Hay que tener en cuenta que el sistema *baseline* de Moses incorpora técnicas de reordenación lexicalizada además de la basada en distancia. En la creación del *baseline* se han llevado a cabo la optimización de los pesos de cada modelo usando BLEU y Minimum Error Rate Training.

Basándonos en este *baseline* hemos incorporado las técnicas explicadas en la sección 2 creando distintos sistemas de traducción. Posteriormente hemos evaluado el impacto que tiene cada técnica. Para incorporar los pseudosintagmas correspondientes al sistema MaTrEx, hay que analizar ambos textos, alinear los sintagmas basándose en los alineamientos palabra por palabra e incorporar los nuevos pares de pseudosintagmas a la tabla de traducción antes de calcular los pesos de los modelos de traducción con los *scripts* proporcionados con Moses. Tras este proceso se continua con el entrenamiento del sistema.

Por otro lado a la hora de usar el texto segmentado, además de preprocesar y post-procesar las oraciones en euskara, para segmentar el texto y volver a generar la forma final, hemos tenido que modificar el proceso de optimización para poder optimizar también el peso del modelo de lenguaje basado en palabras. Como hemos explicado anteriormente, el decodificador utiliza un modelo de lenguaje basado en el texto segmentado, y el modelo de lenguaje basado en palabras se incorpora a la traducción después del post-proceso de generación mediante el reordenamiento de una lista n-best. Por lo que en cada paso de la optimización hay que incorporar tanto la generación como el reordenamiento de las listas basándose en la lista n-best.

Además de los sistemas donde probamos las técnicas presentadas individualmente, también hemos entrenado un sistema donde probamos la combinación de ambas.

4. RESULTADOS

Hemos evaluado los sistemas presentados en la sección 3 sobre el corpus de test usando las métricas automáticas más usuales (BLEU, MBLEU, WER, PER). En la tabla 2 se presentan los resultados para dichos sistemas y métricas.

Lo más destacable de los datos presentados es que ambas técnicas individuales (MaTrEx y segmentación del euskara) mejoran los resultados del sistema *baseline* para todas las métricas utilizadas. A su vez, la combinación de técnicas supera a cada una de ellas considerada individualmente, logrando los mejores resultados.

5. CONCLUSIÓN

Las técnicas que hemos utilizado han dado un resultado satisfactorio mejorando ambas el *baseline*. Además la combinación de las misma supera a las técnicas aplicadas individualmente.

Respecto al trabajo futuro, nos proponemos modificar la segmentación del euskara, buscando una forma alternativa para agrupar los morfemas. Actualmente, todos los morfemas que acompañan al lema se agrupan en un único

corpora	lenguaje	oraciones	tokenes	vocabulario-tokenes
entrenamiento	Español	58202	1284212	50927
	Euskara		1010545	95724
	Euskara-segmentado		1546304	40436
development	Español	1456	32743	7073
	Euskara		25778	9030
	Euskara-segmentado		39420	6191
test	Español	1446	31004	6836
	Euskara		24372	8695
	Euskara-segmentado		37347	5976

Tabla 1. Estadísticas de los corpora utilizados.

	BLEU	MBLEU	NIST	WER	PER
baseline	10.82	10.21	4.51	80.44	61.67
MaTrEx	11.03	10.38	4.54	80.13	61.65
segmentación euskara	11.19	10.49	4.65	79.27	60.60
segmentación + MaTrEx	**11.37**	**10.65**	**4.71**	**78.65**	**60.01**

Tabla 2. Evaluación de los distintos sistemas probados.

token ya que se consiguen mejores resultados que manteniendo cada morfema un token pero pensamos que una agrupación intermedia mejoraría los resultados.

Por otro lado, nos planteamos mejorar el proceso de generación de la forma final, modificando la secuencia de morfemas que devuelve el traductor estadístico en aquellos casos que ésta no sea morfológicamente correcta. Esta modificación puede implicar la reordenación de la secuencia o la eliminación de algunos de los morfemas.

6. BIBLIOGRAFÍA

[1] N. Stroppa y A. Way, "MaTrEx: DCU Machine Translation System for IWSLT 2006," in *Proceedings of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 31–36.

[2] S. Goldwater y D. McClosky, "Improving statistical mt through morphological analysis," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Vancouver, 2005.

[3] Kemal Oflazer y Ilknur Durgar El-Kahlout, "Exploring different representational units in english-to-turkish statistical machine translation," in *Proceedings of Statistical Machine Translation Workshop at ACL 2007*, Prague, Czech Republic, June 2007.

[4] E. Agirre, A. Díaz de Ilarraza, G. Labaka, y K. Sarasola, "Uso de información morfológica en el alineamiento español-euskara," in *XXII Congreso de la SEPLN*, Zaragoza, septiembre 2006.

[5] I. Aduriz y A. Díaz de Ilarraza, "Morphosyntactic disambiguation and shallow parsing in computational porcessing of basque," in *Inquiries into the lexicon-syntax relations in Basque*, Bernarrd Oyharçabal, Ed., Bilbao, 2003.

[6] I. Alegria, A. Díaz de Ilarraza, G. Labaka, M. Lersundi, A. Mayor, K. Sarasola, M. Forcada, S. Ortiz, y L. Padró, "An open architecture for transfer-based machine translation between spanish and basque," in *Workshop on Open-Source Machine Translation*, Asia-Pacific Association for Machine Translation (AAMT), Ed., Phuket, Thailand, September 2005, pp. 7–14.

[7] X. Carreras, I. Chao, L. Padró, y M. Padró, "Freeling: an open-source suite of language analyzers," in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, 2004.

[8] F. Och y H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[9] Andreas Stolcke, "SRILM - An Extensible Language Modeling Toolkit," in *Proc. Intl. Conf. Spoken Language Processing*, Denver, Colorado, September 2002.

[10] 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, y Evan Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic, June 2007.

Uso de información morfológica en el alineamiento español-euskera

E. Agirre, A. Díaz de Ilarraza, G. Labaka, K. Sarasola
Euskal Herriko Univertsitatea/Universidad del Pais Vasco
649 pk - 20080 Donostia
e.agirre@ehu.es, jipdisaa@ehu.es, jiblaing@ehu.es, jipsagak@ehu.es

Resumen: En este artículo presentamos un primer estudio para el alineamiento de un corpus español-euskera mediante un alineador token-a-token en el que se consideran diferentes opciones de preprocesamiento morfológico. Usando GIZA++ conseguimos una reducción del error (Alignment Error Rate) de un 12.48 % respecto el baseline (carente de preproceso alguno), llegando al 23.76 %. Este resultado es comparable al obtenido para otros idiomas aglutinantes como el euskera.
Palabras clave: Traducción automática, alineamiento, idiomas aglutinantes

Abstract: In this paper we present a preliminary study for the alignment of a Spanish-Basque parallel corpus using a token-based aligner (GIZA++). We have studied several morphological pre-processing alternatives, and achieved 23.76 % Alignment Error Rate, with a reduction of 12.48 % over the baseline (no pre-processing). The results are comparable to those obtained for others agglutinative languages.
Keywords: Machine Translation, alignment, agglutinative languages

1. Introducción

En el área de la traducción automática, los sistemas estadísticos basados en corpus alineados están obteniendo buenos resultados en todas las competiciones. Además, el hecho de que gran parte del software y los recursos necesarios para entrenar y aplicar estos modelos estadísticos sean libres, hacen que su popularidad vaya en aumento. Estos sistemas estadísticos suelen tomar como entrada un corpus bilingüe donde las sentencias correspondientes a cada idioma están alineadas. El primer paso que llevan a cabo estos sistemas suele ser la alineación palabra por palabra de los testos.

Los grupos que han desarrollado estos sistemas se han centrado en idiomas con un grado de flexión bajo o medio (p.ej. el inglés, el francés, el español o el chino) para desarrollarlos, por lo que la metodología suele tomar palabras flexionadas como unidad básica.

En el caso de idiomas altamente flexionados (como es el euskera) la metodología de tomar la palabra flexionada como unidad básica provoca una muy alta dispersión de los datos, siendo este un escollo insalvable para la traducción estadística. Por ejemplo, la frase "en la casa" en euskera se traduce en una sola palabra "etxean".

Aunque relativamente menos flexionado, el español también presenta problemas en sus verbos, donde la concordancia compleja hace que la misma raíz verbal tenga muchas formas (p.ej. "venimos" que en inglés seria "we come").

Un estudio reciente (Koehn, 2005) da resultados de hasta 30.1 BLEU para pares como el inglés-español, pero de solo 17.6 para el inglés-alemán. A pesar de que en este estudio se han utilizado corpora equivalente para los distintos pares de idiomas, los resultados varían ostensiblemente. Esto muestra que el nivel de flexión de una lengua repercute en la calidad de los sistemas estadísticos.

En este artículo presentamos un estudio de las diferentes opciones para el alineamiento palabra a palabra de un corpus español-euskera. Específicamente, nos centraremos en las diferentes posibilidades de preprocesamiento morfológico (especialmente para el euskera, pero también para el español) con vistas a salvar la dispersión de datos y realizar un alineamiento óptimo. La importancia del alineamiento como tarea separada, viene avalada por las dos últimas competiciones públicas de alineamiento (Martin, Mihalcea, y Pedersen, 2005) y (Mihalcea y Pedersen, 2003).

Hemos realizado un exhaustivo estudio de diferentes posibilidades, que hemos evaluado sobre un corpus manualmente alineado de 100 sentencias. Para llevar a cabo estos experimentos utilizaremos herramientas que están disponibles libremente, tales como GIZA++

(Och y H. Ney, 2003) para el alineamiento, AlignmentSet[1] para la evaluación del alineamiento, HandAlign[2] para alinear manualmente el corpus de referencia, Freeling para analisis del español (Carreras et al., 2004), y Eustagger (Aduriz et al., 1994) para el del euskera (este ultimo disponible bajo licencia).

La siguiente sección presenta el trabajo relacionado. Seguiremos explicando las distintas técnicas que se han utilizado para preprocesar el euskera en la sección 3, y las utilizadas para preprocesar el español en la sección 4. En la sección 5 se presenta el diseño del experimento, para terminar comentando los resultados obtenidos en la sección 6 y las conclusiones y el trabajo futuro en la sección 7.

2. Trabajo relacionado

Ha habido distintos experimentos que utilizaban el preprocesamiento morfológico con el fin de mejorar el alineamiento de palabras de lenguas altamente flexionados. Así, podemos encontrar publicaciones que presentan distintos métodos probados en distintos pares de lenguas.

Para el par alemán-inglés hay varios trabajos que usan la información morfológica para preprocesar los textos mejorando así la alineación. En (Nießen y H. Ney, 2000) usan la información morfosintáctica para armonizar el orden de las palabras en los dos idiomas. Posteriormente (Nießen y H. Ney, 2004) etiquetaron algunas expresiones especialmente ambiguas con su categoría sintáctica, concretamente, combinaron algunas expresiones multi-palabra en un único token y dividieron ciertas palabras sustituyéndolas por el lema y etiquetas morfológicas. Mediante estas técnicas consiguieron mejorar sustancialmente la calidad del alineado.

En el caso del par inglés-checo (Goldwater y McClosky, 2005) preprocesaban el checo de cuatro modos diferentes; tokenizando tan solo, lematizando, creando etiquetas que se tratarán como tokens diferenciados y creando etiquetas que se unen al lema dando lugar a lo que podríamos identificar como unos lemas modificados. En el artículo mencionado, además del preproceso del texto de entrada, modificaban el algoritmo de alineación por lo que no se puede saber exactamente qué parte de la mejora que consiguen se debe al preproceso y qué parte se debe a la modificación del algoritmo de alineación.

No podemos dejar de mencionar dos publicaciones que al igual que nosotros, alinean el euskera con el español. (Caseli, M. Nunes, y Forcada, 2005) define un nuevo método de alineamiento que pretende superar los problemas que las técnicas más utilizadas tienen con las unidades multipalabra y con las diferencias en el orden de las palabras de los distintos idiomas a alinear. En esta publicación no se hace ningún tipo de preprocesamiento del texto y aunque para el par de lenguas español-portugués no consiguen superar los resultados de GIZA++, en el caso del español-euskera, donde los problemas mencionados tienen más importancia, logran mejorar los resultados conseguidos con GIZA++.

En el caso de (Nevado, Casacuberta, y Landa, 2004) se basan en alineamiento a nivel de palabra conseguido por GIZA++ para lograr segmentos más pequeños que una frase para utilizar en memorias de traducción. Este artículo no trata de mejorar el alineamiento en sí.

3. Preprocesamiento del euskera

El preproceso del euskera incluye tokenización, lematización y segmentación.

3.1. Tokenización

Es el prepocesamiento mínimo que necesita GIZA++ ya que éste tokeniza el texto usando tan sólo los espacios en blanco. Es necesario separar los signos de puntuación de las palabras precedentes, y, aunque cabe la posibilidad de unir los términos multipalabra para que GIZA++ los trate como un sólo token, no hemos utilizado esta opción.

Este es un ejemplo de una frase tokenizada, donde los tokens están separados con blancos:

Edukiaren egiturari dagokionez , funtsezko aldaketaren bat egin da

3.2. Lematización

Para procesar el euskera hemos usado Eustagger (Aduriz et al., 1994). Siguiendo con el ejemplo anterior, ésta sería la sentencia lematizada:

eduki egitura egon , funts aldaketa bat egin izan

[1] http://www.lsi.upc.es/lambert/software/AlignmentSet.html
[2] http://www.isi.edu/hdaume/HandAlign/

3.3. Segmentación

Al lematizar el euskera se pierde la información que dan los sufijos que se añaden al lema. Información como el caso gramatical, el número o información de subordinación. Esta información además se debería alinear con algunos elementos del español como las preposiciones, los artículos o las conjunciones. Por lo cual, además de lematizar el texto, hemos pretendido hacer explícitos estos rasgos mediante etiquetas que GIZA++ tratará como tokens diferentes. Nuestra intención es que GIZA++ alinee estas etiquetas que hemos añadido con los elementos del español anteriormente mencionados.

Hemos probado diferentes conjuntos de rasgos y combinación de ellos en etiquetas para ver cúales son las más adecuadas para este par de lenguas. A continuación presentamos los distintos rasgos y combinaciones en etiquetas que hemos utilizado y de qué modo las hemos combinado.

Número (NUM): Esta etiqueta representa tanto la información de definición (definido o indefinido) como el número en el caso de ser definido. Sus valores posibles son MG (indefinido), M,S (definido singular), M,P (definido plural) y M,PH (definido plural cercano). Creemos que esta etiqueta será alineada con el artículo del español. Este es un ejemplo de codificación según esta etiqueta del ejemplo anterior:

eduki M,S egitura M,S egon , funts MG, aldaketa MG, bat MG, egin izan

Caso (CAS): Esta etiqueta representa el caso, en euskera se diferencian 18 casos. Estos casos creemos que se alinearán con las preposiciones del español. Aunque algunas veces como con los casos gramaticales (ABS y ERG) que son los que toman tanto el sujeto como el objeto la etiqueta creada para el euskera no tendrá ninguna correspondencia en español. En este ejemplo se muestra la misma frase que en ejemplos anteriores codificado mediante esta etiqueta.

eduki GEN egitura DAT egon , funts ABS aldaketa GEN bat ABS egin izan

Declinación (DEC): Mediante esta etiqueta hemos unido las dos anteriores, ya que al empezar con los experimentos nos dimos cuenta que cuando por parte del español se utilizaba el texto lematizado la etiqueta que representa el número no se alineaba con los artículos como pensábamos y se alineaba mal o no se alineaba con nada. Por lo que pensamos en unir las dos etiquetas. En el siguiente ejemplo mostramos la misma frase con este método de segmentacion.

eduki GEN,M,S egitura DAT,M,S egon , funts ABS,MG, aldaketa GEN,MG, bat ABS,MG, egin izan

subordinación (SUB): En euskera la subordinación de las oraciones se representa mediante un sufijo que se añade al verbo subordinado. Hay distintos tipos de subordinaciones que según nuestro planteamiento se deberían alinear con las conjunciones o los pronombres relativos del español. La misma frase de ejemplos anteriores con una nueva segmentacion.

eduki egitura egon KAUS , funts aldaketa bat egin izan

Aspecto (ASP): Esta etiqueta representa el aspecto del verbo, es decir, si el verbo es perfecto o inperfecto. Mediante esta etiqueta y el procesamiento que hemos hecho con el español pretendemos dar el mismo formato a las cadenas verbales de las dos lenguas. La misma frase como ejemplo de este nuevo método de segmentación.

eduki egitura egon , funts aldaketa bat egin BURU izan

Modo y tiempo verbal (MOD): Esta etiqueta representa el tiempo verbal. El analizador de euskera da 16 valores distintos para este rasgo. Al igual que con el aspecto verbal, con esta etiqueta pretendemos igualar el formato de las cadenas verbales de las dos lenguas. Ejemplo de la segmentacion que acabamos de presentar.

eduki egitura egon A1 , funts aldaketa bat egin izan A1

Combinacion de etiquetas: Creemos que cada etiqueta por separado mejorará en parte la calidad de la alineación pero utilizando todas las etiquetas a la vez lograremos

los mejores resultados. Para ello hemos creado dos corpus utilizando todas las etiquetas anteriormente mencionadas.

La diferencia de estos dos corpus reside en que en uno de ellos se utilizan las etiquetas caso y número **(TodoC)** y en el otro la etiqueta declinación **(TodoD)** que es la combinación de estas dos.

Ejemplo de la segmentacion que hemos llamado TodoC:

eduki M,S GEN egitura M,S DAT egon A1 KAUS , funts MG, ABS aldaketa MG, GEN bat MG, ABS egin BURU izan A1

Y la que hemos llamado TodoD:

eduki GEN,M,S egitura DAT,M,S egon A1 KAUS , funts ABS,MG, aldaketa GEN,MG, bat ABS,MG, egin BURU izan A1

4. Preprocesamiento del español

Al igual que hemos hecho con el euskera también hemos preprocesado el español intentando que la apariencia final del texto en las dos lenguas sea lo más parecido posible. Para analizar el español hemos utilizado el analizador de código abierto Freeling.

4.1. Tokenización

El primer paso es tokenizar el texto, ya que, como hemos explicado anteriormente, Giza++ tokeniza el texto usando simplemente los espacios en blanco. Mostramos como Ejemplo de texto tokenizado la traduccion de la frase utilizada como ejemplo para el euskera.

Por lo que se refiere a la estructura del contenido , se ha hecho algún retoque sustantivo

4.2. Lematización

Al igual que con el euskera no hemos unido los terminos multi-palabra en un único token, esperando que Giza++ los reconozca por su cuenta. Seguimos con el mismo ejemplo esta vez lematizado.

por el que él referir a el estructura del contenido , él haber hacer alguno retoque sustantivo

4.3. Segmentación

En el caso del español nos hemos limitado a procesar las cadenas verbales, ya que con el procesamiento del euskera hemos logrado igualar el formato de los dos textos para el resto de casos.

Para prepocesar las cadenas verbales en español hemos analizado éstas con el analizador Freeling y por cada token analizado como verbo hemos creado dos tokens, el lema de dicho token y una etiqueta que expresa el modo y el tiempo de dicho verbo. Hemos creado dos corpus distintos usando este procesamiento. En uno de ellos hemos dejado el resto de tokens con su forma original y en el otro el resto de tokens los hemos remplazado por sus respectivos lemas.

Ejemplo del procesamiento de las cadenas verbales, manteniendo la forma en el resto de casos.

Por lo que se referir IP a la estructura del contenido , se haber IP hacer P0 algún retoque sustantivo

Y ejemplo del procesamiento de las cadenas verbales, lematizando el resto de tokens.

por el que él referir IP a el estructura del contenido , él haber IP hacer P0 alguno retoque sustantivo

5. Diseño del experimento

Para poder evaluar cúal es la técnica de preproceso que da mejores resultados para este par de idiomas, hemos diseñado el siguiente marco experimental. Por un lado hemos utilizado un corpus paralelo concreto español-euskera de aproximadamente un millón de palabras y ochocientas mil palabras para cada idioma respectivamente (seccion 5.1). De este corpus, hemos separado 200 frases para su anotación manual, del que hemos utilizado 100 para desarrollo y 100 para la evaluación final (seccion 5.2). En la anotación manual, se alinea token a token, por lo que en la evaluación los corpus preprocesados alineados automáticamente tienen que ser mapeados a los tokens del corpus original (seccion 5.4). Las distintas estrategias de preproceso fueron evaluadas sobre la parte de desarrollo, ya que guardamos la parte de evaluacion para experimentos futuros.

5.1. Corpus paralelo

El corpus paralelo está extraído de una memoria de traducción donde las sentencias fueron alineadas a mano. Se trata de una coleccion de 6 libros universitarios con temática variada, desde fósiles, música, administración a historia. Contiene aprox. 36.000 frases, dando lugar a un millón de palabras en español y 800.000 palabras en euskera.

5.2. Alineamiento de referencia (gold standard)

La parte del corpus que fue manualmente anotada se tomó completamente al azar de los 6 libros. De las 200 sentencias se separaron, también al azar, 100 para desarrollo y 100 para evaluación.

Para la alineación manual se contó con un lingüista que tardo aproximadamente 24 horas en realizar la tarea. Dado que se alineaba token a token, el alineamineto de referencia incluye un número significativo de alineamientos múltiples (tanto 1:N como M:1 o M:N) y de alineamiento nulos (0:1 o 1:0). La herramienta para la anotación es de libre acceso.

Previamente al etiquetado, preparamos un manual de anotación inspirado en el de Melamed (Melamed, 1998), en el que fuimos incorporando los casos más dudosos para este par de idiomas. Por ejemplo, los pronombres átonos como en "**se** tomó un trago" se alineó con el verbo auxiliar en euskera "tragoa edan **zuen**".

5.3. Software de alineamiento

Hemos utilizado el software de libre distribucion GIZA++, con los parámetros por defecto, que utiliza el modelo de alineamiento IBM-4. Este software es el más utilizado para alinear corpus paralelos, paso imprescindible para la traducción automática estadística.

5.4. Evaluación y mapeo de las alineaciones automáticas

El alineamiento automático de las diferentes posibilidades de preproceso, no produce directamente el alineamiento de los tokens originales, sino que da como salida el alineamiento entre los tokens artificiales introducidos por el preproceso. Por tanto para cada token artificial generado en el preproceso, se guarda la posicion del token original, de forma que del alineamiento automático se puede reproducir el alineamiento para los tokens originales.

Por ejemplo, para la frase que hemos utilizado de ejemplo a lo largo del artículo, habiendo procesado el euskera del modo que hemos llamado TodoD (usando todas las etiquetas expuestas en el artículo pero uniendo el caso y el numero un una sola etiqueta) y el español Lema+V (lematizando todas las palabras y ademas añadiendo a los verbos la información de tiempo), el alineamiento que da GIZA++ es el que se puede ver en la figura 1. En cambio, tras el mapeo, el alineamiento que se consigue esta basado en los tokens originales y es el que se puede ver en la figura 2.

Figura 1: Alineamiento original creado por GIZA++ para el par Lema+V TodoD

Figura 2: Alineamiento tras el enlace para el par Lema+V TodoD

Una vez obtenido el alineamiento de los tokens originales, utilizamos el software libre utilizado en las dos últimas competiciones públicas de alineamiento (Martin, Mihalcea, y Pedersen, 2005) y (Mihalcea y Pedersen, 2003).

6. Análisis de los resultados

La tabla 1 presenta los resultados para una de las combinaciones (LEMA+V para español, y TodoD para el euskera). En las columnas tenemos distintas métricas de calidad segun las devuelve el software de evaluacion. Las tres primeras columnas devuelven la precisión, el cobertura y la combinación armónica de ambas. Un resultado más alto indica más calidad. Estas medidas están disponibles para alineamiento seguros o posibles, pero como en nuestro caso sólo hemos producido alineamientos seguros, las dos versiones producen resultados idénticos. En la última columna se muestra la *Alignment Error Rate*, que es la medida más utilizada y donde el valor inferior denota mejor calidad. En un principio deberíamos dar estos valores para cada combinación probada (ver más abajo), pero

	Precisión	Cobertura	Fmeasure	AER
NULL	50.46	56.62	53.36	46.64
NULL (weighted)	76.87	55.84	64.69	35.31
NO-NULL	73.30	73.79	73.54	26.46
NO-NULL (weighted)	86.64	67.35	75.78	24.22
Unión-NULL	42.06	66.13	51.42	48.58
Unión-NULL (weighted)	63.92	80.08	71.10	28.90
Unión-NO-NULL	63.18	64.42	63.79	36.21
Unión-NO-NULL (weighted)	77.29	75.21	76.24	23.76
Intersección-NULL	60.60	43.88	50.90	53.07
Intersección-NULL (weighted)	80.78	55.10	65.51	31.37
Intersección-NO-NULL	94.23	44.59	60.53	39.47
Intersección-NO-NULL (weighted)	96.20	53.60	68.84	31.16

Cuadro 1: Distintos métodos de evaluación para LEMA+V para el español y TodoD para el euskera

en el caso de nuestros resultados, el Fmeasure y la AER siempre han estado directamente relacionados, por lo que nos centraremos en la AER para ahorrar espacio.

Al evaluar los alineamientos, el software da dos posibilidades: una evaluación que toma en cuenta los alineamientos nulos (0:1 o 1:0, que en la tabla corresponden a las líneas con NULL), o la evaluación que descarta estos alineamientos (que corresponde a NO-NULL en la tabla). Además puede dar la misma importancia a todos los alineamientos o dar un peso inferior a los alineamientos múltiples (indicado por weighted en la tabla).

Dado que GIZA++ produce dos alineamientos, uno para cada dirección, las cuatro primeras líneas corresponden a la media de las métricas para cada dirección. Las siguientes líneas corresponden a dos estrategias para combinar las dos direcciones de alineamiento: en el caso de unión se toman todos los alineamientos indistintamente (con lo que se gana en cobertura, pero se pierde precisión), y en el caso de la intersección solamente se toman aquellos alineamientos que se encuentran en las dos direcciones (con lo que se gana precisión, pero se pierde cobertura).

Al igual que para las métricas de calidad, un análisis detallado de los resultados en las filas para todas las combinaciones probadas, demostró que las mejores técnicas se confirmaban en todas y cada una de las posibilidades de evaluación. Dadas las restricciones de espacio, en adelante sólo mostraremos los resultados para el AER de Union(weighted) para las dos variantes (NULL y NO-NULL), que son los que mejores resultados (menor AER) dan en todos los casos.

En la tabla 2 se muestran los resultados para las 40 posibilidades de combinar los preprocesos. En el caso de que no realizemos ningún tratamiento para el español (primera fila), la mejor técnica de preproceso del euskera es TodoC, que usa todos los rasgos morfológicos sin agruparlos, con lo que los artículos, preposiciones, etc. del castellano encuentran un token correspondiente en euskera. La mejora sobre no procesar nada (forma-forma) es de casi 10 puntos.

En el caso de la lematización del español (segunda fila), se pierde informacion para el castellano, pero el uso de lemas y el rasgo que agrupa el caso y el número es el que mejor se ajusta. Aquí se mejoran los resultados sobre no lematizar el castellano en 8 décimas.

En el caso de que a las formas del castellano les añadamos el rasgo del tiempo y concordancia verbales (tercera fila, Forma+V), la mejor codificación para el euskera vuelve a ser TodoC, mejorando los resultados sobre la forma únicamente en un punto y décima y media sobre la lematización.

Finalmente, en el caso de que el castellano incluya el lema y los rasgos verbales (última fila, Lema+V), los mejores resultados corresponden a TodoD, donde se usan todos los rasgos pero uniendo el caso y el número como en la estrategia DEC. Esta estrategia da los mejores resultados, reduciendo en medio punto la anterior, y haciendo que el error sobre forma-forma descienda en 13 puntos aprox.

Los resultados cuando se toman en cuenta los alineamientos nulos (tabla 3) son análogos, pero en este caso el error es superior, quedando en 28.9.

Nuestro trabajo se diferencia de sus an-

	Forma	Lema	ASP	CAS	DEC	SUB	MOD	NUM	TodoC	TodoD
Forma	36.24	30.21	30.40	26.87	26.72	30.16	30.21	27.71	**25.39**	26.63
Lema	32.74	29.07	28.85	24.90	**24.48**	28.76	28.69	27.37	26.32	25.00
Forma+V	33.80	29.33	29.18	26.15	25.54	29.18	28.41	26.39	**24.32**	25.05
Lema+V	32.17	28.77	28.85	24.92	23.96	28.61	28.12	26.67	25.00	**23.76**

Cuadro 2: AER para las diferentes combinaciones (Union-NO-NULL(weighted))

	Forma	Lema	ASP	CAS	DEC	SUB	MOD	NUM	TodoC	TodoD
Forma	41.28	33.42	33.45	31.57	31.67	33.11	33.22	32.46	**31.44**	31.98
Lema	37.40	31.52	31.44	29.19	**29.00**	31.09	30.97	31.80	31.33	29.75
Forma+V	39.58	33.13	32.96	30.96	30.63	32.75	32.36	31.47	**30.15**	30.26
Lema+V	37.53	32.17	32.02	29.58	29.12	31.74	31.51	31.97	30.22	**28.90**

Cuadro 3: Resultados de los distintos experimentos (Union-NULL(weighted))

tecesores en que es el primero que preprocesa el euskera para intentar mejorar el alineamiento. Como hemos explicado anteriormente en (Caseli, M. Nunes, y Forcada, 2005) no hacían ningún preprocesamiento del corpus e intentaban mejorar la alineacion mediante heurísticos independientes de la lengua. En dicho artículo usaban un corpus diferente para evaluar los resultados, por lo que los resultados no se pueden comparar directamente pero para tener una referencia conseguían mejorar el AER de un 35.52 que conseguia GIZA++ usando los parámetros por defecto, a un 26,52 que conseguían usando la técnica de alineamiento propuesta en el artículo.

Nuestros resultados son comparables a los que se han hecho públicos en las competiciones de alineamiento. Aunque una comparacion directa no sea justa, para tener una referencia queremos citar que nuestros resultados son mejores que los disponibles para los vencedores del rumano-inglés (26.55) y el inglés-hindú (32.12), pero peores que el inglés-francés (5.71) y el inglés-esquimal (9.46).

7. Conclusiones y trabajo futuro

Estos son los primeros pasos para lograr un alineamiento entre el español-euskera de una calidad suficiente para ser utilizada en tareas más complicadas, como la traducción estadística o la basada en ejemplos.

En este primer experimento hemos aumentado sensiblemente la calidad de la alineación, dividiendo las palabras en lemas y etiquetas morfológicas para harmonizar las secuencias de tokens en ambas lenguas y así facilitar que cada token tenga su correspondiente en el otro idioma. Usando GIZA++ conseguimos una reducción del error (Alignment Error Rate) de hasta un 12.5 % sobre el baseline forma-forma, llegando al 23.76 %. Este resultado es comparable al obtenido por otros autores en competiciones públicas para otros pares de idiomas aglutinantes como el euskera.

Para el futuro, hemos observado que el orden de los tokens es distinto en cada uno de los idiomas, y esto crea problemas a los alineadores actuales. Para lo cual pretendemos cambiar el orden de los tokens del euskera para asemejarlo al orden del español, y de este modo facilitar el trabajo del alineador. Creemos que así mejoraremos aún más la calidad de la alineación.

Una vez conseguido un alineamiento de una calidad aceptable, pretendemos, además de usarlo para desarrollar un traductor estadístico, usar el corpus alineado automáticamente como corpus de entrenamiento para aprender la traducción más adecuada de una palabra según el contexto.

Bibliografía

Aduriz, I., I. Alegria, J. Arriola, X. Artola, A. Díaz de Ilarraza, y N. Ezeiza. 1994. EUSLEM: un lematizador/etiquetador de textos en euskara. En *Actas del X Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)*, Córdoba, Spain.

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

Caseli, H., M. Nunes, y M. Forcada. 2005. Evaluating the LIHLA lexical aligner on Spanish, Brazilian Portuguese and Basque parallel texts. En *Proceedings of the XXI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)*, Granada, Spain, September.

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

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. En *MT Summit X*, September.

Martin, J., R. Mihalcea, y T. Pedersen. 2005. Word Alignment for Languages with Scarce Resources. En *Proceedings of the ACL 2005 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Ann Arbor, USA, June.

Melamed, D. 1998. Annotation Style Guide for the Blinker Project. Informe técnico, Institute for Research in Cognitive Science, Philadelphia, USA.

Mihalcea, R. y T. Pedersen. 2003. An Evaluation Exercise for Word Alignment. En *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Canada, May.

Nevado, F., F. Casacuberta, y J. Landa. 2004. Translation Memories Enrichment by Statistical Bilingual Segmentation. En *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*.

Nießen, S. y H. Ney. 2000. Morpho-syntactic analysis for reordering in statistical machine translation.

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

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

Comparing Rule-Based and Data-Driven Approaches to Spanish-to-Basque Machine Translation

Gorka Labaka[1], Nicolas Stroppa[2], Andy Way[2], Kepa Sarasola[1]

1- Informatika Fakultatea
University of the Basque Country
Donostia, Basque Country, Spain
{jiblaing ,kepa.sarasola}@ehu.es

2- National Centre for Language Technology
Dublin City University
Dublin 9, Ireland
{nstroppa,away}@computing.dcu.ie

Abstract

In this paper, we compare the rule-based and data-driven approaches in the context of Spanish-to-Basque Machine Translation. The rule-based system we consider has been developed specifically for Spanish-to-Basque machine translation, and is tuned to this language pair. On the contrary, the data-driven system we use is generic, and has not been specifically designed to deal with Basque. Spanish-to-Basque Machine Translation is a challenge for data-driven approaches for at least two reasons. First, there is lack of bilingual data on which a data-driven MT system can be trained. Second, Basque is a morphologically-rich agglutinative language and translating to Basque requires a huge generation of morphological information, a difficult task for a generic system not specifically tuned to Basque. We present the results of a series of experiments, obtained on two different corpora, one being "in-domain" and the other one "out-of-domain" with respect to the data-driven system. We show that *n*-gram based automatic evaluation and edit-distance-based human evaluation yield two different sets of results. According to BLEU, the data-driven system outperforms the rule-based system on the in-domain data, while according to the human evaluation, the rule-based approach achieves higher scores for both corpora.

1 - Introduction

Data-driven Machine Translation is nowadays the most prevalent approach carried out in Machine Translation (MT) research; translation results obtained with this approach have now reached a high level of accuracy, especially when the target language is English. Data-driven MT systems base their knowledge on bilingually aligned corpora, and the accuracy of their output depends strongly on the quality and the size of these corpora. Consequently, when pointing out the success of data-driven MT, we also need to make two additional remarks: (i) large and reliable bilingual corpora are unavailable for lots of language-pairs, (ii) translating into a morphologically rich target language makes the task of data-driven systems a lot more difficult.

When translating into Basque, we are confronted with both problems at the same time. First, few bilingual corpora are available which include Basque, which obviously limits to some extent the application of data-driven approaches. Second, Basque is a morphologically-rich agglutinative language that is difficult to translate into, in particular because of the morphological information we need to generate.

In this paper, we compare the rule-based and data-driven approaches in the context of Spanish-to-Basque translation. The rule-based system we consider has been developed specifically for Spanish-to-Basque MT, and is tuned to this language pair. On the contrary, the data-driven system we use is generic, and has not been specifically designed to deal with either of these languages. The generation of the Basque morphemes poses a particular problem for a system untuned to this language.

We present the results of a series of experiments, obtained on two different corpora, one being "in-domain" and the other one "out-of-domain" with respect to the data-driven system. We show that *n*-gram based automatic evaluation and edit-distance based human evaluation yield two different sets of results. According to BLEU, the data-driven system outperforms the rule-based system on the in-domain data, while according to the human evaluation, the rule-based approach achieves higher scores for both corpora.

The remainder of this paper is organized as follows. In Section 2, we introduce *Matxin*, a rule-based MT system designed for Spanish-to-Basque translation. In Section 3, we present *MaTrEx*, a data-driven MT system that we trained on a Spanish-to-Basque bilingual corpus extracted from magazines. In Section 4, we describe how to work at the morpheme level for Basque. In Section 5, we evaluate the two approaches mentioned above, and report and discuss our experimental results. Section 6 concludes the paper and gives avenues for future work.

2 - Matxin: a Rule-Based MT System

In this section, we describe *Matxin*, the main rule-based MT system developed at the University of the Basque Country. *Matxin* is an open source RBMT engine, whose first goal is to translate from Spanish into Basque, using the traditional transfer model. The transfer component of the translation system is based on both shallow and dependency parsing.[1]

[1] Note that *Matxin* is part of a more general project, *OpenTrad,* which implements two different translation approaches. The first one, named *Apertium* (Corbí-Bellot et al., 2005), is based on a shallow-transfer engine suited to machine translation between

Matxin is a classical transfer system consisting of three main components: (i) analysis of the source language into a dependency tree structure, (ii) transfer from the source language dependency tree to a target language dependency structure, and (iii) generation of the output translation from the target dependency structure. These three components are described in more detail in what follows.

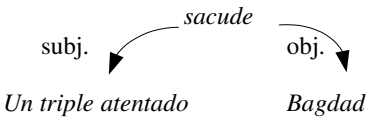
Analysis

The analysis of the Spanish source sentences into dependency trees is performed using an adapted version of the *FreeLing* toolkit (Carreras et al., 2004).[2] *FreeLing* contains a part-of-speech tagger and a shallow parser (or chunker) for Spanish. In *Freeling*, tagging and shallow parsing are performed using the Machine Learning AdaBoost models (Freund & Schapire, 1997). The shallow parses provided by *Freeling* are then augmented with dependency information, using a set of rules that identify the dependencies in the sentence. First, the relationships between chunks is established, based on their labels. As an example, consider the chunked Spanish sentence in (1):

(1) *[np] Un triple atentado* ||| *[verb-chain] sacude* ||| *[np] Bagdad* (a three-pronged attack rocked Baghdad)

Here the dependency parser identifies the verb-chain as the head of the sentence, and the two noun phrases as its children. Then, the dependencies are labelled using a second set of rules. In the previous example "*Un triple atentado*" and "*Bagdad*" are recognised to be the subject and the object respectively of the main verb "sacude". The analysis of this sentence is displayed in (2):

(2)



Transfer

The transfer component consists of lexical transfer and structural transfer.

Lexical transfer is performed using a Spanish-to-Basque dictionary compiled into a finite-state transducer. This dictionary is based on the wide-coverage dictionary *Elhuyar*.[3] This dictionary was enriched with named entities and terms automatically extracted from parallel corpora. This extraction was performed using the Consumer and EITB corpora (see Section 5 for a detailed description of these corpora). Moreover, some Spanish words (such as articles, conjunctions, etc.) do not translate into Basque words, and are translated as morphemes that will be concatenated to other words.

Note that in the actual version of the engine no word-sense disambiguation is performed (we plan to solve semantic ambiguities within a concrete domain in the near future), but a large number of multi-word units representing collocations, named entities and complex terms are included in the bilingual dictionary in order to reduce the influence of this limitation. In the case of prepositions, we adopt another strategy: we decide on the proper translation using some information about verb argument structure extracted automatically from the corpus.

Structural transfer is applied to turn the source dependency tree structure into the target dependency structure. This transformation follows a set of rules that will copy, remove, add, or reorder the nodes in the tree. In addition, specialized modules are included to translate verb chains (Alegria et al., 2005).

Generation

Generation, like transfer, is decomposed into two steps. The first step, referred to as syntactic generation, consists of deciding in which order to generate the target constituents within the sentence, and the order of the words within the constituents. The second step, referred to as morphological generation, consists of generating the target surface forms from the lemmas and their associated morphological information.

In order to determine the order of the constituents in the sentence, a set of rules is defined that state the relative order between a node in the dependency tree and its ancestors. For example, a prepositional phrase is generated before its ancestors if the latter is a noun phrase. The order of the words within the chunks is solely based on the Part-of-Speech information associated with the words.

In Basque, the declension case, number case and other features are assigned to a whole NP as a suffix of the last word of the phrase. Consequently, when generating Basque, the main inflection of a noun phrase is added to its last word. In the case of a verb chain phrase, morphological generation needs to be applied to every word in the phrase.

In order to perform morphological generation, we use the morphological generator for Basque described in (Alegria et al., 1996). This generator makes use of the morphological dictionary developed in *Apertium*, which establishes correspondences between surface forms and lexical forms for Basque. It is used in morphological generation to produce the inflected forms of Basque words. In particular, this dictionary contains:

languages showing syntactic similarities (up to now, Spanish, Catalan and Galician are handled); it can be freely downloaded from http://apertium.sourceforge.net. The second one is *Matxin*, based on a deep-transfer engine, and is focused on the Spanish-Basque language pair; it is a continuation of previous work in the IXA group (Diaz de Ilarraza et al., 2000). *Matxin* can be freely downloaded from http://matxin.sourceforge.net.

[2] Freeling can be freely downloaded from http://www.lsi.upc.edu/~nlp/freeling/.

[3] http://www1.euskadi.net/hizt_el.

- A definition of Basque paradigms (sets of correspondences between partial surface forms and partial lexical forms). Those paradigms are similar to continuation classes in two-level morphology (Koskeniemmi, 1983).
- Lists of surface form to lexical form correspondences for complex lexical units (including multi-word units).

This dictionary is compiled into a finite-state transducer which is used to perform the morphological generation of Basque words. A more detailed description of this process can be found in (Armentano-Oller et al., 2005).

3 - MaTrEx: a Data-Driven System

The *MaTrEx* system (Stroppa & Way, 2006) used in our experiments is a modular data-driven MT engine, which consists of a number of extendible and re-implementable modules, the most important of which are:

- Word Alignment Module: takes as its input an aligned corpus and outputs a set of word alignments.
- Chunking Module: takes in an aligned corpus and produces source and target chunks.
- Chunk Alignment Module: takes the source and target chunks and aligns them on a sentence-by-sentence level.
- Decoder: searches for a translation using the original aligned corpus and derived chunk and word alignments.

The Word Alignment and the Decoder modules are wrappers around existing tools, namely Giza++ (Och & Ney, 2003), and *Moses* (Koehn et al., 2007). The chunking and alignment strategies are described in more detail below.

The translation process can be decomposed as follows: the aligned source-target sentences are passed in turn to the word alignment, chunking and chunk alignment modules, in order to create our chunk and lexical example databases. These databases are then given to the decoder to translate new sentences. These steps are displayed in Figure 1.

Chunking

In the case of Spanish, the extraction of chunks relies on the shallow parser described above (as part of *Freeling*). This shallow parser enables us to identify the main constituents in the sentence: noun phrases, verb phrases, prepositional phrases, etc.

In the case of Basque, we use the toolkit *Eusmg,* which performs POS tagging, lemmatisation and chunking (Adu riz & Díaz de Ilarraza, 2003). It recognizes syntactic structures by means of features assigned to word units, following the constraint grammar formalism (Karlsson, 1995). An example of chunked sentences is given in (3), for Spanish and Basque:

(3)

Spanish:
Un triple atentado sacude Bagdad:
=> [np] Un triple atentado ||| [verb-chain] sacude ||| [np] Bagdad

Basque:
atentatu hirukoitz batek Bagdad astintzen du
=> [np] atentatu hirukoitz batek ||| [np] Bagdad ||| [verb-chain] astintzen du

Note that, since each module of the system can be changed independently of the others, it is possible to use a variety of chunkers, including those of the Marker-based approach, used in other works (Gough & Way, 2004; Stroppa et al., 2006; Stroppa & Way, 2006).

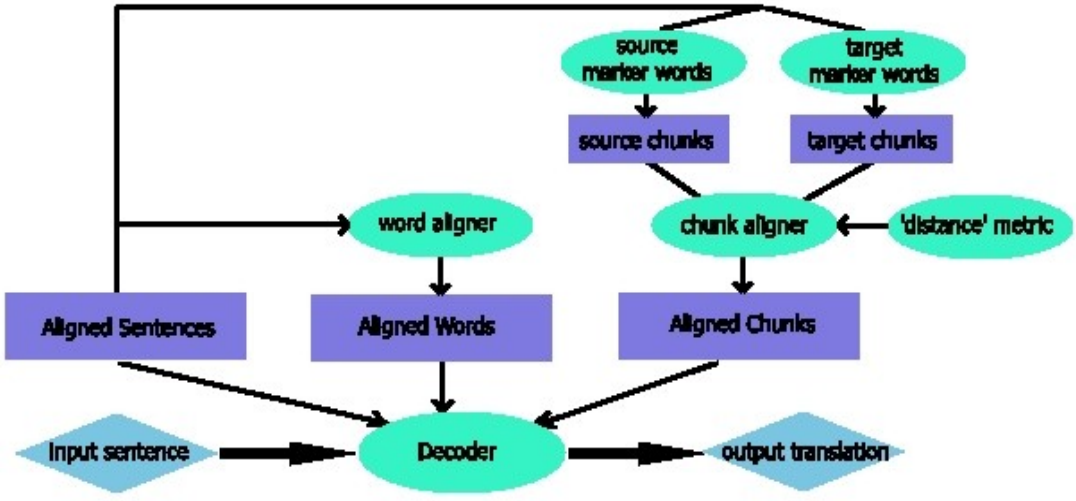


Figure 1: Translation Process in *MaTrEx*

Alignment Strategies

Word alignment
Word alignment is performed using the *Giza++* statistical word alignment toolkit and we followed the "refined" method of (Koehn et al., 2003) to extract a set of high-quality word alignments from the original uni-directional alignment sets. These along with the extracted chunk alignments were passed to the translation decoder.

Chunk alignment
In order to align the chunks obtained by the chunking procedures introduced in Section "Chunking", we make use of an "edit-distance style" dynamic programming alignment algorithm, as described in (Stroppa et al., 2006).

This algorithm works as follows. First, a "similarity" measure is determined for each pair of source-target chunks. Then, given these similarities, we use a modified version of the edit-distance alignment algorithm to find the optimal alignment between the source and the target chunks. The modification consists of allowing for jumps in the alignment process (Leusch et al., 2006), which is a desirable property for translating between languages showing significant syntactic differences. This is the case for Spanish and Basque, where the order of the constituents in a sentence can be very different.

To compute the "similarity" between pair of chunks, we rely on the information contained within the chunks. More precisely, we relate chunks by using the word-to-word probabilities that were extracted from the word alignment module. The relationship between a source chunk and a target chunk is computed thanks to a model similar to IBM model 1 (Stroppa et al., 2006).

Integrating SMT data
Since its inception, EBMT has recommended the use of both lexical and phrasal information (Nagao, 1984); current SMT models now also use phrases in their translation models (Koehn et al., 2003). Actually, it is possible to combine elements from EBMT and SMT to create hybrid data-driven systems capable of outperforming the baseline systems from which they are derived, as shown in (Groves and Way, 2005). Therefore, we also make use of SMT phrasal alignments, which are added to the aligned chunks extracted by the chunk alignment module. The SMT phrasal alignment follows the procedure of (Koehn et al., 2003).

Decoder
The decoding module is capable of retrieving already translated sentences and also provides a wrapper around *Moses*, a phrase-based decoder. This decoder also implements Minimum-Error-Rate Training (Och, 2003) within a log-linear framework (Och & Ney, 2002). The BLEU metric (Papineni et al., 2002) is optimized on a development set. We use a log-linear 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 Basque language model is a simple 3-gram language model trained on the Basque portion of the training data, using the SRI Language Modeling Toolkit,[4] with modified Kneser-Ney smoothing.

4 - Morpheme-Based Machine Translation

Basque is an agglutinative language in which words may be made up of a large number of morphemes. For example, suffixes can be added to the last word of a noun phrase; these suffixes can represent some morpho-syntactic information associated to the noun phrase, such as number, definiteness, grammatical cases and postpositions.

As a consequence, most words only occur once in the training data, leading to serious sparseness problems when extracting statistics from the data. In order to limit this problem, one solution is to working at a different representation level, namely morphemes (cf. (Stroppa et al., 2006)). By segmenting each word into a sequence of morphemes, we reduce the number of tokens that occur only once (cf. (Agirre et al., 2006)). Furthermore, as many Basque words correspond to several Spanish words (for example, the Basque "etxeko" translates to "de la casa" in Spanish), lots of 1-to-n alignments have to be defined when working at the word level. Although 1-to-n alignments are allowed in IBM model 4, training can be harmed when the parallel corpus contains many such cases.

Working at the morpheme level within *MaTrEx* is straightforward: we only need to segment the Basque side of the training (and development) data. The *MaTrEx* system trained on these new data will generate a sequence of morphemes as output.

In the experiments we carried out, we report results obtained when working at both the word and morpheme levels.

From Words to Morphemes

Working at the morpheme level does, however, have some drawbacks. In particular, if we want to be able to generate surface word forms from morphemes, then we need to include some additional information to the morphemes. In (Agirre et al., 2006), a segmentation strategy is proposed, which does not include this information. In this paper, we build upon this strategy,

[4] http://www.speech.sri.com/projects/srilm/

but we also include the required information to recover the surface words from the morphemes.

To obtain the segmented text, the Basque text is analyzed using Eustagger (Aduriz & Díaz de Ilarraza, 2003), a two-level morphology (Koskeniemmi, 1983) analyser/tagger. After this process, each word is replaced with the corresponding lemma accompanied with a list of morphological features. A sentence and the associated segmentation are displayed in (4), where each morpheme is accompanied by the appropriate morphological information:

(4)

Original Basque sentence:
Loe berriak indarrean eusten dio lege horri .

Segmented sentence:
Loe<IZE><IZB> berri<ADJ><ARR>
+<P>+<ABS> indarrean<ADB><ARR>
eutsi<ADI><SIN>+<ADOIN>+<EZBU>
edun +<A1>
+<NR_HURA>+<NI_HARI>+<NK_HARK>
lege<IZE><ARR> hori<DET><ERKARR>
+<S>+<DAT> .

From Morphemes to Words

When working at the morpheme level, the translation of a (source) sentence obtained using *MaTrEx* is a sequence of morphemes. If we want to produce a Basque text, then we need to recover the words from this sequence of morphemes; the output of *MaTrEx* is thus post-processed to produce the final Basque translation.

This post-processing consists of using the morphological generation module of *Matxin*. This module uses the same lexicon and two level rules as *Eustagger*. However, in the context of generation, we are faced with two new additional problems:

- Unknown lemmas: some lemmas do not occur in the *Eustagger* lexicon, such as unknown proper names. To solve this problem, the synthesis component has been enriched to generate words from unknown lemmas using default rules defined for each part of speech.

- Invalid sequences of tags: the output of *MaTrEx* (a sequence of morphemes) is not necessarily a well-formed sequence from a morphological point of view. For example, the correct tags might be generated, but in the wrong order. In some cases, a nominal tag is assigned to verb; sometimes, required tags are missing. In the current work, we do not try to correct these mistakes: we simply output the lemma, and remove the inappropriate tag information. A more refined treatment is left to future work.

5 - Experimental Results

Data and Evaluation

The experiments were carried out using two different test sets.

The first, referred to as *ConsumerTest*, contains 1500 bilingually aligned sentences extracted from the *Consumer* Eroski Parallel Corpus.[5] The *Consumer* Eroski Parallel 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 one million Spanish words for Spanish and more than 800,000 Basque words. This corpus is aligned at the sentence level.

The second, referred to as *EitbTest*, also contains 1500 bilingually aligned sentences extracted from the EITB corpus. This corpus is a collection of news (Basque News and Information Channel, *http://www.eitb24.com/en*), available in Spanish, Basque, and English.[6] This corpus contains approximately 1,500,000 Spanish words and 1,200,000 Basque words.

While *Eitb* is a *general* news corpus (politics, economy, sport, etc.), *Consumer* is a corpus of articles comparing the quality and prices of commercial products and brands. They are consequently from two different terminological "domains". Table 1 summarizes the various statistics related to these corpora.

Since the *Matxin* system is rule-based, it does not need any kind of training, and can be directly applied to translate into Basque the Spanish test sentences. However, Matxin's bilingual lexicon was enriched with 1129 entries (entities and multi-word terms) that were automatically extracted form the *ConsumerTrain* bilingual corpora.

In order to train the *MaTrEx* system, which is data-driven and relies on bilingually aligned training material, we used approximately 50,000 aligned sentences from the *ConsumerTrain* dataset, which was extracted in a similar manner to the *Consumer* dataset. In order to tune the parameters of the *MaTrEx* system, we use an additional development set of 1292 sentence pairs (referred to as *ConsumerDev*). Training *MaTrEx* on *ConsumerTrain* makes the *ConsumerTest* dataset "in-domain", and the *Eitb* dataset "out-of-domain". We thus expect the *MaTrEx* system to perform better on the *ConsumerTest* set than on the *EitbTest* set.

[5] 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).

[6] EITB is the official media group in the Basque Country with four television channels and five radio stations.

	Spanish	Basque
ConsumerTrain		
Sentences	51949	
Running words	976730	786705
Running morphemes	-	910995
Word voc. Size	44715	76292
Morph. Voc. Size	-	29805
ConsumerDev		
Sentences	1292	
Running words	24755	19978
Running morphemes	-	22554
Word voc. Size	5973	7367
Morph. Voc. Size	-	4064
ConsumerTest		
Sentences	1501	
Running words	34231	27278
Running morphemes	-	45480
Word voc. Size	7278	9258
Morph. Voc. Size	-	5999
EitbTest		
Sentences	1500	
Running words	36783	26857
Running morphemes	-	41602
Word voc. Size	7345	7918
Morph. Voc. Size	-	5706

Table 1: Corpus statistics.

In order to assess the quality of the translation obtained using both systems, we used automatic evaluation metrics as well as human evaluation. As for automatic evaluation, we report the following accuracy measures: BLEU (Papineni et al., 2002), and NIST (Doddington, 2002). For each testset, we have access to one Basque reference translation per sentence. Evaluation is performed in a case-insensitive manner. Because of the specific nature of Basque, we perform two types of evaluation: a word-based evaluation, and a morpheme-based evaluation.

Since human evaluation is an expensive process, we selected 50 sentences from the *ConsumerTest* corpus to be human evaluated; this corpus is referred to as *ConsumerTestHuman*. The same applies to *EitbTest*, yielding *EitbTestHuman*. We used the edit-distance metric (Przybocki et al., 2006) called HTER or Translation Error Rate with human-targeted references (Snover et al., 2006). Edit distance is defined as the number of modifications a native Basque professional translator has to make so that the resulting edited translation is an easily understandable Basque sentence that contains the complete meaning of the source sentence. We used the software described in (Snover et al., 2006) to compute HTER. The post-editing work took 6 hours in total.

Automatic Evaluation Results

For the *ConsumerTest* corpus, the results obtained with the *MaTrEx* system are higher than those obtained with the *Matxin* system. With respect to the BLEU score, this difference is 1.58 points absolute for the word-based evaluation (27% relative increase), and 2.47 points absolute for the morpheme-based evaluation (21% relative increase). These differences are statistically significant, with a p-value < 0.002, computed using approximate randomisation (Riezler & Maxwell, 2005).

For the *EitbTest* corpus, the results obtained with the *MaTrEx* system are much lower than those obtained with the *Matxin* system. The differences are also statistically significant, with a p-value < 0.002, for both BLEU and NIST scores. This is consistent with our intuition since with respect to *MaTrEx*, the *EitbTest* corpus is "out-of-domain" (cf. (Koehn & Monz, 2006) for a comparison between in-domain and out-of-domain results of data-driven systems).

These results show that a (generic) data-driven system can be very competitive with a (specialized) rule-based system, if suitable training data is available. The argument in favour of rule-based systems is stronger when no relevant bilingual training data are available.

Given the globally low scores obtained, it is important to make two additional remarks. First, it shows the difficulty of the task of translating to Basque, which is due to the strong syntactic differences with Spanish, and the morphological properties of this language. Second, even if a morpheme-based translation is more appropriate than a word-based translation, *n*-gram based metrics are not suited to the comparison between sequences of morphemes. In particular, the absence of morphological tags that may not affect the global understanding of a sentence are penalised: if such a tag is missing in the system's output, all the *n*-grams that could have contained it would be cut.

	ConsumerTest		EitbTest	
	BLEU	NIST	BLEU	NIST
Matxin-WB	6.31	3.66	9.30	3.13
MaTrEx-WB	8.03	3.69	9.02	2.70
Matxin-MB	12.01	4.62	12.76	3.75
MaTrEx-MB	14.48	4.63	6.25	2.89

Table 2: Automatic evaluation results.

The results obtained for the Spanish-to-Basque translation task using the *ConsumerTest* and *EitbTest* datasets are summarized in Table 2, in which WB and MB denote respectively the word-based evaluation and the morpheme-based evaluation. For the morpheme-based evaluation, we segment the reference sentences into morphemes with which we compare the output of each system (which is also a sequence of morphemes).

Human Evaluation Results

The human evaluation results, obtained using HTER, are reported in Table 3. We conducted a word-based evaluation (WB), as well as a morpheme-based

evaluation (MB). For the morpheme-based evaluation, both the reference and the translated text are divided into morphemes.

	ConsumerTest Human	**EitbTest Human**
	HTER	HTER
Matxin-WB	43.6	40.4
MaTrEx-WB	57.9	71.8
Matxin-MB	39.1	34.9
MaTrEx-MB	49.6	76.3

Table 3: Subjective evaluation results.

For the *ConsumerTestHuman* corpus, we can observe that the error rate obtained by *Matxin* is lower than the one obtained by *MaTrEx*: 14.3 points for the word-based evaluation and 10.5 points for the morpheme-based evaluation.

Concerning the *EitbTestHuman* corpus, i.e. the "out-of-domain" corpus, the difference is even higher. While *Matxin*'s error-rate is quite similar to the one obtained with the *Consumer* corpus (40.4 points), the error-rate for *MaTrEx* becomes quite large (71.8 points).

These results are consistent with the domain independence of the rule-based system, which achieves a comparable translation quality for the two corpora. The data-driven approach is domain-dependent by construction and, as expected, it performs better on the in-domain corpus. According to the subjective evaluation, the translation quality of *Matxin* is better, irrespective of the corpus. However, it must be stressed that *Matxin* has been specifically developed and designed to translate from Spanish to Basque over a number of years, while *MaTrEx* is generic and the cost of adapting it to Spanish-Basque translation is several orders of magnitude lower.

6 - Conclusions and Future Work

In this paper, we have compared a rule-based MT system (*Matxin*) and a data-driven MT system (*MaTrEx*) in the context of Spanish-to-Basque translation. While the rule-based system we consider has been developed specifically for Spanish-to-Basque machine translation, the data-driven system we use is generic, and has not been specifically tuned to Basque.

We have introduced a translation scheme based on morphemes instead of words, in order to be able to deal with the particular agglutinative nature of Basque. This allows for the generation of the morphological information required to recover the full Basque surface word forms.

We have presented experimental results comparing the two types of approaches on two different corpora containing magazine and news articles respectively. Objective evaluation metrics such as BLEU and NIST yield different results to subjective evaluation metrics such as HTER. The automatic metrics indicate that the data-driven system outperforms the rule-based system on the in-domain data. On the contrary, the subjective evaluation indicates that the rule-based system outperforms the data-driven approach for both corpora. Note that these results are also consistent with the findings of (Callison-Burch et al., 2006) concerning objective and subjective evaluation.

Moreover, both types of evaluation confirm that *Matxin*, the rule-based system, is domain-independent while *MaTrEx*, the data-driven system, is more domain -dependent. Accordingly, if a different domain were selected which was quite different from the magazine or news articles used here (weather forecasts, say), then we would expect *MaTrEx* to win out. That said, having invested a large number of person-years in its development, it is encouraging to see the good performance of *Matxin* on out-of-domain data.

Future work consists of building upon the respective strength of both approaches, by exploring various hybridity strategies focused on the problem of Basque translation. One avenue that we would expect to bear fruit is adding into *MaTrEx* the bilingual lexicon from *Matxin.* We also plan to use automatic evaluation metrics that would be more suited to the evaluation of morpheme-based translation (cf. (Owczarzak et al., 2006)).

Acknowledgments

This work is partially supported by Science Foundation Ireland (grant number OS/IN/1732), Spanish M.E.C. (OpenMT project, TIN2006-15307-C03-01), and the Basque Government (AnHitz project, eIE06-185). Our colleagues Iñaki Alegria, Arantza Díaz de Ilarraza, Mikel Lersundi, and Aingeru Mayor are kindly acknowledged for providing their expertise on the *Matxin* system and the evaluation of the output.

References

I. Aduriz and A. Díaz de Ilarraza (2003). Morphosyntactic disambiguation and shallow parsing in Computational Processing of Basque. In *Inquiries into the lexicon-syntax relations in Basque*, B. Oiharcabal (ed.), Univ. of the Basque Country, Donostia, Spain.

I. Alegria, X. Artola Zubillaga, and K. Sarasola. Automatic morphological analysis of Basque (1996). *Literary & Linguistic Computing* **11**(4):193—203.

I. Alegria, A. Díaz de Ilarraza, G. Labaka, M. Lersundi, A. Mayor, and K. Sarasola (2005). An FST grammar for verb chain transfer in a Spanish-Basque MT System. In *Proceedings of Finite-State Methods and Natural Language Processing,* pp.295—96, Helsinki, Finland.

E. Agirre, A. Díaz de Ilarraza, G. Labaka, and K. Sarasola (2006). Uso de información morfológica en el alineamiento. *Español-Euskara XXII Congreso de la SEPLN*, Zaragoza, Spain.

C. Armentano-Oller, A. Corbí-Bellot, M. L. Forcada, M. Ginestí-Rosell, B. Bonev, S. Ortiz-Rojas, J. A. Pérez-Ortiz, G. Ramírez-Sánchez, and F. Sánchez-Martínez (2005). An open-source shallow-transfer Machine Translation toolbox: consequences of its release and availability. In *Proceedings of Open-Source MT workshop, MT Summit X*, Phuket, Thailand.

X. Carreras, I. Chao, L. Padró and M. Padró (2004). FreeLing: An Open-Source Suite of Language Analyzers. In *Proceedings of 4th LREC*, Lisbon, Portugal.

C. Callison-Burch, M. Osborne and P. Koehn (2006). Re-evaluating the Role of Bleu in MT Research. In P*roceedings of EACL 2006*, pp.249—256, Trento, Italy.

A. Corbí-Bellot, M. Forcada, S. Ortiz-Rojas, J. Pérez-Ortiz, G. Ramírez-Sánchez, F. Sánchez-Martínez, I. Alegria, A. Mayor and K. Sarasola (2005). An Open-Source Shallow-Transfer Machine Translation Engine for the Romance Languages of Spain. In *Proceedings of 10th EAMT Conference: Practical Applications of Machine Translation*, Budapest, Hungary, pp.79—86.

G. Doddington (2002). Automatic evaluation of Machine Translation quality using n-gram cooccurrence statistics. In *Proceedings of HLT 2002*, pp. 128—132, San Diego, CA.

N. Gough and A. Way (2004). Robust large-scale EBMT with marker-based segmentation. In *Proceedings of TMI 2004*, pp.95—104, Baltimore, MD.

D. Groves and A. Way (2005). Hybrid data-driven models of MT. *Machine Translation* **19**(3,4):301—323.

F. Karlsson, A. Voutilainen, J. Heikkila, and A. Anttila, editors (1995). *Constraint Grammar: A Language-independent System for Parsing Unrestricted Tex*t. Mouton de Gruyter, Berlin, New-York.

Y. Freund and R. Schapire (1997). A Decision-Theoretic Generalization of On-line Learning and an Application to Boosting. *Journal of Computer and System Sciences* **55**(1):119—139.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst (2007). Moses: Open source toolkit for SMT, in *Proceedings of the ACL 2007 Demo and Poster Session*, Prague, Czech Republic, pp.177—180.

P. Koehn and C. Monz (2006). Manual and Automatic Evaluation of MT. In *Proceedings of HLT-NAACL Workshop on SMT*, pp.102—121, New York.

P. Koehn, F. Och, and D. Marcu (2003). Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003*, pp. 48-54, Edmonton, Canada.

K. Koskenniemi (1983). Two-level Model for Morphological Analysis. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, pp.683—685, Karlsruhe, Germany.

G. Leusch, N. Ueffing, and H. Ney (2006). CDER: Efficient MT evaluation using block movements. In *Proceedings of EACL 2006*, pp.241—248, Trento, Italy.

M. Nagao (1984). Framework of a mechanical translation between Japanese and English by analogy principle. In *Artificial and Human Intelligence*, A. Elithorn and R. Banerji, Eds. Amsterdam, The Netherlands: North-Holland, pp.173—180.

F. Och, (2003). Minimum error rate training in statistical machine translation. In *Proceedings of 41st ACL*, pp. 160—167, Sapporo, Japan.

F. J. Och and H. Ney (2002). Discriminative training and maximum entropy models for SMT. In *Proceedings of 40th ACL*, pp. 295—302, Philadelphia, PA.

K. Owczarzak, D. Groves, J. Van Genabith and A. Way (2006). Contextual Bitext-Derived Paraphrases in Automatic MT Evaluation. In *Proceedings of HLT-NAACL Workshop on SMT*, pp.86—93, New York.

K. Papineni, S. Roukos, T. Ward, and W. Zhu (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of 40th ACL*, pp. 311—318, Philadelphia, PA.

M. Przybocki, G. Sanders, and A. Le (2006). Edit distance: a metric for MT evaluation. In *Proceedings of 5th LREC*, pp. 2038—2043, Genoa, Italy.

S. Riezler and J. Maxwell (2005). On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the the ACL Workshop on Intrisic and Extrinsic Evaluation Measures for MT and/or Summarization*, pp. 57—64, Ann Arbor, MI.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA 2006*, pp.223—231, Cambridge, MA.

N. Stroppa, D. Groves, A. Way, and K. Sarasola (2006). Example-based Machine Translation of the Basque Language. In *Proceedings of AMTA 2006*, pp. 232—241, Cambridge, MA.

Stroppa, N. and A. Way. MaTrEx: DCU Machine Translation System for IWSLT 2006. In *Proceedings of IWSLT 2006*, pp.31—36, Kyoto, Japan.

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

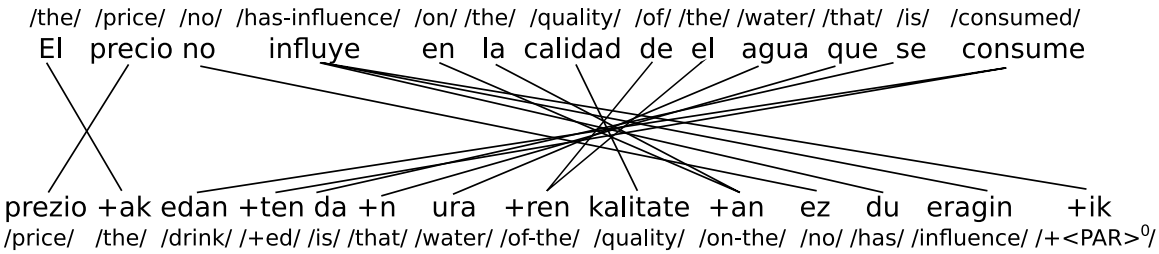


Figure 1: Example of word alignment. */the price does not affect the quality of the drinking water/*

relevant research on the area. Later, we describe the used reordering techniques (Section 3) and the SMT systems developed for this paper (Sections 4). 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. The most commonly used approach is the preprocessing of the source sentence 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

[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

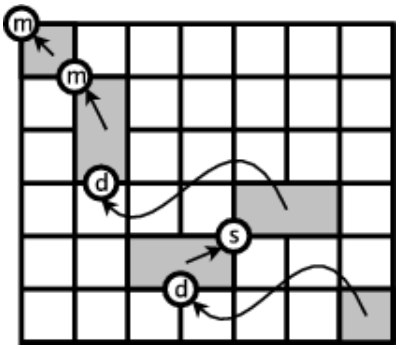


Figure 2: Possible orientation of phrases defined on the lexicalized reordering: monotone (m), swap (s), or discontinuous (d)

to be added to the long-linear framework. Three different orientations are defined (See Figure 2):

- 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.

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

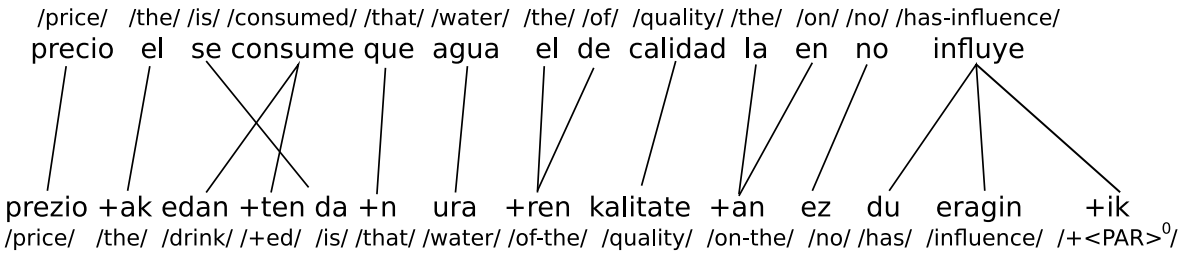


Figure 3: Example of word alignment after Syntax-Based reordering. */the price does not affect the quality of the drinking water/.*

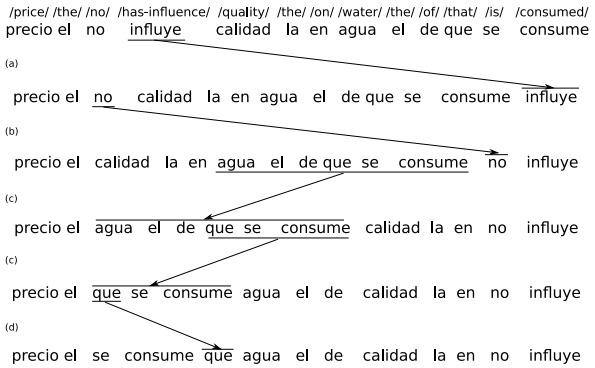


Figure 2: Example of long-range reordering.
/the price does not affect the quality of the drinking water/

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 being 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.

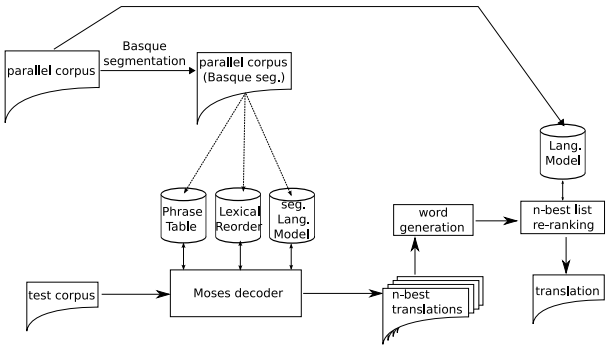


Figure 4: Design of the segmentation-based SMT system

We have trained nine different systems combining three possible source text preprocessing (tokenization, Syntax-Based reordering and statistical reordering) and three reordering configurations at decoding (monotone, distance-based and lexicalized reordering).

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

		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.

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. 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. 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 improvement 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

[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
Tokenized source (without reordering)	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 reordering	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 reordering	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.

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

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.

Spanish-to-Basque MultiEngine Machine Translation for a Restricted Domain

Iñaki Alegria (1), Arantza Casillas (2), Arantza Diaz de Ilarraza (1), Jon Igartua (1), Gorka Labaka (1), Mikel Lersundi (1), Aingeru Mayor (1), Kepa Sarasola (1)
(1) Informatika Fakultatea. IXA group. (2) Zientzia eta Teknologia Fakultatea.
University of the Basque Country. UPV-EHU
i.alegria@ehu.es

Abstract

We present our initial strategy for Spanish-to-Basque MultiEngine Machine Translation, a language pair with very different structure and word order and with no huge parallel corpus available. This hybrid proposal is based on the combination of three different MT paradigms: Example-Based MT, Statistical MT and Rule-Based MT. We have evaluated the system, reporting automatic evaluation metrics for a corpus in a test domain. The first results obtained are encouraging.

1 Introduction

Machine translation for Basque is both a real need and a testing ground for our strategy to develop language tools. The first development was Matxin, a Rule-Based MT system (Mayor, 2007). Later on a Data-Driven Machine Translation system was built and both systems compared (Labaka et al., 2007). As both approaches have their limits, and each deals with a different kind of knowledge, it was decided to try combining them to improve their results. On the one hand, after improvements in 2007 (Labaka et al., 2007) the Spanish-to-Basque RBMT system Matxin proved useful for assimilation, but is still not suitable for unrestricted use in text dissemination. On the other hand, data-driven MT systems base their knowledge on aligned bilingual corpora, and the accuracy of their output depends heavily on the quality and the size of these corpora. When the pair of languages used in translation, such as Spanish and Basque, has very different structures and word orders, the corpus obviously needs to be bigger. However, since Basque is a lesser-used language, large and reliable bilingual corpora are unavailable. At present, domain-specific translation memories for Basque are no bigger than two or three million words, much smaller than corpora used for other languages; for example, Europarl corpus (Koehn, 2005), a standard resource, has 30 million words. So, although domain-restricted corpus-based MT for Basque shows promising results, it is still not ready for general use.

Therefore, it is clear that we should combine the basic techniques for MT (rule-based and corpus-based) in order to build a hybrid system with better performance. Due to the pressing need for translation in public administration and taking into account that huge parallel corpora for Basque are not available, we have tested a first strategy by building a MT engine for a restricted domain related to public administration for which translation memories were available.

The rest of the paper is organized as follows. Section 2 presents some related work. Section 3 describes the corpus we have compiled to carry out the experiments. Section 4 explains the single engines built up for Basque MT and how we have combined them. Section 5 reports our experiments. Finally, we draw conclusions and refer to future work.

2 Related Work

(van Zaanen and Somers, 2005), (Matusov et al., 2006) and (Macherey and Och, 2007) review a set of references about MEMT (Multi-Engine MT) including the first attempt by (Frederking and Nirenburg, 1994). All the papers on MEMT reach the same

conclusion: combining the outputs results in a better translation. Most of the approaches generate a new consensus translation combining different SMT systems using different language models and in some cases combining also with RBMT systems. Some of the approaches require confidence scores for each of the outputs. The improvement in translation quality is always lower than 18% relative increasing in BLEU score.

(Chen et al., 2007) reports 18% relative increment for in-domain evaluation and 8% for out-domain, by incorporating phrases (extracted from alignments from one or more RBMT systems with the source texts) into the phrase table of the SMT system and use the open-source decoder Moses to find good combinations of phrases from SMT training data with the phrases derived from RBMT.

(Matusov et al., 2006) reports 15% relative increment in BLEU score using consensus translation computed by voting on a confusion network. Pairwise word alignments of the original translation hypotheses were estimated for an enhanced statistical alignment model in order to explicitly capture reordering.

(Macherey and Och, 2007) presented an empirical study on how different selections of translation outputs affect translation quality in system combination. Composite translations were computed using (i) a candidate selection method based on inter-system BLEU score matrices, (ii) a ROVER-like combination scheme, and (iii) a novel two-pass search algorithm which determines and re-orders bags of words that build the constituents of the final consensus hypothesis. All methods gave statistically significant relative improvements of up to 10% BLEU score. They combine large numbers of different research systems.

(Mellebeek et al., 2006) reports improvements of up to 9% BLEU score. Their experiment is based in the recursive decomposition of the input sentence into smaller chunks, and a selection procedure based on majority voting that finds the best translation hypothesis for each input chunk using a language model score and a confidence score assigned to each MT engine.

(Huang and Papineni, 2007) and (Rosti et al., 2007) combines multiple MT systems output at word-, phrase- and sentence-levels. They report improvements of up to 10% BLEU score.

3 The Corpus

Our aim was to improve the precision of the existing Spanish-to-Basque MT system by trying to translate texts in a restricted domain, because reliable Spanish-Basque corpora are not sufficiently available for a general domain. Also, we were interested in a kind of domain where a formal language would be used and in which many public organizations and private companies would be interested.

The Basque Institute of Public Administration (IVAP[1]) collaborated with us in this selection by examining some possible domains, available parallel corpora, and translation needs. We selected the domain related to labor agreements. Then, we built the Labor Agreements Corpus using a bilingual parallel corpus with 585,785 words in Basque and 839,003 in Spanish.

To build the test corpus, we randomly chose the full text of several labor agreements. We chose full texts because we wanted to ensure that several significant but short elements, such as headers and footers, would be represented, and also because it is important to measure the coverage and precision we get when translating the whole text in one document and not only some parts of it. First, we automatically aligned the corpus at sentence level, and then we performed manual revision. We did not allow system developers to see the test corpus.

As we have said, our goal was to combine different MT approaches: Rule-Based (RBMT), Example-Based (EBMT) and Statistical (SMT). Once we had the corpus, we split it into three parts for SMT (training, development and test corpus) and into two parts for EBMT (development and test corpus). In SMT we used the training corpus to learn the models (translation and language model), the development corpus to tune the parameters, and the test corpus to evaluate the system. In RBMT and EBMT there are no parameters to optimize, and so we considered only two corpora: one for development (combining the training and development parts used in SMT) and one for the test.

Table 1 shows the size, number of documents, sentences and words in the training, development,

[1] *http://www.ivap.euskadi.net*

Subset	Lang.	Doc.	Senten.	Words
Train	Basque	81	51,740	839,393
	Spanish	81		585,361
Develop	Basque	5	2,366	41,408
	Spanish	5		28,189
Test	Basque	5	1,945	39,350
	Spanish	5		27,214

Table 1: Labor Agreements Corpus

and test subsets of each language.

4 The MultiEngine MT system

In the next subsections we explain the three single MT strategies we have developed: Example-Based Approach, Statistical Machine Translation Approach and Rule-Based Machine Translation Approach. Finally, we explain how we have combined these three approaches.

4.1 Example Based Approach

In this subsection we explain how we automatically extract translation patterns from the bilingual parallel corpus and how we exploit them.

Translation patterns are generalizations of sentences that are translations of each other, replacing various sequences of one or more words by variables (McTait, 1999).

Starting from the aligned corpus we carry out two steps to automatically extract translation patterns. First, we detect some concrete units (mainly entities) in the aligned sentences and then we replace these units by variables. To detect the units, due to the morphosyntactic differences between Spanish and Basque, we need to execute particular algorithms for each language. We have developed algorithms to determine the boundaries of dates, numbers, named entities, abbreviations and enumerations.

After detecting the units, they must be aligned, relating the Spanish and Basque units of the same type that have the same meaning. For numbers, abbreviations, and enumerations, the alignment is almost trivial; however, the alignment algorithm for named entities is more complex. It is explained in more detail in (Martínez et al., 1998). Finally, to align the dates, we use their canonical form. Table 2 shows an example of how a translation pattern is extracted.

ES-EU Sentences	Sentences with generalized units	Translation Pattern
En Vitoria-Gasteiz, a 22 de Diciembre de 2003	En<rs type=loc> Vitoria -Gasteiz</rs> , a<date date=22/12/2003> 22 de Diciembre de 2003</date>	En<rs1> , a<date1>.
Vitoria Gasteiz, 2003ko Abenduaren 22.	<rs type=loc> Vitoria -Gasteiz </rs>,<date date=22/12/2003> 2003ko 22 Abenduaren </date>	<rs1> <date1>

Table 2: Example of Translation Pattern extraction

Once we have automatically extracted all the possible translation patterns from the training set, we store them in a hash table for use in the translation process.

When we want to translate a source sentence, we check if that sentence matches any pattern in the hash table. If the source sentence matches a sentence in the hash table with no variable, the translation process will immediately return its translation. A Word Error Rate (WER) metric was used to compare the two sentences. Otherwise, if the source sentence does not match anything in the hash table, the translation process will try to generalize that sentence and will check the hash table again for a generalized template. To generalize the source sentence, the translation process will apply the same detection algorithms used in the extraction process.

In a preliminary experiment using a training corpus of 54,106 sentence pairs we automatically extracted 7,599 translation patterns at the sentence level. These translation patterns covered 35,450 sentence pairs of the training corpus. We also consider an aligned pair of sentences as a translation pattern if it does not have any generalized unit but appears at least twice in the training set.

As this example-based system has very high precision but very low coverage, it is interesting to com-

bine it with the other MT engines, especially in this kind of domain where a formal and quite sublanguage is used.

4.2 Statistical Machine Translation Approaches

Two different approaches have been implemented: a conventional SMT system and a morpheme-based system. These corpus-based approaches have been carried out in collaboration with the National Center for Language Technology in Dublin. The system exploits SMT technology to extract a dataset of aligned chunks. Based on a training corpus, we conducted Spanish-to-Basque translation experiments (Labaka et al., 2007).

We used freely available tools to develop the SMT systems:

- GIZA++ toolkit (Och, 2003) for training the word/morpheme alignment.
- SRILM toolkit (Stolcke, 2002) for building the language model.
- Moses Decoder (Koehn et al., 2007) for translating the sentences.

Due to the morphological richness of Basque, some Spanish words, like prepositions or articles, correspond to one or more suffixes in Basque. In order to deal with this problem, we built a morpheme-based SMT system.

Adapting the SMT system to work at the morpheme level consists of training the basic SMT on the segmented text. The translation system trained on this data will generate a sequence of morphemes as output. In order to obtain the final Basque text, words have to be generated from those morphemes.

To obtain the segmented text, we analyzed Basque texts using Eustagger (Aduriz et al., 2003). This process replaces each word with the corresponding lemma followed by a list of morphological tags. The segmentation is based on the strategy proposed in (Agirre et al., 2006).

We optimized both systems (the conventional SMT and the morpheme-based) by decoding parameters using Minimum Error Rate Training. The metric used to carry out the optimization is BLEU. Table 3 shows: the conventional SMT system reported 9.51 for BLEU accuracy measure and 3.73 for NIST; the morpheme-based SMT system reported 8.98 BLEU and 3.87 NIST accuracy measures.

	BLEU	**NIST**	**WER**	**PER**
SMT	9.51	3.73	83.94	66.09
Morpheme based SMT	8.98	3.87	80.18	63.88

Table 3: Evaluation for SMT Systems

4.3 Rule-Based Machine Translation Approach

In this subsection we present the main architecture of an open-source RBMT engine named Matxin (Alegria et al., 2007), the first implementation of which translates from Spanish to Basque using traditional transfer, based on shallow and dependency parsing.

The design and the programs of the Matxin system are independent from this pair of languages, so the software can be used for other projects in MT. Depending on the languages included in the adaptation, it will be necessary to add, reorder and change some modules, but this will not be difficult because a unique XML format is used for communication among all the modules.

The project has been integrated in the OpenTrad[2] initiative, a government-funded project shared among different universities and small companies, which includes MT engines for translation among the main languages in Spain. The main objective of this initiative is the construction of an open, reusable and interoperable framework.

In the OpenTrad project, two different but coordinated architectures have been developed:

- A shallow-transfer-based MT engine for similar languages (Spanish, Catalan and Galician).
- A deeper-transfer-based MT engine for the Spanish-Basque and English-Basque pair. It is named Matxin, and stored it in *matxin.sourceforge.net*. It is an extension of previous work by the IXA group.

For the second engine, following the strategy of reusing resources, another open-source engine,

[2] *http://www.opentrad.org*

FreeLing (Carreras et al., 2004), was integrated for parsing Spanish sentences.

The transfer module is divided into three phases which match the three main objects in the translation process: words or nodes, chunks or phrases, and sentences.

1. First, it carries out lexical transfer using a bilingual dictionary compiled into a finite-state transducer.

2. Then, it applies structural transfer at sentence level, transferring information from some chunks to others, and making some chunks disappear. For example, in the Spanish-Basque transfer, person and number information for the object is imported from other chunks to the verbal chunk. As in Basque the verb also agrees in person and number with the object, later on the generation of the verb in Basque will require this information.

3. Finally, the module carries out the structural transfer at chunk level. This process can be quite simple (e.g. noun chains between Spanish and Basque) or more complex (e.g. verb chains between these same languages).

Then the XML file coming from the transfer module is passed on to the generation module.

- In the first step, this module performs syntactic generation in order to decide the order of chunks in the sentence and the order of words in the chunks. It uses several grammars for this purpose.

- The last step is morphological generation. In generating Basque, the main inflection is added to the last word in the phrase (the declension case, the article and other features are added to the whole noun phrase at the end of the last word), but in verb chains other words need morphological generation. We adapted a previous morphological analyzer/generator for Basque (Alegria et al., 1996) and transformed it according to the format used in *Apertium*.

The results for the Spanish-Basque system using FreeLing and Matxin are promising. The quantitative evaluation uses the open-source evaluation tool IQMT and we give figures using BLEU and NIST measures (Giménez et al., 2005). We also carried out an additional user-based evaluation, using Translation Error Rate (Snover et al., 2006). (Mayor, 2007) shows the results of the RBMT system's evaluation: 9.30, using the BLEU accuracy measure. In interpreting the results, we need to keep in mind that the development of this RBMT system was based on texts of newspapers.

We adapted this RBMT system to the domain of Labor Agreements in three main ways:

1. Terminology. Semiautomatic extraction of terminology using Elexbi, a bilingual terminology extractor for noun phrases (Alegria et al., 2006). Additionally, we carried out an automatic format conversion to the monolingual and bilingual lexicons for the selected terms. We extracted more than 1,600 terms from the development corpus, examined them manually, and selected nearly 807 to be include in the domain-adapted lexicon.

2. Lexical selection. Matxin does not address the lexical selection problem for lexical units (only for the preposition-suffix translation); it always selects the first translation in the dictionary (other possible lexical translations are stored for the post-edition process). For the domain adaptation, we calculated a new order for the possible translations based on the parallel corpus using GIZA++.

3. Resolution of format and typographical variants found frequently in the administrative domain.

After these improvements, the RMBT engine was ready to process sentences from this domain.

4.4 Approaches Combination

We experimented with a simple mixing alternative approach up to now used only for languages with huge corpus resources: selecting the best output in a multi-engine system (MEMT, Multi-engine MT). In our case, we combined RBMT, EBMT, and SMT approaches. In our design we took into account the following points:

1. Combination of MT paradigms: RBMT and data-driven MT.

2. Absence of large and reliable Spanish-Basque corpora.

3. Reusability of previous resources, such as translation memories, lexical resources, morphology of Basque and others.

4. Standardization and collaboration: using a more general framework in collaboration with other groups working in NLP.

5. Open-source: this means that anyone having the necessary computational and linguistic skills will be able to adapt or enhance it to produce a new MT system,

For this first attempt, we combined the three approaches in a very simple hierarchical way, processing each sentence with the three engines (RBMT, EBMT and SMT) and then trying to choose the best translation among them. First, we divided the text into sentences, then processed each sentence using each engine (parallel processing when possible). Finally, we selected one of the translations, dealing with the following facts:

- Precision of the EBMT approach is very high, but its coverage is low.
- The SMT engine gives a confidence score.
- RBMT translations are more adequate for human postedition than those of the SMT engine, but SMT gets better scores when BLEU and NIST are used with only one reference (Labaka et al., 2007). Table 4[3] summarizes the results of the automatic evaluation (BLEU) with one reference and those of the user-driven evaluation (HTER). Those evaluations where perfomed with two more general corpora related to news in the Basque Public Radio-Television (EiTB) and to articles in a magazine for consumers (Consumer).

With these results for the single approaches we decided to apply the following combinatory strategy:

1. If the EBMT engine covers the sentence, we chose its translation.

[3]*The Consumer corpus used for evaluation is the one referenced in Table 3 but before a cleaning process.*

Corpus	**BLEU RBMT**	**BLEU SMT**	**HTER RBMT**	**HTER SMT**
EiTB corpus	9.30	9.02	40.41	71.87
Consumer	6.31	8.03	43.60	57.97

Table 4: Evaluation using BLEU and HTER for single SMT and RBMT systems

	Coverage	**BLEU**	**NIST**
EBMT	EBMT 100%	29.02	4.70
RBMT	RBMT 100%	7.97	3.21
SMT	SMT 100%	14.37	4.43
EBMT +RBMT	EBMT 46.42% RBMT 53.58%	35.57	6.19
EBMT +SMT	EBMT 46.42% SMT 53.58%	38.31	6.82
EBMT +SMT +RBMT	EBMT 46.42% SMT 31.22% RBMT 22.36%	37.84	6.68

Table 5: Evaluation for MEMT systems using the development corpus

2. We chose the translation from the SMT engine if its confidence score was higher than a given threshold.

3. Otherwise, we chose the output from the RBMT engine.

5 Evaluation

In order to assess the quality of the resulting translation, we used automatic evaluation metrics. We report the following accuracy measures: BLEU (Papineni et al. , 2002) and NIST (Doddington, 2002).

The results using the development corpus for this second approach appear in Table 5.

Table 6 shows the results using the test corpus.

The best results, evaluated by using automatic metrics with only one reference, came from combining the two Data-Driven approaches: EBMT and SMT. Taking into account the single approaches, the best results are returned with EBMT strategy.

The results of the initial automatic evaluation showed very significant improvements. For example, a 193% relative increase for BLEU when comparing the EBMT+SMT+RBMT combination

	Coverage	BLEU	NIST
EBMT	EBMT 100%	32.42	5.76
RBMT	RBMT 100%	5.16	3.08
SMT	SMT 100%	12.71	4.69
EBMT +RBMT	EBMT 64.92% RBMT 35.08%	36.10	6.84
EBMT +SMT	EBMT 64.92% SMT 35.08%	37.31	7.20
EBMT +SMT +RBMT	EBMT 64.92% SMT 23.40% RBMT 11.68%	37.24	7.17

Table 6: Evaluation for MEMT systems using the test corpus

to the SMT system alone. Furthermore, we realized a 193.55% relative increase for BLEU when comparing the EBMT+SMT combination with the SMT system alone and 15.08% relative increase when comparing EBMT+SMT combination with the EBMT single strategy.

The consequence of the inclusion of a final RBMT engine (to translate just the sentences not covered by EBMT and with low confidence score for SMT) is a small negative contribution of 1% relative decrease for BLEU. Of course, bearing in mind our previous evaluation trials with human translators (Table 4), we think that a deeper evaluation using user-driven evaluation is necessary to confirm similar improvements for the MEMT combination including a final RBMT engine.

For example in the translation of the next sentence in Spanish (it is taken from the development corpus) *"La Empresa concederá préstamos a sus Empleados para la adquisición de vehículos y viviendas, en las siguientes condiciones"* the RBMT system generates *"Enpresak maileguak emango dizkio haren Empleados-i ibilgailuen erosketarentzat eta etxebizitzak, hurrengo baldintzetan"* and the SMT system *"Enpresak mailegu ibilgailuak bertako langileei emango, eta etxebizitza erosteko baldintzak"*. The figures using BLEU and NIST are higher for the SMT translation, but only the RBMT translation can be understood.

The results of the MEMT systems are very similar in the development and test corpora. Although the percentage of coverage of the EBMT single system is lower for the development corpus, its precission is higher.

Most of the references about Multi-Engine MT do not use EBMT strategy, SMT+RBMT is the most used combination in the bibliography. One of our main contributions is the inclusion of EBMT strategy in our Multi-Engine proposal; our methodology is straightforward, but useful.

6 Conclusions and Future Work

We applied Spanish-to-Basque MultiEngine Machine Translation to a specific domain to select the best output from three single MT engines we have developed. Because of previous results, we decided to apply a hierarchical strategy: first, application of EBMT (translation patterns), then SMT (if its confidence score is higher than a given threshold), and then RBMT.

It has carried out an important improvement in translation quality for BLEU in connection with the improvements obtained by other systems. We obtain 193.55% relative increase for BLEU when comparing the EBMT+SMT combination with the SMT system alone, and 15.08% relative increase when comparing EBMT+SMT combination with the EBMT single strategy.

Those improvements would be dificult to get for single engine systems. RBMT contribution seems to be very small with automatic evaluation, but we expect that HTER evaluation will show better results.

In spite of trying the strategy for a domain, we think that our translation system is a major advance in the field of language tools for Basque. However the restriction in using a corpus in a domain is given by the absence of large and reliable Spanish-Basque corpora.

For the near future, we plan to carry out new experiments using a combination of the outputs based on a language model. We also plan to define confidence scores for the RBMT engine (including penalties when suspicious or very complex syntactic structures are present in the analysis; penalties for high proportion of ignored word senses; and promoting translations that recognize multiword lexical units). Furthermore, we are planning to detect other types of translation patterns, especially at the phrase or chunk level.

Acknowledgements

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). Andy Way, from Dublin City University, kindly provided his expertise on data-driven MT systems. Consumer corpus has been kindly supplied by Asier Alcázar from the University of Missouri-Columbia and by Eroski Fundazioa.

References

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

Eneko Agirre, Arantza Díaz de Ilarraza, Gorka Labaka, Kepa Sarasola. 2006. *Uso de información morfológica en el alineamiento Español-Euskara.* XXII Congreso de la SEPLN.

Iñaki Alegria, Xabier Artola, Kepa Sarasola. 1996. *Automatic morphological analysis of Basque.* Literary & Linguistic Computing Vol. 11, No. 4, 193-203. Oxford University Press.

Iñaki Alegria, Antton Gurrutxaga, Xabier Saralegi, Sahats Ugartetxea. 2006. *ELeXBi, A Basic Tool for Bilingual Term Extraction from Spanish-Basque Parallel Corpora.* Proceedings of the 12th EURALEX International Congress. pp 159-165

Iñaki Alegria, Antantza Díaz de Ilarraza, Gorka Labaka, Mikel Lersundi, Aingeru Mayor, Kepa Sarasola. 2007. *Transfer-based MT from Spanish into Basque: reusability, standardization and open source.* LNCS 4394. pp. 374-384.

Xavier Carreras, Isaac Chao, Lluis Padró, Muntsa Padrò. 2004. *FreeLing: An open source Suite of Language Analyzers.* Proceedings of the 4th International Conference on Language Resources and Evaluation.

Yu Chen, Andreas Eisele, Christian Federmann, Eva Hasler, Michael Jellinghaus, Silke Theison. 2007. *Multi-engine machine translation with an open-source decoder for statistical machine translation.* proceedings of the Second Workshop on Statistical Machine Translation, pp. 193-196.

George Doddington. 2002. *Automatic evaluation of Machine Translation quality using n-fram cooccurrence statics* Proceedings of HLT, pp. 128-132.

Robert Frederking and Sergei Nirenburg. 1994. *Three heads are better than one.* Proceedings of the fourth ANLP

Jesús Giménez, Enrique Amigó, Chiori Hori. 2005. *Machine Translation Evaluation Inside QARLA.* Proceedings of the International Workshop on Spoken Language Technology.

Fei Huang and Kishore Papineni. 2007 *Hierarchical system combination for machine translation.* Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 277-286.

Philipp Koehn. 2005. *Europarl: A parallel corpus for statistical machine translation.* MT Summit X.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcelo Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. 2007. *Moses: Open Source Toolkit for Statistical Machine Translation.* Annual Meeting of the ACL.

Gorka Labaka, Nicolas Stroppa, Andy Way, Kepa Sarasola. 2007. *Comparing Rule-Based and Data-Driven Approaches to Spanish-to-Basque Machine Translation* Proceedings of MT-Summit XI.

Wolfgang Macherey and Franz J. Och. 2007. *An Empirical Study on Computing Consensus Translations from Multiple Machine Translation Systems.* Proceedings of the EMNLP and CONLL 2007.

Raquel Martínez, Joseba Abaitua, Arantza Casillas. 1998. *Alingning Tagged Bitext.* Proceedings of the Sixth Workshop on Very Large Corpora.

Evgeny Matusov, Nicola Ueffing, Hermann Ney. 2006. *Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment.* Proceedings of EACL 2006.

Aingeru Mayor. 2007. *Matxin: erregeletan oinarritutako itzulpen automatikoko sistema.* Ph. Thesis. Euskal Herriko Unibertsitatea.

Kevin M. McTait. 1999. *A Language-Neutral Sparse-Data.* Algorithm for Extracting Translation Patterns. Proceedings of 8th International Conference on Theorical and Mathodological Issues in Machine Translation.

Bart Mellebeek, Karolina Owczarzak, Josef Van Genabith, Andy Way. 2006. *Multi-engine machine translation by recursive sentence decomposition.* Proceedings of the 7^{th} Conference of the Association for Machine Translation in the Americas, Visions for the Future of Machine Translation, pp.110-118.

Franz J. Och and Hermann Ney. 2003. *A Systematic Comparison of Various Statistical Alignment Models.* Computational Linguistics, 29(1): 19-51.

Kishore Papineni, Salim Roukos, Tod Ward, Wei-Jing Zhu. 2002. *BLEU: a method for automatic evaluation of machine translation.* Proceedings of 40^{th} ACL, pp. 311-318.

Antti-Veikko I. Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, Bonnie J.Dorr. 2007. *Combining outputs from multiple machine translation systems.* NAACL-HLT-2007 Human Language Technology: the conference of the North American Chapter of the Association for Computational Linguistics, pp.228-235.

K. Sim, W. Byrne, M. Gales, H. Sahbi, P. Woodland. 2007. *Consensus network decoding for statistical machine translation system combination.* Proceedings of ICASSP, 2007.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, John Makhoul. 2006. *A study of translation edit rate with targeted human annotation.* Proceedings of AMTA'2006.

Andreas Stolcke. 2002. *SRILM - An Extensible Language Modeling Toolkit.* Proceedings Intl. Conerence. Spoken Language Processing.

Nicolas Stroppa, Declan Groves, Andy Way, Kepa Sarasola. 2006. *Example-Based Machine Translation of the Basque Language.* Proceeding of the 7^{th} Conference of the AMTA.

Menno van Zaanen and Harold Somers. 2005. *DEMOCRAT: Deciding between Multiple Outputs Created by Automatic Translation.* MT Summit X.

Andy Way and Nano Gough. 2005. *Comparing Example-Based and Statistical Machine Translation.* Natural Language Engineering, 11(3):295-309.

Mixing Approaches to MT for Basque: Selecting the best output from RBMT, EBMT and SMT

I. Alegria, A. Casillas, A. Diaz de Ilarraza, J. Igartua, G. Labaka,
M. Lersundi, A. Mayor, K. Sarasola
Ixa taldea. University of the Basque Country.
X. Saralegi
Elhuyar Fundazioa
B. Laskurain
Eleka S.L.
i.alegria@ehu.es

Abstract

We present the first steps in the definition of a mixing approach to MT for Basque based on combining single engines that follow to three different MT paradigms. After describing each engine we present the hierarchical strategy we use in order to select the best output, and a first evaluation.

1 Introduction and Basque Language

Basque is a highly inflected language with free order of sentence constituents.

It is an agglutinative language, with a rich flexional morphology. In fact for nouns, for example, at least 360 word forms are possible for each lemma. Each of the declension cases such as absolutive, dative, associative… has four different suffixes to be added to the last word of the noun phrase. These four suffix variants correspond to indefinite, definite singular, definite plural and "close" definite plural. Basque syntax and word order is very different compared with other languages as Spanish, French or English.

Machine translation is both, a real need, and a test bed for our strategy to develop NLP tools for Basque. We have developed corpus based and rule based MT systems, but they are limited.

On the one hand, corpus based MT systems base their knowledge on aligned bilingual corpora, and the accuracy their output depends heavily on the quality and the size of these corpora. When the pair of languages used in translation have very different structure and word order, obviously, the corpus needed should be bigger.

Being Basque a lesser resourced language, nowadays large and reliable bilingual corpora are unavailable for Basque. Domain specific translation memories for Basque are not bigger than two-three millions words, so they are still far away from the size of the present corpora for languages; e.g., Europarl corpus (Koehn, 2005), that is becoming a quite standard corpus resource, has 30 million words. So, the results obtained in corpus based MT to Basque are promising, but they are still not ready for public use.

On the other hand, the Spanish->Basque RBMT system Matxin's performance, after new improvements in 2007 (Labaka et al., 2007), is becoming useful for assimilation, but it is still not suitable enough to allow unrestricted use for text dissemination.

Therefore it is clear that we should combine our basic hes for MT (rule-based and corpus-based) in order to build a hybrid system with better performance. As the first steps on that way, we are experimenting with two simple mixing alternative approaches used up to now for languages with huge corpus resources:

- Selecting the best output in a multi engine system (MEMT, Multi-engine MT), in our case combining RBMT, EBMT and SMT approaches.
- Statistical post-editing(SPE) after RBMT.

This paper deals with the first approach. Our design has been carried out bearing in mind the following concepts:

- Combination of MT paradigms.
- Reusability of previous resources, such as

translation memories, lexical resources, morphology of Basque and others.

- Standardization and collaboration: using a more general framework in collaboration with other groups working in NLP.
- Open-source: this means that anyone having the necessary computational and linguistic skills will be able to adapt or enhance it to produce a new MT system,

Due to the real necessity for translation in our environment the involved languages would be Basque, Spanish, French and English.

The first strategy we are testing when we want to build a MT engine for a domain, is translating each sentence using each of our three single engines (rule-based, example-based and statistical) and then choosing the best translation among them (see section 4).

In section 2 we present the corpus that we will use in our experiments, while in section 3 we explain the single engines built up for Basque MT following the three traditional paradigms: rule-based, example-based and statistical. In section 4, we report on our experiment to combine those three single engines. We finish this paper with some conclusions.

2 The corpus

Our aim was to improve the precision of the MT system trying to translate texts from a domain. We were interested in a kind of domain where a formal and quite controlled language would be used and where any public organization or private company would be interested in.

Finally the domain related to *labor agreements* was selected. The Basque Institute of Public Administration (IVAP[1]) collaborated with us in this selection, by examining some possible domains, parallel corpora available and their translation needs. The Labor Agreements Corpus is a bilingual parallel corpus (Basque and Spanish) with 585,785 words for Basque and 839,003 for Spanish. We automatically aligned it at sentence level and then manual revision was performed.

As said before, our goal is to combine different MT approaches: Rule-Based (RBMT), Example Based (EBMT) and Statistical (SMT). Once we had the corpus, we split it in three for SMT (training, development and test corpus) and in two for EBMT (development and test corpus).

To build the test corpus the full text of several labor agreements was randomly chosen. We chose full texts because we wanted to ensure that several significant but short elements as the header or the footer of those agreements would be represented, and because it is important to measure the coverage and precision we get when translating the whole text in one agreement document and not only some sentences of parts of it. System developers are not allowed to see the test corpus.

In SMT we use the training corpus to learn the models (translation and language model); the development corpus to tune the parameters; and the test corpus to evaluate the system.

In RBMT and EBMT there are not parameters to optimize, and so, we consider only two corpora: one for the development (joining the training and development ones) and one for the test.

The size of each subset is shown in Table 1 (eu= Basque, es = Spanish).

		Doc	Sentences	Words
Training	es	81	51,740	839.393
	eu	81		585,361
Development	es	5	2,366	41,508
	eu	5		28,189
Test	es	5	1,945	39,350
	eu	5		27,214

Table 1. Labor Agreements Corpus

3 Single MT engines for Basque

In this section we present three single engines for Spanish-Basque translation following the three traditional paradigms: rule-based, example-based and statistical. The first one has been adapted to the domain corpus, and the other two engines have been trained with it.

3.1 The rule-based approach

In this subsection we present the main architecture of an open source MT engine, named *Matxin* (Alegria et al., 2007), the first implementation of which translates from Spanish into Basque using the traditional transfer model and based on shallow and dependency parsing. Later on, in a second step, we have specialized it to the domain.

1 http://www.ivap.euskadi.net

The design and the programs of Matxin system are independent from the pair of languages, so the software can be used for other projects in MT. Depending on the languages included in the adaptation, it will be necessary to add, reorder and change some modules, but this will not be difficult because a unique XML format is used for the communication among all the modules.

The project has been integrated in the *OpenTrad*[2] initiative, a government-funded project shared among different universities and small companies, which include MT engines for translation among the main languages in Spain. The main objective of this initiative is the construction of an open, reusable and interoperable framework.

In the *OpenTrad* project, two different but coordinated architectures have been carried out:

- A shallow-transfer based MT engine for similar languages (Spanish, Catalan and Galician).
- A deeper-transfer based MT engine for the Spanish-Basque and English-Basque pair. It is named *Matxin* and it is stored in *matxin.sourceforge.net.* It is an extension of previous work in IXA group.

In the second engine, following the strategy of reusing resources, another open source engine, *FreeLing* (Carreras et al., 2004), was used for analysis.

The transfer module is divided into three phases dealing at the level of the three main objects in the translation process: words or nodes, chunks or phrases, and sentences.

- First, lexical transfer is carried out using a bilingual dictionary compiled into a finite-state transducer.
- Then, structural transfer at sentence level is applied, some information is transferred from some chunks to others, and some chunks may disappear. For example, in the Spanish-Basque transfer, person and number information of the object and the type of subordination are imported from other chunks to the chunk corresponding to the verb chain.
- Finally the structural transfer at chunk level is carried out. This process can be quite simple (e.g. noun chains between Spanish and Basque) or more complex (e.g. verb chains between these same languages).

The XML file coming from the transfer module is passed on the generation module.

- In the first step, syntactic generation is performed in order to decide the order of chunks in the sentence and the order of words in the chunks. Several grammars are used for this purpose.
- Morphological generation is carried out in the last step. In the generation of Basque, the main inflection is added to the last word in the phrase (in Basque, the declension case, the article and other features are added to the whole noun phrase at the end of the last word), but in verb chains other words need morphological generation. A previous morphological analyzer/generator for Basque (Alegria et al., 1996) has been adapted and transformed to the format used in *Apertium*.

	BLEU	Edit-distance TER
Corpus1 (newspapers)	9.30	40.41
Corpus2 (web magazine)	6.31	43.60

Table 2. Evaluation for the RBMT system

The results for the Spanish-Basque system using *FreeLing* and *Matxin* are promising. The quantitative evaluation uses the open source evaluation tool IQMT and figures are given using Bleu and NIST measures (Giménez et al., 2005). An additional user based evaluation has been carried out too, using Translation Error Rate (Snover, 2006). The results using two corpora without very long sentences are shown in Table 2 (Mayor, 2007).

We have to interpret the results having in mind that the development of this RBMT system was based on texts of newspapers.

Adaptation to the domain

The adaptation to the domain has been out in three main ways:

2 www.opentrad.org

- Terminology. Semiautomatic extraction of terminology using Elexbi, a bilingual terminology extractor for noun phrases (Alegria et al., 2006). Additionally, an automatic format conversion to the monolingual and bilingual lexicons is carried out for the selected terms. More than 1,600 terms were extracted from the development corpus, manually examined, and near to 807 were selected to be included in the domain adapted lexicon.
- Lexical selection. Matxin does not face the lexical selection problem for lexical units (Matxin only does it for the preposition-suffix translation); just the first translation in the dictionary is always selected (the other possible lexical translations are stored for the post-edition). For the domain adaptation, a new order for the possible translations has been calculated in the dictionary, based on the parallel corpus and using GIZA++.
- Resolution of format and typographical variants which are found frequently in the administrative domain.

After this improvements this engine is ready to process the sentences from this domain.

3.2 The example-based approach

In this subsection we explain how we automatically extract translation patterns from the bilingual parallel corpus and how we exploit it in a simple way.

Translation patterns are generalizations of sentences that are translations of each other in that various sequences of one or more words are replaced by variables (McTait, 1999).

Starting from the aligned corpus we carry out two steps to automatically extract translation patterns.

First, we detect some concrete units (entities mainly) in the aligned sentences and then we replace these units by variables. Due to the morphosyntactic differences between Spanish and Basque, it was necessary to execute particular algorithms for each language in the detection process of the units. We have developed algorithms to determine the boundaries of dates, numbers, named entities, abbreviations and enumerations.

After detecting the units, they must be aligned, to relate the Spanish and Basque units of the same type that have the same meaning. While in the case of numbers, abbreviations and enumerations the alignment is almost trivial, in the case of named entities, the alignment algorithm is more complex. It is explained in more detail in (Martinez et al., 1998). Finally, to align the dates, we use their canonical form.

Table 3 shows an example of how a translation pattern is extracted.

Once we have extracted automatically all the possible translation patterns from the training set, we store them in a hash table and we can use them in the translation process. When we want to translate a source sentence, we just have to check if that sentence matches any translation pattern in the hash table. If the source sentence matches a sentence of the hash table that has not any variable, the translation process will immediately return its translation. Otherwise, if the source sentence does not exactly match any sentence in the hash table, the translation process will try to generalize that sentence and will check again in the hash if it finds a generalized template. To generalize the source sentence, the translation process will apply the same detection algorithms used in the extraction process.

In a preliminary experiment using a training corpus of 54.106 sentence pairs we have extracted automatically 7.599 translation patterns at sentence level.

Aligned sentences	**Aligned sentences with generalized units**	**Translation pattern**
En Vitoria-Gasteiz, a 22 de Diciembre de 2003.	En <rs type=loc> Vitoria-Gasteiz </rs> , a <date date=22/12/2003> 22 de Diciembre de 2003</date> .	En <rs1> , a <date1>.
Vitoria-Gasteiz, 2003ko Abenduaren 22.	<rs type=loc> Vitoria-Gasteiz </rs> , <date date=22/12/2003> 2003ko Abenduaren 22</date>.	<rs1>, <date1>.

Table 3. Pattern extraction process

These translation patterns cover 35.450 sentence pairs of the training corpus. We also think that an aligned pair of sentences can be a transla-

tion pattern if it does not have any generalized unit but it appears at least twice in the training set.

As this example based system has a very high precision but quite low coverage (see Table 6 and Table 7), it is very interesting to combine with the other engines specially in this kind of domain where a formal and quite controlled language is used.

3.3 The SMT approach

The corpus-based approach has been carried out in collaboration with the National Center for Language Technology in Dublin.

The system exploits SMT technology to extract a dataset of aligned chunks. We have conducted Basque to English (Stroppa et al., 2006) and Spanish to Basque (Labaka et al., 2007) translation experiments, based on a quite large corpus (270,000 sentence pairs for English and 50,000 for Spanish).

Freely available tools are used to develop the SMT systems:

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

Due to the morphological richness of Basque, in translation from Spanish to Basque some Spanish words, like prepositions or articles, correspond to Basque , and, in case of ellipsis, more than one of those suffixes can be added to the same word. In order to deal with this features a morpheme-based SMT system has been built.

Adapting the SMT system to work at morpheme level consists on training the basic SMT on the segmented text. The system trained on these data will generate a sequence of morphemes as output. In order to obtain the final Basque text, we have to generate words from those morphemes.

To obtain the segmented text, Basque texts are previously analyzed using *Eustagger* (Aduriz and Díaz de Ilarraza, 2003). After this process, each word is replaced with the corresponding lemma followed by a list of morphological tags. The segmentation is based on the strategy proposed on (Agirre et al., 2006).

Both systems (the conventional SMT system and the morpheme based), were optimized decoding parameters using a Minimum Error Rate Training. The metric used to carry out the optimization is BLEU.

The evaluation results in a quite general domain (for the same type of texts) are in Table 4.

	BLEU	NIST	WER	PER
SMT	9.51	3.73	83.94	66.09
morpheme-based SMT	8.98	3.87	80.18	63.88

Table 4. Evaluation for SMT systems

Details about the system and its evaluation can be consulted in (Díaz de Ilarraza et al., 2008).

4 Combining the approaches and evaluation

van Zaanen and Somers (2005) and Matusov et al. (2006) review a set of references about MEMT (Multi-engine MT) including the first attempt by Frederking and Nirenburg (1994), Macherey and Och (2007)

All those papers reach the same conclusion: combining the outputs results in a better translation.

Most of the approaches generate a new consensus translation using different language models. They have to train the system on those language models. Some of the approaches require confidence scores for each of the outputs. This approach is being used in several works (Macheret&Och, 2007; Sim et al., 2007), and some of them are used inside the GALE research program.

MEMT for Basque

Bearing in mind that huge parallel corpora for Basque are not available we decided to combine the different methods in a domain where translation memories were available.

Because confidence scores are not still available for the RBMT engine, we decided, for a first attempt, to combine the three approaches in a very simple hierarchical way: processing each sentence by the three engines (RBMT, EBMT and SMT) and then trying to choose the best

translation among them.

In a first step the text is divided into sentences, then each sentence is processed using each engine (parallel processing is possible). Finally one of the translations is selected.

In order to make this selection the facts we can deal with are the followings:

- Precision for the EBMT approach is very high, but its coverage low.
- The SMT engine gives a confidence score.
- RBMT translations are more adequate for human postedition than those of the SMT engine, but SMT gets better scores when BLEU and NIST are used with only one reference (Labaka et al., 2007).

	BLEU RBMT	BLEU SMT	HTER RBMT	HTER SMT
EiTB corpus (news)	9.30	9.02	40.41	71.87
Consumer (magazine)	6.31	8.03	43.60	57.97

Table 5. Evaluation using Bleu and HTER for RBMT and SMT (Labaka et al., 2007)

We can see in Table 5 that automatic evaluation (BLEU) with one reference and user-driven evaluation (HTER) yield different results.

Bearing this in mind, in this first attempt, we decided to apply a hierarchical strategy:

- If the EBMT engine covers the sentence its translation is selected.
- Else we chose the translation from the SMT engine if its confidence score is higher than a given threshold.
- Otherwise the output from the RBMT engine will be taken.

The results on the development corpus appear in Table 6.

The best results, evaluated using automatic metrics with only one reference, are obtained combining EBMT and SMT. But bearing in mind our previous evaluation trials with human translators (Table 5), we think that a deeper evaluation is necessary.

Table 7 shows the results on the test corpora.

	Coverage	BLEU	NIST
RBMT (domain adapted)	100%	7.97	3.21
SMT	100%	14.37	4.43
EBMT+RBMT	EBMT 42% RBMT 58%	26.85	5.15
EBMT+SMT	EBMT 42% SMT 58%	30.44	5.93
EBMT+SMT+ RBMT	EBMT 42% SMT 33% RBMT 25%	29.41	5.68

Table 6. Results for the MEMT system using the development corpus

	Coverage	BLEU	NIST
RBMT (domain adapted)	100%	5.16	3.08
SMT	100%	12.71	4.69
EBMT+RBMT	EBMT 58% RBMT 42%	26.29	5.40
EBMT+SMT	EBMT 58% SMT 42%	29.11	6.25
EBMT+SMT+ RBMT	EBMT 58% SMT 28% RBMT 14%	28.50	6.02

Table 7. Results for the MEMT system using the test corpus

5 Conclusions

We have presented a hierarchical strategy to select the best output from three MT engines we have developed for Spanish-Basque translation.

In this first attempt, we decided to apply a hierarchical strategy: First application of EBMT (translation patterns), then SMT (if its confidence score is higher than a given threshold), and then RBMT.

The results of the initial automatic evaluation showed very significant improvements. For example, 129% relative increase for BLEU when comparing. EBMT+SMT combination with SMT single system. Or 124% relative increase for BLEU when comparing. EBMT+SMT+RBMT combination with SMT single system.

Anyway the best results, evaluated using automatic metrics with only one reference, are obtained combining just EBMT and SMT.

The consequence of the inclusion of a final RBMT engine (to translate just the sentences not covered by EBMT and with low confidence score

for SMT) has a small negative contribution of 2% relative decrease for BLEU. But based on previous evaluations we think that a deeper evaluation based on human judgements is necessary.

For the near future we plan to carry out new experiments using combination of the outputs based on a language model. We are also plan defining confidence scores for the RBMT engine (penalties when suspicious or very complex syntactic structures are present in the analysis, penalties for high proportion of ignored word senses, promoting translations that recognize multiword lexical units, …)

Acknowledgments

This work has been partially funded by the Spanish of Education and Science (OpenMT: Open Source Machine Translation using hybrid methods,TIN2006-15307-C03-01) and the Local Government of the Basque Country (AnHITZ 2006: Language Technologies for Multingual Interaction in Intelligent Environments., IE06-185). Gorka Labaka is supported by a PhD grant from the Basque Government (grant code, BFI05.326)

Reference

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

Agirre, E., D´ de Ilarraza, A., Labaka, G., and Sarasola, K. (2006). Uso de información morfológica en el alineamiento Español-Euskara. In *XXII Congreso de la SEPLN*.

Alegria I., Artola X., Sarasola K. 1996. Automatic morphological analysis of Basque. *Literary & Linguistic Computing* Vol. 11, No. 4, 193-203. Oxford University Press. Oxford. 1996.

Alegria I., Gurrutxaga A., Saralegi X., Ugartetxea S. 2006. ELeXBi, A Basic Tool for Bilingual Term Extraction from Spanish-Basque Parallel Corpora. *Proc. of the 12th EURALEX International Congress*. pp 159-165

Alegria I., Díaz de Ilarraza A., Labaka G., Lersundi M., Mayor A., Sarasola K. 2007. Transfer-based MT from Spanish into Basque: reusability, standardization and open source. *LNCS 4394*. 374-384. Cicling 2007.

Carreras X.,, Chao I., Padró L., Padró M. 2004. FreeLing: An open source Suite of Language Analyzers, in P*roceedings of the 4th International Conference on Language Resources and Evaluation* (LREC'04).

Díaz de Ilarraza A., Labaka G., Sarasola K.. 2008. Spanish-Basque SMT system: statistical translation into an agglutinative language. (Submitted to LREC 2008)

Frederking R., Nirenburg S. 1994. Three heads are better than one. *Proc. of the fourth ANLP*. Stuttgart,

Giménez J., Amigó E., Hori C. 2005. Machine Translation Evaluation Inside QARLA. In *Proceedings of the International Workshop on Spoken Language Technology* (IWSLT'05)

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Annual Meeting of the ACL*, Prague.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. *MT Summit X.* Phuket.

Labaka G., Stroppa N., Way A., Sarasola K. 2007 Comparing Rule-Based and Data-Driven Approaches to Spanish-to-Basque Machine Translation *Proc. of MT-Summit XI,* Copenhagen

Macherey W., Och F, 2007. An Empirical Study on Computing Consensus Translations from Multiple Machine Translation Systems. *Proc. of the EMNLP and CONLL* 2007. Prague.

Martínez R., Abaitua J., Casillas A. Alingning Tagged Bitext. *Proceedings of the Sixth Workshop on Very Large Corpora,* 1998.

Mayor A. 2007. *Matxin: erregeletan oinarritutako itzulpen automatikoko sistema*. Ph. Thesis. Euskal Herriko Unibertsitatea.

Matusov E., Ueffing, N, Ney H. 2006. Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment. *Proc. of EACL 2006*, Trento.

McTait K. A Language-Neutral Sparse-Data 1999. Algorithm for Extracting Translation Patterns". *Proceedings of 8th International Conference on Theorical and Mathodological Issues in Machine Translation*,.

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

Sim K., Byrne W., Gales M., Sahbi H. 2007., Wood-

land P. Consensus network decoding for statistical machine translation system combination. Proc. of ICASSP, 2007

Snover M., Dorr B., Schwartz R., Micciulla L., and Makhoul J.. 2006. A study of translation edit rate with targeted human annotation. In Proceedings of AMTA-2006, Cambridge, USA.

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. In P*roc. Intl. Conf. Spoken Language Processing*, Denver, Colorado.

Stroppa N., Groves D., Way A., Sarasola K. 2006.Example-Based Machine Translation of the Basque Language. *7th conf. of the AMTA*.

van Zaanen M. and Somers H. 2005. DEMOCRAT: Deciding between Multiple Outputs Created by Automatic Translation. *MT Summit X*. Phuket.

Way A. and Gough N. 2005. Comparing Example-Based and Statistical Machine Translation. *Natural Language Engineering*, 11(3):295–309.

Statistical Post-Editing: A Valuable Method in Domain Adaptation of RBMT Systems for Less-Resourced Languages

A. Diaz de Ilarraza, G. Labaka, K. Sarasola
Ixa taldea
University of the Basque Country
{jipdisaa, jiblaing, jipsagak}@ehu.es

Abstract

We present two experiments with Basque to verify the improvement obtained for other languages by using statistical post editing. The small size of available corpora and the use a morphological component in both RBMT and SMT translations make different our experiments from hose presented for similar works. Our results confirm the improvements when using a restricted domain, but they are doubtful for more general domains.

1 Introduction

Corpus based MT systems base their knowledge on aligned bilingual corpora, and the accuracy of their output depends heavily on the quality and the size of these corpora. When the two languages used in translation have very different structure and word order, the corpus needed to obtain similar results should be bigger.

Basque is a highly inflected language with free constituent order. Its structure and word order is different compared with languages as Spanish, French or English.

Being Basque a lesser used language, nowadays large and reliable bilingual corpora are unavailable. At present, domain specific translation memories for Basque are not bigger than two-three millions words, so they are still far away from the size of the corpora used for other languages; for example, Europarl corpus (Koehn, 2005), that is becoming a quite standard corpus resource, has 30 million words. So, although domain restricted corpus based MT for Basque shows promising results, it is still not ready for general use.

Moreover, the Spanish>Basque RBMT system Matxin's performance, after new improvements in 2007 (Alegria et al., 2007), is becoming useful for content assimilation, but it is still not suitable enough to allow unrestricted use for text dissemination.

Therefore, it is clear that we should experiment combining our basic approaches for MT (rule-based and corpus-based) to get a better performance. As the first steps on that way, we are experimenting with two simple alternative approaches to combining RBMT, SMT and EBMT:

- Selecting the best output in a multi engine system combining RBMT, EBMT and SMT approaches. (Alegría, et al., 2008)
- Statistical post-editing (SPE) on RBMT systems.

This paper deals with the second approach, where significant improvements have been recently published (Dugast et al., 2007; Ehara, 2007; Elming, 2006; Isabelle et al., 2007; Simard et al., 2007a and 2007b).

We don't have large corpus on post editing for Basque as proposed in (Isabelle et al., 2007), because our RBMT system has recently been created. However, we could manage to get parallel

corpus on some domains with a few million of words,

We will show that the issue of domain adaptation of the MT systems for Basque can be performed via the serial combination of a vanilla RBMT system and a domain specific statistical post-editing system even when the training corpus is not very big (half a million words). Unfortunately, we could not show that RBMT+SPE combination improves the result of RBMT systems when the corpus used is not related to a restricted domain.

The rest of this paper is arranged as follows: In section 2, we position the present work with respect to our ongoing research on SMT and SPE. In section 3 we present the corpora that will be used in our experiments. Section 4 describes the basic RBMT and statistical translation systems. In section 5, we report on our experiments comparing translation results under a range of different MT conditions: SMT versus RBMT, RBMT+SPE versus RBMT, and RBMT+SPE versus SMT. We finish this paper with some conclusions and future work.

2 Related work

In the experiments related by (Simard et al., 2007a) and (Isabelle et al., 2007) SPE task is viewed as translation from the language of RBMT outputs into the language of their manually post-edited counterparts. So they don't use a parallel corpus created by human translation. Their RBMT system is SYSTRAN and their SMT system PORTAGE. (Simard et al., 2007a) reports a reduction in post-editing effort of up to a third when compared to the output of the rule-based system, i.e., the input to the SPE, and as much as 5 BLEU points improvement over the direct SMT approach. (Isabelle et al., 2007) concludes that such a RBMT+SPE system appears to be an excellent way to improve the output of a vanilla RBMT system and constitutes a worthwhile alternative to costly manual adaptation efforts for such systems. So a SPE system using a corpus with no more than 100.000 words of post-edited translations is enough to outperform an expensive lexicon enriched baseline RBMT system.

The same group recognizes (Simard et al., 2007b) that this sort of training data is seldom available, and they conclude that the training data for the post-editing component does not need to be manually post-edited translations, that can be generated even from standard parallel corpora. Their new RBMT+SPE system outperforms both the RBMT and SMT systems again. The experiments show that while post-editing is more effective when little training data is available, it remains competitive with SMT translation even when larger amounts of data. After a linguistic analysis they conclude that the main improvement is due to lexical selection.

In (Dugast et al., 2007), the authors of SYSTRAN's RBMT system present a huge improvement of the BLEU score for a SPE system when comparing to raw translation output. They get an improvement of around 10 BLEU points for German-English using the Europarl test set of WMT2007.

(Ehara, 2007) presents two experiments to compare RBMT and RBMT+SPE systems. Two different corpora are issued, one is the reference translation (PAJ, Patent Abstracts of Japan), the other is a large scaled target language corpus. In the former case, RBMT+SPE wins, in the later case RBMT wins. Evaluation is performed using NIST scores and a new evaluation measure NMG that counts the number of words in the longest sequence matched between the test sentence and the target language reference corpus.

Finally, (Elming, 2006) works in the more general field called as Automatic Post-Processing (APE). They use transformation-based learning (TBL), a learning algorithm for extracting rules to correct MT output by means of a post-processing module. The algorithm learns from a parallel corpus of MT output and human-corrected versions of this output. The machine translations are provided by a commercial MT system, PaTrans, which is based on Eurotra. Elming reports a 4.6 point increase in BLEU score.

3 The corpora

Our aim was to improve the precision of the MT

system trying to translate texts from a restricted domain. We were interested in a kind of domain where a formal and quite controlled language would be used and where any public organization or private company would be interested in automatic translation on this domain. We also wanted to compare the results between the restricted domain and a more general domain such as news.

Specific domain: Labor Agreements Corpus

The domain related to *Labor Agreements* was selected. The Basque Institute of Public Administration (IVAP[1]) collaborated with us in this selection, after examining some domains, available parallel corpora and their translation needs. The Labor Agreements Corpus is a bilingual parallel corpus (Basque and Spanish) with 640,764 words for Basque and 920,251 for Spanish. We automatically aligned it at sentence level and then manual revision was performed.

To build the test corpus the full text of several labor agreements was randomly chosen. We chose full texts because we wanted to ensure that several significant but short elements as the header or the footer of those agreements would be represented. Besides it is important to measure the coverage and precision we get when translating the whole text in one agreement document and not only those of parts of it. System developers are not allowed to see the test corpus.

In SMT we use the training corpus to learn the models (translation and language model); the development corpus to tune the parameters; and the test corpus to evaluate the system.

The size of each subset is shown in Table 1.

		Sentences	Words
Training	Spanish	51,740	839.393
	Basque		585,361
Development	Spanish	2,366	41,508
	Basque		28,189
Test	Spanish	1,945	39,350
	Basque		27,214

Table 1. Statistics of Labor Agreements Corpus

1 http://www.ivap.euskadi.net

General domain: Consumer Eroski Corpus

As general domain corpus, we used the *Consumer Eroski* parallel corpus. The *Consumer Eroski* parallel corpus is a collection of 1,036 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 one million Spanish words for Spanish and more than 800,000 Basque words. This corpus is aligned at sentence level.

In order to train the data-driven systems (both SMT and SPE systems), we used approximately 55,000 aligned sentences extracted from the Consumer dataset. Two additional sentence sets are used; 1501 sentences for parameter tuning and 1515 sentences for evaluation (see Table 2).

		Sentences	Words
Training	Spanish	54,661	1,056,864
	Basque		824350
Development	Spanish	1,501	34,333
	Basque		27,235
Test	Spanish	1,515	32,820
	Basque		34,333

Table 2. Statistics of Consumer Eroski corpus

4 Basic translation systems

Rule based system: Matxin

In this subsection we present the main architecture of an open source MT engine, named *Matxin* (Alegria et al., 2007). the first implementation of Matxin translates from Spanish into Basque using the traditional transfer model and based on shallow and dependency parsing.

Matxin is a classical transfer system consisting of three main components: (i) analysis of the source language into a dependency tree structure, (ii) transfer from the source language dependency tree to a target language dependency structure, and (iii) generation of the output translation from the target dependency structure. These three components are described in more detail in what follows.

The analysis of the Spanish source sentences into dependency trees is performed using an adapted version of the Freeling toolkit (Carreras et al., 2004). The shallow parser provided by Freeling is augmented with dependency information between chunks.

In the transfer module the Spanish analysis tree is transformed into Basque dependency tree. In this step, a very simple lexical selection is carried out, the Spanish lemma is translated by most frequent equivalent.

Finally, the dependency tree coming from the transfer module is passed on the generation module, in order to get the target language sentence. The order of the words in the final sentence is decided and morphological generation is carried out when it is necessary (in Basque: the declension case, the article and other features are added to the whole noun phrase at the end of the last word). We reused a previous morphological analyzer/generator developed for Basque (Alegria et al., 1996) adapted and transformed to our purposes.

Corpus based system

The corpus-based approach has been carried out in collaboration with the National Center for Language Technology in Dublin City University (DCU).

The system is based on a baseline phrase-based SMT system, but the dataset of aligned phrases is enriched with linguistically motivated phrase alignments. We have carried out Basque to English (Stroppa et al., 2006) and Spanish to Basque (Labaka et al., 2007) translation experiments.

Freely available tools are used to develop the SMT systems:

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

Due to the morphological richness of Basque, when translating from Spanish to Basque some Spanish words, like prepositions or articles, correspond to Basque suffixes, and, in case of ellipsis, more than one of those suffix can be added to the same word. Example of concatenation of two case suffixes:

```
  puntuarenean =
= puntu  +  aren    +  ean  =
= point  +  of the  +  in the  =
= in the one(ellipsis) of the point
```

In order to deal with these features a morpheme-based SMT system was developed.

Adapting the SMT system to work at the morpheme level consists on training the basic SMT on the segmented text. The system trained on these data will generate a sequence of morphemes as output. In order to obtain the final Basque text, we have to generate words from those morphemes.

To get the segmented text, Basque texts are previously analyzed using Eustagger (Aduriz & Díaz de Ilarraza, 2003). After this process, each word is replaced with the corresponding lemma followed by a list of morphological tags. The segmentation is based on the strategy proposed on (Agirre et al., 2006).

Both systems (the conventional SMT system and the morpheme based), were optimized decoding parameters using a Minimum Error Rate Training. The metric used to carry out the optimization is BLEU.

The evaluation results for the general domain Consumer corpus (also used in this paper) are in Table 3. The morpheme based MT system gets better results for all the measures except BLEU.

	BLEU	NIST	WER	PER
SMT	**9.85**	4,28	82,72	63,78
Morpheme-based SMT	9,63	**4,43**	**80.92**	**62,27**

Table 3. Evaluation for SMT systems

RBMT and Statistical Post-Editing

In order to carry out experiments with statistical post-editing, we have first translated Spanish sentences in the parallel corpus using our rule-based translator (Matxin). Using these automatically translated sentences and their

corresponding Basque sentences in the parallel corpus, we have built a new parallel corpus to be used in training our statistical post-editor.
The statistical post-editor is the same corpus-based system explained before. This system is based on freely available tools but enhanced in two main ways:

- In order to deal morphological richness of Basque, the system works on morpheme-level, so a generation phase is necessary after SPE is applied.
- Following the work did in collaboration with the DCU, the phrases statistically extracted are enriched with linguistically motivated chunk alignments.

5 Results

We used automatic evaluation metrics to assess the quality of the translation obtained using each system. For each system, we calculated BLEU (Papineni et al., 2002), NIST (Doddington, 2002), Word Error Rate (WER) and Position independent Error Rate (PER).

Besides, our aim was to evaluate performance using different corpora types, so we tested the output of all systems applied to two corpora: one domain specific (Labor Agreements Corpus), and a general domain corpus (Consumer corpus).

	BLEU	NIST	WER	PER
Rule-based	4,27	2,76	89,17	74,18
Corpus-based	12,27	4,63	77,44	58,17
Rule-based + SPE	**17,11**	**5,01**	**75,53**	**57,24**

Table 4. Evaluation on domain specific corpus

Results obtained on the Labor Agreements Corpus (see Table 4) shows that the rule-based gets a very low performance (rule-based system is not adapted to the restricted domain), and the corpus-based system gets a much higher score (8 BLEU points higher, a 200% relative improvement). But if we combine both systems using the corpus-based system as a statistical post-editor, the improvement is even higher outperforming corpus-based system in 4.48 BLEU point (40% relative improvement).

	BLEU	NIST	WER	PER
Rule-based MT	6,78	3,72	81,89	66,72
Corpus-based MT	**9,63**	**4,43**	80,92	**62,27**
Rule-based + SPE	8,93	4,23	**80,34**	63,49

Table 5. Evaluation on general domain corpus

Otherwise, results on the general domain corpus (see Table 5) do not indicate the same. Being a general domain corpus, the vanilla rule-based system gets better results, and those approaches based on the corpus (corpus-based MT and RBMT+SPE) get lower ones. Furthermore, the improvement achieved by the statistical post-editor over the rule-based system is much smaller and it does not outperforms the corpus-based translator.

6 Conclusion

We performed two experiments to verify the improvement obtained for other languages by using statistical post editing. Our experiments differ from other similar works because we use a morphological component in both RBMT and SMT translations, and because the size of the available corpora is small.

Our results are coherent with huge improvements when using a RBMT+SPE approach on a restricted domain presented by (Dugast eta al., 2007; Ehara, 2007; Simard et al., 2007b). We obtain 200% improvement in the BLEU score for a RBMT+SPE system working with Matxin RBMT system, when comparing to raw translation output, and 40% when comparing to SMT system.

Our results also are coherent with a smaller improvement when using more general corpora as presented by (Ehara, 2007; Simard et al., 2007b).

We can not work with manually post-edited corpora as (Simard et al., 2007a) and (Isabelle et al., 2007) because there is no such a big corpus for Basque, but we plan to collect it and compare results obtained using a real post-edition corpus and the results presented here.

We also plan automatic extracting rules to

correct MT output by means of a post-processing module (Elming, 2006).

Acknowledgments

This work has been partially funded by the Spanish Ministry of Education and Science (OpenMT: Open Source Machine Translation using hybrid methods,TIN2006-15307-C03-01) and the Local Government of the Basque Country (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 Díaz de Ilarraza, A. (2003). Morphosyntactic disambiguation and shallow parsing in Computational Processing of Basque. In Inquiries into the lexicon-syntax relations in Basque. Bernard Oyharçabal (Ed.), Bilbao.

Alegria, I., Artola Zubillaga, X. and Sarasola, X. (1996). Automatic morphological analysis of Basque. Literary & Linguistic Computing 11(4):193—203.

Alegria I., Díaz de Ilarraza A., Labaka G., Lersundi M., Mayor A., Sarasola K. (2007) Transfer-based MT from Spanish into Basque: reusability, standardization and open source. Cicling.

Alegria, I., Casillas, A., Díaz de Ilarraza, A., Igartua, J., Labaka, G., Lersundi, M., Mayor, A., Sarasola, K., Saralegi, X., Laskurain, B. (2008). A Simple Mixing Approach to MT for Basque. To be presented in MATMT08 workshop: Mixing Approaches to Machine Translation. Donostia.

Agirre, E., Díaz de Ilarraza, A., Labaka, G., and Sarasola, K. (2006). Uso de información morfológica en el alineamiento Español-Euskara. In XXII congreso de la SEPLN, Zaragoza, Spain.

Carreras, X., Chao, I., Padró, L., Padró, M. (2004). FreeLing: An Open-Source Suite of Language Analyzers. In Proceedings of 4th LREC, Lisbon, Portugal.

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

Dugast, L., Senellart, J., & Koehn, P. (2007). Statistical post-editing on SYSTRAN's rule-based translation system. ACL 2007: proceedings of the Second Workshop on Statistical Machine Translation, 23, 2007, Prague, Czech Republic; pp. 220-223

Elming, J. (2006). Transformation-based correction of rule-based MT. 11th Annual Conference of the European Association for Machine Translation, Oslo, Norway.

Isabelle, P., Goutte, C., & Simard, M. (2007). Domain adaptation of MT systems through automatic post-editing. MT Summit XI, 10-14 September 2007, Copenhagen, Denmark. pp.255-261

Koehn, Ph. (2005). Europarl: A parallel corpus for statistical machine translation. Proc. of the MT Summit X ,pp. 79–86, September.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan N., Shen, W. Moran, C. Zens, R. Dyer, C. Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In Annual Meeting of the Association for Computational Linguistics (ACL), Prague, Czech Republic.

Labaka, G., Stroppa, N., Way, A. and Sarasola, K. (2007). Comparing Rule-based and Data-driven Approaches to Spanish-to-Basque Machine Translation. In Proceedings of the MT-Summit XI, Copenhagen, Denmark.

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

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

Simard, M., Goutte, C., and Isabelle, P.. (2007a). Statistical phrase-based post-editing. In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pages 508–515, Rochester, USA, April. Association for Computational Linguistics.

Simard, M., Ueffing, N., Isabelle, P., & Kuhn, R.

(2007b). Rule-based translation with statistical phrase-based post-editing. ACL 2007: proceedings of the Second Workshop on Statistical Machine Translation, June 23, 2007, Prague, Czech Republic; pp. 203-206

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. In Proc. Intl. Conf. Spoken Language Processing, Denver, Colorado.

Stroppa, N., Groves, D., Way, A., and Sarasola, K. (2006). Example-base Machine Translation of the Basque Language. In Proceedings of AMTA 2006, pp. 232—241, Cambridge, MA.

Ehara Terumasa (2007). Rule based machine translation combined with statistical post editor for Japanese to English patent translation. MT Summit XI Workshop on patent translation, 11 September 2007, Copenhagen, Denmark; pp.13-18.