Lengoaia eta Sistema Informatikoak Saila


Informatika Fakultatea

# EUSMT: Incorporating Linguistic Information to SMT for a Morphologically Rich Language.
## Its use in SMT-RBMT-EBMT hybridation

Gorka Labaka Intxauspek Arantza Diaz de Ilarrazaren eta Kepa Sarasolaren zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Informatikan Doktore titulua eskuratzeko aurkeztua

Donostia, 2009ko apirila.

*Lan hau Eusko Jaurlaritzaren ikertzaileak prestatzeko beka batekin* (BFI05.326) *egin dut*

*Bertsoak izan ahal du itzulpen askea*
*zaila baita denari erabat eustea*
*zaila baino gehiago, zer buruhaustea*
*nahi baduzu bertsoak zu gozaraztea*
*hobe duzu euskara ikasten hastea.*

Xabier Paya

*Un bertso puede ser de traducción somera*
*pues es cosa difícil de otra manera*
*digo difícil como si posible fuera*
*si quieres sentir lo que el bertso genera*
*será mejor que empieces a aprender euskara.*

Xabier Paya

*Etxekoei, bereziki amari eta aitari*

# Eskerrak/Acknowledgements

Lehenik eta behin tesi honen zuzendari izan diren Arantzari eta Kepari eman nahi nieke eskerrak. Beraien gidaritzapean egin dut lan hau eta beraiek gabe ez dakit honaino iritsiko nintzen.

Tesi zehar izan ditudan arazoei aurre egiten lagundu didazuten guztiei. Katearekin edo makinekin arazoren bat izan dudan bakoitzean laguntzeko prest izan zarete.

IXA talde osoari, taldean dagoen giroak lanera etortzea errezagoa egiten baitu. *Kafe* baten aurrean, *tupper*retik bazkaltzerakoan ala *afari* baten erdian, inoiz ez dakizu noiz eta nola, baina momenturen batean lankideak lagun bihurtzen dira.

Larramendikoei, egoera beretxuan egoteagatik nire pozak eta atsekabeak hobekien ulertu duzuenak izan zaretelako. Pazientzia gehixeago izan, gutxiago falta baita egindako esfortzuak merezi duen ospakizunerako.

Gracias también a aquellos que me ayudáis a desconectar tras cada semana de trabajo, para poder volver cada lunes a *eso de la tesis* (que aunque no sepáis muy bien en qué consiste habeis ayudado a que se haga realidad).

Eta bukatzeko, nola ez, familiakoei. Batzuk dioten bezela tesi bat aurrera ateratzea mendi bat igotzea bezalakoa bada, zuen laguntzarekin *K2*ra astero igotzeko gai izango nintzen.

# Contents

# List of Figures

# List of Tables

# INTRODUCTION

# CHAPTER I

# Introduction

This thesis is defined in the framework of machine translation for Basque. Having developed a Rule-Based Machine Translation (RBMT) system for Basque in the IXA group (Mayor, 2007), we decided to tackle the Statistical Machine Translation (SMT) approach and experiment on how we could adapt it to the peculiarities of the Basque language. Once we had achieved a minimal quality SMT system, we used it in preliminary hybridization experiments.

The nowadays globalized society means that Human Language Technologies and Machine Translation have become essential for the survival of minority languages such as Basque. Even so, the lower economic interest in these languages prevents much research being carried out on their particular characteristics. Basque has to face up to the problems arising from it being a minority language (lack of funding and resources), as well as several linguistic peculiarities (both morphological and syntactic) that make translation a truly challenging issue.

## I.1 Machine Translation

The information society we live in is undoubtedly multilingual. Every day, hundreds of thousands of documents are 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 1.8 million 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 that are freely available 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 effort 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. Researchers claimed that within three to five years, machine translation issues would be fully resolved. The actual progress was much slower, however, and after 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; Arnold et al., 1993). 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 example-based (Nagao, 1984) approaches. Over the years the statistical approach achieved a prominent status at the expense of the linguistic approach, until it became the dominant paradigm with little research work on older paradigms. Nowadays, on the other hand, most research efforts are focused on enhancing SMT by incorporating any type of linguistic knowledge.

In its pure form, Statistical Machine Translation (Brown et al., 1993) 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 (named phrases) and individual words of the parallel texts and then calculate the

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


Figure I.1: Illustration of the evolution of the funding in MT (Arnold et al., 1993).

probabilities of each phrase in the source sentence being translated into a phrase with which it 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 (Koehn, 2005) the results obtained for different language pairs varies drastically depending on the language pair, with the lowest scores being achieved by the Dutch-Finnish translation system. Although automatic evaluation metrics could not be used to compare systems trained in such different environments the great difference between scores certainly shows the difference in the complexity of the task.

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) in 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).

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.

In addition, Basque is a less-resourced 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 (1.3 million Spanish words) to 7 million Basque words (9 million Spanish words) during the development of this thesis, it is still far below the corpora available for other languages. In Europarl, the corpus used to develop much of the SMT research, there are at least 30 million words for each language.

## I.2 Basque Language

Basque is an isolated language, and little is known of its origins. It is likely that an early form of the Basque language was already present in Western Europe before the arrival of the Indo-European languages.

Basque is an agglutinative language, with a rich inflectional morphology; for nouns, for example, there are at least 360 word forms possible for each lemma. Each one of the grammar cases, absolutive, dative, associative, etc, has four different suffixes to be added to the last word of the noun phrase. These four suffix variants correspond to undetermined, singular determined, plural determined and proximity plural determined. Furthermore, in the case of ellipsis more than one suffix can be added to the same lemma, increasing the word forms that can be generated from a noun (from a unique lemma up to 1 million different forms can be generated). 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.

Basque is also an ergative-absolutive language. The subject of an intransitive verb is in the absolutive case (which is unmarked), and the same case is used for the direct object of a transitive verb. The subject of the transitive

| | | | |
|---|---|---|---|
| etxe | */house/* | | |
| | etxea | */the house/* | |
| | etxeak | */the houses/* | |
| | etxeok | */these houses/* | |
| | [edozein] etxetara | */to [any] house/* | |
| | etxera | */to the house/* | |
| | etxeetara | */to the houses/* | |
| | etxeotara | */to these houses/* | |
| | [edozein] etxetatik | */from [any] house/* | |
| | etxetik | */from the house/* | |
| | etxeetatik | */from the houses/* | |
| | etxeotatik | */from these houses/* | |
| | [edozein] etxerekin | */with [any] house/* | |
| | etxearekin | */with the house/* | |
| | etxeekin | */with the houses/* | |
| | etxeokin | */with these houses/* | |
| | etxeko | */of the house/* | |
| | | etxekoa | */the one of the house/* |
| | | etxekoak | */the ones of the house/* |
| | | etxekook | */these ones of the house/* |
| | | etxekora | */to the one of the house/* |
| | | etxekoetara | */to the ones of the house/* |
| | | etxekootara | */to these ones of the house/* |
| | | ... | |
| | etxeetako | */of the houses/* | |
| | | etxeetakoa | */the one of the houses/* |
| | | etxeetakoak | */the ones of the houses/* |
| | | etxeetakook | */these ones of the houses/* |
| | | etxeetakora | */to the one of the houses/* |
| | | etxeetakoetara | */to the ones of the houses/* |
| | | etxeetakootara | */to these ones of the houses/* |
| | | ... | |
| | etxeotako | */of these houses/* | |
| | | etxeotakoa | */the one of these houses/* |
| | ... | | |

Figure I.2: Illustration of the Basque inflectional morphology. Up to 1 million different forms can be generated from a unique lemma.

verb (i.e., the agent) is marked differently, with the ergative case (shown by the suffix '-k'). This also triggers main and auxiliary verbal agreement.

The auxiliary verb which accompanies most main verbs, agrees not only

with the subject, but with the direct object and the indirect object, if present. Among European languages, this polypersonal system (multiple verb agreement) is only found in Basque, some Caucasian languages, and Hungarian. The ergative-absolutive alignment is rare among European languages, but not worldwide. Since Statistical Machine Translation works on a word basis, all these morpho-syntactic features, which modify the word forms, have a negative effect on all the steps of the SMT (from alignment to the decoding).

In addition, there are syntactic differences related to the word order that have a negative impact on the translation. Modifiers of both verbs and noun phrases are ordered differently in Basque than in Spanish or English. For example, prepositional phrases attached to noun phrases are placed before the modified noun phrase instead of after it. Furthermore, 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. In the word alignment presented in Figure I.3 we can see the great word order difference between Spanish and Basque. Since SMT does not use syntactic information (reordering in basic SMT is based on distance), word order differences like the ones presented in Spanish-Basque translation seriously harm the translation.

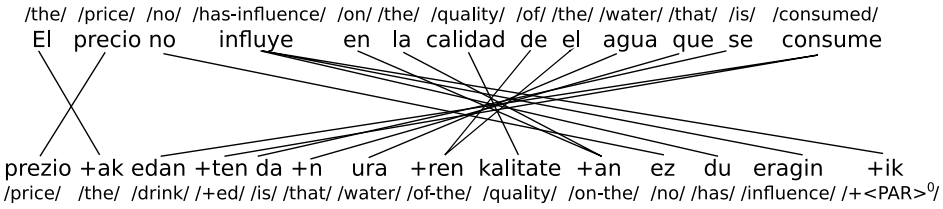



Figure I.3: Example of word alignment.

It remains alive but over the last few centuries Basque has suffered continuous regression. The region in which Basque is spoken is smaller than what is known as the Basque Country and the distribution of Basque speakers is not homogeneous there. The main reasons of this regression (Amorrortu, 2002) are that Basque was not an official language and was not included in the educational system, or used in the media and for industrial environments. In addition, the fact that there are six different dialects has made the wider development of written Basque difficult.

However, since 1980 some of those features have changed and many citizens and some local governments have promoted the recovery of the Basque

Language.

Today Basque holds co-official language status in the Basque regions of Spain: the fully autonomous community of the Basque Country and some parts of Navarre. Basque has no official standing in the Northern Basque Country.

In the past, Basque was associated with a lack of education and with people stigmatized as uneducated, rural, and having little wealth or power. There is no such association today, Basque speakers are no different from Spanish or French monolinguals in any of these characteristics.

Standard Basque, called *Batua* (unified) in Basque, was defined by the Academy of the Basque Language (*Euskaltzaindia*)[2] in 1966. At present, the morphology is completely standardized, but the lexical standardization process is ongoing. Nowadays this is the language model taught in most schools and used in some media and official papers published in Basque.

There are around 700,000 Basque speakers, around 25% of the total population of the Basque Country, and they are not evenly distributed. However, the use of Basque in industry and especially in Information and Communication Technology is still not widespread. A language that seeks to survive in the modern information society must also be present in such fields and this requires language technology products. Basque, along with other minority languages, has to make a great effort to face this challenge (Petek, 2000; Nadeu et al., 2001). In this context, the use of applications based on MT will be a significant help promoting the use of Basque and ensuring its presence in today's multilingual society. That is why Machine Translation for Basque is thus both a real need and a test bed for our strategy for developing NLP tools for Basque.

## I.3 Motivation

This Ph.D. thesis has been carried out within the Ixa research group, which was created in 1986 by 5 university lecturers in the Computer Science Faculty of the University of the Basque Country with the aim of laying the foundations for research and development of NLP software, mainly for Basque, and facing the challenge of adapting Basque to language technology. The subject

[2]http://www.euskaltzaindia.net/

of this work is closely related to the history and works of the group, so we will briefly explain these.

The first of the group's projects was to develop a Spanish-Basque Machine translation system but, after a preliminary study, the necessity of building basic resources and tools (such as a morphological analyzer/generator, parser and so on) was established as a must before moving on to the development of a machine translation system.

This thought was the seed for the design of the strategy followed during subsequent years. This strategy deals with two main issues:

1. The need for the standardization of resources to be used in future research, tools and applications.
2. The need for incremental design and development of language foundations, tools and applications in a parallel and coordinated manner in order to get the best benefit from them.

The strategy followed has been developed in four phases:

1. **Foundations**. Collection of raw text without any tagging marks. Creation of machine-readable dictionaries, a robust lexical data base and a general and scalable description of morphology.
2. **Basic tools and applications**. Statistical tools for the morphological tagging of a corpus. Morphological analyzer/generator, lemmatizer/tagger, spelling checker and corrector.
3. **Advanced tools and applications**. Beginning with syntax and semantics. Traditional search machines that integrate lemmatization and language identification. Surface syntax. Grammar and style checkers. Structured versions of dictionaries (these allow enhanced functionality not available for printed or raw electronic versions), creation of a concept taxonomy (e.g.: Wordnet), word sense disambiguation, Computer Aided Language Learning (CALL) systems.
4. **Multilingualism and general applications**. Nowadays this is structured in 3 main layers: Content management, learning and Machine Translation.

Within this general strategy, and after years working on basic resources and tools, we decided it was time to tackle the MT task (Hutchins and Somers, 1992). The languages involved would be Basque, Spanish and English, because of the real necessity of translation in our environment.

The first attempt consisted of the development of an RBMT system to translate from Spanish into Basque (Mayor, 2007), which makes use of the resources previously developed for Basque (such as bilingual dictionaries and morphological analyzers/generators). On the basis of this development, we focused on researching SMT and hybridization of Machine Translation paradigms. We wanted to combine the two basic approaches for MT (rule-based and corpus-based) in order to build a hybrid system, because it was generally agreed that both approaches have limitations that could be overcome with some kind of hybridization.

Data-driven Machine Translation (example-based or statistical) is nowadays the most prevalent trend in Machine Translation research. Translation results obtained with this approach have now reached a high level of accuracy, especially when the target language is English. However, these 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. Unfortunately, large and reliable bilingual corpora are not available for many language pairs.

## I.4 Objectives of this Ph.D. Thesis

This thesis has been developed in the context presented above and our general objective is to improve the quality of MT for Basque. Thus, we will investigate the difficulties found in translating into Basque and investigate different techniques to overcome them. Once we have achieved a minimal quality SMT system, we want to start examining initial hybridization attempts. This general objective can be detailed as follows:

- **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 expect to reduce the number of tokens that

occur only once and, at the same time, to reduce the number of 1-to-n alignments. 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 will try different segmentation options and we will analyze their effects on the translation quality.

- **Implement different techniques to deal with word order differences in statistical machine translation**. We will test different techniques to overcome the errors derived from the great word order differences between the two languages. These techniques cover the most common research, applying them both at decoding (a lexicalized reordering model has been integrated) and pre-processing. At the same time, we will test another two new methods of carrying out this pre-processing, 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 define two different hybridization experiments. In the first experiment, we will translate each sentence using the three systems we have available (SMT, RBMT and EBMT systems) and the most appropriate translation will be chosen for each sentence. In the second experiment, we will build a Statistical Post-Editing system in order to correct 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, to "translate" from the output of the RBMT system to the real target language.

- **Enlarge bilingual corpora**. Basque, as a less-resourced language, has few corpora available and this is one of the biggest obstacles to the success of SMT. Thus, we will make a constant effort to collect as many corpora as possible, in order to overcome this obstacle.

- **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 want to rerun our experiments using corpora from different domains and of different sizes.
- **Carry out a human evaluation based on the HTER metric**. In order to perform a final general evaluation of the work done in this thesis, and taking into account the doubts that have arisen around BLEU (Melamed et al., 2003; Koehn and Monz, 2006), we decided to use HTER (Snover et al., 2006). This evaluation, based on manual post-edition, will allow us to contrast the results obtained by the automatic evaluation based on the BLEU metric. Although the generation of post-edited versions of the MT outputs is expensive, which prevents this kind of evaluation from being carried out at development, the difficulty in interpreting the BLEU scores and the necessity to create several references to get more accurate BLEU scores (we just have one reference in our test set), leads us to consider HTER more reliable and less expensive.

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 dissertation, we present the work performed and the results obtained.

## I.5 Thesis Organization

This Ph.D. thesis dissertation is composed of seven chapters. Following this introductory chapter, the second chapter contains a general overview of the state of the art in Statistical Machine Translation and approaches to combining different MT paradigms. The next three chapters present the thesis contributions. Finally, the last two chapters present a final deeper and homogeneous evaluation of the different techniques developed in the thesis and the conclusions drawn from it.

Outline of the thesis dissertation:

- Chapter II reviews state of the art research on Machine Translation, focusing more specifically on Statistical Machine Translation (Section II.1), hybridization (Section II.4) and evaluation (Section II.5).

- Chapter III presents the work done to adapt a baseline SMT system to carry out translation into a morphologically-rich, agglutinative language, such as Basque. We deal with the agglutinative nature of Basque by splitting words into smaller tokens, which allows a better statistical translation. In order to determine the most appropriate segmentation for a Spanish-Basque system, we explore different segmentation options and analyze their effects on the translation quality.

- Chapter IV deals with the different word orders in Spanish and Basque. We implement three different techniques to deal with these word order differences, which cover the most common techniques (these are applied both at decoding and pre-processing). By combining decoding time techniques and pre-processing techniques we achieve significant improvements.

- Chapter V is devoted to investigating two methods of combining MT paradigms, where the SMT system developed in this thesis is combined with two other systems we have available (RBMT and EBMT systems developed by our own research team). In the first experiment, we translated each sentence using the three systems we have available and the most appropriate translation was chosen for each sentence. Even using such a simple hybridization technique we achieved positive results. In our second attempt, we used Statistical Machine Translation for post-editing the output of the RBMT system. Thus, the SMT system "translates" from the output of the RBMT system to real Basque.

- Chapter VI presents the overall evaluation performed for all the systems developed in this PhD thesis. All the systems are evaluated in the same framework and using a training corpus 7 times larger than those used in the partial evaluations. As well as the automatic metrics used until this point, in this chapter we defined a human evaluation based on HTER.

- Chapter VII draws the main conclusions from this Ph.D. thesis dissertation and details possible future lines of research.

## I.6 Research Contributions

In this section we list all our publications related to this thesis, organizing them according to the dissertation chapters[3]:

**Chapter III**

- Agirre E., Díaz de Ilarraza, Labaka G. and Sarasola K. *Uso de información morfológica en el alineamiento Español-Euskara.* In *Journal of the Spanish Association for Natural Language Processing*. Vol 37, pp. 257-265. 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. Copenhagen, 2007*

- 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 of 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.

**Chapter IV**

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

**Chapter V**

[3]All these papers are available in the web page of the IXA research group (ixa.si.ehu.es/argitalpenak)

- 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, USA, 2008.

- Alegria I., Casillas A., Díaz de Ilarraza A., Igartua J., Labaka G., Laskurain B., Lersundi M., Mayor A., Sarasola K. and Saralegi X. *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, 2008

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

***Other publications***

The following paper are not strictly related to this PhD thesis, but they include other work done on Machine Translation:

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

- 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.

- Alegira 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*. Helsinki, Finland, 2005

# CHAPTER II

# State of the Art

This chapter presents the state of the art on Machine Translation, focusing more specifically on Statistical Machine Translation (Section II.1), hybridization (Section II.4) and evaluation (Section II.5).

## II.1 Statistical Machine Translation

Statistical Machine Translation has been evolving very fast in the last ten years and, specially, in the last three. Koehn (2010) introduces the major established methods in SMT and gives pointers to most of the recent researches.

Statistical machine translation is based on the assumption that every sentence $e$ in a target language is a possible translation of a given sentence $f$ in a source language. The main difference between two possible translations of a given sentence is a probability assigned to each, which is to be learned from a bilingual text corpus. The first SMT models applied these probabilities to words, therefore considering words to be the translation units of the process.

### II.1.1 IBM translation models

Supposing we want to translate a source sentence $f$ into a target sentence $e$, we can follow a noisy-channel approach (regarding the translation process as a channel which distorts the target sentence and outputs the source sentence) as introduced in Brown et al. (1988), defining statistical machine translation as the optimisation problem expressed by:

$$e = \arg\max_{e} Pr(e|f) \tag{II.1}$$

Typically, Bayes rule is applied, obtaining the following expression:

$$e = \arg\max_{e} Pr(f|e)Pr(e) \tag{II.2}$$

Thus, translating $f$ becomes the problem of detecting which $e$ (among all possible target sentences) scores best given the product of two models: $Pr(e)$, the target language model, and $Pr(f|e)$, the translation model. Although it may seem less appropriate to estimate two models instead of just one (considering that $Pr(e|f)$ and $Pr(f|e)$ are equally difficult to estimate), the use of such a target language model justifies the application of Bayes rule, as this model penalise non-grammatical target sentences during the search.

Although language model, typically implemented using Ngrams, was already being used in other fields, such as speech processing, the translation model was first presented in Brown et al. (1993). In order to automatically learn this huge number of parameters, authors used the EM algorithm with increasingly complex models. Those models are widely known as the five IBM models, and are inspired in the generative process described in Figure II.1.

Conceptually, this process states that for each target word, we first find how many source words will be generated (following a model denoted as fertility); then, we find which source words are generated from each target word (lexicon or word translation probabilities); and finally, we reorder the source words (according to a distortion model) to obtain the source sentence.

These models are expressed as:

- $n(\phi|e)$ or Fertility model, which accounts for the probability that a target word $e_i$ generates $\phi_i$ words in the source sentence

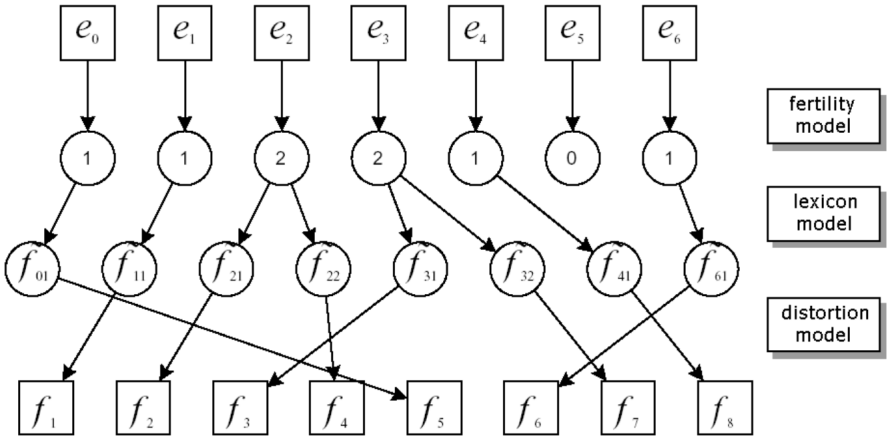



Figure II.1: Illustration of the generative process underlying IBM models.

- $t(f|e)$ or Lexicon model, representing the probability of producing a source word $f_j$ given a target word $e_i$
- $d(\pi|\tau, \phi, e)$ or Distortion model, which models the probability of placing a source word in position j given that the target word is placed in position i in the target sentence.

IBM models 1 and 2 do not include fertility parameters so that the likelihood distributions are guaranteed to achieve a global maximum. Their difference is that Model 1 assigns a uniform distribution to alignment probabilities, whereas Model 2 introduces a zero-order dependency with the position in the source. Vogel et al. (1996) presented a modification of Model 2 that introduced first-order dependencies in alignment probabilities, the so-called HMM alignment model, with successful results. Model 3 introduces fertility and Model 4 and 5 introduce more detailed dependencies in the alignment model to allow for jumps, so that all of them must be numerically approximated and not even a local maximum can be guaranteed.

A detailed description of IBM models and their estimation from a parallel corpus can be found in Brown et al. (1993) and an informal yet clarifying tutorial on IBM models can be found in Knight (1999).

With regard to freely available tools for training and decoding of IBM models: in 1999, the John Hopkins University summer workshop research

team on SMT released GIZA (as part of the EGYPT toolkit), a tool implementing IBM models training from parallel corpora and best-alignment Viterbi search, as reported in Al-Onaizan et al. (1999). This was a breakthrough in that it enabled many other teams to easily join SMT research. In 2001 and 2003 improved versions of this tool were released, and named GIZA++ (Och and H. Ney, 2003).

### II.1.2 Phrase-based Statistical Machine Translation

By the turn of the century it became clear that in many cases specifying translation models at the level of words turned out to be inappropriate, as much local context seemed to be lost during translation. Novel approaches were needed to describe their models according to longer units, typically sequences of consecutive words (or phrases).

The first approach using longer translation units was presented in Och et al. (1999) and named Alignment Templates, which are pairs of generalized phrases that allow word classes and include an internal word alignment. A simplified version of the previous approach is the so-called phrase-based statistical machine translation presented in Zens et al. (2002). Under this framework, word classes are not used (but the actual words from the text instead), and the translation unit looses internal alignment information, turning into so-called bilingual phrases. Mathematically, the new translation model is expressed by:

$$Pr(f_1^J|e_1^I) = \alpha(e_1^I) \cdot \sum_B Pr(\bar{f}_k|\bar{e}_k) \tag{II.3}$$

where the hidden variable $B$ is the segmentation of the sentence pair in $K$ bilingual phrases $(\bar{f}_1^K\ ,\ \bar{e}_1^K\ )$, and $\alpha(e_1^I)$ is assuming the same probability for all segmentations.

The phrase translation probabilities are usually estimated, over all bilingual phrases in the corpus, by relative frequency of the target sequence given the source sequence, as in:

$$Pr(\bar{f}_k|\bar{e}_k) = \frac{N(\bar{f}_k, \bar{e}_k)}{N(\bar{e}_k)} \tag{II.4}$$

where bilingual phrases are defined as any pair of source and target phrases that have consecutive words and are consistent with the word alignment matrix. According to this criterion, any sequence of consecutive source words and consecutive target words which are aligned to each other and not aligned to any other token in the sentence, become a phrase. This is exemplified in Figure II.2, where eight different phrases are extracted, but it is worth noting that $AB \rightarrow WY$ is not extracted, given the definition constraint. For more details on this criterion, see Och et al. (1999) or Zens et al. (2002).

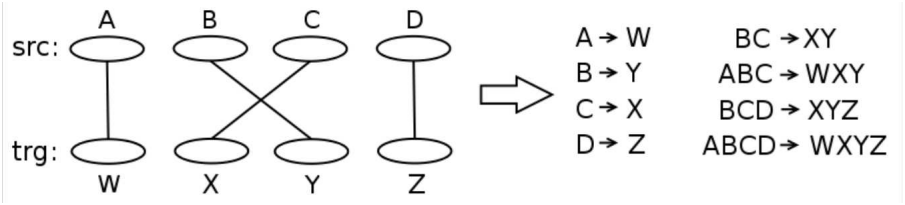



Figure II.2: Phrase extraction from a certain word aligned pair of sentences.

### II.1.3 Feature-based model combination

Another alternative to the noisy-channel approach is to directly model the posterior probability $Pr(e_1^I|f_1^I)$, a well-founded approach in the framework of maximum entropy, as shown in Berger et al. (1996). By treating many different knowledge sources as feature functions, a log-linear combination of models can be performed, allowing an extension of a baseline translation system with the addition of new feature functions. In this case, the decision rule responds to the following expression:

$$\hat{e}_1^I = \arg\max_{e_1^I} \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \tag{II.5}$$

so that the noisy-channel approach can be obtained as a special case if we consider only two feature functions: the target language model $h_1(e_1^I, f_1^J) = log\ p(e_1^I)$ and the translation model of the source sentence given the target $h_2(e_1^I, f_1^J) = log\ p(f_1^J|e_1^I)$.

Typically, this log-linear combination includes, apart from a translation model, other feature functions, such as:

- sentence length models, also called word bonuses
- lexical models (such as IBM model 1 from source to target and from target to source)
- phrase penalties
- others (regarding information on manual lexicon entries or other grammatical features)

In order to determine the weight of each model in the featured-based combination, the Minimum Error Rate training is widely used. This approach, which was introduced in Papineni et al. (1998) for a natural language understanding task, suggests that the training optimization task becomes finding out the $\lambda_m$ which weight each model according to a certain criterion. In Och and Ney (2002), minimum error training is introduced for statistical machine translation, stating that these weights need to be settled by directly minimizing the translation error on a development set, as measured by a certain automatic measure.

Nowadays, the state-of-the-art SMT system uses a log-linear combination of feature models, optimized according to a certain automatic measure on the development data.

The availability of many free tools makes it easier for a beginner to become quickly acquainted with phrase-based SMT, and even run preliminary experiments in one day. Probably the most widely used of these tools is Moses (Koehn et al., 2007), a complete SMT toolkit which enables you to develop a state-of-the-art SMT system.

### II.1.4 Use of linguistic knowledge in SMT

Although SMT systems did not initially incorporate any linguistic analysis and worked at the surface level of word forms, an increasing number of research efforts are introducing a certain degree of linguistic knowledge into their statistical framework.

At this point, the pair of languages involved and their respective linguistic properties are crucial to justify a certain approach and explain its results. Therefore, the idea that a good statistical translation model for a certain

pair of languages can be used for any other pair is faced against the view that the goodness of such a model may be, at least in part, dependent on the specific language pair. Of course, conclusions will easily hold for languages sharing many linguistic properties.

Even for the same language pair some modifications can entail some improvement in a translation direction and no in the opposite. For example, a certain vocabulary reduction of the French may be useful at translating into English, since many French words may translate to the same English word (due to morphological derivations which are not present in English), but the same technique can be useless when translating from English to French.

Many researchers have tried to use morphological information in improving machine translation quality. In Koehn and Knight (2003), the authors achieved improvements by splitting compounds in German. Nießen and H. Ney (2004) achieved a similar level of alignment quality with smaller corpora by restructuring the source based on morpho-syntactic information when translating from German to English.

An alternative approach to integrate the treatment of morphology into the SMT decoder (instead of treating it by means of pre-processing and post-processing stages), consists in the use of the factored models integrated in Moses (Koehn and Hoang, 2007). From a search perspective it is desirable to integrate these pre-processing and post-processing stages into one model.

Integrated search makes it easier to find the global optimal translation, which is less likely to be found when passing along one-best or n-best choices between the stages.

In the factored model approach, each text token is tagged in different levels (or factors), such as lemma, PoS or morphological tags. This way, the translation can be carried out for each of those factors independently, and then, in a final generation step, the information about the different factors is combined to generate the final translation. However, due to implementation issues, it is required that all steps operate in the same phrase segmentation, so that translation options for all input phrases can be efficiently precomputed in an expansion process. And that requirement of operating in the same phrase segmentation reduces the degree of generalization of this approach.

Factored translation models have been used to translate morphologically rich languages, such us Czech (Bojar, 2007) or German (Holmqvist et al., 2007). They have been used to integrate not only morphological information,

but also other kind of linguistic information, such as PoS language models (Koehn and Hoang, 2007) or many other kinds of syntactic information (Birch et al., 2007; Avramidis and Koehn, 2008).

However, due to efficiency problems, the use of complex paths in factored models may be become unmanageable. This way, the experiments that have deal with rich morphology by means of complex translation paths have been carried out using reduced corpora. For example, interesting improvements have been obtained for German-English translation (Koehn and Hoang, 2007) in a one-million-word corpus, but when similar decoding paths have been trained on the Europarl corpus (Holmqvist et al., 2007) the decoding time turned out to be unmanageable.

Regarding to syntax, and still for the German-to-English translation, a sentence reordering as pre-processing is presented in (Collins et al., 2005). They define a small amount of rules to reorder verbal clauses in German, obtaining a English-like word order. In this way, they get a significant improvement both in BLEU and human judgments.

In addition, a number of researchers have proposed other translation models where the translation process involves syntactic representations of the source and/or target languages. These models have radically different structures and parameterization from phrase-based models for SMT. Without aiming at completeness, some of the relevant works are mentioned here.

An approach to phrasal SMT based on a parsed dependency tree representation of the source language is introduced in (Quirk et al., 2005). This approach, named Treelet translation, uses a source dependency parser and projects a target dependency tree using word alignment. After this projection, tree-based phrases are extracted and a tree-based ordering model can be trained.

Related to that, hierarchical phrases (Chiang, 2005) also remove the limitation to contiguous phrases and allow phrases to include indexed placeholders, thus turning phrase-based SMT into a parallel parsing problem over a grammar with one non-terminal symbol. This improves the global reordering search.

### II.1.5 SMT for Basque

Regarding previous work on SMT for Basque there are some systems that deserve to be mentioned. First of all, there are a few works that directly deal with the creation of bilingual corpora (Nevado et al., 2004; Casillas et al., 2007) and with their alignment (Martínez et al., 1998; Martínez et al., 1998; Casillas et al., 2000).

Pérez et al. (2008) created an Spanish-Basque SMT system, but in the context of speech translation and for a very reduced domain. They carried out the translation from Spanish speech into Basque text, making use of an integrated architecture of stochastic finite-state transducers. The models were assessed under a very restricted domain task (short descriptions of weather forecast), and thus the corpus was very repetitive and had a reduced lexicon. It consists of 14,615 sentences with an average length of 13.1 words in Spanish and 12.8 in Basque. The amount of words in the lexicon was only 702 for Spanish and 1135 for Basque, out of which 162 were singletons, words with a single occurrence in Spanish and 302 in Basque. There are some previous contributions related to the model of stochastic finite-state transducer presented in (Pérez et al., 2008). Ortíz et al. (2003) presented some first experiments on the use of SMT that revealed the complexities of translating from Spanish into Basque, and the need for better alignment methods. González et al. (2004) compared different SMT approaches: IBM translation models, Phrase-Based alignment models, Pharaoh, and the architecture called GIATI that, based stochastic finite-state transducers. Phrase based models presented the best results in experiments performed with the mentioned corpura: the above mentioned corpus related to weather forecast, a tourist corpus, and a corpus of administrative documents. The tourist corpus was artificially created. It was an adaptation to Basque of a series of Spanish-German grammars that generated sentences pairs for both languages. Sanchís and Casacuberta (2007) evaluates reordering via n-best list for Spanish-Basque translation. The results showed to be promising, but again the corpus used in training and evaluation was the semi-synthetic small corpus presented in (González et al., 2004). As future work the author reported they were planning on obtaining results with other non-synthetic, richer and more complex corpora.

## II.2 Rule-Based Machine Translation

Rule-based MT (RBMT) systems use knowledge in the form of rules explicitly coded by human experts that try to describe the translation process. This kind of MT system relies heavily on linguistic knowledge such as morphological, bilingual dictionaries (containing lexical, syntactic and even semantic information), part-of-speech (PoS) disambiguation rules or manually disambiguated corpora, and a large set of rules. The process of building a RBMT system involves a huge human effort for building the necessary linguistic resources.

Generally, RBMT systems work by parsing (or analyzing) the source language text, usually creating an intermediate (symbolic) representation, from which the text in the target language is generated. According to the nature of the intermediate representation used, a RBMT system may be said to be either an interlingua or a transfer-based MT system.

An interlingua MT system uses a single, language-independent intermediate representation. The advantage of using a language-independent intermediate representation is that no bilingual information (dictionaries or rules) are needed; as a disadvantage we have that the definition of a language-independent intermediate representation is very difficult, perhaps impossible for open-domain translations.

In transfer-based MT the intermediate representation depends on the languages involved in the translation. These systems usually work by applying a set of structural transfer rules to the source language intermediate representation created during the analysis in order to transform it into the target language intermediate representation from which the target language text is finally generated. The level of analysis, and therefore the degree of abstraction provided by the intermediate representation, varies depending on how closely related the languages involved are. Translating between "distant" languages (such as English-Japanese) requires a deep analysis (syntactic and semantic), while translation between closely related languages (for example between Romance languages) can be achieved with shallow parsing.

Although during the last few years the growing availability of machine-readable monolingual and parallel corpora has led to an increased interest in corpus-based approaches, RBMT systems are still being actively developed, mainly because:

1. Corpus-based MT systems require large parallel corpora, in the order of tens of millions of words, to achieve a reasonable translation quality in open-domain tasks. Such vast parallel corpora are not available for most less-resourced language pairs demanding MT services, such as Spanish-Basque, French-Catalan or English-Afrikaans, among others.

2. RBMT systems are easier to diagnose during development and the translation errors they produce usually have a repetitive nature, making them more predictable and easier to post-edit, and therefore, better suited for dissemination purposes.

There have been several attempts to build general RBMT system for the Spanish-Basque language pair, but there are only three systems currently available and useful: *Matxin*[1], the main RBMT system developed at the University of the Basque Country (Mayor, 2007), which translates from Spanish to Basque, the system created by *AutomaticTrans* available in the website of the *Instituto Cervantes*[2], and *Erdaratu*[3] that is a prototype (Ginestí-Rosell et al., 2009) to translate in the opposite sense, from Basque to Spanish. Matxin performs deep transfer translation and Erdaratu operates following a shallow transfer model. All the three systems can be used in the web, and two of them (Matxin and Erdaratu) are free/open systems.

There follows a description of Matxin, the RBMT system used in the hybrid systems developed in this PhD thesis.

### II.2.1 Matxin: a Spanish-Basque RBMT system.

Matxin is an open-source RBMT engine, whose main goal is to translate from Spanish into Basque using the traditional transfer model. Matxin consists 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 below.

[1] http://www.opentrad.org
[2] http://oesi.cervantes.es/traduccionAutomatica.html
[3] http://www.erdaratu.eu

#### II.2.1.1 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). 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 and 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.

#### II.2.1.2 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. The bilingual dictionary is based on the Elhuyar wide-coverage dictionary. This dictionary was enriched with named entities and terms automatically extracted from parallel corpora. 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., 2006).

#### II.2.1.3 Generation

Generation, like transfer, is decomposed into two steps. The first step, referred to as syntactic generation, consists in 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 in 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 order to perform morphological generation, the morphological generator for Basque described in (Alegria et al., 1996) is used. This generator makes use of the morphological dictionary developed in Apertium, which establishes correspondences between surface forms and lexical forms for Basque. 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).

## II.3 Example-Based Machine Translation

The idea of EBMT dates form the early 80s (Nagao, 1984). The essence of EBMT, called "machine translation by example-guided inference, or machine translation by the analogy principle" by Nagao, is succinctly captured by his much quoted statement:

> "Man does not translate a simple sentence by doing deep linguistic analysis, rather, man does translation, first, by properly decomposing an input sentence into certain fragmental phrases, ... then by translating these phrases into other language phrases, and finally by properly composing these fragmental translation into one long sentence. The translation of each fragmental phrase will be done by the analogy translation principle with proper examples as its reference." (Nagao, 1984:178f.)

Nagao correctly identified the three main components of EBMT, namely (i) matching fragments against a database of real examples, (ii) identifying the corresponding translation fragments, and (iii) recombining these to give the target text.

Just as Somers did (Carl and Way, 2003, chapter 1), it is instructive to take the familiar pyramid diagram and superimpose the task of EBMT (Figure II.3). The source text analysis in conventional MT is replaced by the

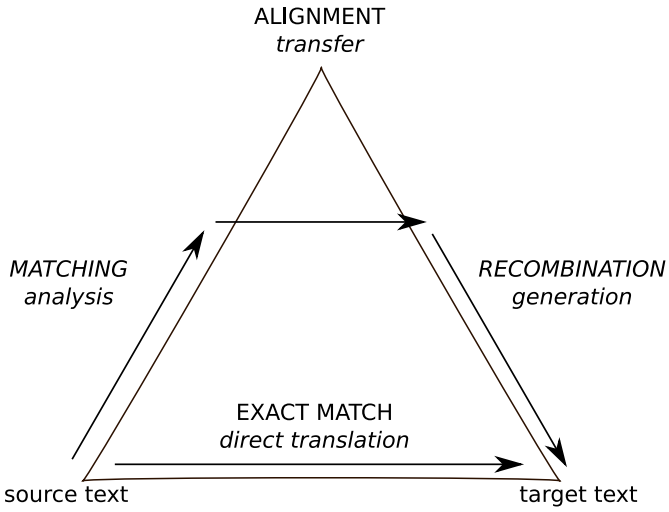



Figure II.3: The 'Vauquois pyramid' adopted for EBMT(Sommers, 2003)

**matching** of the input against the example set. Once the relevant examples have been selected the corresponding fragments in the target test must be selected. This has been termed **alignment** and, like transfer in conventional MT, involves contrastive comparison of both languages. Once the appropriate fragments have been selected, they must be **combined** to form a legal target text, just as for the generation stage in conventional MT.

An historical introduction to EBMT, up to 2003, is given in Sommers (2003). It is included in the general overview of the EBMT research given by Carl and Way (2003).

Current progress in EBMT has recently been reflected in the 3rd Workshop on Example-Based Machine Translation. Its organizers recognize a decline in EBMT research, and that SMT has almost completely taken over the corpus-based machine translation arena, with many EBMT practitioners moving into hybrid approaches integrating EBMT with other approaches, mostly (but not only) SMT. But, despite of this, EBMT practitioners are working about putting together their tools, their engines and their data and releasing them under open licenses to extend their use both in academia and industry.

In the same way as it has been done for SMT, new free/open-source EBMT software is being developed. Cunei MT platform (Phillips and Brown,

2009) is distinguished from a traditional SMT system in that it delays as much computation as possible until run-time; in particular, translations are not retrieved from a wholly pre-built phrase table, but rather generated at run-time working dynamically with the bilingual corpus. Besides, an extension of the widely used Freeling analyser suite for EBMT (Farwell and Padró, 2009) has been proposed, and also a tool for sampling-based alignment (Lardilleux et al., 2009).

Inside what is called "Pure EBMT" the three main current trends are the use of proportional analogies (Somers et al., 2009; Lepage and Denoual, 2005), the extension of memory-based translation formalisms to deal with phrases (van Gompel et al., 2009), and the use of a top-down transfer strategy for EBMT (Vandeghinste and Martens, 2009).

### II.3.1 Example-based Machine Translation for Basque

The Example-Based Machine Translation system created to translate from Spanish to Basque (Alegria et al., 2008) is based on the use of translation patterns representing generalizations of sentences that are translations of one another, replacing various sequences of one or more words with variables (McTait and Trujillo, 1999).

Starting from the aligned corpus, the following steps were carried out to automatically extract translation patterns. The basic idea is that, first, the system detects a number of concrete units (mainly named entities) in the aligned sentences and then, these units are replaced with variables. In order to detect the units, and due to the morphosyntactic differences between Spanish and Basque, specific algorithms are executed for each language. In the system, algorithms to determine the boundaries of dates, numbers, named entities, abbreviations and enumerations were developed.

After detecting the units, they must be aligned, pairing up Spanish and Basque units of the same type and 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 they use their canonical form.

Once all variables were aligned, it was necessary to extract the Basque morphemes for named entities in Basque translation patterns. To do this,

they use the morphosyntactic analyzerAduriz and Díaz de Ilarraza (2003), replacing each named entity by its corresponding lemma and suffixes (declension case). Thus, when a translation is proposed, the Basque translation pattern had to be modified according to the new characteristics of the named entity that was in the Spanish source sentence. In other words, when a <rs> tag appears in the Basque translation pattern this tag is replaced by the new content of the Spanish <rs> tag, and declined according to the declension case stored in the Basque translation pattern. For example in Table II.1 which shows the process of pattern extraction,**-<ERG-S-M>** represents the suffix for the declension case ergative (with number-determination in singular(S)-determinate(M)), and **-<ERG-S-M>** represents the suffix for the declension case inesive

| **ES-EU sentences** | **Sentences with generalized units** | **Morpheme extraction** | **Translation Pattern** |
|---|---|---|---|
| El Departamento de Educación ha decidido lo siguiente el 5 de noviembre de 2006 | <rs type=org> El Departamento de Educación </rs> ha decidido lo siguiente el <date date=05/11/2006> 5 de noviembre de 2006 </date>. | | <rs1> ha decidido lo siguiente el <date1> |
| 2006ko azaroaren 5ean honakoa erabaki du Hezkuntza-Sailak | <date date=05/11/2006> 2006ko azaroaren 5ean </date> honakoa erabaki du <rs type=org> Hezkunta-Sailak </rs>. | <date date=05/11/2006> 2006ko azaroaren 5 </date> -<INE-S-M> honakoa erabaki du <rs type=org> Hezkunta-Saila </rs> -<ERG-S-M>. | <date1> -<INE-S-M> honakoa erabaki du <rs1> -<ERG-S-M>. |

Table II.1: Example of Translation Pattern extraction

The number of translation patterns extracted varies greatly from one

corpus to another, depending on the relative frequency of those patterns. For example, when using a corpus of a restricted domain such as labor agreements (where there is a lot of very similar sentences) the number of patterns they extracted was bigger than when they used a general domain corpus.

Once the system has automatically extracted all the possible translation patterns from the training set, the patterns are stored in a hash table for use in the translation process. When a source sentence has to be translated, the system checks the hash table, looking for patterns matching the sentence. If the source sentence matches a pattern without variables, the translation process will immediately return its translation. A Word Error Rate (WER) metric is used to compare the two sentences. Otherwise, if the source sentence does not match any pattern in the hash table, the translation process will try to generalize the sentence, and then check the hash table again looking for a generalized template. To generalize the source sentence, the translation process will apply the same detection algorithms used in the process for extracting patterns from the corpus.

## II.4 Hybrid Approaches

There have been several attempts to combine different MT approaches in order to improve translation results. Most of those attempts can be classified in one of the following categories (see Figure II.4):

- Enrichment of translation resources: This kind of hybridization in the main translation process is carried out by a common MT architecture, but the resources available are obtained (or enriched) by means of other MT techniques. For example, in this category we can find RBMT systems whose lexicon or grammars are enriched using bilingual corpora (Dugast et al., 2009; Sánchez-Martínez et al., 2009) or SMT systems whose phrase tables are modified using RBMT systems (Stroppa and Way, 2006; Tyers, 2009).

- Multi-Engine hybridization: Multi-Engine systems directly use different MT systems to translate each sentence and, after these first translations, the distinct outputs are combined in order to obtain a final translation. The simplest Multi-Engine systems directly select the most

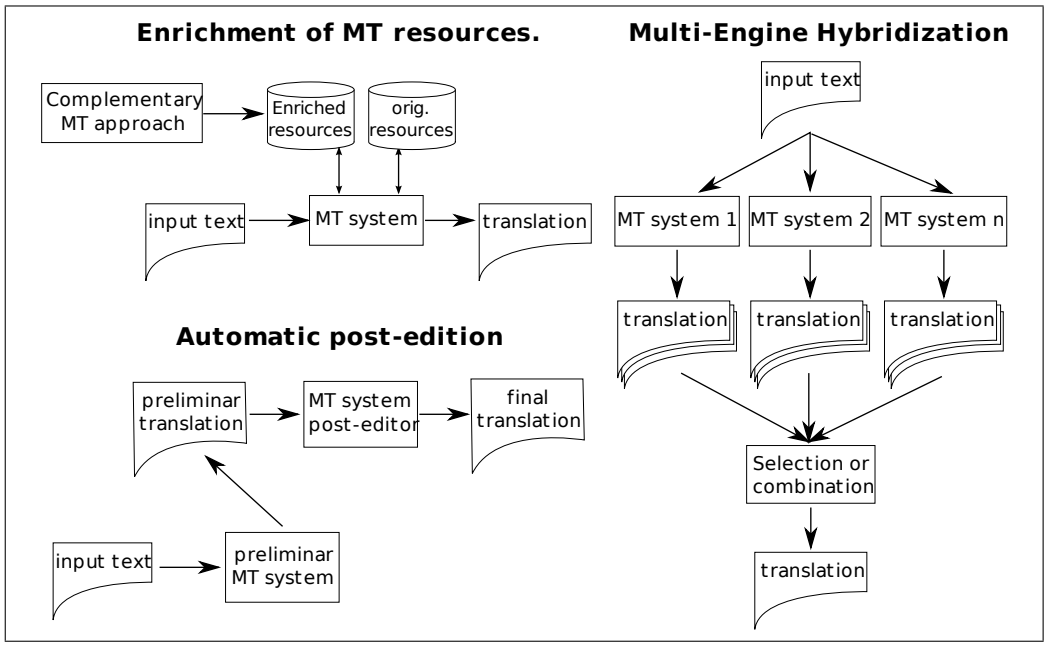



Figure II.4: Overview of Major Hybrid Architectures

appropriate output from the ones it has available (Eisele, 2005; Chen et al., 2009; Du et al., 2009).

- Automatic Post-Editing: This hybridization consists in using MT techniques to correct the output of a different MT system. This way, you could use rules to correct SMT output (Elming, 2006) or train a full SMT system to post-edit the output of RBMT systems so that the target-language output is corrected (Simard et al., 2007a; Isabelle et al., 2007).

In this work, we have experimented with Multi-Engine hybridization and Automatic Post-Editing. Thus, we now present an overview of the work done in those areas.

### II.4.1 Multi-Engine combination

Combinations of MT systems into multi-engine architectures have a long tradition, starting perhaps with Frederking and Nirenburg (1994) and being well

stated in Rosti et al. (2007). Multi-engine systems can be roughly divided into *simple architectures*, on the one hand, which try to select the best output from a number of systems doing no modification in those individual hypotheses (Tidhar and Küssner, 2000; Callison-Burch and Flournoy, 2001; Akiba et al., 2002; Eisele, 2005), and *more sophisticated setups*, on the other hand, which try to recombine the best parts in each of the multiple hypotheses into a new utterance that can be better than the best of the given candidates, as described in Macherey and Och (2007); Chen et al. (2009); Leusch et al. (2009); Du et al. (2009).

In those researches where an individual hypothesis is selected from the output of a number of systems, the main issue consists in defining the suitable selection criteria. Those criteria use different metrics to measure the adequacy of each candidate sentence, selecting the most fluent translation or the one that best suits the source sentence. Alternatively, some other researches measure the level of agreement among the candidate translations and then selects the one that is closest to the consensus.

Recombining multiple MT results requires finding the correspondences among alternative renderings of a source language expression proposed by different MT systems. Like the systems that directly select one of the candidates, a recombination system also needs a way to chose the best combination of alternative building blocks.

### II.4.2 Automatic Post-Editing

In those cases where the MT system is used to create a preliminary version that will be published after a post-edition process, those post-editions could be collected generating a corpus of post-editions. Making use of this kind of corpora, some researchers have developed systems that automatically post-edit the output of the MT systems.

Most of those researches (Simard et al., 2007a; Isabelle et al., 2007) use an SMT system to learn to post-edit the output of RBMT systems. Incorporating, in such a way, some of the benefits of the SMT (bigger fluency and more accurate lexical selection) to the translation generated by a RBMT system.

A different approach is used in Elming (2006). In this paper, instead of using an SMT system as automatic post-editor, the author employs a transformation-based learning (TBL) algorithm (Lager, 1999) for extracting

rules to correct the RBMT output by means of a post-processing module. The translation resulting after post-edition achieves a relative improvement of 4.6% compared to the original RBMT system.

### II.4.3 MaTrEx: EBMT-SMT hybrid MT system

As a result of the collaboration with the National Centre for Language Technology (NCLT) in Dublin, we have incorporated Matrex (Stroppa and Way, 2006) in the final evaluation of this PhD thesis (Chapter VI). Thus, we now present a brief description of MaTrEx a data-driven MT system which combines both EBMT and SMT techniques. MaTrEx consists of a number of extensible and re-implementable modules, including the word alignment, chunking, chunk alignment and decoder modules. The word alignment, chunking and decoder modules are wrappers around existing tools, namely Giza++ (Och and H. Ney, 2003) and Moses (Koehn et al., 2007).

The translation process can be decomposed as follows: the aligned source-target sentences are passed to the word alignment module, chunking module and the chunk alignment module in turns, 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 II.5.

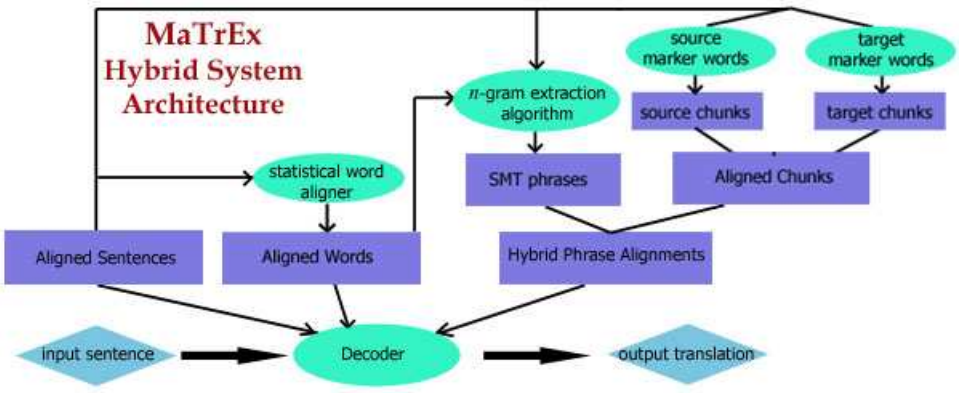



Figure II.5: General design of the Matrex system (Stroppa and Way, 2006).

Word alignment is performed using the Giza++ statistical word alignment toolkit and followed by the Koehn's "refined" method (Koehn et al.,

2003) to extract a set of high-quality word alignments from the original unidirectional alignment sets. These are passed, along with the extracted chunk alignments, to the translation decoder.

In order to align the chunks obtained by the chunking procedures, an "edit-distance style" is used as described in Stroppa and Way (2006). This is a dynamic programming alignment algorithm which works as follows: First, a "similarity" measure is determined for each pair of source-target chunks. Then, given these similarities, a modified version of the edit-distance alignment algorithm is used to find the optimal alignment between the source and the target chunks. The modification consists in 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 may differ considerably.

In order to compute the "similarity" between a pair of chunks, the system makes use of information contained within the chunks. More precisely, chunks are related by using the word-to-word probabilities previously 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 and Way, 2006).

MaTreX also makes use of SMT phrasal alignments, which are added to the aligned chunks extracted by the chunk alignment module. Combining phrases from EBMT and SMT in this way, the system is able to create hybrid data-driven systems that outperform the baseline systems from which they are derived, as shown in Groves and Way (2005). The SMT phrasal alignment follows Koehn's classic procedure (Koehn et al., 2003).

The decoding module 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 and Ney, 2002). The BLEU metric (Papineni et al., 2002) is optimized on a development set. Matrex, as SMT decoders do, uses a log-linear combination of several feature functions. The most common ones are the following: phrase translation probabilities (in both directions), word-based translation probabilities (lexicon model, in both directions), a phrase length penalty, and a target language model.

As a result of some collaboration between NCLT and Ixa (Stroppa et al., 2006; Labaka et al., 2007), we developed two new wrappers that allow Ma-

trex to deal with Basque. Those new wrappers, around Freeling (Carreras et al., 2004) and Eustagger (Aduriz and Díaz de Ilarraza, 2003), are used to carry out chunking for Spanish and Basque before applying the usual Matrex process (alignment of chunks and their incorporation into the SMT phrase-table).

## II.5 Evaluation in Machine Translation

It is well known that Machine Translation is a very hard task to evaluate automatically. Usually, this task is performed by producing some kind of similarity measure between the translation hypothesis and a set of human reference translations, which represent the expected solution of the system.

The fact that there are several correct alternative translations for any input sentence adds complexity to this task, and although the higher the correlation with the human references the better the quality, theoretically we cannot guarantee that poor correlation with the available set of references means poor translation quality, unless we have all possible correct translations available.

Therefore, in general, it is accepted that all automatic metrics comparing hypotheses with a limited set of manual reference translations are pessimistic. However, instead of an absolute quality score, automatic measures are claimed to capture progress during system development and to correlate well statistically with human intuition.

Next the most widely-used MT evaluation measures are introduced, including BLEU, NIST, mWER, mPER and HTER.

### II.5.1 Lexical Similarity-Based Automatic Evaluation Metrics

#### II.5.1.1 BLEU score

The most widely-used evaluation measure, BLEU (acronym for BiLingual Evaluation Understudy), was introduced by IBM in Papineni et al. (2002), and is usually referred to with a given n-gram order (BLEUn, n usually being 4).

The metric works by measuring the n-gram co-occurrence between a given translation and the set of reference translations and then taking the weighted geometric mean. BLEU is specifically designed to approximate human judgement on a corpus level and can perform badly if used to evaluate the quality of isolated sentences.

$BLEU_n$ is defined as:

$$BLEU_n = \exp\left(\frac{\sum_{i=1}^{n} bleu_i}{n} + length\_penalty\right) \tag{II.6}$$

where $bleu_i$ and $length\_penalty$ are cumulative counts (updated sentence by sentence) referred to the whole evaluation corpus (test and reference sets). Even though these matching counts are computed on a sentence-by-sentence basis, the final score is not computed as a cumulative score, ie. it is not computed by accumulating a given sentence score.

Equations II.7 and II.8 show $bleu_n$ and $length\_penalty$ definitions, respectively:

$$bleu_n = \log\left(\frac{Nmatched_n}{Ntest_n}\right) \tag{II.7}$$

$$length\_penalty = \min\left\{0, 1 - \frac{shortest\_ref\_length}{Ntest_1}\right\} \tag{II.8}$$

Finally, $Nmatched_i$, $Ntest_i$ and $shortest\_ref\_length$ are also cumulative counts (updated sentence by sentence), defined as:

$$Nmatched_i = \sum_{n=1}^{N} \sum_{ngr \in S} \min\left\{N(test_n, ngr), \max_r\{N(ref_{n,r}, ngr)\}\right\} \tag{II.9}$$

where $S$ is the set of Ngrams of size $i$ in sentence $test_n$, $N(sent, ngr)$ is the number of occurrences of the Ngram $ngr$ in sentence $sent$, $N$ is the number of sentences to evaluate, $test_i$ is the $i^{th}$ sentence of the test set, $R$ is the number of different references for each test sentence and $ref_{n,r}$ is the $r^{th}$ reference of the $n^{th}$ test sentence.

$$Ntest_i = \sum_{n=1}^{N} length(test_n) - i + 1 \tag{II.10}$$

$$shortest\_ref\_length = \sum_{n=1}^{N} \min_{r}\{length(ref_{n,r})\} \tag{II.11}$$

Note that slight variations of these definitions have led to alternative versions of BLEU score, although in the literature is considered BLEU as a unique evaluation measure and no distinction is made between versions.

Several doubts have recently emerged around *BLEU*, which has become the most commonly used evaluation metric in the last decade. In addition to the fact that it is extremely difficult to interpret what is being expressed in *BLEU* (Melamed et al., 2003), improving *BLEU* does not guarantee an improvement in the translation quality (Callison-Burch et al., 2006) and it does not offer as much correlation with human judgement as was believed (Koehn and Monz, 2006).

#### II.5.1.2 NIST score

The NIST evaluation metric, introduced in Doddington (2002), is based on the BLEU metric, but with some alterations. Whereas BLEU simply calculates n-gram precision, considering each n-gram to be of equal importance, NIST calculates how informative a particular n-gram is, and the rarer a correct n-gram is, the more weight it will be given. NIST also differs from BLEU in its calculation of the brevity penalty, and small variations in translation length do not have as much impact on the overall score.

Again, the NIST score should always be referred to with a given n-gram order ($NIST_n$ , with $n$ usually being 4), and it is defined as:

$$NIST_n = \left( \sum_{i=1}^{n} nist_i \right) \cdot nist\_penalty \left( \frac{test_1}{\frac{ref_1}{R}} \right) \tag{II.12}$$

where $nist_n$ and $nist\_penalty(ratio)$ are cumulative counts (updated sentence by sentence) referred to the whole evaluation corpus (test and reference

sets). Even though these matching counts are computed on a sentence-by-sentence basis, the final score is not computed as a cumulative score.

The ratio value computed using $test_1$, $ref_1$ and $R$ shows the relation between the number of words of the test set ($test_1$) and the average number of words of the reference sets ($ref_1/R$). In other words, the relation between the number of words translated and the expected number of words for the whole test set.

Equations II.13 and II.14 show $nist_n$ and $nist\_penalty$ definitions, respectively.

$$nist_n = \frac{Nmatch\_weight_n}{Ntest_n} \tag{II.13}$$

$$nist\_penalty(ratio) = \exp\left(\frac{\log(0.5)}{\log(1.5)^2} \cdot \log(ratio)^2\right) \tag{II.14}$$

Finally, $Nmatch\_weight_i$ is also a cumulative count (updated sentence by sentence), defined as:

$$Nmatch\_weight_i = \sum_{n=1}^{N} \sum_{ngr \in S} (\min\{N(test_n, ngr), max_r\{N(ref_{n,r}, ngr)\}\} \cdot weight(ngr)) \tag{II.15}$$

where $weight(ngr)$ is used to weight every n-gram according to the identity of the words it contains, expressed as follows:

$$weight(ngr) = \begin{cases} -\log_2\left(\frac{N(ngr)}{N(mgr)}\right) & \text{if mgr exists;} \\ -\log_2\left(\frac{N(ngr)}{N(words)}\right) & \text{otherwise;} \end{cases} \tag{II.16}$$

where $mgr$ is the same N-gram of words contained in $ngr$ except for the last word. $N(ngram)$ is the number of occurrences of the Ngram $ngram$ in the reference sets. $Nwords$ is the total number of words of the reference sets.

The NIST score is a quality score ranging from 0 (worst translation) to an unlimited positive value. In practice, this score ranges between 4 and 12, depending on the difficulty of the task (languages involved and test set length).

From its definition, we can conclude that NIST favours those translations that have the same length as the average reference translation. If the translation provided is perfect but 'short' (for example, it is the result of choosing the shortest reference for each sentence), the resultant NIST score is much lower than another translation with a length more similar to that of the average reference.

#### II.5.1.3 mWER score

Word Error Rate (WER) is a standard speech recognition evaluation metric, where the problem of multiple references does not exist. For translation, its multiple-reference version (mWER) (Nießen et al., 2000) is computed on a sentence-by-sentence basis, so that the final measure for a given corpus is based on the cumulative WER for each sentence.

$$mWER = \frac{\sum_{n=1}^{N} WER_n}{\sum_{n=1}^{N} Avrf\_Ref\_Lenght_n} \cdot 100 \tag{II.17}$$

where $N$ is the number of sentences to be evaluated. Assuming we have $R$ different references for each sentence, the $average\_reference\_length$ for a given sentence $n$ is defined as:

$$Avrg\_Ref\_Lenght_n = \frac{\sum_{r=1}^{R} Length(Ref_{n,r})}{R} \tag{II.18}$$

Finally, the $WER$ cost for a given sentence $n$ is defined as:

$$WER_n = \min_r LevDist(Test_n, Ref_{n,r}) \tag{II.19}$$

where $LevDist$ is the Levenshtein Distance between the test sentence and the reference being evaluated, assigning an equal cost of 1 for deletions, insertions and substitutions. All lengths are computed in number of words.

mWER is a percentage error metric, thus defined in the range of 0 to 100, with 0 meaning the *perfect* translation (matching at least one reference for each test sentence).

From the description of mWER, we can conclude that the score tends to slightly favour shorter translations over longer translations. This can be explained by considering that the absolute number of errors (found as the Levenshtein distance) is being divided by the average sentence length of the references, so that a mistake of one word for a long reference is overweighted in contrast to a mistake of one word for a short reference.

#### II.5.1.4 mPER score

Similar to WER, the so-called Position-Independent Error Rate (mPER) (Tillmann et al., 1997) is computed on a sentence-by-sentence basis, so that the final measure for a given corpus is based on the cumulative PER for each sentence. This is expressed as follows:

$$mPER = \frac{\sum_{n=1}^{N} PER_n}{\sum_{n=1}^{N} Avrg\_ref\_Length_n} \cdot 100 \tag{II.20}$$

where $N$ is the number of sentences to be evaluated. Assuming we have $R$ different references for each sentence, the $average\_reference\_length$ for a given sentence $n$ is defined as in equation II.18.

Finally, the $PER$ cost for a given sentence n is defined as:

$$PER_n = min_r(Pmax(Test_n, Ref_{n,r})) \tag{II.21}$$

where $Pmax$ is the greater of:

- POS = number of words in the REF that are not found in the TST sentence (recall)
- NEG = number of words in the TST that are not found in the REF sentence (precision)

In this case, the number of words includes repetitions. This means that if a certain word appears twice in the reference but only once in the test, then POS=1.

### II.5.2 Linguistically Informed Similarity

Due to the doubts surrounding BLEU and other metrics that are based on the lexical similarity between the automatically generated translation and a set of reference sentences, some studies have tried to define new metrics to measure the quality of translations by means of dealing with deeper linguistic information. Among these new metrics, we can find those dealing with either constituent-based or dependency-based parsing (Liu and Gildea, 2005; Amigó et al., 2006; Mehay and Brew, 2007), and those that take into account semantic information (Giménez and Màrquez, 2007) such as semantic roles and named entities. In most of these researches, linguistic representations of the hypothesis and the references are constructed making use of available analyzers, and their similarity is computed by means of a matching function, or a distance measure such as cosine.

That way, Giménez and Màrquez (2007) compared the behavior of a wide set of metrics for automatic MT evaluation at different evaluation levels (lexical, shallow syntactic, syntactic, and shallow semantic) and under different scenarios (single-reference scenario and multiple-reference scenario). Based on the obtained results, the authors concluded that linguistic features at a more abstract level may provide more reliable rankings of the systems, specially when the systems under evaluation do not share the same lexicon.

Nevertheless, none of current metrics provides, in isolation, a global measure of quality; indeed, all metrics focus on partial aspects. The main problem of relying on partial metrics is that we may obtain biased evaluations, which may lead us to reach inaccurate conclusions. So that, in order to perform more robust, i.e. less biased, automatic MT evaluations, different quality dimensions should be jointly taken into account. In the last few years, several approaches to metric combination have been suggested (Shieber, 2004; Liu and Gildea, 2007; Albrecht and Hwa, 2007; Padó et al., 2007). In spite of working on a limited set of quality aspects, mostly lexical features, these approaches have provided effective means of combining different metrics into a single measure of quality. All these methods implement a parametric combination scheme. Their models involve a number of parameters whose weight must be adjusted and can be a source of overfitting.

On the other hand, Giménez and Màrquez (2008) explore the possibility of relying on non-parametric combination schema, in which metrics are combined without having to adjust their relative importance. They show that

non-parametric schema offer a valid means of putting different quality dimensions together, effectively yielding a significantly improved evaluation, both in terms of human likeness and human acceptability. They have also verified that the performance of these methods works well across test beds.

### II.5.3 Human Evaluation

Human evaluation metrics require a certain degree of human intervention in order to obtain the quality score. This is a very costly evaluation strategy that can seldom be conducted.

Usually, the tendency has been to evaluate adequacy and fluency (or other relevant aspects of translation) according to a 1 to 5 quality scale (White and O'Connell, 1994). Fluency indicates how natural the hypothesis sounds to a native speaker of the target language, usually with these possible scores: 5 for Flawless, 4 for Good, 3 for Non-native, 2 for Disfluent and 1 for Incomprehensible.

Adequacy is assessed after the fluency judgement is made. The evaluator is presented with a certain reference translation and has to judge how much of the information from the original translation is expressed in the translation by selecting one of the following grades: 5 for all of the information, 4 for most of the information, 3 for much of the information, 2 for little information, and 1 for none of it.

#### II.5.3.1 Human-targeted scores

Evaluation methods based on measuring the cost of human post-editing to achieve an acceptable translation are being increasingly recognized. Using these methods, differences between translation and reference only account for real errors and or example the final score is not influenced by the effects of synonymy. The human-targeted reference is obtained by editing the output with two main constraints: the resultant post-edited translation preserves the meaning and it is fluent. If we make use of those references to calculate usual MT evaluation measures, we refer to the resulting measures as the human-targeted variants of the originals, for example, BLEU, NIST and TER give rise to HBLEU, HNIST, as cited in Snover et al. (2006).

Specifically, the HTER (*Human-targeted Translation Edit Rate*) measure (also called edit distance in Przybocki et al. (2006) and post-editing cost in Goutte (2006)) was presented in Snover et al. (2006) based on the TER metric. The combination of human post-edited references together with the TER metric is specially interesting since it allows to measure the general effort needed to correct the MT systems' output.

TER is defined as the minimum number of edits needed to change a hypothesis so that it exactly matches one of the references, normalized by the average length of the references. Since we are concerned with the minimum number of edits needed to modify the hypothesis, we only measure the number of edits to the closest reference. Specifically:

$$HTER = \frac{\text{\# of edits}}{\text{average \# of reference words}} \tag{II.22}$$

Possible edits include the insertion, deletion, and substitution of single words as well as shifts of word sequences. A shift moves a contiguous sequence of words within the hypothesis to another location within the hypothesis. All edits, including shifts of any number of words, by any distance, have an equal cost.

The acceptability of a hypothesis is not entirely indicated by the TER score, which ignores notions of semantic equivalence. So HTER involves a procedure for creating targeted references, using human annotators who are fluent speakers of the target language, to generate a new targeted reference. We then compute the minimum TER using this single targeted reference as a new human reference.

This intuitive evaluation metric aims to measure translation quality in a realistic way, showing how usable the MT output is. Furthermore, it offers greater correlation with human judgements than *BLEU* (Przybocki et al., 2006; Snover et al., 2006).

Furthermore, within this approach it is possible to reduce the cost for development within a system, since it is not necessary to create new references for each run. For many systems, most translations do not change from run to run. New targeted references only need to be created for those sentences whose translations have changed. Moreover, we probably only need to create new references for sentences with a significantly increased edit rate (since the last run).

# SMT FOR BASQUE

# CHAPTER III

## Adaptation of SMT to Basque Morphology

In this chapter we present the work done to adapt a baseline SMT system to carry out 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 the case of ellipsis, more than one of these suffixes can be added to the same word. 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.

As a consequence most words occur only once in the training data leading to a serious sparseness problem which is exacerbated by the few corpora available for the Spanish—Basque language pair.

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 by the inflection. By dividing each Basque word in this way, we aim to reduce the sparseness produced by the agglutinative nature of Basque and the small amount of training data.

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 intermediate combinations. In order to define the most appropiate segmentation for our Spanish-

Basque system, we have tried some of these segmentation options and have measured their impact on the translation quality.

This chapter is organized as follows: In Section III.1, we present a brief analysis of related works adapting SMT to highly inflected languages. In Section III.2, we describe the systems developed for this research (the baseline and the morpheme-based systems) and the different segmentation used by those systems. In Section III.3, we evaluate the different systems, and report and discuss our experimental results. Section III.4 presents our conclusion and gives avenues for future work.

## III.1 Related Work

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

In general, most experiments are 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 some other approaches have also been tried. In Bojar (2007), factored translation is used in English-Czech translation. Words of both languages are tagged with extra linguistic information, creating different factors which are translated independently and later combined in a later generation stage. In the experiment the author uses three different factors (word form, lemma and morphological information). He also experiments with different translation paths: (1) direct translation, (2) generation of the morphological information for each word along with the use of a second language model for this level, and (3) complex translation path where the form is generated from the independently-translated lemma and morphological information. The use of additional language models over the linguistic information entails a substantial BLEU increase. But, on the contrary, the use of complex translation paths, where the final form is generated from the lemma and the morphological information (as in the third scenario), does not imply any improvement over the use of linguistically informed language models (as in the second scenario).

Finally, Minkov et al. (2007) present a morphology generation model. The authors perform the process of translation in two steps. In the first step they use a traditional SMT system to translate into target lemmas and then, in the second step, they determine the inflection of each lemma, making use of bilingual information. They do not present any evaluation over translation quality, instead they just evaluat the generation step in isolation. As continuation of this work, Toutanova et al. (2008) tests different methods to integrate both steps (translation and morphological generation). For all the integration methods, they obtain significant improvements over the baseline, in both automatic and manual evaluations. On the other hand, the differences measured by means of BLEU when evaluating different integration methods are not confirmed by the manual evaluation.

Regarding Basque, there are few researches that directly deal with its agglutinative nature. For example, Pérez et al. (2008) incorporates morphology generation model in a Spanish-Basque speech translation system for a much reduced domain (short descriptions of weather forecast). In this work, they carry out the translation from Spanish speech into Basque text making use of an integrated architecture of stochastic finite-state transducers. In order to deal with Basque morphology, the authors decompose the translation into two steps; in the first step they translate into Basque lemmas, and

then, they generate the inflection for each word. Unlike Minkov et al. (2007); Toutanova et al. (2008), the morphological generation is carried out on the basis of monolingual information. The morphological generation model has been tested in three different scenarios (text translations, decoupled speech translation and integrated speech translation) and in all of them, the use of morphological generation implies BLEU score improvements. However, the scores obtained for WER and PER are not so consistent and show small quality decreases for some of the scenarios.

## III.2 Treatment of Basque Morphology

The main aim 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.

### III.2.1 Baseline

We have trained Moses on the tokenized corpus (without any segmentation) as the baseline system. Moses and the scripts provided with it allow a state-of-the-art phrase-based SMT system to be easilly trained. 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 3-gram language model trained on the Basque portion of the training data, using the SRILM 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 the *'msd-bidirectional-fe'* option. The general design of the baseline system is presented in Figure III.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).

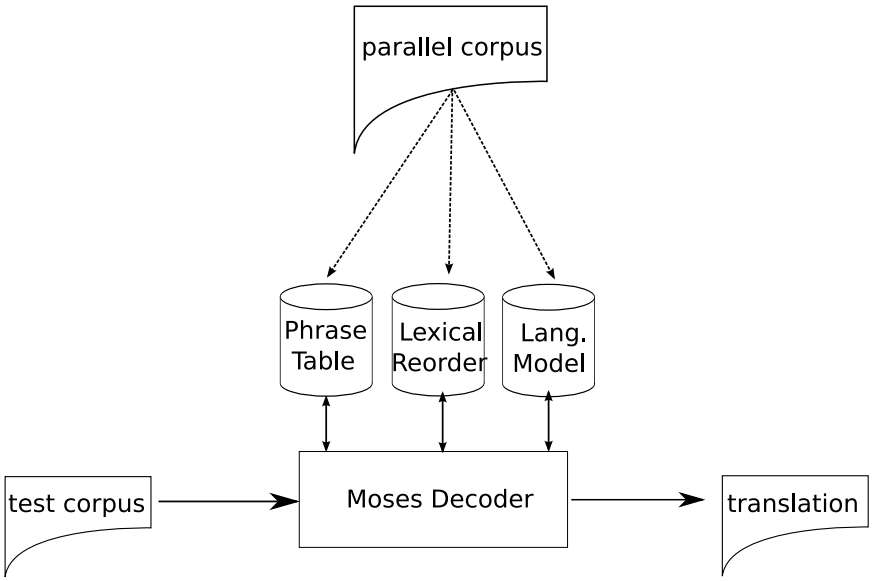



Figure III.1: Basic design of an SMT system, where all models are directly trained on the original parallel corpus.

## III.2.2 Morpheme-based statistical machine translation

In Basque the morphemes are added as suffixes to the last word of noun phrases and verb chains. Suffixes represent the morpho-syntactic information associated with the phrase, such as number, definiteness, grammar case and postposition.

[1]http://www.statmt.org/moses/?n=Moses.AdvancedFeatures#ntoc1

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 overcome this problem, we segmented each word into a sequence of morphemes, and then we applied SMT techniques 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 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.

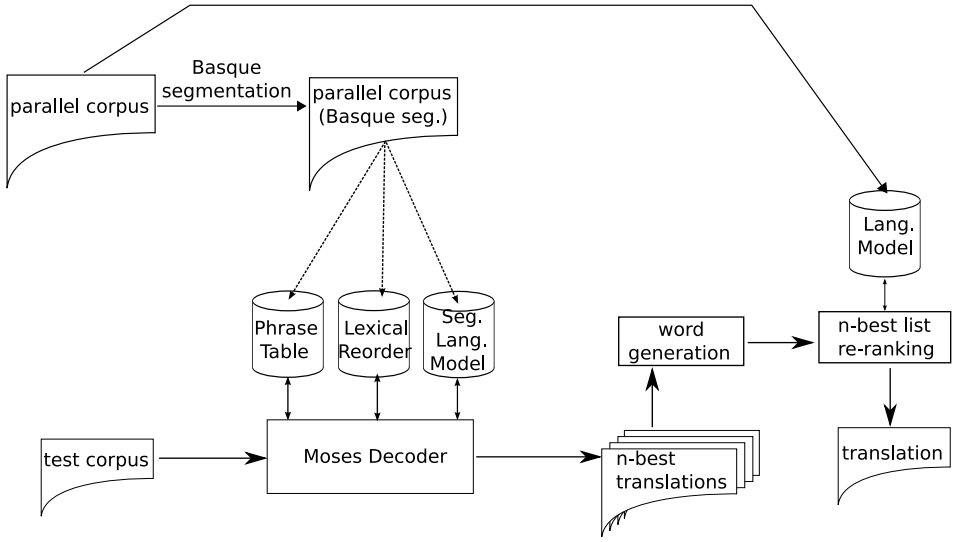



Figure III.2: Design of a morpheme-based SMT system, where the models used in decoding are trained in the segmented target text, and the final word language model is incorporated using an nbest list.

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 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 in Figure III.2.

### III.2.2.1 Segmentation options for Basque

Segmentation of Basque words can be done in different ways and we want to measure the impact these 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 option.

The different segmentations 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 these analyses we have split out 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.

Following this we have defined the four segmentation options we are experimenting with. Figure III.3 shows the analysis obtained by Eustagger and segmentation of the word 'aukeratzerakoan' */at the election time/* according to each of the four segmentation options. The morphemes are delimited using the '+' character and each segmentation is numerated according the enumeration.

| | | | | | | |
|---|---|---|---|---|---|---|
| | Analysis: | aukeratu+<adize>+<ala>+<gel>+<ine> | | | | |
| | | aukeratu+tze | +ra | +ko | +an | |
| 1 | Eustagger: | aukeratu | +<adize> | +<ala> | +<gel> | +<ine> |
| 2 | Automatic: | aukeratu | +<adize><ala> | +<gel> | +<ine> | |
| 3 | Manually defined: | aukeratu<adize> | +<ala><gel><ine> | | | |
| 4 | OneSuffix: | aukeratu | +<adize><ala><gel><ine> | | | |

Figure III.3: Analysis obtained by Eustagger for word 'aukeratzerakoan' */at the election time/*, and the four possible segmentations inferred from it.

**1. Eustagger Segmentation**: Our first approach is strictly based on the Eustagger lexicon, and we have created a separate token for each morpheme recognized by the analyzer. This lexicon has been created from a linguistic perspective and, although it has been proved to be very useful for the development of several applications, the granularity used is probably not the most

appropiate for the translation task. As the lexicon is very fine-grained, some suffixes that could also be considered as a unique morpheme are represented as a concatenation of several fine-grained morphemes in the Eustagger lexicon. Furthermore, some of these morphemes do not have any effect on the word form. They are null suffixes that only add some morphological features.

**2. 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 of putting together all the suffixes linked to a lemma in one token. Thus, on splitting one Basque word we will generate, at most, three tokens (prefixes, lemma and suffixes).

**3. Manual morpheme grouping**: After realizing the impact of the segmentation on translation, we tried to obtain an intermediate segmentation which would optimize the translation quality. Our first attempt consists of manually defining which morphemes can be grouped together into one token and which ones can be considered a token on their own. In order to decide which morphemes to group, we have analyzed the alignment errors that occurred in previous segmentation experiments, defining a small number of rules for grouping morphemes. For instance, the '+<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.

**4. Automatic morpheme grouping**: The manually defined morpheme grouping depends on the language and if we wanted to apply it to another language, we would have to redefine the grouping criteria after again analyzing the errors detected. Thus, in order to find a language-independent way of defining the most appropriate segmentation, we have focused our research on establishing a statistical method to decide which morphemes have to be put into the same token. We have observed that the morphemes that generate most of the errors are those which do not have their own *meaning*, those that *need* another morpheme to complete their meaning. We thought about using the *mutual information* metric (Kenneth W. Church, 1989) in order to measure the statistical interdependence between two morphemes. Then we grouped those morphemes that are more interdependent than a threshold value. In this experiment we experimented with different thresholds and obtained the best results with a value of 0.5 (a value that involves a high deegre of grouping of morphemes). Table III.1 shows the evolution of the size of the

[2]suffix for verb nominalization

| Threshold | Tokens | Vocabulary | BLEU |
|---|---|---|---|
| -1 | 1,572,530 | 35,639 | 11.08 |
| 0 | 1,574,305 | 35,581 | 10.89 |
| 0.5 | 1,580,551 | 35,549 | 11.24 |
| 1 | 1,583,373 | 35,516 | 10.94 |
| 3 | 1,594,845 | 35,409 | 10.65 |
| fully segmented | 1,699,988 | 35,316 | 10.52 |

Table III.1: Evolution of the size of the Basque training corpus and BLEU score depending on different thresholds for *mutual information*

Basque training corpus and BLEU score depending on different thresholds for *mutual information*.

### III.2.2.2 Generating words from morphemes

As previously stated, when working at 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 corresponding to this sequence, so the output of the SMT system is post-processed to produce the final Basque translation.

To develop this generation post-processing, we reuse the lexicon and two-level rules of our Eustagger morphological tool. The same generation engine is useful for all the segmentation options defined in section III.2.2.1 since we have produced all of them using the same fine-grained segmentation. However, we have to face two main problems:

- Unknown lemmas: some lemmas such as proper names are not in the Eustagger lexicon and cannot be generated by it. To solve this problem and to be able to generate inflection of these 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 placed in the 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 do not try to correct these mistakes, and when the generation module can not generate a word it outputs the

lemma without inflection. A more refined treatment is left for future work.

#### III.2.2.3 Incorporation of a word-level language model

When training our SMT system over the segmented text the language model used in decoding is a language model based on morphemes (or groups of morphemes depending on the segmentation option). Real words are not available at decoding, but, after morphological generation we can incorporate a second language model based on words. The most appropriate way of incorporating the word-level language model is using an n-best list, as was done in (Oflazer and El-Kahlout, 2007). So, 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 a word-level language model. Once the 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. The Minimum Error Rate Training procedure has been modified to post-process Moses output and to include word-level language model weight at optimization process.

## III.3 Experimental results

### III.3.1 Data and evaluation

In order to carry out this experiment we used the *Consumer Eroski* parallel corpus (Alcázar, 2005). This corpus is a collection of 1036 articles written in Spanish (January 1998 to May 2005, Consumer Eroski magazine[3], 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[4] and it is accessible online via

[3] http://revista.consumer.es

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

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

Table III.2: Some statistics of the Corpus (Eroski Consumer).

University of Vigo[5] (public access) and the University of Deusto[6] for research. Consumer Eroski magazine is composed of the articles that compare the quality and prices of commercial products and brands.

We have divided this corpus into three sets: a training set (≈60,000 sentences); a development set (≈1,500 sentences); and a test set (≈1,500 sentences). Table III.2 shows more detailed statistics.

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.

### III.3.2 Results

The evaluation results for the test corpus are reported in Table III.3. These results show that the differences in segmentation have a significant impact on the translation quality.

Segmenting words according to the morpheme boundaries of the Eustagger lexicon does not give any improvement. Compared to the baseline, which did not use any segmentation, the results obtained for the evaluation metrics are not consistent and vary depending on the metric. According to BLEU,

[5]http://sli.uvigo.es/CLUVI/
[6]http://www.deli.deusto.es

| | 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-ManualGrouping** | **11.36** | 4.67 | 78.92 | 60.23 |

Table III.3: Evaluation of SMT systems with five different segmentation options.

segmentation harms translation. However, according to the rest of the metrics segmentation slightly improves translation, although this improvement is probably not statistically significant.

The rest of the segmentation options, which are based on the same Eustagger analysis and contain the same morpheme sequences, consistently outperform the baseline according to all the metrics. The best results are obtained using the manually defined criteria (based on the alignment errors), but automatically defined segmentation criteria obtain similar results.

Due to the small differences in the results obtained for the evaluation metrics we have carried out a statistical significance test (Zhang et al., 2004) over BLEU. According to this, the system using manually defined segmentation significantly outperforms the baseline, the system using OneSuffix segmentation and the system using segmentation based on mutual information. The difference between the system using OneSuffix segmentation and the system based on mutual information is not statistically significant.

Finally, given the low scores obtained for BLEU ($\leq$11.36) and NIST ($\leq$4.74) in all systems, 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 marked 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.

### III.3.3 Correlation between segmentation and BLEU

Analyzing the obtained results, we have realized that there is a correlation between the number of tokens generated at segmentation and the results

| Segmentation option | Running tokens | Vocabulary size | BLEU |
|---|---|---|---|
| Source text | 1,284,089 | 46,636 | - |
| No Segmentation | 1,010,545 | 87,763 | 10.78 |
| Manually 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 III.4: Correlation between token number in the training corpus and BLEU evaluation results

obtained at evaluation. Before segmentation, there are 1 million words for Basque, which, taken together with the 1.2 million words for Spanish, makes the word alignment more difficult (due to the number of 1-to-n alignment). However, after segmenting the Basque words according to the morpheme boundaries of Eustagger, the Basque text contains 1.7 million tokens (the same alignment problem is generated but in the opposite direction) see Table III.4.

Similarly, the difference in vocabulary size between the unsegmented Basque text (87,763 words) and the Spanish text (46,636 words) is very high. When segmenting Basque according the Eustagger lexicon the differece in the size of the two vocabularies (46,636 tokens in Spanish and 35,316 tokens in Basque) is reduced but is still high.

Intermediate segmentations, where morphemes marked by Eustagger are grouped in different ways, achieve better results when the target vocabulary size is closer to the vocabulary size we have in Spanish part. The best BLEU results are obtained with the smaller difference in vocabulary size (40,288 tokens). We leave for future work to research ways of reducing the difference between the number of tokens of the two languages.

## III.4 Chapter Summary and Conclusions

We have proved that the quality of the translation varies significantly when applying different options for word segmentation. Based on the same output from 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. The other segmentation options outperform the baseline, the best results being obtained with a manually defined intermediate grouping based on an alignment error analysis of the word alignments.

In any case, the work done manually is language dependent and could not be reused for a different pair of languages. Therefore, we also tried a statistical way of determining the morpheme grouping criteria which gets almost as accurate results as those obtained with the 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 interdependence of the morphemes, such as $\chi^2$. Furthermore, as the interdependence 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).

Taking into account the correlation obtained 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 the two languages would be reduced.

# CHAPTER IV

## Adaptation of SMT to Basque Syntax

As we said before, Basque language has many particularities which differences it from most European languages. Those differences make the translation between Spanish (or English) and Basque an interesting challenge which involves both morphology and syntax features. Besides the morphological differences discussed in Chapter III, there are also syntactic differences which are related to the word order into the sentence, that have a negative impact on the translation. As we have already explained, the agglutinative nature of Basque entails that prepositions, placed at the beginning of the phrase in Spanish, are 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. Those prepositional phrases attached to noun phrases are placed preceding the noun phrase instead of following it. Besides, 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 I.3 in chapter I showed an example of a sentence's word alignment between two sentences in Spanish and Basque. The figure illustrates clearly the word-order differences in the two languages.

Those differences on word order have an extremely negative impact on most of the steps of the Statistical Machine Translation, such as word align-

ment, phrase extraction and decoding. We have explored different approaches to deal with the problem of word order when translating from Spanish to Basque using the SMT approach, and we have tried to determine the strength and the weakness of each approach.

This chapter is structured as follows: In Section IV.1, we do a quick revision of the most relevant research on the area. Later, we describe the used reordering techniques (Section IV.2) and the SMT systems developed for this work (Section IV.3). We continue presenting and analyzing the results on Section IV.4. Finally, Section IV.5 presents conclusions and future work.

## IV.1 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 pre-processing of the source sentence in order to obtain a word-order which matches with the word-order of the target language, allowing an almost monotonous translation. Two main approaches are found on the bibliography; those based on the use of hand-defined reordering rules that usually defined looking at the linguistic analysis of the source, and those in which the reordering is automatically inferred from the training corpus.

In 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 a significant improvement both in BLEU and human judgments. Later, similar attempts are carried out for several language-pairs. For example, Popović and Ney (2006) proposed different reordering rules depending on the languages involved on the translation. They defined long-range reordering when translating into German and some local reordering for English-Spanish and German-Spanish language pairs. More recently, in (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

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

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 (Sanchís 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.

## IV.2 Reordering techniques

The main deal of this work is to analyze 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 three reordering techniques following described.

### IV.2.1 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 among the different methods we have tried which does not consist on the pre-processing 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 their probability distribution is estimated in order to be added to the log-linear framework. Three different orientations are defined (See Figure IV.1):

- monotone: continuous phrases appear in same order in both languages.
- swap: continuous phrases are swapped in the target language.
- discontinuous: continuous phrases in the source language are not continuous in the target language.

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

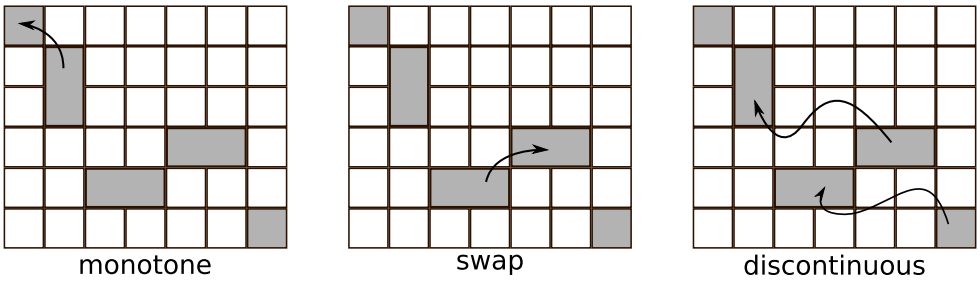


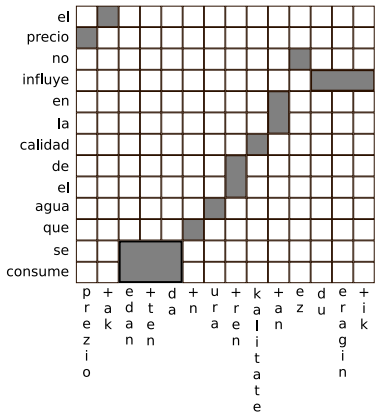

Figure IV.1: Possible orientations of phrases defined on the lexicalized reordering: monotone, swap, or discontinuous

According to that, and following with the same sentence pair presented above, Figure IV.2 shows the word alignment and some of the phrases extracted from it. For each phrase pair, we show probability distribution for each orientation. Those probabilities are inferred from the whole corpus (the probability corresponding to the orientation present in the current example are mark in bold). We can see that in most of the cases the orientation of the phrases in the current example is the most probable one. There are just two exceptions to this: on the one hand, for the alignment of (el precio, prezio+ak, /the price/) the orientation of the example sentence is 'monotonous' while the most probable orientation in the whole corpus is 'swap' (monotonous probability =0.17; swap probability =0.43), and on the other hand, for the alignment of (calidad de el agua, ura +ren kalitate, /quality of the water/) the orientation extracted from the example is 'swap' but the most probable orientation is 'discontinuous' (swap probability =0.31; discontinuous probability =0.68).

Finally, at decoding, automatically inferred reordering models are used to score each hypothesis according the orientation of the phrases used in each hypothesis. In such a way, the reordering model penalizes the phrase pairs that are used in a orientation that was not corresponds to those saw in the training corpus.

### IV.2.2 Syntax-Based reordering

The second method presented here consists on the pre-processing of the Spanish sentences to adapt their word order to the order in Basque. This pre-processing is based on the dependency tree obtained with the morpho-




| | | | mon. | swap | disc. |
|---|---|---|---|---|---|
| /the/ | el | +ak | 0.01 | **0.79** | 0.20 |
| /the price/ | el precio | prezio +ak | **0.17** | 0.43 | 0.40 |
| /not/ | no | ez | 0.30 | 0.10 | **0.60** |
| /does not influence/ | no influye | ez du eragin +nik | 0.20 | 0.20 | **0.60** |
| /does not influence in the/ | no influye en la | +an ez du eraginik | 0.08 | **0.79** | 0.13 |
| /influence/ | influye | du eragin +nik | **0.60** | 0.20 | 0.20 |
| /in the/ | en la | +an | 0.01 | **0.83** | 0.16 |
| /in the quality/ | en la calidad | kalitate +an | 0.04 | **0.56** | 0.40 |
| /in the quality of the/ | en la calidad de el | +ren kalitate +an | 0.14 | **0.71** | 0.15 |
| /quality of the water/ | calidad de el agua | ura +ren kalitate | 0.01 | **0.31** | 0.68 |
| /quality of the water that/ | calidad de el agua que | +n ura +ren kalitate | 0.03 | **0.86** | 0.11 |
| /water that is consumed/ | agua que se consume | edan +ten da +n ura | 0.20 | 0.20 | **0.60** |
| /that is consumed/ | que se consume | edan +ten da +n | 0.09 | 0.27 | **0.64** |
| /is consumed/ | se consume | edan +ten da | 0.07 | 0.46 | **0.47** |

Figure IV.2: Word alignment and lexicalized reordering probabilities.

logical analyzer 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).

#### IV.2.2.1 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 explained, prepositions are translated 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.

| /the/ /price/ | /no/ | /has-influence/ | /on/ /the/ /quality/ | /of/ /the/ /water | /that/ | /is/ /consumed/ |
|---|---|---|---|---|---|---|
| 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 |

### IV.2.2.2 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.

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

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

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

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

Figure IV.3 shows an example of the application of these rules. Figure I.3 in chapter I and Figure IV.4 show the word alignments before and after syntax-based reordering; note that the final alignment is almost monotonous.

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

### IV.2.3 Statistical Reordering

The Statistical Reordering considers the reordering pre-processing 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 rules.

The training process consists on the following steps; (1) align source and target training corpora in both directions and combine word alignments to obtain many-to-many word alignments, (2) Modify the many-to-many word alignments to many-to-one (keeping for each source word only the alignment with a higher IBM-1 probability), (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 the reordered source language to the target one.

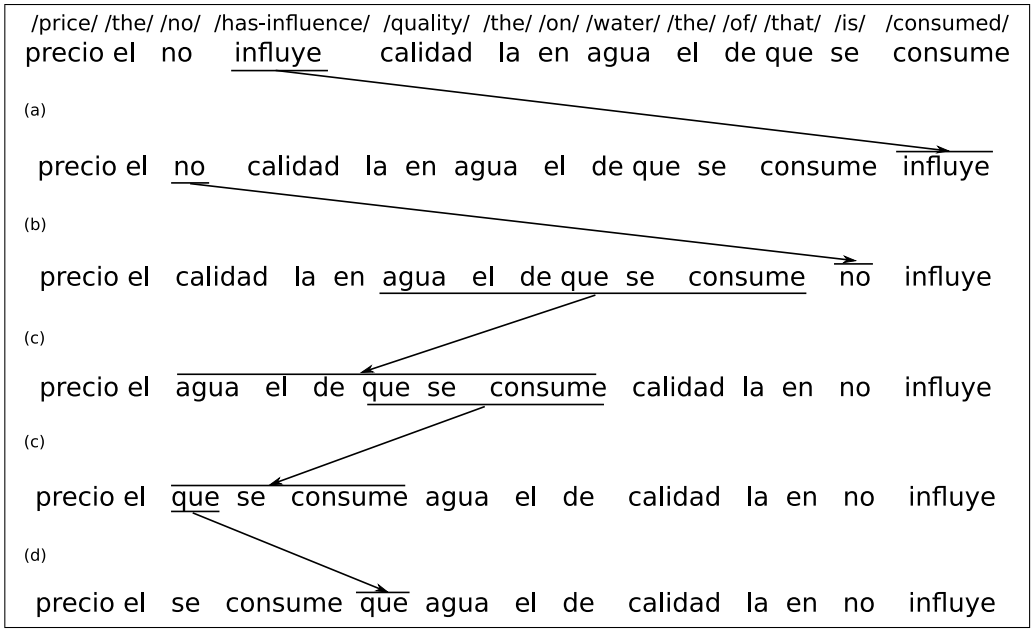



Figure IV.3: Example of long-range reordering.

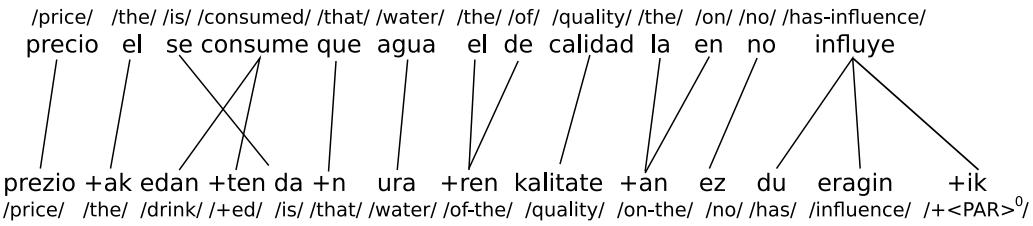



Figure IV.4: Example of word alignment after syntax-based reordering.

## IV.3 Reordering experiments with Basque

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

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

In order to deal with the agglutinative nature of Basque, and reusing the previous work, we have used segmented Basque text, where words are split into different tokens, before training all our systems. After translation, a word generation post-processing has carried out to generate the final translation based on the segmented output of the decoder. After generation, the word-based language model is incorporated using n-best lists reranking. So two language models are applied: the segment-based language model inside the Moses decoder, and word-based language model after generation. Figure IV.5 shows the general design of the system used in this work.

We have trained six different systems: a baseline (the default Moses system that lexicalized reordering has been disabled), the systems that implements each individual technique (Statistical reordering, Syntax-based reordering and lexicalized reordering) and two systems that combines the lexicalized reordering (which is applied at decoding) with each of the techniques

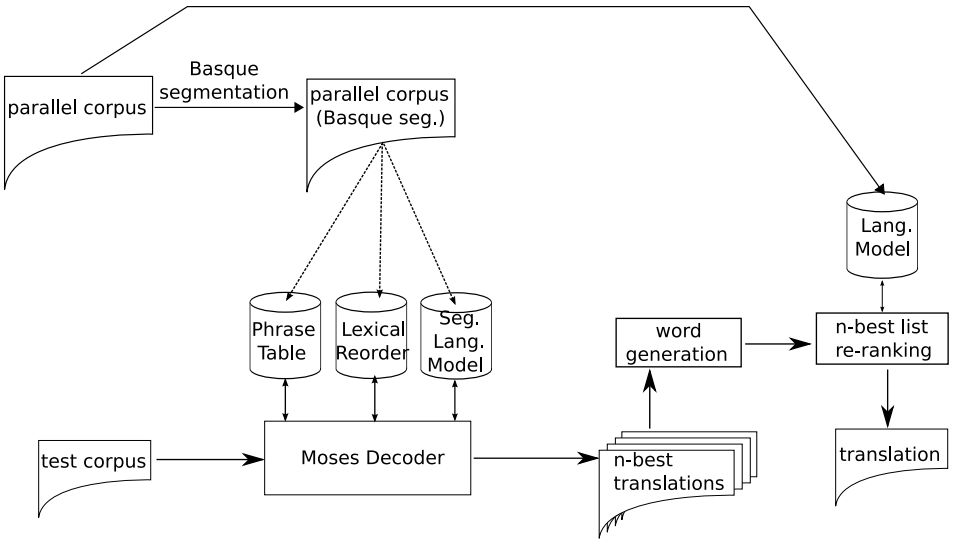



Figure IV.5: Design of the segmentation-based SMT system

applied as a pre-processing (Statistical reordering and Syntax-based reordering).

We should point out that the baseline used in this chapter is not the same used at the Chapter III. The differences between both baselines are mainly two: the baseline use in this chapter has the lexicalized reordering disabled (in order to measure it impact in the translation), while it was activated in the previous chapter; and the current baseline make use of the segmentation while the previous one has been trained on the unsegmented text. We want to point also that the system called "Manual grouping" the previous chapter corresponds to the system called "Lexicalized" in the current one, as both uses the same segmentation option (manual grouping) and reordering technique (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 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).

## IV.4 Experimental results

### IV.4.1 Data and evaluation

In order to carry out this experiment we used the *Consumer Eroski* parallel corpus. This corpus is already used in the previous chapter as is divided in three sets, training set (≈60,000 sentences), development set (≈1,500 sentences) and test set (≈1,500 sentences) (see Table III.2 in chapter III). Note that the Basque singletons (words with just one occurrence) are much more in Basque than the Spanish ones. Otherwise, when Basque is segmented both figures are similar.

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.

### IV.4.2 Results

The evaluation results for the test corpus are reported on Table IV.1. According to BLEU scores all single reordering methods outperforms the baseline ($10.37 < 11.03 < 11.13 < 11.36$), 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 pre-processing and the lexicalized reordering at decoding (11.51 BLEU score).

Considering those systems which uses single reordering methods, lexicalized reordering get the best results (11.36 BLEU), followed by the statistical reordering (11.13 BLEU). Finally, the Syntax-Based reordering (11.03

| | BLEU | NIST | WER | PER |
|---|---|---|---|---|
| Baseline | 10.37 | 4.54 | 79.47 | 60.59 |
| Statistical reord. | 11.13 | **4.69** | 78.21 | **59.66** |
| Syntax-based reord. | 11.03 | 4.60 | 78.79 | 61.35 |
| Lexicalized | 11.36 | 4.67 | 78.92 | 60.23 |
| Statistical + Lexicalized | 11.12 | 4.66 | 78.69 | 60.19 |
| Syntax-based + Lexicalized | **11.51** | **4.69** | **77.94** | 60.45 |

Table IV.1: BLEU, NIST, WER and PER evaluation metrics.

BLEU) get the smaller improvement over the baseline. In all the cases, the improvement using sophisticated reordering methods is substantial.

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

## IV.5 Chapter Summary and Conclusions

Results obtained in this chapter allow us to compare different reordering methods on a specially demanding task as the Spanish-Basque translation. According to 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. Because of which we consider that the distance-based reordering, which is the technique used by the baseline system, does not have enough information to properly handle big word order differences.

We have also tested different combinations 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 treated 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 pre-processing 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 pre-processing (using a n-best list of reordering or a word-graph as input of the decoder).

# HYBRID APPROACHES

# CHAPTER V

# Hybridization

Once we had achieved a minimal quality SMT system, we wanted to use it in preliminary hybridization experiments. we expect to improve the 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 this purpose we define two different hybridization experiments. In the first experiment, we will translate each sentence using the three systems we have available (SMT, RBMT and EBMT systems) and the most appropriate translation will be chosen for each sentence. In the second experiment, we will build a Statistical Post-Editing system in order to correct 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, to "translate" from the output of the RBMT system to the real target language.

## V.1 Related Work

### V.1.1 Multiengine systems

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 translation 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.

Rosti et al. (2007) describes three different approaches to SMT system combination. These combination methods operate on sentence, phrase and word level exploiting information from n-best lists, system scores and target-to-source phrase alignments. The most robust gains are provided by the word-level combination, since the phrase-level combination achieves good results at tuning but does not properly generalize to the test set. Finally, sentence level combination yields slight gains on the tuning set, and very small ones, if any, on the test sets.

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.

In the last years some research works have tried to recombine the translation provided by different MT systems instead choosing the most suitable one (Macherey and Och, 2007; Chen et al., 2009; Leusch et al., 2009; Du et al., 2009). Recombining multiple MT results requires finding the correspondences between alternative renderings of a source language expression proposed by different MT systems. Additionally, a recombination system needs a way to pick the best combination of alternative building blocks. when judging the quality of a particular configuration, both the plausibility of the building blocks as such and their relation to the context need to be taken into account.

### V.1.2 Statistical PostEditing

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[1] 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

[1] http://www.systran.co.uk/

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-Editing (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.

## V.2 The corpora

Due to the still low quality of the translation obtained, our aim has been to improve the precision of the MT system trying to translate texts from a restricted domain. We are 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 want to compare the results between the restricted domain and a more general domain such as news.

### V.2.1 Specific domain: Labor Agreements Corpus

The domain related to Labor Agreements has been selected. The Basque Institute of Public Administration (IVAP[2]) has collaborated with us in this selection, after examining different domains; the parallel corpora available and the 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 a 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.

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 V.1.

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

Table V.1: Some statistics of the Labor Agreements Corpus

[2]http://www.ivap.euskadi.net

### V.2.2 General domain: Consumer Eroski Corpus

As general domain corpus, we used the same Consumer Eroski parallel corpus already used in previous chapters (see Table III.2 in chapter III). In order to train the data-driven systems (both SMT and SPE systems), we used approximately 60,000 aligned sentences extracted from the Consumer dataset. Two additional sentence sets are used; approximately 1,500 sentences for parameter tuning and 1500 sentences for evaluation.

## V.3 Multi Engine MT

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). Unlike the most of the state-of-the-art MEMT works, in our case, we combined the main three different MT approaches (RBMT, EBMT, and SMT) instead of combining different SMT systems Rosti et al. (2007).

Our aim in MT system combining was simple: we wanted to verify that even for a morphologicaly rich language as Basque combining RBMT and SMT systems the outputs result in a better translation. So our system is not directly comparable with other MEMT systems because the language pair is different, and because most of the approaches generate a new consensus translation by combining different SMT systems using different language models and in a few cases also combining with RBMT systems.

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). In this paper, the results RBMT and SMT systems are evaluated using automatic metrics (BLEU) and user-driven metrics (HTER). Those evaluations where performed with two more general corpora related to news in the Basque Public Radio-Television (EiTB[3]) and to articles in a magazine for consumers (Consumer[4]). And in both corpora the RBMT gets significantly better HTER scores than the SMT, while the BLEU metric assigns a higher score to the SMT system.

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.
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.

[3]http://www.eitb.com/telebista/etb1/

[4]The Consumer corpus used for evaluation is the one referenced in Table III.2 in chapter III but before a cleaning process.

| | **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 V.2: Evaluation for MEMT systems using the Labour Agreements corpus

### V.3.1 Evaluation

In order to assess the quality of the resulting translation, we used automatic evaluation metrics on the specific domain. We report the following accuracy measures: BLEU (Papineni et al., 2002) and NIST (Doddington, 2002). Table V.2 shows the results using the test corpus. The coverage percentages given for each approach show the relative participation of each subsystem in the provided output.

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 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 (Labaka et al., 2007), 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.

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.

## V.4 Statistical Postediting

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 corpus to be used in training our statistical post-editor (Figure V.1 shows a diagram of the general architecture).

Using a SMT system to post-edit the RBMT output we expect obtain the best of both approaches. First, the RBMT system will deal with the syntax and long-range reordering issues, and, after that, the post-editing system will correct fluency and lexical selection issues. Making use of the bilingual phrases, the Statistical Post-Edition system will change the lexical selections of the RBMT system according to the near context. Similarly, thanks to the language model, the SMT system will make local reorderings getting a more fluent final translation.

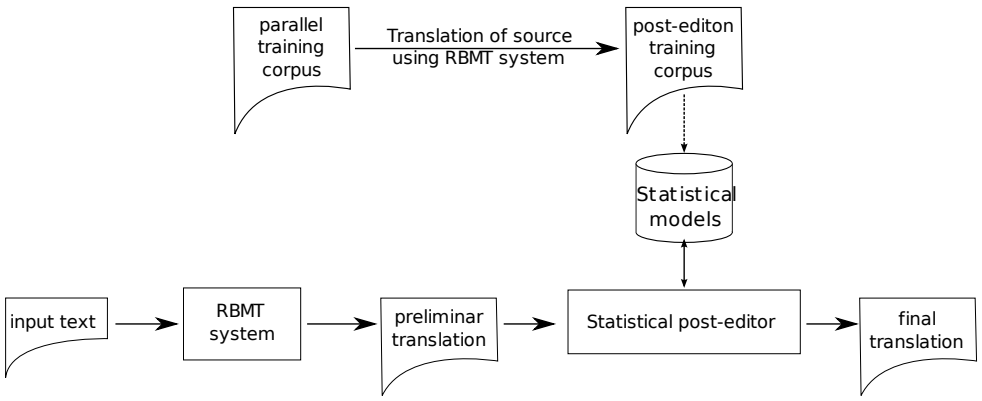



Figure V.1: General architecture of an Statistical Post-Editing system

In our experiments, the corpus-based system used as statistical post-editor will be the SMT system developed in the Chapter III of this thesis. Where, in order to deal morphological richness of Basque, the target sentences are segmented and the system works at the morpheme-level. So a generation phase is necessary after SPE is applied. Since the input of the SMT system is not Spanish anymore, the Syntax-based reordering techniques developed in Chapter IV are not used.

Finally, and following the work did in collaboration with the Dublin City University (DCU), the SMT phrases are enriched with phrases extracted using EBMT techniques (a deeper description can be found in Section II.4.3).

### V.4.1 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: the domain specific corpus (Labor Agreements Corpus), and the 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 V.3: Evaluation on domain specific corpus.

Results obtained on the Labor Agreements Corpus (see Table V.3) 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). 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, the best of the individual systems, in 4.48 BLEU points (40% relative improvement).

| | BLEU | NIST | WER | PER |
|---|---|---|---|---|
| **Rule-Based** | 6.78 | 3.72 | 81.89 | 66.72 |
| **Corpus-based** | **11.51** | **4.69** | **77.94** | **60.23** |
| **Rule-Based + SPE** | 10.14 | 4.57 | 78.23 | 60.89 |

Table V.4: Evaluation on general domain corpus.

Otherwise, results on the general domain corpus (see Table V.4) 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.

Again, due to the bias of the automatic evaluation metrics towards the SMT systems, we think that a deeper evaluation is need before draw firm conclusions (see Chapter VI).

## V.5 Chapter Summary and Conclusions

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 difficult 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.

We also 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 statistical post-editing 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 statistical post-editing system working with Matxin RBMT system, when comparing to raw RBMT, 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).

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.

# OVERALL EVALUATION

# CHAPTER VI

## Overall evaluation

In this chapter we present the final, global evaluation of all the systems developed in this PhD thesis and presented in the previous chapters. Until now, we have presented the evaluations of each system in isolation, emphasizing their improvement in terms of automatic metrics (BLEU, NIST, WER and PER) with respect to a baseline. In this chapter we will evaluate all these systems in a new framework (the same for all of them) using the same training corpus (seven times larger than the corpora used in previous experiments) and measuring their improvements by means of two different kinds of metrics (automatic and manual-based).

As previously stated, as Basque is a less-resourced language, one of our main difficulties is obtaining a larger enough bilingual corpus. The experiments presented so far have been performed with a small corpus (the one we had available at the beginning of this work); however, during recent years, we have been actively working on enlarging our corpus collection for machine translation. We have increased the size of the bitext from 1 million Basque words in the bilingual corpus (1.3 million in Spanish) to 7 million (9 million in Spanish). We have also collected up to 28 million words of monolingual Basque text to be used for the training of the language model (initially we only had the target side of the bilingual corpus available to train the language model). With this new material, we decided to rerun the evaluation of all the systems trained with these new larger corpora. We wanted to corroborate the conclusions drawn from the previous partial evaluations.

In addition, during recent years some doubts have arisen about the validity of BLEU (Callison-Burch et al., 2006; Melamed et al., 2003; Koehn and Monz, 2006) as a metric for MT evaluation. In order to overcome those problems, a set of metrics that measure the linguistic similarity between the automatically generated translation and a set of reference sentences have been presented (a deeper description is provided in Section II.5.2). However, the applicability of these evaluation techniques is deeply conditioned by the accessibility to the required linguistic processors for Basque. Furthermore, just like BLEU does, these metrics compare the automatic translations with human-defined references, and the evaluation is not so precise when we have only one reference (as our test set has).

Taking these difficulties into account, we decided to use human-targeted evaluation metrics to perform a final assessment of the work done in this thesis. Human-targeted evaluation, based on manual post-editions, would give us a more confident score, as the output of the MT system is compared with the closest correct translation, thus avoiding the low scores obtained by those correct sentences that differ from the available referent translations. We can use the post-edited references to calculate any evaluation metrics, such as BLEU, NIST or TER (giving rise to human-targeted metrics HBLEU, HNIST or HTER). Since the corpus we have available for evaluation contains only one reference, the use of human post-edited references would provide us with a more reliable evaluation that would allow us to validate the partial results obtained by the automatic evaluation throughout the development of this PhD thesis.

Two facts have led us to consider human-targeted evaluation more reliable and cheaper than the usual evaluation based on automatic metrics and general references. On the one hand, generating post-edited versions of the MT outputs is expensive, and prevents us from carrying human-targeted evaluation at development, but we consider that creating the necessary references to get accurate automatic scores (we have to remind that we only have one reference sentence in our test set) is also expensive. On the other hand, we consider that the results obtained by means of HTER (Human-targeted Translation Error Rate) provides us a reliable and easily interpretable evaluation of the general quality of the system, as provides a realistic measure of the work necessary to correct the MT output of each system.

Considering that, due to its economic cost, all experiments done in this PhD thesis could not be evaluated using human post-editions, the number

of systems to be evaluated in this way was restricted to five: two baseline systems (the RBMT and the baseline SMT systems) and the three systems developed in this PhD thesis (the enhanced SMT, the MultiEngine system and the statistical post-editing).

Through the human-targeted evaluation of these five systems, we wanted to measure the overall improvement reached with the most promising systems built in this PhD thesis. Thus, as it would have been too expensive to evaluate them by means of HTER, the intermediate systems that have led us to the enhanced SMT have only been evaluated by means of the automatic metrics and using only one reference. Additionally, at carrying out the evaluation by means of fully automatic metrics and human-targeted metrics, we would have the chance to compare both evaluation procedures.

## VI.1 Enlarged corpora

After much work we have compiled a heterogeneous set of bilingual corpora that we have combined for this experiment. The collected corpora are from very different sources and genres; from news to university reference books. This diversity could harm the SMT performance, since this type of system fits the training corpus. The genre and style differences present in our corpus could interfere with the extraction of statistics, affecting the translation quality. In any case, this is what we were able to collect, but, for the future, we consider that it is important to keep looking for parallel corpora, in order to compile more coherent bilingual corpora.

### VI.1.1 Parallel corpus

Below, we briefly describe the different corpora we have combined for this work. Table VI.1 shows some statistics of the corpora, showing the number of sentences, tokens, vocabulary and singletons. Let us remark the big difference in vocabulary and singletons figures in Spanish and Basque, which is a consequence of the agglutinative nature of Basque.

- EITB[1]: This corpus is a collection of news from the Basque News and

[1]EITB is the official media group in Basque country, with four television channels and five radio stations

| | | sentences | tokens | vocabulary | singletons |
|---|---|---|---|---|---|
| **EITB** | Spanish | 16,634 | 544,978 | 36,776 | 17,408 |
| | Basque | | 377,253 | 51,809 | 29,168 |
| **EHUBooks** | Spanish | 39,583 | 1,036,605 | 47,987 | 21,761 |
| | Basque | | 794,284 | 84,150 | 47,723 |
| **Consumer** | Spanish | 61,104 | 1,347,831 | 54,457 | 23,552 |
| | Basque | | 1,060,695 | 103,152 | 56,769 |
| **ElhuyarTM** | Spanish | 186,003 | 3,160,494 | 109,035 | 49,259 |
| | Basque | | 2,291,388 | 200,403 | 109,654 |
| **EuskaltelTB** | Spanish | 222,070 | 3,078,079 | 110,201 | 48,535 |
| | Basque | | 2,405,287 | 189,290 | 99,974 |
| **Total** | Spanish | 525,394 | 9,167,987 | 219,472 | 97,576 |
| | Basque | | 6,928,907 | 438,491 | 236,238 |

Table VI.1: Statistics on the final collection of parallel corpora.

Information Channel [2], available in Spanish, Basque, and English. This corpus was automatically aligned at sentence level.

- EHUBooks: Six different reference books translated by the translation service of the University of the Basque Country. These books discuss very different topics (from economics to biology) and were manually translated from Spanish into Basque. This corpus was manually aligned at sentence level.

- Consumer: As we said before, this corpus is a collection of 1036 articles written in Spanish (Consumer Eroski magazine[3] along with their Basque translation. This corpus was automatically aligned at sentence level.

- ElhuyarTM: Translation memories developed by Elhuyar, a translation service company with a heterogeneous client list, from town councils to international companies. Most of the texts use administrative language.

- EuskaltelTB: Translation memory including short descriptions of TV programmes (about 3-4 sentences for each description).

[2] http://www.eitb24.com/en
[3] http://revista.consumer.es

### VI.1.2 Basque monolingual

At the same time, we have collected a 21 million words monolingual Basque corpus, which together with the Basque side of the parallel bilingual corpora, build up a 28 million word corpus to be used for the training of a Basque language model. This monolingual corpus is also heterogeneous, and includes text from three different sources:

| | tokens | vocabulary | singletons |
|---|---|---|---|
| **Bilingual** | 6,928,907 | 438,491 | 236,238 |
| **ZT** | 9,908,312 | 563,109 | 317,949 |
| **Egunkaria** | 11,112,894 | 415,532 | 216,723 |
| **Total** | 27,950,113 | 1,057,237 | 580,477 |

Table VI.2: Statistics on the collection of monolingual Basque texts available for training.

- Bilingual: This consists of the Basque side of the bilingual corpora presented above. As we have already explained, it is made up of very heterogeneous texts.
- ZT corpus (Basque Corpus of Science and Technology): This is a specific or specialized corpus that covers a wide range of topics (mathematics, life sciences, technology...) and genres (schoolbooks, popular science articles, specialists' texts...). The texts included in the corpus were published from 1990 to 2002.
- Egunkaria: This consists of all the articles published by Berria (the only daily newspaper written entirely in Basque) in 2004. The articles cover an assorted range of topics (economics, culture, entertainment, international, local, opinion, politics, sports...).

## VI.2 Automatic evaluation

Since we will carry out the manual evaluation on a new corpus (training the SMT systems on this whole corpus instead of using only the Consumer corpus), and in order to compare the results of the human-targeted evaluation with those obtained using the automatic metrics, we have rerun the automatic

evaluation on the whole corpus. In the same way, before presenting the new evaluation we will summarize the results reported in the previous chapters.

The systems evaluated in this section are a selection of those presented in the previous chapters. We define two baselines: Matxin (the RBMT system previously developed by our research team) and an SMT baseline trained using Moses (Koehn et al., 2007). We used the original text (without any pre-processing) and the original Moses configuration (including lexicalized reordering and MERT weight optimization) to train the SMT Moses system.

Furthermore, we have also evaluated the improvements incorporating them in an incremental way. Thus, SMT-Segmented incorporates segmentation, training the SMT on the segmented Basque text (this system makes use of the segmentation that gets the best results in Chapter III, the one we called ManualGrouping), and SMT-Segmented+Reorder is the system that gets the best result in Chapter IV, which incorporates Syntax-based reordering (this system, as well as the rest of the SMT systems presented in this evaluation, also uses the lexical reordering available in Moses).

Finally, we have also selected the two hybrid systems defined in Chapter V for this final evaluation. Since we have not trained and tested the EBMT system on the Consumer corpus, the MultiEngine system evaluated here only combines the RBMT and SMT systems. On the other hand, the SMT system used for post-editing the output of the RBMT system is equivalent to that used in the SMT-Segmented; where the input of the SMT is not pre-processed (as the RBMT system has already reordered words) and the output is segmented using the ManualGrouping segmentation.

### VI.2.1 Previous evaluation: using small training corpora

In order to contextualize the figures we are providing, let us summarize the evaluation previously presented in Chapters III, IV and V. Table VI.3 shows the scores obtained by systems trained and tested on the Consumer Corpus.

As we showed previously, the two techniques developed in this work helped to increase the translation quality (they obtained 11.36 and 11.51 BLEU scores in comparison to 10.78 achieved by the SMT-baseline). In contrast, automatic metrics did not show any improvement (hybrid systems obtained 11.16 and 10.14, while the SMT system used to develop them obtained 11.51) in using hybridization techniques on a general domain corpus such as

| | **BLEU** | **NIST** | **WER** | **PER** |
|---|---|---|---|---|
| Matxin (RBMT) | 6.87 | 3.78 | 81.68 | 66.06 |
| SMT-baseline | 10.78 | 4.52 | 80.46 | 61.34 |
| SMT-Segmented | 11.36 | 4.67 | 78.92 | **60.23** |
| SMT-Segmented+Reorder | **11.51** | **4.69** | **77.94** | 60.45 |
| Multi-Engine Hybridization | 11.16 | 4.56 | 79.83 | 62.31 |
| Statistical Post-Editing | 10.14 | 4.57 | 78.23 | 60.89 |

Table VI.3: Scores for the automatic metrics for systems trained on the Consumer corpus.

the Consumer corpus (although they achieve significant improvements in a specific domain, such as Labour Agreements, as we saw in Chapter V).

Matxin, the RBMT system we used in our hybridization experiments, was strongly penalized by all the automatic evaluation metrics, but, as we have already explained in a preliminary HTER evaluation (Labaka et al., 2007), this huge difference in automatic metrics was not corroborated by human evaluation. In the aforementioned paper the SMT system obtained better BLEU scores (8.03) than RBMT scores (6.31), but worse HTER scores (57.97 vs 43.40).

All the research presented above, has been performed in the framework of state-of-the-art SMT. But now, we also want to test them incorporated into a more advanced corpus-based system, the one based on the MaTrEx system. The MaTrEx system is an EBMT-SMT hybrid system that uses an SMT decoder to guide the translation process, but which incorporates bilingual phrases extracted using EBMT techniques into the phrase table (a deeper description can be found in Section II.4.3 or in Stroppa and Way (2006)). Since the techniques presented here are based on pre- or post-processing it wasn't difficult to incorporate them into the MaTrEx system.

| | **BLEU** | **NIST** | **WER** | **PER** |
|---|---|---|---|---|
| MaTrEx-baseline | **11.41** (+0.63) | 4.60 (+0.08) | 79.80 (-0.66) | 61.08 (-0.26) |
| MaTrEx-Segmented | **11.41** (+0.05) | **4.73** (+0.06) | **78.54** (-0.38) | **59.87** (-0.36) |
| MaTrEx-Segmented+Reorder | 11.21 (-0.30) | 4.57 (-0.12) | 79.64 (+1.30) | 62.21 (+1.76) |

Table VI.4: Scores for the automatic metrics for MaTrEx systems trained on the Consumer corpus and improvement compared to SMT.

Table VI.4 shows the scores obtained by the different techniques incorporated into the MaTrEx system, together with the score obtained by each system, and we show the difference with respect to the equivalent SMT system in brackets. As we can observe, those systems that did not use reordering

(MaTrEx-baseline and MaTrEx-Segmented) outperformed their SMT equivalents. The system that used reordering, on the other hand, obtain worse scores than those obtained by the SMT system for all the metrics.

The techniques used to enhance SMT did not have the same behaviour when they were incorporated to MaTrEx. When using the segmentation and the reordering techniques with MaTrEx the gain obtained by means of the reordering disappeared and worse results were obtained than the baseline (and the system that uses segmentation) for all the metrics. On the other hand, the results obtained by segmentation were still better than the baseline for most of the metrics (NIST, WER and PER); but not for BLEU, where the scores were equal.

### VI.2.2 Evaluation using enlarged corpora

Below, we present the evaluation of the same systems using the new enlarged corpora (see Table VI.5). We retrained all the corpus-based systems (SMT and EBMT systems) on the new corpora, and evaluated them with the same test set used previously (which was extracted from the Consumer corpus). Although the RBMT system could be tuned to the corpus (by adapting the dictionary to the terminology used), we did not perform any adaptation; hence the RBMT scores maintain the same values.

| | **BLEU** | **NIST** | **WER** | **PER** |
|---|---|---|---|---|
| Matxin (RBMT) * | 6.87 (=) | 3.78 (=) | 81.68 (=) | 66.06 (=) |
| SMT-baseline | 11.12 (+0.34) | 4.71 (+0.19) | 78.13 (-2.33) | 59.48 (-1.86) |
| SMT-Segmented | 11.56 (+0.20) | **4.83**(+0.16) | 77.83 (-1.09) | 58.94 (-1.29) |
| SMT-Segmented+Reorder | 11.19 (-0.32) | 4.69 (=) | 77.44 (-0.50) | 60.09 (-0.36) |
| MaTrEx-baseline * | 11.23 (-0.18) | 4.75 (+0.15) | 78.21 (-1.59) | 59.66 (-1.42) |
| MaTrEx-Segmented | **11.71**(+0.30) | 4.82 (+0.09) | 77.69 (-0.85) | 58.99 (-0.88) |
| MaTrEx-Segmented+Reorder * | 11.52 (+0.31) | 4.82 (+0.25) | **76.35**(-3.29) | **58.94**(-3.27) |
| Multi-Engine Hybridization * | 11.29 (+0.13) | 4.73 (+0.17) | 76.99 (-2.84) | 59.63 (-2.68) |
| Statistical Post-Editing * | 10.85 (+0.71) | 4.67 (+0.10) | 77.45 (-0.78) | 60.42 (-0.47) |

Table VI.5: Scores for the automatic metrics for all systems trained on the larger corpus.

When we enlarge the training corpus, the overall scores slightly increase. However, there are two exceptions, the MaTrEx-baseline and the SMT-Segmented+Reorder systems now obtain worse scores that when they are trained on the Consumer corpus. These inconsistencies of the scores make it extremely difficult to draw

conclusions. However, we want to remark some trends that remain (more or less) constant for the different evaluations performed.

- The MaTrEx system (which incorporates EBMT phrases into the SMT phrase-table) outperforms the state-of-the-art SMT. Excluding the MaTrEx-Segmented+Reorder system trained on the Consumer corpus (See Table VI.4), which clearly obtains worse results, the rest of the MaTrEx systems outperform (or equal, depending on the metric) their SMT versions.

- The use the segmentation improves the result for the baseline. This conclusion is supported by the different evaluations carried out, with the system that uses segmentation obtaining better scores, for all the metrics, than the system that does not use it.

- Regarding the reordering, the original conclusion drawn from the first evaluation, that combining the lexicalized-reordering with syntax-based reordering outperforms the isolated use of lexicalized reordering, is not supported by the later evaluations. The results obtained vary too much for the different evaluations. However, we would like to remark that according to the metric that most severely penalizes word order differences (WER) the system using reordering outperforms those that do not use it.

- All the automatic metrics severely penalize the RBMT system, and, consequently, the hybrid systems that make use of it are also penalized obtaining worse scores than the SMT system. However, in our opinion these Hybrid systems are the systems that obtain the best translation, and we expected to verify this perception by means of the HTER evaluation.

### VI.2.3 Albayzin open evaluation task

In 2008, jointly to the *Jornadas de Tecnologia del Habla* (JTH) conference, there was organized the Albayzin open evaluation, where, among other speech related task, a Spanish-Basque machine translation task was organized. The teams which took place on the evaluation were three, being our system the one that consider the best. According to the official results (see Table VI.6),

the difference between our system (called EHU-IXA) and the system called Avivavoz were not significant regarding to BLEU, but the scores obtained in the rest of the metrics (NIST, WER and PER) were the deciding factor.

| | **BLEU** | **NIST** | **WER** | **PER** |
|---|---|---|---|---|
| UPV-PRHLT | 7.11 | 3.65 | 82.64 | 65.56 |
| Avivavoz | **8.12** | 3.90 | 81.60 | 64.22 |
| EHU-IXA (MaTrEx-Segmented) | **8.10** | **3.98** | **78.70** | **62.25** |

Table VI.6: Official results provided by the Albayzin evaluation organizers.

The system that we present to the evaluation (Labaka et al., 2008) was equivalent to the system called "MaTrEx-Segmented" along this chapter. It was trained on the bilingual corpus provided by the organizers (a variation of the Consumer corpus used in this work) and was evaluated against a blind test set extracted from more recent articles of the same magazine (Consumer[4]).

The system presented by the partners of the Avivavoz project (Henríquez et al., 2008), also makes use of the segmentation of Basque, what in our opinion is the reason because both systems obtains so similar BLEU scores. Finally, the third system (called UPV-PRHLT) uses is based on a finite-state transducer which carries out the translation in a monotone way, what highly penalized its results (Sanchís-Trilles and Sánchez, 2008). Furthermore, the authors did not carried out any linguistic processing of Basque.

## VI.3 Human-targeted evaluation

Evaluating MT outputs is a complex task and several doubts have recently emerged concerning *BLEU*, which had become the most commonly used evaluation metric in the last decade. In addition to the fact that it is extremely difficult to interpret what is being expressed in *BLEU* (Melamed et al., 2003), recent works have shown that improving *BLEU* does not guarantee an improvement in the translation quality (Callison-Burch et al., 2006) and that BLEU scores do not offer as much correlation with human judgement as was believed (Koehn and Monz, 2006). Moreover, these objections are intensified by the fact that we only dispose of one reference to compute BLEU.

[4] www.consumer.es

In order to face those problems, we have decided to use human-targeted measures (Snover et al., 2006) and above all HTER (*Human-targeted Translation Error Rate*) (also called *edit distance* by Przybocki et al. (2006) and *post-editing cost* by Goutte (2006)) to evaluate a selection of our systems. Due to the high cost of human post-editions, we have had to make a limited selection of the implemented systems (those that are marked with an asterisk in Table VI.5). We have selected two SMT systems: the basic MaTrEx system as baseline and the enhanced one with reordering and segmentation; we have also evaluated the two hybrid systems developed in this work: Multiengine and Statistical Post-edition. In addition, we have evaluated the Spanish-Basque RBMT system, Matxin (Mayor, 2007), in order to carry out a better comparison and to know, more precisely, the real impact of all the systems in our multi-engine approach.

To perform the human-targeted evaluation, we have extracted texts from the corpora. We have chosen, at random, 200 sentences (made up of between 4 and 40 words) to be corrected in a post-edition phase done by a bilingual translator. These new references are used to calculate the human-targeted score. Table VI.7 shows some statistics of the test set used for human-targeted evaluation.

| | | **sentence** | **tokens** | **vocabulary** | **singletons** |
|---|---|---|---|---|---|
| **test** | Spanish | 200 | 4,190 | 1,769 | 1,445 |
| | Basque (token) | | 3,172 | 1,949 | 1,643 |

Table VI.7: Some statistics of the test set used for human-targeted evaluation.

Table VI.8 shows the results of this manual evaluation. Additionally, we carried out statistical significant test (by means of *Paired Bootstrap Resampling*) over the scores obtained by HBLEU. According to those test, we have verified (with a 95% confidence level) the following order relations between the systems:

$$\begin{aligned} &\text{HBLEU(Matxin)} &&< \text{HBLEU(Enhanced-MaTrEx)} \\ &\text{HBLEU(MaTrEx-baseline)} &&< \text{HBLEU(Enhanced-MaTrEx)} \\ &\text{HBLEU(Enhanced-MaTrEx)} &&< \text{HBLEU(Multiengine)} \end{aligned}$$

On the other hand, although the overall evaluation score is higher, the different between the Enhanced-MaTrEx system and the Statistical Post-edition could not be with a 95% confidence level, and statistical significance is only corroborated when the confidence level is reduced to 85%. Similarly,

the difference between Matxin system and the Matrex-baseline as well as the one between the Multiengine system and Statistical Post-edition system only can be verified in even lower confidence levels, which does not allow us to establish a preference between them.

Those figures illustrate the fact that all the systems proposed in this PhD thesis outperform the SMT baseline consistently for all human-targeted evaluation measures. So that, we can conclude that the techniques applied to enhance SMT make an important contribution at translating into a highly inflected language like Basque. In addition, the two hybridization attempts obtain even better results, showing up as an interesting field in which to continue our investigation.

| | **HTER** | **HBLEU** | **HNIST** | **HWER** | **HPER** | **BLEU** |
|---|---|---|---|---|---|---|
| Matxin | 54.74 | 26.88 | 6.84 | 58.51 | 42.98 | 6.87 |
| MaTrEx-baseline | 53.59 | 27.86 | 7.23 | 58.48 | 40.23 | 11.23 |
| Enhanced-MaTrEx | 48.10 | 33.29 | 7.60 | 54.52 | 35.45 | **11.52** |
| Multiengine | 47.62 | 34.71 | 7.64 | 53.74 | **35.27** | 11.29 |
| Statistical-Postedition | **47.41** | **34.80** | **7.74** | **52.04** | 36.05 | 10.85 |

Table VI.8: Scores for the human-targeted metrics for selected systems. BLEU scores obtained in the automatic evaluation are also included.

We have to point out that the Matxin RBMT system was not properly tuned when the evaluation was performed and that it obtained the worse score. Nevertheless, we can observe that RBMT positively contributes in the Multiengine's performance. On the other hand, although the EBMT system only completes translation for a small number of sentences, the translation obtained for those sentences is almost perfect. Referring to EBMT, for those few sentences for which the EBMT system obtains a translation (5.50% of all sentences) the HTER error rate is very low (5.29).

The techniques presented to adapt the SMT to a morphologically rich language are effective (achieving a 10% gain in HTER and a 16% gain in HBLEU over the baseline), even though the baseline used already incorporates some improvements (such as the lexicalized reordering and MaTrEx's enriched phrase-tables). However, we can not measure the improvement achieved by each technique, since we evaluated both of them jointly. The results obtained by means of the automatic metrics are contradictory and do not enable us to draw clear conclusions. Therefore, we would like to evaluate the system that

makes use of segmentation, but not reordering, by means of human-targeted evaluation, in order to measure the contribution of each technique.

Based on the results for HTER and in order to delimit the upper bound of the MultiEngine approach, we calculated the results for two oracle systems. These oracle systems select the best translation from the output of different systems, based on their HTER scores. In that way, we want to measure the maximal improvement we could obtain using a MultiEngine system by changing only the method of selection, without changing the individual translation systems. We measured two different oracle systems: the first one makes use of the same systems used in the previous MultiEngine experiments (that is, RBMT, EBMT and the SMT system developed in the thesis). The second one also incorporates, besides these three systems, the post-edition system. Table VI.9 shows the scores obtained by the oracle systems.

| | **HTER** |
|---|---|
| Oracle-MultiEngine1 | 42.10 |
| Oracle-MultiEngine2 | 37.857 |

Table VI.9: HTER scores for the oracle MultiEngine systems.

The first conclusion we can draw from these scores is that there is still room for improvement via MultiEngine hybridization. The HTER scores obtained by the oracle systems are significantly better than the scores for the rest of the systems, with even better results when the Statistical Post-Editing system is incorporated.

The usefulness of the RBMT system for assimilation was checked (69% of the users found RBMT translation useful[5] when integrated in a Multi-Lingual Information Retrieval system) in a previous work (Leturia et al., 2009). But if we were able to achieve the translation quality obtained by the oracle system (37.85 HTER score) the Spanish-Basque MT would be also useful for Computer-Aided Translation system (since when HTER goes down to 40%, post-editing an MT output is faster than creating a new translation).

[5]In 30.00% of the cases the users found the translations of the MT system "very good", "good" or "quite good" and in another 38.89% of the cases they found them "comprehensible"

### VI.3.1 Examples

In this subsection we include running examples of the translations obtained by all the MT systems evaluated. All the translations are obtained from the same Spanish sentence ('*Procure que la temperatura de la calefacción se mantenga alrededor de 20 C, nivel ideal para una vivienda.*' /Maintain the temperature of the heater around 20C, the ideal temperature for a house/).

The EBMT engine covers the translation of this sentence, which will be the translation selected by the MultiEngine system. In Figure VI.1 we can see from its output that it is correct and does not need any post-edition.

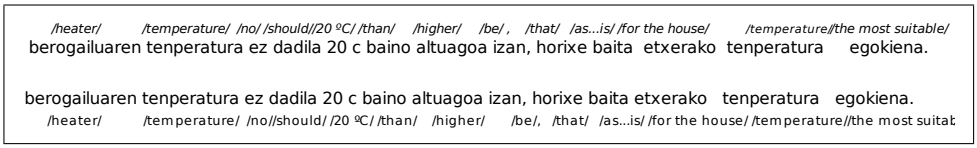



Figure VI.1: Post-edition performed on the EBMT system's output.

We have also translated the same sentence using the other translation approaches. Figure VI.2 shows how the RBMT system translates the Spanish source sentence and its post-edition; we can observe that 4 corrections are needed (all of them are word substitutions).

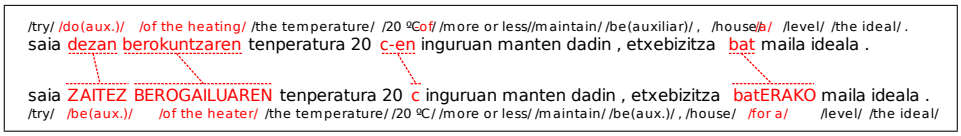



Figure VI.2: Post-edition performed on the RBMT system's output.

In Figure VI.3 we can observe the translation for the source sentence created by the MaTrEx-baseline system. In order to correct this translation we need 6 post-editions (3 substitutions and 3 shifts).

In Figure VI.4, we can see the output of the Enhanced-MaTrEx system. This translation needed 6 corrections (2 deletions and 4 substitutions) in order to obtain a correct translation.

Finally, Figure VI.5 shows the translation of the Statistical Post-Editing system. Starting from the automatic output, we need 6 post-editions (4 substitutions, 1 insertion and 1 deletion) to generate a correct translation.

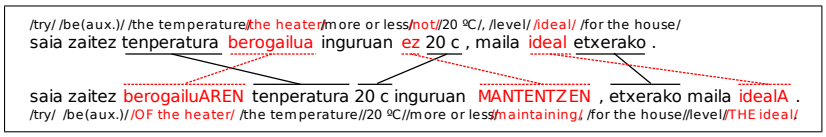



Figure VI.3: Post-edition performed on the MaTrEx-baseline system's output.

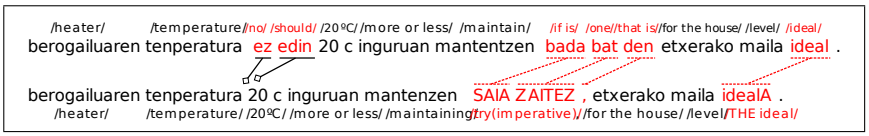



Figure VI.4: Post-edition performed on the Enhanced-MaTrEx system's output.

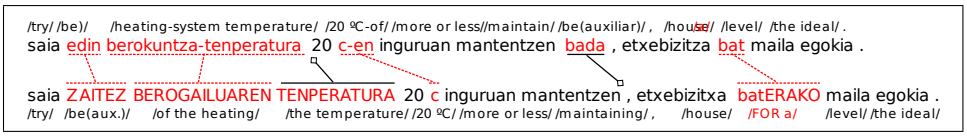



Figure VI.5: Post-edition performed on the Statistical Post-editing system's output.

These examples are representative of the systems' behaviour. When the EBMT system finds a translation, it is almost perfect (as in this example) and it is used as the final translation of the MEMT system.

The other four systems always obtain a translation, but they are not as accurate. The baseline SMT system usually uses the right lemmas but the word order and the inflection are not correct, producing a translation that is difficult to understand. However, when the SMT system incorporates the use of morphology and reordering, it generally gives better results (in relation both to the inflection and word order), but, sometimes (as in this example) it adds some "noise", which does not correspond to anything in the source, making the translation difficult to understand.

Finally, although the RBMT system gets the worst overall results of all the evaluated systems it obtains a really high quality translation for some sentences, which helps the MEMT system to improve its results. Due to the particularly good translation of the RBMT system for this sentence, Statistical Post-Editing is not able to correct the RBMT output (in contrast

to its usual performance).

## VI.4 Chapter Summary and Conclusions

The results of the manual evaluation performed clearly confirm that the new techniques proposed in this PhD thesis are definitely valuable. Although the baseline we choose as reference is quite high (the MaTrEx system always obtains best results than Moses), the use of segmentation and reordering allows our system to achieve a relative improvement of 10% in the HTER metric.

Furthermore, the hybrid systems we have implemented are also effective, even though they are preliminary experiments. The two techniques developed in this work helped to increase the translation quality. It is not so clear when using automatic metrics, as they show significant improvements using hybridization techniques in a specific domain, such as Labour Agreement, but they don't show any improvement in a general domain corpus, such as the Consumer corpus. However, HTER evaluation definitely proves their reliability, showing a clear improvement even in a general domain.

Finally, we want to remark on the possibility of improvement in Multi-Engine hybridization. As we can see in the results for the oracle systems, we can still make a significant improvement working on the selection of the individual MT outputs.

# CONCLUSION AND FURTHER WORK

# CHAPTER VII

## Conclusion and Further Work

Due to the multilingual nature of the present information society, Human Language Technologies and Machine Translation have become essential for the survival of minority languages such as Basque. Even so, the lower economic interest in these languages prevents much research being carried out on their particular characteristics. Basque has to face up not only the problems arising from it being a minority language (lack of funding and resources), but also several linguistic peculiarities (both morphological and syntactical) that make translation a truly challenging issue.

In this context, we have adapted SMT to translate into Basque. First, we analyzed the impact of the agglutinative nature of Basque and the best way to deal with it. Similarly, we also studied the differences in word order between Spanish and Basque, examining different techniques for dealing with them.

Once we had obtained a more accurate SMT system, we started the first attempts to combine different MT systems into a hybrid one that would allow us to get the best of the different paradigms. The hybridization attempts carried out in this PhD dissertation are preliminaries, but, even so, this work can help us to determine the ongoing steps.

## VII.1 Conclusions

In **chapter III** we prove that the quality of the translation varies significantly when applying different options for word segmentation, even if they are based on the output produced by the morphological analyzer. In order to do so, we segmented Basque 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. The criteria based on considering each morpheme as a separate token obtain worse results than the system in which segmentation is not applied. The other segmentation options outperform the baseline, the best results being obtained with a manually-defined intermediate morpheme grouping criterium based on an error analysis of word alignments.

Analyzing the results obtained, we realized that there is a correlation between the size of the vocabulary generated at segmentation and the results obtained at evaluation. Intermediate segmentations, where morphemes marked by the morphological analyzer are grouped in different ways, achieve better results when the target vocabulary size is closer to the size of the vocabulary in the source language.

In **chapter IV** we confirm the weakness of the basic SMT in dealing with great word order differences in the source and target languages. Distance-based reordering, which is the technique used by the baseline system, does not have enough information to properly handle big word order differences, so any of the techniques tested in this work (based on both statistics and manually generated rules) outperforms the baseline.

In addition, as the combination of statistical techniques at decoding and syntax-based pre-processing gets the best results, we conclude that both techniques handle a different kind of reordering. Lexicalized reordering is limited to continuous phrases, which allows the decoder a medium-range reordering, but long-range reordering still remains untreated. Instead, syntax-based pre-processing mainly treats long-range reordering, moving whole phrases along the sentence. Thus, we consider that for language pairs which need long-range reordering, such as Spanish-Basque, the incorporation of syntactical information in the reordering process is helpful. We also observe that techniques that are incorporated into the decoding process get better results, that is why we consider this an interesting direction for further research.

In **chapter V** we explore two ways of performing hybridization of dif-

ferent MT approaches. On the one hand, we develop a Multi-Engine system which translates each sentence using three different MT systems (the SMT system developed in this PhD thesis, and the previously systems, RBMT and EBMT, developed in our group). Even using a quite simple hierarchical selection strategy, we achieve a significant improvement with automatic metrics such as BLEU.

In the other hand, we use Statistical Machine Translation for post-editing the output of the RBMT system. Thus, the SMT system "translates" from the output of the RBMT system to real Basque. Statistical Post Editing turned out to be a valuable method for improving the results obtained by RBMT. In any case, the behavior differs according to the domain. Thus, using SPE after RBMT clearly outperforms SMT in a domain specific corpus, but scores obtained in general domain corpus do not outperform SMT, even though it improves the results achieved by RMBT.

In **chapter VI** we perform an final overall human evaluation in order to evaluate all systems in the same framework and to verify the conclusions drawn from automatic evaluation. Besides, to make a more reliable evaluation, we build, for training, a seven times larger bilingual corpus collected during the last three years. The human evaluation confirms that the new techniques proposed in this PhD thesis are definitely valid. The use of segmentation and reordering allows our system to make a relative improvement of 10% in the HTER metrics.

Furthermore, the hybrid systems we have implemented are also effective, even though they are preliminary experiments. The two techniques developed in this work helped to increase the translation quality. This is not so clear when using automatic metrics, as they show significant improvement by using hybridization techniques in a specific domain, such as Labour Agreement, but they bring no gain in a general domain corpus, such as Consumer. However, the final HTER evaluation proves their reliability, showing, for the Multi-Engine System, a statistical significant improvement even in the general domain.

Finally, we want to remark on the possibility of improvement in Multi-Engine hybridization. As we can see in the results for the oracle systems, we can still make a significant improvement working on the selection of the individual MT outputs.

Taking into account the work presented in this dissertation, we consider

we have achieved all our objectives, getting a significant improvement for SMT and pave the way for a better translation quality by means of hybridization. We consider that we have established a basis for SMT and hybridization research for Basque. These results allow us to identify the main research priorities in the coming years. Although this work has been specifically developed for Basque, we consider that these findings can be extended to other agglutinative languages with a highly free order of sentence components.

## VII.2 Further Work

There are some open research lines in this work that can be explored further. We will describe the main experiments and paths that we would like to explore in the future.

- Further experiments in automatic segmentation criteria. We want to experiment with a different method, such as $\chi^2$, to determine the statistical interdependence between consecutive morphemes and use the correlation between the vocabulary size of both languages as a criterion when defining this segmentation. In addition, we want to test this statistical grouping criterion with a different language pair such as English-Basque, in order to test its language independence.

- Enhanced syntax-based reordering. We are planning to add more specific reordering rules, since the present ones do not exhaustively cover all the order differences between the two languages. Furthermore, we are considering a way to allow the decoder to chose from among different reorderings proposed by the syntax-based pre-processing (using an n-best list of reordering alternatives, or using a word-graph as decoder input).

- Going deeper into multi-engine hybridization. We have not considered elements smaller than the whole sentence when combining the translation output of different MT engines. By splitting the translations into phrases and merging the phrases proposed by different engines, we expect to make the most of each engine.

- Further post-edition experiments. We are planning to automatically learn post-editing rules to correct SMT translation in the way Elming

(2006) does. In order to carry out this kind of experiment, we will have to collect a real post-edition corpus.

- We are also interested in analyzing the suitability of n-gram based evaluation metrics when translating into languages like Basque. The agglutinative nature of Basque together with its word order freedom entails the need of more references to cover different translation possibilities (something that is not easily obtained in less-resourced languages like Basque). On the other hand, evaluation metrics, based on human post-edition achieves more accurate results, since each translation is compared with the closest reference possible.

# Bibliography


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. Bernarrd Oyharçabal (Ed.)*, Bilbao.

Akiba, Y., Watanabe, T., and Sumita, E. (2002). Using Language and Translation Models to Select the Best among Outputs from Multiple MT Systems. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–7, Morristown, NJ, USA. Association for Computational Linguistics.

Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, D., Och, F.-J., Purdy, D., Smith, N. A., and Yarowsky, D. (1999). Statistical Machine Translation. Final Report. Technical report, JHU Summer Workshop.

Albrecht, J. S. and Hwa, R. (2007). A Re-examination of Machine Learning Approaches for Sentence-level MT Evaluation. In *Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 880–887, Prague, Czech Republic.

Alcázar, A. (2005). Towards Linguistically Searchable Text. In *Proceedings of BIDE (Bilbao-Deusto) Summer School of Linguistics*, Bilbao.

Alegria, I., Artola, X., Sarasola, K., and Urkia, M. (1996). Automatic Morphological Analysis of Basque. *Literary & Linguistic Computing*, 11(4):193–203.

Alegria, I., Casillas, A., Díaz de Ilarraza, A., Igartua, J., Labaka, G., Lersundi, M., Mayor, A., and Sarasola, K. (2008). 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*, Hawaii, USA.

Alegria, I., Díaz de Ilarraza, A., Labaka, G., Lersundi, M., Mayor, A., and Sarasola, K. (2006). An FST Grammar for Verb Chain Transfer in a Spanish-Basque MT System. In Yli-Jyrä, A., Karttunen, L., and Karhumäki, J., editors, *Finite-State Methods and Natural Language Processing, 5th International Workshop, FSMNLP 2005, Helsinki, Finland, September 1-2, 2005. Revised Papers*, volume 4002 of *Lecture Notes in Computer Science*, pages 87–98. Springer.

Amigó, E., Giménez, J., Gonzalo, J., and Màrquez, L. (2006). MT Evaluation: Human-like vs Human-acceptable. In *Coling-ACL 2006: Proceedings of the Coling/ACL 2006 Main Conference: Poster Sessions*, pages 17–24.

Amorrortu, E. (2002). Bilingual Education in the Basque Country: Achievements and Challenges after Four Decades of Acquisition Planning. *Journal of Iberian and Latin American Literary and Cultural Studies*, 2.

Armentano-Oller, C., Corbí-Bellot, A. M., Forcada, M. L., Ginestí-Rosell, M., Bonev, B., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G., and Sánchez-Martínez, F. (2005). An Open-Source Shallow-Transfer Machine Translation Toolbox: Consequences of its Release and Availability. In *OSMaTran: Open-Source Machine Translation, A workshop at Machine Translation Summit X*, pages 23–30.

Arnold, D., Balkan, L., Meijer, S., Humphreys, R., and Sadler, L. (1993). *Machine Translation: an Introductory Guide*. Blackwells-NCC, London.

Avramidis, E. and Koehn, W. B. P. (2008). Enriching Morphologically Poor Languages for Statistical Machine Translation. In *Proceedings of Human Language Technologies: The Annual Conference of the Association for Computational Linguistics (ACL-HLT'08)*, pages 763–770.

Berger, A. L., Della Pietra, V. J., and Della Pietra, S. A. (1996). A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39–71.

Birch, A., Osborne, M., and Koehn, P. (2007). CCG Supertags in Factored Statistical Machine Translation. In *StatMT '07: Proceedings of the Second Workshop on Statistical Machine Translation*, pages 9–16, Morristown, NJ, USA. Association for Computational Linguistics.

Bojar, O. (2007). English-to-Czech Factored Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 232–239, Prague, Czech Republic. Association for Computational Linguistics.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Mercer, R. L., and Roossin, P. S. (1988). A Statistical Approach to Language Translation. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING)*, pages 71–76.

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

Callison-Burch, C. and Flournoy, R. S. (2001). A Program for Automatically Selecting the Best Output from Multiple Machine Translation Engines. In *In Proceedings of the Machine Translation Summit VIII*, pages 63–66.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the International Conference of European Chapter of the Association for Computational Linguistics (EACL)*, pages 249–256.

Carl, M. and Way, A. (2003). *Recent Advances in Example-Based Machine Translation*, volume 21 of *Text, Speech and Language Technology*. Kluwer Academic Publishers, Dordrecht.

Carreras, X., Chao, I., Padró, L., and Padró, M. (2004). Freeling: an Open-Source Suite of Language Analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pages 239–242.

Casillas, A., Abaitua, J., and Martínez, R. (2000). Recycling Annotated Parallel Corpora for Bilingual Document Composition. In White, J. S., editor, *Envisioning machine translation in the information future: 4th conference*

*of the Association for Machine Translation in the Americas, AMTA 2000*, pages 117–126, Cuernavaca, Mexico. Springer Verlag.

Casillas, A., Díaz de Ilarraza, A., Igartua, J., Sarasola, K., Sologaistoa, A., and Martínez, R. (2007). Spanish-Basque Parallel Corpus Structure: Linguistic Annotations and Translation Units. In *Text, Speech and Dialogue*, volume 4629 of *Lecture Notes in Computer Science*, pages 230–237. Springer.

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

Chen, Y., Eisele, A., Federmann, C., Hasler, E., Jellinghaus, M., and Theison, S. (2007). Multi-engine Machine Translation with an Open-Source Decoder for Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 193–196.

Chen, Y., Jellinghaus, M., Eisele, A., Zhang, Y., Hunsicker, S., Theison, S., Federmann, C., and Uszkoreit, H. (2009). Combining Multi-Engine Translations with Moses. In *StatMT '09: Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 42–46, Morristown, NJ, USA. Association for Computational Linguistics.

Chiang, D. (2005). A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *ACL*, pages 263–270.

Collins, M., Koehn, P., and Kucerova, I. (2005). Clause Restructuring for Statistical Machine Translation. In *ACL*, pages 531–540.

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

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality using N-gram Co-Occurrence Statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Du, J., Ma, Y., and Way, A. (2009). Source-Side Context-Informed Hypothesis Alignment for Combining Outputs from Machine Translation Systems.

In *Proceedings of the Twelfth Machine Translation Summit*, pages 230–237, Ottawa, Canada.

Dugast, L., Senellart, J., and Koehn, P. (2007). Statistical Post-Editing on SYSTRAN's Rule-Based Translation System. In *StatMT '07: Proceedings of the Second Workshop on Statistical Machine Translation*, pages 220–223, Morristown, NJ, USA. Association for Computational Linguistics.

Dugast, L., Senellart, J., and Koehn, P. (2009). Selective Addition of Corpus-Extracted Phrasal Lexical Rules to a Rule-Based Machine Translation System. In *Proceedings of the Twelfth Machine Translation Summit*, Ottawa, Canada.

Ehara, T. (2007). Rule Based Machine Translation Combined with Statistical Post Editor for Japanese to English Patent Translation. In *Proceedings of MT-Summit XI: Workshop on Patent Translation*, pages 13–18, Copenhagen, Denmark.

Eisele, A. (2005). First Steps towards Multi-Engine Machine Translation. In *In Proceedings of the ACL Workshop on Building and Using Parallel Texts*.

Elming, J. (2006). Transformation-Based Correction of Rule-Based MT. In *Proceedings of the 11th Annual Conference of the European Association for Machine Translation*, pages 219–226, Oslo, Norway.

Farwell, D. and Padró, L. (2009). FreeLing: From a Multilingual Open-Source Analyzer Suite to an EBMT Platform. In *Proceedings of 3rd International Workshop on Example-Based Machine Translation*, pages 37–45.

Frederking, R. and Nirenburg, S. (1994). Three heads are better than one. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, pages 95–100, Morristown, NJ, USA. Association for Computational Linguistics.

Freund, Y. and Schapire, R. E. (1997). A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Giménez, J. and Màrquez, L. (2007). Linguistic Features for Automatic Evaluation of Heterogeneous MT Systems. In *Annual Meeting of the Association for Computational Linguistics (ACL): Second Workshop on Statistical Machine Translation*, pages 256–264, Prague, Czech Republic.

Giménez, J. and Màrquez, L. (2008). Heterogeneous Automatic MT Evaluation through Non-Parametric Metric Combinations. In *Proceedings of the IJCNLP 2008: Third International Joint Conference on Natural Language Processing*, pages 319–326, Hyderabab, India.

Ginestí-Rosell, M., Ramírez-Sánchez, G., Ortiz-Rojas, S., Tyers, F. M., and Forcada, M. L. (2009). Development of a Free Basque to Spanish Machine Translation System. *Journal of the Spanish Association for Natural Language Processing*, 43:187–195.

Goldwater, S. and McClosky, D. (2005). Improving Statistical MT through Morphological Analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 676–683, Vancouver, Canada.

González, J., Ortíz, D., Tomás, J., and Casacuberta, F. (2004). A comparison of Different Statistical Machine Translation Techniques for SpanishtoBasque Translation. In *Proceedings of III Jornadas en Tecnología del Habla (VJTH'2004)*.

Goutte, C. (2006). Automatic Evaluation of Machine Translation Quality. *A Xerox Research Centre Europe Publication*.

Groves, D. and Way, A. (2005). Hybrid Data-Driven Models of Machine Translation. *Machine Translation*, 19(3-4):301–323.

Henríquez, C. A., Khalilov, M., Mariño, J. B., and Ezeiza, N. (2008). The AVIVAVOZ Phrase-Based Statistical Machine Translation System for ALBAYZIN 2008. In *Proceedings of V Jornadas en Tecnología del Habla (VJTH'2008)*.

Holmqvist, M., Stymne, S., and Ahrenberg, L. (2007). Getting to Know Moses: Initial Experiments on German–English Factored Translation. In *StatMT '07: Proceedings of the Second Workshop on Statistical Machine Translation*, pages 181–184, Morristown, NJ, USA. Association for Computational Linguistics.

Hutchins, W. J. (1986). *Machine Translation: Past, Present, Future*. Ellis Horwood Limited, Chichester, England.

Hutchins, W. J. and Somers, H. L. (1992). *An Introduction to Machine Translation*. Academic Press, London.

Isabelle, P., Goutte, C., and Simard, M. (2007). Domain Adaption of MT Systems through Automatic Post-editing. In *Proceedings of MT-Summit XI*, pages 255–261, Copenhagen, Denmark.

Kenneth W. Church, P. H. (1989). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Knight, K. (1999). Decoding Complexity in Word-Replacement Translation Models. *Computational Linguistics*, 25:607–615.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit X*, pages 79–86, Phuket, Thailand.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press, Cambridge.

Koehn, P. and Hoang, H. (2007). Factored Translation Models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Processing and Computational Natural Language Learning*, pages 868–876.

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., 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.

Koehn, P. and Knight, K. (2003). Empirical Methods for Compound Splitting. In *Proceedings of the International Conference of European Chapter of the Association for Computational Linguistics (EACL'03)*, Budapest, Hungary.

Koehn, P. and Monz, C. (2006). Manual and Automatic ´Evaluation of Machine Translation between European Languages. In *In Proceedings of NAACL 2006 Workshop on Statistical Machine Translation*, pages 102–121.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical Phrase-Based Translation. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT'03)*.

Koskeniemmi, K. (1983). Two-level Model for Morphological Analysis. In *Proceedings of the Eigth International Joint Conference on Artificial Intelligence*, pages 683–685, Karlsruhe, Germany.

Labaka, G., Díaz de Ilarraza, A., and Sarasola, K. (2008). Descripción de los sistemas presentados por IXA-EHU a la evaluación ALBAYCIN08. In *Proceedings of V Jornadas en Tecnología del Habla (VJTH'2008)*.

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 MT-Summit XI*, pages 297–304.

Lager, T. (1999). The $\nu$-TBL System: Logic Programming Tools for Transformation-Based Learning. In *Proceedings of the Third International Workshop on Computational Natural Language Learning (CoNLL'99)*.

Lardilleux, A., Chevelu, J., Lepage, Y., Putois, G., and Gosme, J. (2009). Lexicons or phrase tables? An Investigation in Sampling-Based Multilingual Alignment. In *Proceedings of 3rd International Workshop on Example-Based Machine Translation*, pages 45–53.

Lepage, Y. and Denoual, E. (2005). Purest ever Example-Based Machine Translation: Detailed Presentation and Assessment. *Machine Translation*, 19:251–282.

Leturia, I., del Pozo, A., Arrieta, K., Iturraspe, U., Sarasola, K., Díaz de Ilarraza, A., Navas, E., and Odriozola, I. (2009). Development and Evaluation of AnHitz, a Prototype of a Basque-Speaking Virtual 3D Expert on Science and Technology. In *Computational Linguistics - Applications Workshop (CLA'09)*.

Leusch, G., Matusov, E., and Ney, H. (2009). The RWTH System Combination System for WMT 2009. In *StatMT '09: Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 56–60, Morristown, NJ, USA. Association for Computational Linguistics.

Leusch, G., Ueffing, N., and Ney, H. (2006). CDER: Efficient MT Evaluation Using Block Movements. In *Proceedings of the International Conference of European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 242–248.

Liu, D. and Gildea, D. (2005). Syntactic Features for Evaluation of Machine Translation. In *Annual Meeting of the Association for Computational Linguistics (ACL'05): Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32, Prague, Czech Republic.

Liu, D. and Gildea, D. (2007). Source-Language Features and Maximum Correlation Training for Machine Translation Evaluation. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT'07)*, pages 41–48.

Macherey, W. and Och, F. J. (2007). An Empirical Study on Computing Consensus Translations from Multiple Machine Translation Systems. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 2007*, pages 986–995, Morristown, NJ, USA. Association for Computational Linguistics.

Martínez, R., Abaitua, J., and Casillas, A. (1998). Alingning Tagged Bitext. In *Coling-ACL 1998: Proceedings of the Sixth Workshop on Very Large Corpora*, pages 102–109, Montreal, Quebec, Canada.

Martínez, R., Abaitua, J., and Casillas, A. (1998). Bitext Correspondences through Rich Mark-up. In Charniak, E., editor, *Proceedings of Coling-ACL 1998: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 812–818, Montreal, Quebec, Canada.

Matusov, E., Ueffing, N., and Ney, H. (2006). Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment. In *Proceedings of the International Conference of European Chapter of the Association for Computational Linguistics. EACL'06*, pages 33–40.

Mayor, A. (2007). *Matxin: erregeletan oinarritutako itzulpen automatikoko sistema*. PhD thesis, Euskal Herriko Unibertsitatea.

McTait, K. M. and Trujillo, A. (1999). A language-neutral sparse-data algorithm for extracting translation patterns. In *Proceedings of 8th International Conference on Theorical and Methodological Issues in Machine Translation.*, pages 23–30, Chester, England.

Mehay, D. N. and Brew, C. (2007). BLEUÂTRE: Flattening Syntactic Dependencies for MT Evaluation. In *Proceedings of TMI-2007: The 11th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 122–131.

Melamed, I. D., Green, R., and Turian, J. P. (2003). Precision and Recall of Machine Translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 61–63, Morristown, NJ, USA. Association for Computational Linguistics.

Mellebeek, B., Owczarzak, K., Genabith, J. V., and Way, A. (2006). 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*, pages 110–118.

Minkov, E., Toutanova, K., and Suzuki, H. (2007). Generating Complex Morphology for Machine Translation. In *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 128–135, Prague, Czech Republic.

Nadeu, C., Ó'Cróinín, D., Petek, B., Sarasola, K., and Williams, B. (2001). ISCA SALTMIL SIG: Speech and Language Technology for Minority Languages. In *Proc. EUROSPEECH*, pages 1353–1360.

Nagao, M. (1984). A Framework of a Mechanical Translation Between Japanese and English by Analogy Principle. In *Proceedings of the International NATO Symposium on Artificial and Human Intelligence*, pages 173–180, New York, NY, USA. Elsevier North-Holland, Inc.

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

Nießen, S. and H. Ney (2004). Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information. *Computational Linguistics*, 30(2):181–204.

Nießen, S., Och, F. J., Leusch, G., and Ney", H. (2000). An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of LREC-2000: Second International Conference on Language Resources and Evaluation*, pages 39–45.

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

Och, F. J. (2003). Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167.

Och, F. J. and Ney, H. (2002). Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 295–302, Morristown, NJ, USA. Association for Computational Linguistics.

Och, F. J., Tillmann, C., Ney, H., and Informatik, L. F. (1999). Improved Alignment Models for Statistical Machine Translation. In *University of Maryland, College Park, MD*, pages 20–28.

Oflazer, K. and El-Kahlout, I. D. (2007). Exploring Different Representation Units in English-to-Turkish Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 25–32, Prague, Czech Republic.

Ortíz, D., GarcíaVarea, I., Casacuberta, F., Lagarda, A., and González, J. (2003). On the Use of Statistical MachineTranslation Techniques within a MemoryBased Translation System (AMETRA). In *Proceedings of the Machine Translation Summit: MT Summit IX*, pages 299–306.

Padó, S., Galley, M., Jurafsky, D., and Manning, C. (2007). Robust Machine Translation Evaluation with Entailment Features. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP (ACL-IJCNLP-2009)*, pages 297–305, Prague, Czech Republic.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th ACL*, pages 311–318, Philadelphia, PA.

Papineni, K. A., Roukos, S., and Ward, R. T. (1998). Maximum Likelihood and Discriminative Training of Direct Translation Models. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 189–192, Seattle, Washington, USA.

Pérez, A., Inés Torres, M., and Casacuberta, F. (2008). Joining linguistic and statistical methods for spanish-to-basque speech translation. *Speech Communication*, 50(11-12):1021–1033.

Petek, B. (2000). Funding for Research into Human Language Technologies for Less Prevalent Languages. In *Second Internacional Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece.

Phillips, A. and Brown, R. (2009). Cunei Machine Translation Platform: System Description. In *Proceedings of 3rd International Workshop on Example-Based Machine Translation*, pages 29–37.

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

Przybocki, M., Sanders, G., and Le, A. (2006). Edit Distance: A Metric for Machine Translation Evaluation. In *LREC-2006: Fifth International Conference on Language Resources and Evaluation*, pages 2038–2043, Genoa, Italy.

Quirk, C., Menezes, A., and Cherry, C. (2005). Dependency Treelet Translation: Syntactically Informed Phrasal SMT. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 271–279, Morristown, NJ, USA. Association for Computational Linguistics.

Ramanathan, A., Bhattacharya, P., Hegde, J., M.Shah, R., and M, S. (2008). Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation. In *Third International Joint Conference on Natural Language Processing (JCNLP'08)*, pages 513–520, Hyderabad, India.

Rosti, A.-V. I., Ayan, N. F., Xiang, B., Matsoukas, S., Schwartz, R., and Dorr, B. J. (2007). Combining Outputs from Multiple Machine Translation Systems. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT'07)*, pages 228–235.

Sánchez-Martínez, F., Forcada, M. L., and Way, A. (2009). Hybrid Rule-Based - Example-Based MT: Feeding Apertium with Sub-Sentential Translation Units. In *EBMT 2009 - 3rd Workshop on Example-Based Machine Translation*, pages 11–18, Dublin, Ireland. DORAS.

Sanchís, G. and Casacuberta, F. (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.

Sanchís-Trilles, G. and Sánchez, J. A. (2008). Phrase Segments Obtained with Stochastic Inversion Transduction Grammars for Spanish-Basque Translation. In *Proceedings of V Jornadas en Tecnología del Habla (VJTH'2008)*.

Shieber, A. K. . S. M. (2004). A learning Approach to Improving Sentence-level MT Evaluation. In *TMI-2004: proceedings of the Tenth Conference on Theoretical and Methodological Issues in Machine Translation*, pages 75–84.

Simard, M., Goutte, C., and Isabelle, P. (2007a). Statistical Phrase-Based Post-Editing. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT'07)*, pages 508–515.

Simard, M., Ueffing, N., Isabelle, P., and Kuhn, R. (2007b). Rule-Based Translation with Statistical Phrase-Based Post-Editing. In *StatMT '07: Proceedings of the Second Workshop on Statistical Machine Translation*,

pages 203–206, Morristown, NJ, USA. Association for Computational Linguistics.

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*, pages 223–231, Columbus, Ohio.

Somers, H., Dandapat, S., and Naskar, S. (2009). A review of EBMT using Proportional Analogies. In *Proceedings of 3rd International Workshop on Example-Based Machine Translation*, pages 53–61.

Sommers, H. (2003). An overview of ebmt. In Carl, M. and Way, A., editors, *Recent Advances in Example-Based Machine Translation*, volume 21 of *Text, Speech and Language Technology*, pages 3–57. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference of Spoken Language Processing*, pages 901–904, Denver, Colorado.

Stroppa, N. and Way, A. (2006). MaTrEx: DCU Machine Translation System for IWSLT 2006. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 31–36, Kyoto, Japan.

Stroppa, N., Way, A., and Sarasola, K. (2006). Example-Based Machine Translation of the Basque Language. In *Proceedings of the 7th Conference of the Association for Machine Translation in the, Visions for the Future of Machine Translation*, pages 232–241.

Tidhar, D. and Küssner, U. (2000). Learning to Select a Good Translation. In *Proceedings of the 18th conference on Computational linguistics*, pages 843–849, Morristown, NJ, USA. Association for Computational Linguistics.

Tillmann, C., Vogel, S., and Zubiaga, A. (1997). A DP Based Search Using Monotone Alignments in Statistical Translation. In *Proceedings of the EACL-EACL-1997: 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 289–296.

Toutanova, K., Suzuki, H., and Ruopp, A. (2008). Applying Morphology Generation Models to Machine Translation. In *Proceedings of Human Language Technologies: The Annual Conference of the Association for Computational Linguistics (ACL-HLT'08)*, pages 514–522.

Tyers, F. M. (2009). Rule-based Augmentation of Training Data in Breton-French Statistical Machine Translation. In Màrquez, L. and Somers, H., editors, *EAMT-2009: Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, pages 213–217.

van Gompel, M., van den Bosch, A., and Berck, P. (2009). Extending Memory-Based Machine Translation to Phrases. In *Proceedings of 3rd International Workshop on Example-Based Machine Translation*, pages 61–69.

van Zaanen, M. and Somers, H. (2005). DEMOCRAT: Deciding between Multiple Outputs Created by Automatic Translation. In *MT Summit X: Asia-Pacific Association for Machine Translation*, pages 173–180, Bangkok, Thailand.

Vandeghinste, V. and Martens, S. (2009). Top-down Transfer in Example-based MT. In *Proceedings of 3rd International Workshop on Example-Based Machine Translation*, pages 69–77.

Vogel, S., Ney, H., and Tillmann, C. (1996). HMM-Based Word Alignment in Statistical Translation. In *Proceedings of the 16th conference on Computational linguistics*, pages 836–841, Morristown, NJ, USA. Association for Computational Linguistics.

White, J. S. and O'Connell, T. A. (1994). Evaluation in the ARPA Machine Translation Program: 1993 Methodology. In *Proceedings of Human Language Technologies (HLT'94)*.

Zens, R., Och, F. J., and Ney, H. (2002). Phrase-Based Statistical Machine Translation. In *Proceedings of the 25th Annual German Conference on AI*, pages 18–32, London, UK. Springer-Verlag.

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

Zhang, Y., Zens, R., and Ney, H. (2007). Chunk-Level Reordering of Source Language Sentences with Automatically Learned Rules for Statistical Machine Translation. In *SSST '07: Proceedings of the NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.