IMPROVING WORD-TO-WORD ALIGNMENTS USING MORPHOLOGICAL INFORMATION

A Thesis

Presented to the

Faculty of

San Diego State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

in

Linguistics

---

by

Pablo Malvar Fernández

May 2008

SAN DIEGO STATE UNIVERSITY

The Undersigned Faculty Committee Approves the

Thesis of Pablo Malvar Fernández:

IMPROVING WORD-TO-WORD ALIGNMENTS USING MORPHOLOGICAL INFORMATION

Jean Mark Gawron, Chair
Department of Linguistics and Asian/Middle Eastern Languages

Robert Malouf
Department of Linguistics and Asian/Middle Eastern Languages

Joseph Lewis
Department of Computer Science

Approval Date

DEDICATION

Dedicated to my mom and to my two other favorite girls, Jenny and Sophi.

Da minha aldeia vejo quando da terra se pode ver no Universo....
Por isso a minha aldeia é grande como outra qualquer
Porque eu sou do tamanho do que vejo
E não do tamanho da minha altura...

Nas cidades a vida é mais pequena
Que aqui na minha casa no cimo deste outeiro.
Na cidade as grandes casas fecham a vista a chave,
Escondem o horizonte, empurram nosso olhar para longe de todo o céu,
Tornam-nos pequenos porque nos tiram o que os nossos olhos nos podem dar,
E tornam-nos pobres porque a única riqueza é ver.

– Alberto Caeiro, in “O Guardador de Rebanhos”.

ABSTRACT OF THE THESIS

IMPROVING WORD-TO-WORD ALIGNMENTS USING MORPHOLOGICAL INFORMATION

by
Pablo Malvar Fernández
Master of Arts in Linguistics
San Diego State University, 2008

All current Statistical Machine Translation systems rely on an initial layer of word-to-word alignment; not surprisingly, alignment quality has been shown to be a key factor in performance at higher levels of abstraction. This thesis is an investigation into the effects of part-of-speech tagging and morphological transformations of a parallel corpus on aligment quality.

Contents

List of Tables

PAGE

List of Figures

ACKNOWLEDGEMENTS

This research was able to be conducted thanks to a shared grant from the Spanish Fulbright Comission and the Ministerio de Educaión y Ciencia of Spain. I also would like to thank Professor Jean Mark Gawron and Professor Robert Malouf for their advice and guidance.

Chapter 1

PREFACE

From the point of view of Hallidayan Functional-Systemic Linguistics (FSL) theory, language serves, according to Gee (1999, 1) “as both a tool for action and a scaffolding for ‘human affiliation within cultures and social groups and institutions”’. In other words, language works as a tool not only for communication but for negotiating the relationships and the structures of the society itself. It is, precisely, through this social dimension that language manages to play a extremely crucial symbolic role.

In developing computational tools for particular languages, computational linguists[1], whether they are primarily computer scientists or linguists, have a responsibility to the language(s) they are working with. It is possible that for languages with a large speaker population and a high prestige this responsibility is not obvious. In these cases, decisions about which linguistic phenomena are to be studied and (more importantly from the point of view of this thesis) which tools are developed; may seem trivial, because they may be regarded as not implying any ideological position. However, for those scientists who have decided to work with and for minority or minoritized[2] languages, especially if they are speakers of that language, their decisions are never innocuous or empty of ideological content.

It is with this responsibility as a language researcher and as a speaker firmly in mind that this work is undertaken.

[1] As well as the rest of Linguists and other professionals who deal with languages.

[2] Minoritized languages are languages, although not spoken by a minority, are relegated to a low position in terms of use and prestige by another language, which symbolically and pragmatically works as the language of culture and (economical and/or political) power.

Chapter 2

INTRODUCTION

The motivation for this Master's thesis comes from Sociolinguistic and Language Planning considerations, as much as from a purely Computational Linguistics perspective. The corpus used in this research is a domain-specific English-Galician parallel corpus. The choice of this particular pair of languages is directly related to the fact that the autonomous government of Galicia approved on January the 2nd of 2005 a *Plan Xeral de Normalización Lingüística* (General Normalization Plan for Galician) that states as one of its main objectives in its section dedicated to the *New Technologies* that:

"Sector B. New Technologies: *B.4. Potentiate the research in Machine Translation, Speech Recognition, and other new techniques that may appear and facilitate the positive option in the information and communication market and that ensure the free circulation of Galician language in the advanced systems of the current society.*" Xunta de Galicia (2005, 51)[1]

Galician is a Romance language spoken in the Northwestern region of Spain. Galician is intimately related to Portuguese but, after more than eight centuries of political domination, it has been massively influenced by Spanish.

Galician's current sociolinguistic situation is roughly defined, as Monteagudo (2002) states, by two main phenomena: diglossia and language replacement. The term diglossia refers to a linguistic context, such as Galician, in which two linguistic system in coexist with one being considered more prestigious, that is, High; and the other one stigmatized, that is, Low. In Galicia this situation was, for centuries, related to social factors. Thus, the High system, Spanish, was associated with dominant urban classes, while Galician, the Low system, was associated with rural and subaltern layers. This was, therefore, a diglossic situation. In addition to this, language replacement was taking place, so that Galician was losing speakers in favor of Spanish on three fronts:

"a) the massive emigration from rural to urban Galicia as well as the emigration to other countries; b) intergenerational language switch, that is, Galician-speaking progenitors preferred to transmit Spanish to their descendants; c) the biographical language switch, that is,

[1] "Sector B. Novas Tecnoloxías: *B.4. Potenciar a investigación da tradución automática, o recoñecemento de voz, e outras novas técnicas que vaian aparecendo e que faciliten a opción positiva no mercado da información e da comunicación e que aseguren a libre circulación do galego nos sistemas avanzados da sociedade actual.*"

the adoption of Spanish as language for professional interactions, as well as the language of public speeches and writing." (10)

After Spanish dictator General Franco's death in the late 1970's, Spain began a period of transition to democracy. Within this new political frame, in 1978 a new Constitution was approved that recognized the co-officiality of Spanish —official of the entire Spanish territory— and three peripheral languages, Basque, Galician and Catalan, but only in the territory of their respective autonomous regions.

Under this new political situation and after the approval of the new Constitution, the historical self-government claims of the Basque, Galician and Catalan nations obliged the Spanish authorities to re-establish the autonomous status they had before the civil war began in 1936. Thus, in 1980 a new *Estatuto de Autonomia de Galicia* (Statute of Autonomy of Galicia) was approved. This new legal framework recognized the status of Galician as the language of Galicia, as well as its co-officiality with Spanish. The Galician autonomous government received, then, non-exclusive authority over education, cultural advancement, autonomous administration and mass media. To linguistically regulate these new domains, the Galician Parliament approved in 1983 a *Lei de Normalización Lingüística* (Linguistic Normalization Law). In addition, this law also imposed *Real Academia Galega* (RAG) and *Instituto da Lingua Galega* (ILG) orthographical norms as the only officially recognized orthographical norms of Galician.

Under this legal and political frame, a gradual normalization of Galician has been in progress. However, after a full twenty years of normalization, it was becoming clear that the political, academic and social changes instituted in 1983 were insufficient to subvert the above mentioned process of language replacement. Thus, in 2005 the Galician Parliament approved the above cited General Plan of Normalization (GPN). The GPN tries to readjust normalization efforts according to the new political, social and technological situation of Galician. It is therefore within this context that this thesis is situated.

As it can be seen from the *New Technologies* section of the GPN (above cited), new technologies and, in particular, Information Technologies (IT) are regarded as powerful tools to integrate minoritized languages into the so called Information Society.

It is, therefore, clear that for Galician institutions the implementation of computational tools in Galician is now an essential part of the task of promoting the usage of Galician within the IT sector. However, the lack of a proper technical vocabulary and an extensive inventory of Galician localized computational applications only reinforces the impression that Galician is the language of old technology. In other words, Galician appears to be the language of what is ancient and rural, and, therefore, is simply seen as an inadequate tool for new and modern technologies. In fact, since the overwhelming majority of new technologies are primarily

released in English, the work of promotion of Galician within the IT sector is not only a matter of planning the *status* of Galician —as the GPN mandates—, but also entails a substantial amount of corpus planning (*vid.* Cooper (1990)) which must priorize two interrelated tasks: *i)* the localization into Galician of software originally released in English and *ii)* the development of all the technical apparatus and infrastructure necessary to implement that localization.

The reason for prioritizing these two particular tasks when promoting is that Galician can only be successfully integrated within the IT sector if, at the same time as positive attitudes reinforced, negative attitudes and expectations associated with Galician are subverted.

Drawing from this particular legal and sociolinguistic context, this research project has been designed so that its object of study, methodology and objectives can contribute to the goal of successfully integrating Galician within the realm of ICT.

Thus, automatically generated word-to-word alignments, which fall within the scope of Machine translation and Terminology Extraction and whose current state-of-art research entirely relies on the usage of corpora, are the object of study of this research project. In order to accomplish the objective of this research, which is the improvement of word-to-word alignment quality and, as direct side effects, the improvement of bilingual resources for translation of IT-related texts, and the improvement of Phrase-based Machine Translation (MT) quality, a corpus-based methodology has been designed. Thus, a software localization English-Galician parallel corpus has been compiled, as well as two English and Galician domain-specific POS-taggers and lemmatizers.

Chapter 3

Software Localization English-Galician Parallel Corpus

As it has already been described, the corpus used in this research project is an English-Galician parallel corpus in the domain of software localization. Taking advantage of the fact that the open source projects used in this research have been released under the *General Public License* (GPL), this software localization English-Galician parallel corpus has been compiled by downloading localization files of several Linux distribution and applications from the Internet. Localization data of this kind are stored in files with *.po* extension that contain all dialogs presented to users when interacting with the Operating System or any particular application.

Although the original idea of compiling this kind of corpus has been developed over the course of several years, the effective compilation of the entire corpus used in this project took place between the May 16th and May 26th 2007. For this project the Linux distributions and applications whose localization *.po* files have been manipulated are: *GNU Project, Debian, openSuse, Mandriva, KDE, Gnome, OpenOffice, AbiWord, eMule, Drupal, Gazpacho, Inkscape, Pidgin, KmyMoney, Tenes Empanadas Graciela (TEG), wxWidgets, xChat, Anti Bot Question (ABQ), Fireboard 1.0.3, Joomla 1.5 RC3 and phpBB3*. Table 1 presents a complete relation of the Linux distributions and applications and the websites that host their localization files.

Table 3.1. Relation of Linux distributions and applications and their hosting project website.

Software	**Hosting project websites**
*G*NU Project	**http://translationproject.org/team/gl.html**
*D*ebian	**http://www.debian.org/intl/l10n/po-debconf/gl**
*O*penSuse	**http://i18n.opensuse.org/stats/**
*M*andriva	**http://www.trasno.net/mandriva:inicio**
*K*DE	**http://l10n.kde.org/team-infos.php?teamcode=gl**
*G*nome	**http://l10n.gnome.org/languages/gl**
*O*penOffice	**http://gl.openoffice.org/source/browse/gl/src/**

(table continues)

Table 3.1 (continued)

Software	Hosting project websites
*A*biWord, eMule, Drupal, Gazpacho, Inkscape, Pidgin, KmyMoney, TEG, wxWidgets, xChat	**http://www.trasno.net/outros:inicio**
*A*BQ, Fireboard, Joomla, phpBB3	**http://www.ciberirmandade.org/traduCIF/main.php**

Due to the fact that this corpus has been compiled for two different purposes, namely the development, first, of two English and Galician POS-taggers and lemmatizers and, second, the experimentation with corpus-based word-to-word alignment algorithms, the preprocessing of this corpus was carried out in two stages. Furthermore, as it will be explained in detail later, the versions of the corpus used in these two stages ended up having different sizes.

3.1 SHARED CORPUS PREPROCESSING TASKS

Despite their differences, these two different preprocessing stages shared the very first preprocessing steps: *i)* the extraction of the corpus from the raw *.po* files downloaded from the Internet and *ii)* the tokenization process.

3.1.1 Corpus Extraction

After downloading all the *.po* files to be used as source files for this corpus, the first preprocessing step was the extraction all the pairs of the translation units[1] contained in the *.po* files. This task was accomplished by developing a general extraction algorithm that had to deal with formal properties of *.po* files.

Thus, the first problem this algorithm had to deal with was that the first lines of every *.po* file are usually, but not always, occupied by heading comments that provide information, such as the name of the file, the date of creation and revision and the name of the translator (see Figure 3.1).

Since for this project the only relevant information is the pairs of Galician and English translation units, these headers did not need to be extracted. However, in these heading lines, there is a line that provides a specially relevant piece of information, i.e.: “Content-Type:

[1]Following Delisle et al. (1999, 295), *Translation Unit* is understood as a concept in which the process, carried out by a translator, of interpreting the meaning of the source text and reproducing it as a target text in another language establishes an equivalence between both source and target texts.

```
"Project-Id-Version: debian-installer\n"
"Report-Msgid-Bugs-To: \n"
"POT-Creation-Date: 2007-04-19 04:30+0200\n"
"PO-Revision-Date: 2005-12-11 16:22+0100\n"
"Last-Translator: Jacobo Tarrio <jtarrio@debian.org>\n"
"Language-Team: Galego <trasno@ceu.fi.udc.es>\n"
```

Figure 3.1. Sample of a typical *.po* file heading lines.

text/plain; charset=UTF-8\n". This is the line that defines what is the character encoding of each file. This information helps to solve problems of characters display which every corpus compilation project has to struggle. In fact those problems came up during the compilation of this corpus. Since the localization of most of the open-source projects was carried out by volunteers who did not all follow the same guidelines, not all of the files in a particular localization project were encoded using the same character encoding set. In addition, another serious problem was also encountered. Despite their character encoding definition, some files *.po* files were not encoded as defined. Luckily, this did not represent an unsolvable problem because, since the desired orthography for Galician uses the same characters as Spanish, the character encoding set of those wrongly defined files was ISO-8859-1 —default encoding of Spanish.

After having guessed the character encoding of all the files of each localization project, a simple shell script, ran on an Unix-like Operating System, converted that ISO-8859-1 character encoding into UTF-8, which was the character encoding chosen for the corpus used in this project.

Once the normalization into UTF-8 was completed, the next problem encountered was the extraction of the Galician and English translation units. Every *.po* file stores English entries —source language of all the software manipulated— marking them up with the string *msgid* and target language translations, in this case Galician, marking them up with the string *msgstr*. In this sense, the systematical use of these two markup tags simplified enormously the process of extraction because those marks were used as identifiers of the language of the translation unit for the extraction algorithm. As an example, I present below, Figure 3.2, a pair of translation units as stored in *.po* files:

Despite the use of marks for each entry not all the translation units do occupy the same number of lines. Thus, it is quite common that the extension of the English and Galician entries occupy more than one line. The way *.po* files store these dialogs is determined by the

```
msgid "Aboot installation failed.  Continue anyway?"
msgstr "Non se puido instalar aboot. ¿Continuar igualmente?"
```

Figure 3.2. Sample of translation units as stored in .po files.

size of the window in which these dialogs are intended to be displayed. Therefore, depending on the size of the window sentences are split in lines. As an example, Figure 3.3 shows a sample pair of translation units that occupy more than one line:

```
msgid ""
"The aboot package failed to install into /target/.  Installing aboot as a "
"boot loader is a required step.  The install problem might however be "
"unrelated to aboot, so continuing the installation may be possible."
msgstr ""
"Non se puido instalar o paquete aboot en /target/. Instalar aboot coma "
"cargador de inicio é un paso necesario. O problema coa instalación, "
"nembargantes, pode non estar relacionado con aboot, así que pode ser posible "
"continuar a instalación."
```

Figure 3.3. Sample of a pair of translation units that occupy more than one line.

As can be observed, although these two relatively long translation units are still marked by the strings that define them as being the source or the target language, it is important to note that the source and target translation units are stored in a different number of lines.

Another problem that needed to be dealt with had to do with the fact that some open-source projects were, by the time they were downloaded, not completely localized. Incomplete localizations were full of provisional translations that had not been revised by a human translator. In this cases, the pairs of translation units have been automatically translated by a machine translation system using fuzzy logic strategies to capture string similarity. According to this special status, these translations are marked as *fuzzy* localized. Because this type of fuzzy localizations are not reliable, in that they are usually incomplete or, at least, completely wrong, it was decided not to extract them to avoid an unnecessary source of noise. As an example, I present in Figure 3.4, below, a pair of fuzzy localized translation units:

```
#, fuzzy
msgid "Saturday"
msgstr "bandexa"
```

Figure 3.4. Sample of a pair of fuzzy localized translation units.

Since this type of entries are clearly marked as fuzzy localized, it was not difficult to incorporate this feature into the extraction algorithm, so that the source language entry and its fuzzy localized target language entry were not extracted.

The last problem that I encountered during the extraction process was derived from the fact that Galician is a minoritized language and until recently there was no interest shown by the Galician political authorities in promoting the use and localization of open-source software. Thus, the pace of localization into Galician is, in comparison to other languages, very slow; in addition, the process has suffered from more discontinuities. Therefore, in an important number of *.po* files there were many entries that had not been localized yet. As an example, Figure 3.5 shows the statistics of progress made by June 1, 2007 by four different teams —Galician, Portuguese (from Portugal and Brazil) and Spanish— in relation to openSuse-10.2 localization project:

Language	Translated	%	Fuzzy	%	Untranslated	%	Total	Graph
Galician (gl)	3196	12 %	9627	37 %	12697	49 %	25520	
Spanish (es)	25197	98 %	185	0 %	138	0 %	25520	
Portuguese (pt)	23229	91 %	1264	4 %	1027	4 %	25520	
Brazilian Portuguese (pt_BR)	25199	98 %	178	0 %	143	0 %	25520	

Figure 3.5. openSuse-10.2 localization statistics for Galician, Portuguese (from Portugal and Brazil) and Spanish teams by June 1, 2007.

As can be observed in Figure 3.5, it is clear that Galician team is far behind the rhythm of localization of the rest of teams taken into account. Given this, Galician-translated *.po* files usually have entries that have been not localized yet. In these cases, the space reserved for the target language translation unit is simply empty, as it shown in Figure 3.6.

```
msgid "Cote D'Ivoire"
msgstr ""
```

Figure 3.6. Example of a non-localized English translation unit.

If it had been decided to include this incomplete entries, the result would have been that the corpus would be unbalanced in terms of number of words for English and Galician. The extraction algorithm did not take into account incomplete entries and, therefore, they were not extracted.

The final product of the extraction algorithm were two files, one containing all the English entries and another containing all the Galician translation. The correspondence between English and Galician translation units is made by means of the position, that is, the line they occupy in each of the files. Thus, an English translation unit occupying, for example, the third line of the English file corresponds to the Galician translation unit that also occupies the third line of the Galician file.

3.1.2 Corpus tokenization

After the extraction of all the pairs of translation units from the raw *.po* files, the next step was the tokenization of each of the resulting files.

Pre-tokenization cleanup:

As a preliminary step before tokenization, several elements needed to be removed, so that they were not part of both the POS-taggers training corpus and the final corpus used for word-to-word alignment experimentation. Thus, all the xml tags were deleted, as well as all "&_" and "_&" symbols, used as anchor-marks for the keyboard shortcuts; and all the quotation marks used to enclose translation units in *.po* files. The presence of these three elements would have represented another unnecessary source of noise that would have lowered the performance of the English and Galician POS-taggers and both word-to-word alignment and Phrase-based MT algorithms. Thus, on the one hand, xml tags and keyboard shortcut marks, because they are attached to word forms, would have multiplied the presence of low frequency units and, therefore, would have introduced serious sparseness issues in all the subsequent stages of this research project. On the other hand, the presence of quotations would have introduced strings in the corpus that do not belong to translation units. In addition, the use of these quotations as markers of sentence ending would would have brought a serious source of noise into the sentence-to-sentence alignment process, which, as it will be explained, was a problematic step *per se*.

Tokenizer customization:

Finally, the tokenization process was done using the tokenizer included within Philipp Koehn's sentence-to-sentence alignment package. Since this tokenizer was developed taking into account general properties of a few languages, namely English, French, Spanish and

German, this tokenizer was extended so it could take into account the existence of domain specific abbreviations both for English and Galician sides of the corpus.

3.2 NON-SHARED CORPUS PREPROCESSING TASKS

As pointed out before, the two differentiated preprocessing stages of this corpus did not share some preprocessing tasks. In this section, all the different tasks carried out in these two stages will be explained in detail.

3.2.1 Preprocessing of the Software Localization POS-taggers and Lemmatizers Training Corpora

The amount of preprocessing required for the domain specific English and Galician POS-taggers training corpora was significant, due to the size of the corpus and its particularities as a domain specific corpus.

In this stage, both sides of the parallel corpus were treated as independent corpora because they were used as training corpora of two differentiated POS-taggers, one for English and another for Galician.

As a starting point for the preprocessing of the each of theses two POS-taggers, English and Galician text files were provisionally POS-tagged using the currently available parameters. Thus, for English the POS-tagging parameters that are included with the freely downloadable version of TreeTagger[2] Schmid (1994) were used. For Galician, the parameters developed by Pablo Gamallo[3] were used. Using these parameters two provisional POS-tagged and lemmatized versions of the corpora were generated. Since provisional English and Galician parameters were trained on corpora of very different sizes, the POS-tagged and lemmatized versions of the corpora were of very different quality. Despite the fact that this represented a much longer re-tagging process of Galician corpus —in terms of POS-tagging and lemmatizing by hand many unknown common nouns, verbs, contractions, adjectives and adverbs—, both Galician and English corpora needed substantial revision due mainly to the particularities of these domain specific corpora including

- many non-recognized proper names that are exclusive to the domain of this corpus
- a huge number of non-recognized variables that are set by each particular piece of software according to the values they take within the software framework as a whole (i.e., user names, contextual menus, etc.)

[2]TreeTagger can be downloaded at http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[3]Galician TreeTagger parameters developed by Pablo Gamallo can be downloaded at http://gramatica.usc.es/~gamallo/

Besides correctly hand-tagging and lemmatizing unknown proper names and variables —for which the POS-tag VAR was made up—, another step was necessary so that the resulting POS-tagging parameters could be more efficiently reused during the experimentation with word-to-word alignments algorithms. Thus, the different tagsets used for English and Galician were normalized so that the same tag was used to refer to the same type of morphological entities. Whereas the English parameters included in the freely downloadable version of Treetagger used a tagset derived from the Brown corpus tagset, Francis (1964) and Francis and Kučera (1982), Galician parameters used a completely different tagset based on traditional Romance grammar. Table 3.2 presents the correspondences between Galician and English tagsets used by the provisional POS-tagging parameters and the normalized tagset used to train the POS-tagger that were built in this project.

Table 3.2. Correspondences between Galician and English tagsets used by the provisional POS-tagging parameters and the normalized tagset used by the POS-tagger developed in this project. *N/A* means that a particular distinction is not applicable for that language. *Not used* means that a particular tag also pertinent due to the morphosyntactic properties of a language is not used by the original tagset.

English	Galician	Normalization
JJ	ADJ	ADJ
JJR	ADJ	ADJ
JJS	ADJ	ADJ
RB	ADV	ADV
RBR	ADV	ADV
RBS	ADV	ADV
DT	DET	DET
CD	CARD	CARD
NNS	NOM	N
NN	NOM	N
CC	CONJ	CCONJ
IN	CONJSUB	SUBCONJ
IN	PRP	PRP
PDT	*N/U*	PDET (for *all* and *both*)
UH	I	I
EX	*N/A*	EX

(table continues)

Table 3.2 (continued)

English	Galician	Normalization
FW	FW	FW
LS	*N/U*	LS
SYM	*N/U*	SYM
SENT	SENT	SENT
N/U	QUOTE	QUOTE
PP	*N/U*	P
POS	*N/U*	POSE
PP$	*N/U*	ADJ
NPS	NOM	NP
NP	NOM	NP
RP	*N/A*	RP
MD	*N/A*	MD
VBN	V	V
VBG	V	V
VBD	V	V
VBP	V	V
VB	V	V
VBZ	V	V
TO	*N/A*	TO
N/U	*N/A*	MD+ADV (for *cannot*)
WP	*N/U*	INTEX
WDT	*N/U*	REL
WP$	*N/U*	P
WRB	*N/U*	INTEX
,	VIRG	COMMA
N/A	PRP+DET	PRP+DET
N/A	V+P	V+P
N/A	*N/U*	V+DET
N/A	*N/U*	ADV+DET
N/A	*N/U*	PDET+DET
N/A	*N/U*	SUBCONJ+DET
N/U	*N/U*	VAR

(table continues)

Table 3.2 (continued)

English	Galician	Normalization
N/U	*N/U*	LOC
N/A	*N/U*	PER (for verbal periphrases)
N/A	*N/U*	V+P+P
N/A	*N/U*	P+P
N/A	*N/U*	PRP+PDET+DET
SYM	SYM	SYM

In addition, the tags of some special character were normalized, so that they were consistently tagged in both English and Galician. Table 3.3, below, shows a relation of special characters and their normalized tags:

Table 3.3. Relation of relation of special characters and their tags.

Symbol	Tag
^	SYM
=	SYM
~	SYM
*	SYM
+	SYM
:	SENT
[	QUOTE
]	QUOTE
|	QUOTE
/	QUOTE
-	QUOTE
“	QUOTE
’	QUOTE
#	QUOTE

These two fully preprocessed corpora, which included all the Linux distributions and applications listed above, have the following size: English 3379134 tokens and Galician 3447541. As mentioned, English and Galician corpora were, finally, used to train new English and Galician POS-tagging and lemmatizing parameters for TreeTagger[4].

[4]Both English and Galician training corpora and POS-tagging and lemmatizing parameters can be freely downloaded from http://paulomalvar.dyndns.org:8080/Paulo%20Malvar%20personal%20webpage/Resources.html

3.2.2 Preprocessing of the Software Localization English-Galician Parallel Corpus

As has been already pointed out, the preprocessing of the Software Localization parallel corpus was carried out in a different stage from the two “independent” English and Galician corpora described in subsection 3.2.1. Two main factors affected this decision. On the one hand, the parallel corpus intended to be reused in subsequent stages of this research project needed to be POS-tagged and lemmatized. Since this task can be only performed by TreeTagger placing every token on a different line, it was necessary to enrich the corpus with sentence ending markers to help recovering sentences after the corpus had been effectively POS-tagged and lemmatized. On the other hand, since recovering sentences by means of using sentence-ending markers did not ensure that a parallel sentence-to-sentence alignment could be recovered, another necessary step needed to be carried out: sentence-to-sentence realignment. In order to perform this second task Philipp Koehn’s sentence aligner was used. This sentence aligner has, however, the limitation, not mentioned in the documentation that accompanies the software, that it can only process pieces of text that do not exceed 2500 lines without making 2GB of RAM computers run out of memory.

For these reasons, the preprocessing of the parallel corpus had to be redone after the extraction of English and Galician translation units. Thus, the extraction was constrained to ensure that no extracted file exceeded that maximum number of lines.

In addition to these two problems, another problem had to be resurfaced again. As I pointed out at the beginning of this chapter, there were problems related to character encoding of *.po* files. Some projects proved to have a mixture of character encoding for their *.po* files. Although, some of these problems could be resolved easily there was a project, namely *GNU Project*, that exhibited enormous variation in character encodings in its 362 files. Due to time limitations, this package was excluded from the final parallel corpus. Otherwise, since each file had to be individually processed to ensure a successful sentence-to-sentence realignment, this tasks might have compromised the realization of this research project.

Thus, each individual file was re-extracted and tokenized following the same methodology explained in subsections 3.1.1 and 3.1.2. Each of these files was, after tokenization, enriched with sentence-ending markers. Thus, if none the following sentence-ending markers was found at the end of each line, “. : ; ? !”, a period was appended to serve as an explicit sentence ending marker. Once each line was ensured to have a

sentence-ending marker, the corpus was POS-tagged and lemmatized using TreeTagger and the English and Galician parameters trained during in the previous stage[5].

Once the previous process was carried out, POS-tagged and lemmatized word forms were processed by attaching to them their correspondent POS-tag and lemma. As an example, I present in Figure 7, below, an extract of an English sentence with POS-tags and lemmas attached to their correspondent word forms:

The_DET_the
aboot_N_aboot
package_N_package
failed_V_fail
to_TO_to
install_V_install
[...]

Figure 3.7. Pair of English and Galician parallel sentences illustrating the attachment of POS-tags and lemmas to word forms.

After the attachment of POS-tags and lemmas to word forms, it was still necessary to recover sentences because, as it can be observed in Figure 7 every particular set of *word form + POS-tag + lemma* occupied a single line. In order to solve this problem, sentences were recovered by means of using ". : ; ? !" as sentence ending markers. This process had, however, the result that the corpora no longer had the same number of lines. Given this, it was necessary to do sentence-to-sentence alignment on the resulting corpus. To accomplish this task Philipp Koehn's sentence aligner, *sen_align*[6], was used for this step. Sentence-to-sentence alignment is not an error-free process. However, given the absolute necessity of realigning the both sides of the corpus to recover parallelization information, this step was regarded as an unavoidable source of noise. And, in fact, it was discovered during the development of the small gold standard corpus used to evaluate the performance of the different word-to-word alignment algorithms it was a serious source of noise. Although, sentence realignment was not systematically evaluated in this research project, it is worth noting that roughly 20% of the sentence seen during the development of that gold standard were misaligned.

[5]Errors in the automatic POS-tagging and lemmatization process, as an inherent problem of any automatic language processing, were simply assumed as a potentially minimizable, but unavoidable source of noise for the word-to-word alignment algorithms.

[6]sen_align can be found and downloaded from http://www-rohan.sdsu.edu/~gawron/mt_plus/mt/course_core/course_outline.html

Another property of the sentence-to-sentence aligner used in this research project is that, if no suitable sentence or group of sentences is found for a particular sentence, the algorithm favors correspondences of the types 1:0 and 0:1. In other words, some sentences, independently of the side of the corpus they belong to, are aligned with an empty sentence in the other side of the corpus. This type of correspondences are problematic for word-to-word alignment algorithms because no explicit word-to-word alignment can be made. Therefore, since Phrased-based MT system are totally dependent on word-to-word alignment quality, it was also necessary to prune this type of sentence alignments out of the corpus.

Finally, in order to avoid another unnecessary source of noise, long sentences were pruned out of the parallel corpus —in this case sentences with more than 40 words, as it is the normal procedure when working with word-to-word alignment algorithms and Phrased-based MT systems.

After all the preprocessing steps were carried out, the resulting English side of the parallel corpus has 2535425 tokens, whereas the Galician side has 2628140 tokens.

Chapter 4

Statistical Machine Translation

As is commonly accepted by translation researchers and professionals, translating from one language to another is a task in which the main challenge is to find a balance between fidelity to the meaning expressed in the source language and fluency of the equivalent text in the target language. According to Jurafsky and Martin (To be published in spring 2008), "Statistical MT is the name for a class of approaches that do just this, by building probabilistic models of faithfulness and fluency, and then combining these models to choose the most probable translation"(18). Thus, the best translation $\hat{T}$ of a particular source language sentence S can be formalized, following Jurafsky and Martin (To be published in spring 2008, 18), as:

$$\hat{T} = \underset{T}{\operatorname{argmax}} \mathit{faithfulness}(T, S)\mathit{fluency}(T) \tag{4.1}$$

This intuitive informal definition of best translation can be mathematically defined as the conditional probability of a possible translation given a particular source language sentence:

$$\hat{T} = \underset{T}{\operatorname{argmax}} P(T|S) \tag{4.2}$$

Using Bayes Rule this conditional probability can be rewritten as:

$$\hat{T} = \underset{T}{\operatorname{argmax}} \frac{P(S|T)P(T)}{P(S)} \tag{4.3}$$

$$\hat{T} = \underset{T}{\operatorname{argmax}} P(S|T)P(T) \tag{4.4}$$

The final equation obtained after applying Bayes Rules is justified because $P(S)$ does not depend on T and so remains constant across all T (this is exactly the move made in Noisy Channel approaches to Natural Language Processing). Thus, although our intuitive formalization had made the translation T conditional on the source sentence S, equation 4.4 makes the source S conditional on the translation T. This backwards way of formalizing statistical problems is usual in Noisy Channel models. From this point of view the translation is metaphorically turned into the source sentence of the translation problem. This metaphorical new source sentence is seen as having, somehow, become corrupted after having passed through a noisy channel. The result if this corruption is what we see as the original source sentence S. The advantage of formalizing translation from the Noisy Channel

perspective is that now we have an equation that can perfectly parallelized with our informal definition of the problem of finding the best translation:

$$P(S|T) = faithfulness(T, S) \tag{4.5}$$

$$fluency(T) = P(T) \tag{4.6}$$

In other words, now we can model by means of using a perfectly mathematical formulation both the translation probability of T given S and the fluency of translation T.

4.1 WORD-TO-WORD ALIGNMENT ALGORITHMS

Back in the 1990's IBM research group at Yorktown Heights (NY), Brown et al. (1990) and Brown et al. (1993), started publishing algorithms that, with relative success, used a Bayesian derivation of the Noisy Channel Model to perform SMT. The IBM approach began by establishing word-to-word alignments across aligned sentences in a parallel corpus. Word-to-word alignments simply formalize the idea that there is an explicit, yet not perfect, mapping between the words of source and target sentences of parallel corpora. Following the same Noisy Channel approach, word-to-word alignment algorithms model the conditional probability of a source sentence S given a translation T, by word-to-word aligning these S and T sentences:

$$P(S|T) = \sum_{A} P(S, A|T) \tag{4.7}$$

In other words, for a particular pair of aligned sentences, S and T, the conditional probability of S given T is found by summing over all the possible word-to-word alignments A between S and T.

Since no hand-labeled parallel corpora are usually available[1], it is necessary to use an algorithm able to calculate word-to-word correspondence probabilities using the information given by the co-occurrence of words in a set of parallel sentences. For this task, Expectation Maximization (EM) is used. The EM algorithm is usually initialized with a uniform distribution for all possible word-to-word alignments; it then iterates for a given number of times reestimating the parameters of every particular model and recomputing word-to-word alignment probabilities (and other parameters) on the basis of the model from the previous iteration —for a more detailed explanation of basic EM function (*vid.* Jurafsky and Martin (To be published in spring 2008, 28-31)).

[1] In fact, it would be to expensive in terms of economical and human resources to label by hand word-to-word correspondences in parallel corpora of the size it is necessary to achieve MT performance of quality.

Since publication of the paper describing the first IBM models in the early 1990's, many word-to-word alignment models have been proposed during the two subsequent decades. However, the most used and well-known word-to-word alignment models remain the 5 IBM models and the non IBM models known as, HMM Model, and, confusingly, IBM Model 6 Och and Ney (2003).

A detailed description of IBM Models 1-6 and HMM properties can be found in Och and Ney (2003)). A summary of their properties is presented below:

- IBM Model 1: IBM Mode 1 is the simplest and, therefore, the most unrealistic IBM model. IBM Model 1 simply models $P(S, A|T)$ as the sequential product of the probability of all the possible word-to-word alignments for a given pair of sentences:

$$P(S, A|T) = P(S|E, A) = \prod_{j=1}^{J} t(f_j|e_{a_j}) \tag{4.8}$$

  where f = S and e = T[2]. In other words, the probability of a word-to-word alignment between a pair of sentences is simply the product of the conditional translation probabilities —known as *translation probability model*— of every pair of aligned words .

- IBM Model 2: Model 2, which is usually initialized by transferring Model 1 word alignment probabilities as estimated at its termination iteration, is based on a generative story that incorporates a distortion model, which is based on the concept that source words are usually aligned with target words that occupy a similar absolute position. By adding this feature Model 2 tries to address issues that have to do, for example, with the alignment of repeated source words. Between languages such as French, Spanish, Portuguese or Galician and English, which use articles to introduce nouns in Noun Phrases, it usually occurs that several occurrences of the same article are found in the same sentence. Thus, in order to avoid that an article occupying the first position of the source sentence is aligned with an article occupying, for example, the tenth position of the target language sentence, Model 2 penalizes long-distance alignments.
  Therefore, Model 2 general equation is:

$$P(S|E, A) = \prod_{j=1}^{J} t(f_j|e_{a_j}) d(a_j|j, I, J) \tag{4.9}$$

  where a_j represents the absolute position the target language word aligned with the word occupying the *jth* position of the source sentence and *I* and *J* refers to the absolute length of the target and source sentences, respectively.

- HMM: HMM Model could be thought of as a slightly more complex non-IBM version of Model 2. Thus, HMM Model uses the same translation model as Model 2 but a distortion probability model. Instead of modeling the distortion between the position of

[2]To be consistent with the traditional notion, from now on, f will be used to refer to target language and e to refer to source language.

source and target words, the HMM models the distortion between the absolute position of target language words. In this sense, the HMM is an order 1 model because it makes the position of target word aligned with source word occupying the *jth* position depend on the the position of the target word aligned with the source word occupying the (*j-1)th* position. HMM general equation is:

$$P(S|E,A) = \sum_{a_1^j} \prod_{j=1}^{J} t(f_j|e_{a_j}) d(a_j|a_{j-1}, I) \tag{4.10}$$

- IBM Model 3: Model 3, introduced by Brown et al. (1993), represents a watershed moment on the history of probabilistic word alignment models. Thus, Model 3's generative story incorporates new elements into the picture, making word alignment models, somewhat, more realistic. In this sense, Model 3's generative story introduces the concept of *fertility*. Each source word s is assigned a probability distribution over possible fertilities n, $P(n \mid s)$. Thus, the probability of an alignment takes into account the probability of multiple target words being generated by one source word. Model 3 general equation is:

$$P(S|E,A) = \prod_{i=1}^{I} \Phi_i! n(\Phi|e_j) * \prod_{j=1}^{J} t(f_j|e_{a_j}) * \prod_{j:a(j)\neq 0}^{J} d(j|a_j, I, J) * \binom{J-\Phi_0}{\Phi_0} p_0^{\Phi_0} p_1^{J-2\Phi_0} \tag{4.11}$$

  where Φ_i represents the fertility of e_i. Note that the IBM and HMM models model not only the alignment probability between source and target words, that is, mappings of type one-to-one or many-to-one, but also mappings of the type zero-to-one and one-to-zero. In other words, word-to-word alignment models take into account the fact that source words are often left untranslated and that target words sometimes appear without a direct correspondence to any of the source words. In order to achieve this, IBM and HMM models assume that at the beginning of both the source and the target sentences there is a NULL word with which source and target words can be aligned. Thus, in the formula above Φ_0 represents the fertility of the NULL word, p_0 represents the probability that a source word will be aligned with NULL and p_1 the probability that a source word is not aligned with NULL.

- IBM Model 4: Model 4, as HMM, is an order 1 word alignment algorithm. Thus, Model 4 modifies Model 3 formula by updating the distortion probability model and making it dependent on the relative position of the source words. Therefore, although both HMM and Model 4 are order 1 models, they differ in which axis is modeled by the distortion probability model. Model 4 distortion model is:

$$d(B_{i_k}, \overline{B}_{i-1}) \tag{4.12}$$

  where B_{i_k} represents the *jth* position of the source word aligned with the *ith* target word; and $\overline{B}_{i-1}$ represents the average position of the *jth* words aligned with the (*i -1)th* target word. In practice, $\overline{B}_{i-1}$ the is just assigned the highest j-index value of the source words aligned with the previous target word.

- IBM Model 5: As noted in Brown et al. (1993), Models 3 and 4 are deficient. Strictly speaking, they do not manipulate probability models because each parameter does not have an overall value that equals 1. In this sense, as explained by Och and Ney (2003), "Model 5 is a reformulation of Model 4 with a refined alignment model to avoid deficiency." (27)
- IBM Model 6: Model 6 proposed by Och and Ney (2003) is just a proposal that combines the results of different models, namely HMM and Model 4, so that the resulting algorithm can benefit from particularities of both algorithms. Model 6, in effect, an creates a two ways order 1 distortion model, which is the result of combining i-axis HMM and j-axis Model 4 distortion models. Thus, HMM and Model 4 are combined using a weighted log-linear interpolation:

$$p_6(f|e,a) = \frac{p_4(f|e,a)^{\alpha} \cdot p_{HMM}(f|e,a)}{\sum_{a',f'} p_4(f'|e,a')^{\alpha} \cdot p_{HMM}(f'|e,a')} \tag{4.13}$$

4.1.1 Word-to-word alignments symmetrization

As has been noted before, word alignments are functions that map target words to source words. Thanks to this one way conception, it has been possible to develop models, such as Model 3 and above, where several target words are mapped to a single source word. This kind of mapping is, therefore, many-to-one. However, hand annotation of parallel sentence pairs demonstrates that this conception is not sufficient to account for real translation complexity, for which many-to-many word alignments are usually needed. Thus, in order to achieve many-to-many relations, symmetrization algorithms have to be run on word aligning algorithms output.

The basic idea behind symmetrization algorithms is that the output of word alignment algorithms, which have been ran in both directions —that is, using alternating source and target languages in both sides of the word alignment algorithms—, can be merged.

From a simplistic point of view, this task can be performed by just finding either the intersection or the union of the two-directional output of word alignment algorithms. However, by extending these two approaches with a slightly more complex search algorithm more reliable results can be achieved. Och and Ney (2003, 32-33) propose a *refined* method, currently known as *Neighboring*. Neighboring is a technique that simply explores the words surrounding word alignments contained, for example, in the intersection.

To perform this symmetrization task, the version of the refined method that has been chosen for this research implements the following heuristic methodology:

- First, the algorithm finds the intersection of the bidirectional alignments produced by a particular word alignment algorithm.

- In a second step, it searches for additional alignment points to be added. Only alignments included in the union of the bidirectional alignments are taken into consideration. In this second step, known as *grow-diag step*, only words that are not aligned in the intersection but that, at the same time, are in the union and neighbor current alignments are considered. Formally, neighboring is defined as being directly to the left, right, top, bottom or in the diagonal of alignment included in the intersection.
- In the last step, alignments that do not neighbor established alignment points are considered. This step, known as *final-and step*, is carried out by explicitly aligning pairs of words, only if both of them are unaligned.

This symmetrizing method is called *grow-diag-final-and* algorithm and its pseudo-code extracted from Koehn et al. (2005) is shown in Figure 4.1 below:

```
GROW-DIAG-FINAL(e2f,f2e):
  neighbouring = ((-1,0),(0,-1),(1,0),(0,1),
                  (-1,-1),(-1,1),(1,-1),(1,1))
  alignment = intersect(e2f,f2e);
  GROW-DIAG(); FINAL(e2f); FINAL(f2e);

GROW-DIAG():
  iterate until no new points added
    for english word e = 0 ... en
      for foreign word f = 0 ... fn
        if ( e aligned with f )
          for each neighbouring point ( e-new, f-new ):
            if ( ( e-new not aligned and f-new not aligned ) and
                 ( e-new, f-new ) in union( e2f, f2e ) )
              add alignment point ( e-new, f-new )
FINAL(a):
  for english word e-new = 0 ... en
    for foreign word f-new = 0 ... fn
      if ( ( e-new not aligned or f-new not aligned ) and
           ( e-new, f-new ) in alignment a )
        add alignment point ( e-new, f-new )
```

Figure 4.1. *grow-diag-final-and* symmetrizing algorithm.

4.2 PHRASE-BASED MT SYSTEMS

In Phrase-based MT system, as in any other SMT systems, translation is formalized by the same basic equation 4.4, as described at the beginning of chapter 3, repeated here as equation4.14:

$$\hat{T} = \underset{T}{\operatorname{argmax}}\, P(S|T)P(T) \tag{4.14}$$

However, Phrase-based MT systems are different from other SMT system in terms of what is understood as being the basic unit of translation. Thus, the main intuition behind this type of MT is that words are not always the best translation unit because the correspondence between

languages is not usually 1:1. It could be argued that this problem has already been solved by word-to-word algorithms because, as explained in subsection 4.1.1, *fertility* is incorporated as one of the modeling parameters from IBM Model 3 through 6, and also because many-to-many mappings are made possible by the symmetrization of bidirectional word alignments. This is true but only in part. In fact, Word-based MT systems, such as Brown et al. (1993), which are conceptually prepared to deal with this type of translation mappings, treat these mappings as the conversion of, for example, one unit into several.

Phrase-based MT systems, take one step further and convert word alignments into higher order units, known as phrases —although from a linguistic point of view it is usual that they don't have a linguistic motivation whatsoever. Thus, Phrase-based MT systems do not perform translational mappings between several units, but rather from one unit to another, that is, from a source language chunk or group of words to a target language chunk.

The Phrase-based MT model adopted in this research project is Koehn et al. (2003). For this reason the explanation that will follow will mainly apply to their proposal.

As sketched before, Phrase-based MT systems basic translation unit are phrases or rather, chunks, that is, groups of words that are treated as a single unit. In order to find such chunks Phrase-based MT systems need to perform an additional operation in order to convert word-to-word alignments into phrasal alignments. Thus, symmetrized word alignments are processed by a phrase-extraction algorithm that that pairs chunks.

For every particular pair of sentences, the Phrase-extraction algorithm initial step finds all possible partitions of the target sentence. For each of these partitions the algorithm finds all aligned source sentence words.

Then, for each of these pairs of potentially aligned source and target chunks, the algorithm verifies that all target words aligned with a particular source are, in fact, contained within the target chunk aligned with that source chunk. If there exists any target word that is not contained within the target chunk, the entire pair is ruled out. Otherwise, it is accepted and stored.

The last verification the algorithm performs before including a pair of source and target chunks in the final *phrasikon* has to do with unaligned neighbors. Thus, if there is any unaligned word right next to the source chunk this unaligned word is included within the source chunk. Otherwise, the source chunk remains as it entered this step.

Finally, every pair of source and target chunks that passed step previous verifications are enriched with the conditional probability of the source chunk given the target chunk, and stored in what it is known as a *Phrase Table*.

Phrase tables are, therefore, the piece of information that will encode from the point of view Phrase-based MT, what I called in the beginning of chapter 4 the *translation probability*

model. Thus, Phrase alignments effectively substitute, as pointed out, word alignments as the basic translation units. It is worth noting, however, that there is nothing in the phrase-extraction algorithm that prevents single words from forming their own phrase. In fact, since the algorithm finds all the possible partitions of the target sentence in its initial step, single words partitions are actually considered. As long as single word partitions are consistent with the constraints imposed by the algorithm they are legal phrases and, therefore, are included in the phrasikon.

Before moving on to the detailed explanation of the rest of components of Koehn et al. (2003) Phrase-based MT model, it needs to remarked that, as it can be induced from the explanation of the algorithm above, phrases are not forced to fulfill any linguistically motivated constraint. Therefore, the type of phrases stored in phrase tables do not have to correspond with syntactically motivated units such as constituents. In fact, as pointed out in Koehn et al. (2003), experiments conducted forcing phrases to fulfill syntactic constraints performed much worse in terms of MT output quality.

As an SMT approach, the rest of the components of Koehn et al. (2003)'s Phrase-based MT model are consistent with the equation 4.14. Thus, in addition to the phrasikon, the translation model is completed with a relative distortion probability model, whose function is to model the reordering of target language output phrases. The equation of the distortion model is:

$$d(a_i - b_{i-1}), \tag{4.15}$$

"where a_i denotes the start position of the [source] phrase that was translated into the *ith* [target] phrase, and b_{i-1}denotes the end position of the [source] phrase translated into the *(i – 1)th* target phrase" Koehn et al. (2003, 2).

Both parameters all together account, therefore, for the translation model of equation 4.14 as follows:

$$P(S|T) = p(\bar{f}_1^I|\bar{e^i}) = \prod_{i=1}^{I} \Phi(\bar{f}_i|\bar{e_i})d(a_i - b_{i-1}) \tag{4.16}$$

The second component of equation 4.14, which was informally introduced as the parameter modeling the fluency of the target language translation, is usually known as the *language model*. The language model in Koehn et al. (2003) is a trigram model. Its functions are check the fluency of three-words output chunks and, by an additional factor, ω , calibrate the length of the target language sentence, so that it is neither too long or too short relative to the source sentence.

Therefore, best translation is found in accord with the following general equation:

$$\hat{T} = \underset{T}{\operatorname{argmax}} \prod_{i=1}^{I} \Phi(\bar{f}_i|\bar{e}_i) d(a_i - b_{i-1}) * p_{LM}(T) \omega^{length(T)} \tag{4.17}$$

In order to solve this equation so that the maximum argument, that is, the translation with the highest probability is found, Koehn et al. (2003) model implements a beam search decoder that expands an initial hypothesis by translating sequences of untranslated source words.

Basically, this type of decoder stores all possible incremental hypotheses in stacks, whose size is reduced by pruning out weak hypotheses. Weak hypotheses are defined as hypotheses whose cost is so expensive, that is, that have such a low probability that the final output translation would have an overall low probability. Thus, cost of incremental hypothesis is found by combining their actual cost and the, so called, *future cost* of translating new untranslated source words.

Finally, the output of this search algorithm is the hypothesis with the highest probability that has no untranslated source words.

Chapter 5

Research Project Justification

As pointed out at the beginning of chapter 4, current SMT systems still make use of the word-to-word alignment models that IBM started releasing in the beginning of the 90s, Brown et al. (1990) and Brown et al. (1993). Since then there has been more than two decades of intensive research among researchers to take advantage of those algorithms in order to apply them to SMT.

Improving the quality of those word-to-word alignments was for a long time the only way of also improving the quality of the final tasks those alignments were applied to. The reason is that, regardless of the nature of the basic unit of translation, the same equation, 4.4, originally defined for the early IBM models, is used, in different degrees, to model the currently most important SMT approaches, such as: word substitution models, such as Brown et al. (1993), phrase-based substitution models, such as Koehn et al. (2003), Och and Ney (2004), and also syntax-based substitution models, such as May and Knight (2007) and DeNeefe et al. (2007). Thus, out of the two components of that general equation, the translation model, $P(T|S)$ —as the primary source of information, in the case of the word substitution approach, and on a lower level of abstraction, in the case of the phrase-based and the syntax-based approaches—, is used to model word-to-word alignments between both sides of a parallel corpus.

Therefore, although nowadays the main research efforts in SMT are focused on modeling higher level layers of the phrase-based or syntax-based translation models, the improvement of word-to-word alignments is still an issue of capital importance, because word-to-word alignments are the starting point.

5.1 Previous Approaches to Improve Word-to-Word Alignments Quality

There have been several previous approaches to improve word alignment quality from two different approaches. First, by modifying or using new the word-to-word alignment algorithms and second, by simply modifying the input corpus enriching it with morpho-syntactic information.

Nießen and Ney (2004), Toutanova and Tolga (2002), Lee (2004) and Liu et al. (2005) are approaches that fall within the scope of that first approach. Toutanova and Tolga (2002), who used English-French as their pair of languages, extended HMM alignment model, among

other extensions, using POS-tags of both languages to introducing a new probability function, $P(fT_j|eT_aj)$ —where fT and eT stand for source and target words POS-tag—, which is combined with regular HMM translation and distortion probability functions. Nießen and Ney (2004), whose source and target languages are German and English, combined a series of techniques that modified the German side of the corpus —such as, question inversion, separation of verb prefixes and merging multiword phrases— with the use an MaxEnt-based log-linear model to combine IBM Model 4 with information provided by a conventional bilingual dictionary, a hierarchical lemma-tag lexicon and a POS-tag transition model. Lee (2004), who used Arabic and English as source and target languages respectively, reported improvements on phrase-based SMT by using IBM Model 1 to word align a POS-tagged and segmented version of the source language to a POS-tagged version of the target language. Thus, IBM Model 1 is extended to learn an English tag given an Arabic tag translation probability, an English tag given an Arabic prefix/suffix and their POS-tag translation probability, an English tag given an Arabic suffix and its stem-tag context translation probability, and an English POS-tag given an Arabic prefix and its stem-tag context translation probability. Liu et al. (2005), who used Chinese-English as their pair of languages, implemented a MaxEnt-based log-linear model that combined IBM Model 3, among other features, with a POS-tag transition model leaned from held-out data.

Previous approaches to improve word-to-word alignment quality modifying input corpora are Al-Onaizan et al. (1999), Nießen and Ney (2000), Goldwater and McClosky (2005) and Nguyen and Shimazu (2006). Al-Onaizan et al. (1999), who used Czech and English as source and target languages respectively, reported improvements on word substitution SMT quality by word aligning —using IBM Model 3— several modified versions of the Czech side of the parallel corpus. Thus, they replaced word forms of several POS classes —such as nouns, verbs, adjectives, etc.— by their lemmas, assigning different lemmas to some of the inflected forms of those POS classes; they also treated inflectional suffixes of some POS classes as pseudo-words that were added to the corpus. Nießen and Ney (2000), who used German as source language and English as target language, used IBM Model 1 to word-to-word align a modified version of the source language side by separating verb prefixes, splitting compound words, adding POS-tags to the word forms of some frequent ambiguous words and by merging idiomatic phrases. Goldwater and McClosky (2005), who used Czech and English as their source and target languages respectively, reported improvements on word substitution SMT by word-to-word aligning —using GIZA++ default configuration, that is, m1[5] hmm[5] m3[3] m4[31]— several modified versions of Czech with a

[1]This notation for GIZA++ configurations is taken from Och and Ney (2003). In the this, what this notation means is that GIZA++ will run IBM Model 1 for five iterations, HMM Model for 5 five iterations, Model 3 for three iterations and Model 4 for another three iterations.

non-modified version of English. Thus, for Czech word forms of some POS classes was replaced by their lemma; suffixes and prefixes were treated as pseudo-words; lemmas of some POS classes were enriched with tags representing morphological features, such as Person or Tense. They also experimented with a modified version of GIZA algorithms to make them model associations between lemmas and morphemes. Nguyen and Shimazu (2006), who used English as source language and Vietnamese and French as target languages, reported improvements on phrase-based SMT by word aligning —using GIZA++ default configuration— a modified version of the source language with a non-modified version of the target languages. Thus, they modified the English side of the corpora by replacing English word forms by their lemmas and treating English inflectional suffixes as pseudo-words —and reordering them, if necessary, according to target language word order.

5.2 Exploiting Morphological Information to Improve Word-to-Word Alignment Quality

Whereas all the above-mentioned previous approaches to exploit morphological information to improve word-to-word alignment quality proposed the morphological enrichment of the source language side of the corpus, only three of them, Toutanova and Tolga (2002), Lee (2004) and Liu et al. (2005), proposed also enriching the target language side of the corpus. In my opinion, the reason for this unbalanced proposals is that in most of the cases in which only the source language was morphologically enriched, this was a language explicitly described as having a richer inflectional morphology than the target language. Thus, their attempt was to modify the side of the corpus with a richer morphology so that some of its idiosyncrasies could be faded away or, at least, approached to forms closed to the target language patterns.

In this research project, no *a priori* assumption is made regarding the usefulness of morphological enrichment of parallel corpora. Thus, in my opinion, morphological transformation should not only be applied to that language of the corpus regarded as richer in terms of morphological inflection. However, unlike those three approaches that also transformed the target language side of the parallel corpus, my approach to morphological enrichment does not imply the modification nor the development of a new word alignment algorithm.

5.2.1 Corpus Morphological Enrichment Methods

Once the parallel corpus was preprocessed, following the methodology explained in chapter 3, three basic different versions of the corpus were generated. One that, from now on, will be called *Baseline* and that only contains word forms, as the originally downloaded corpus;

another one that I will call *Tag-Lemmas*, in which word forms are replaced by their *POS-tag + lemma* representation; and another one that I will call *Lemmas*, in which word forms are replaced by their lemmas. Figure 5.1 shows a sample English sentences for each of the versions and Table 5.1 shows the basic statistics of these three different version of the parallel corpus.

Table 5.1. Basic statistics of each of the versions of the parallel corpus.

	Words		**Tag-lemmas**		**Lemmas**	
	Tokens	Types	Tokens	Types	Tokens	Types
English	2535385	42394	2535385	43659	2535385	42104
Galician	2628116	54883	2628116	45523	2628116	43194

Baseline:

the aboot package failed to install into / target / .

Tag-Lemmas:

DET_the N_aboot N_package V_fail TO_to V_install PRP_into QUOTE_/ N_target QUOTE_/ SENT_.

Lemmas:

the aboot package fail to install into / target / .

Figure 5.1. Three different versions of the same English sentence. Two of them are morphologically enriched and another one, Baseline, has not been modified.

Chapter 6

Experiments

6.1 EXPERIMENTAL SETTINGS

As pointed out in subsection 5.2, in this research no *a priori* assumption about which side of the corpus should be, due to its morphological inflection properties, morphologically transformed. Thus, different experiments combining the different (non-)modified versions of each side of the corpus were conducted.

All the different experiments followed the same methodology. In this sense, for each of the combinations of the (non-)modified versions of the parallel corpus a word-to-word alignment model was trained in both directions using GIZA++[1], Och and Ney (2000b). Both directions of word-to-word alignments produced by GIZA++ were symmetrized using the *grow-diag-final-and* algorithm, as described in subsection 4.1.1. Reusing this symmetrized versions of GIZA++'s word alignments, a Phrase-based SMT system was trained using Moses[2], Koehn et al. (2007). The language model used in my experiments, trained with SRILM Toolkit[3], is a backoff fivegram model as opposed to the trigram language model of the Phrase-based MT system described in section 4.2.

For developing and testing Moses SMT system two small parallel corpora of unseen data were compiled. Finally, for MT evaluation it was used BLEU score Papineni et al. (2001) calculated by the NIST script version 11b[4]. As proven by Fraser and Marcu (2007), given the high degree to which phrase-based SMT systems rely on word-to-word alignments, it can be found, in practice, a high correlation between word-to-word alignment quality and

[1]GIZA ++, first developed during John Hopkins University 1999 Summer Workshop, is a F. Och and H. Ney implementation of all the IBM word aligning models as well as the HMM model.

[2]Moses is Phillip Koehn's currently implementation of his 2003 Phrase-based MT proposal.

[3]SRILM is a freely available language modeling toolkit that can be downloaded from http://www.speech.sri.com/projects/srilm/

[4]BLEU score is a MT evaluation measure that measures the closeness of a machine translation to a professional human translation, assuming that the closer the machine translation is to the human translation the better the machine translation is. From this perspective, what BLEU score roughly does is counting the number of n-grams of the machine translation that overlap with n-grams of the human translation, which is used as a reference translation. In practice BLEU score works by combining weighted overlapping n-grams of different sizes — four-grams, trigrams, bigrams and unigrams. In addition to this backoff overlapping n-gram model, BLEU score also implements a *brevity penalty* factor that prevents translations from being too short with respect to the human reference translation.

phrase-based MT quality. Therefore, BLEU score can be used as an indirect way of measuring word-to-word alignment quality.

6.2 EMPIRICAL RESULTS

The empirical results that will be shown in this section correspond to experiments conducted using two different GIZA++ configurations: *i) m1^5 hmm^5 m3^3 m4^3* and *ii) m1^5 hmm^5 m4^3 m6^3*, in which Galician is always the source language and English the target language.

Since no assumption was made about the language that should be *a priori* morphologically enriched, a first series of experiments was conducted, in which, using GIZA++ configuration *i)*, three SMT systems were developed: one using the baseline corpus, another one replacing word forms in both sides of the corpus by their Tag-Lemma representation and another one replacing word forms in both sides of the corpus by their Lemma representation (see Table 5)[5].

As observed in Table 6.1, both morphologically enriched versions of the corpus, *Lemmas to Lemmas* and *Tag-lemmas to Tag-lemmas*, performed better than the Baseline.

[5]As can be observed in this table, as well as in the tables showing BLEU scores for the systems trained for this project, the BLEU scores achieved by the different Phrase-based SMT systems built during the course of this research are considerably lower than those of the *state-of-the-art* Phrase-based SMT systems. A couple of considerations need to be made at this respect. On the one hand, the corpus used in this research project is relatively small in comparison to the corpora, namely Europarl corpora, used in most MT research projects. However, the use of a relatively small training corpus does not seem to fully explain such low BLEU scores. In this sense, it is necessary to point out that a parallel experiment has been conducted to test the impact that a reduction in size would have on a Phrase-based SMT system trained on a Spanish-English Europarl corpus of comparable size to corpus used in this project. Thus, the BLEU score achieved by a Europarl corpus with 2,770,183 tokens on the Spanish side and 2,699,321 on the English side was 0.2851, instead of the BLEU score 0.3136 achieved by a similar SMT system trained on the full Spanish-English Europarl corpus. Therefore, the small size of the corpus used in this research cannot be by any means the only explanation for the low BLEU scores achieved. On the other hand, as it has been already pointed out in section 3.2.2, the sentence-to-sentence realignment process has introduced a considerably high degree of noise in the corpus. Thus, even though no systematic evaluation of the sentence realignment precision has been conducted, it results evident that a rough observation of 20% of misaligned sentences during the compilation of the gold-standard corpus is a clear indication of the degree of noise introduced during this process. Besides these two factors, further evidence would be necessary to irrefutably explain why higher BLEU scores were not achieved in this research. However, an additional experiment was conducted to verify the degree of influence that this sentence misalignment noise had on the BLEU scores achieved by the systems trained in this project. Thus, a Baseline version of the corpus, that is, *Words to Words*, that did not undergo the tagging process and, therefore, did not need to be sentence realigned, was used to train a word alignment model using GIZA++ configuration *m1^{5}hmm^{5}m3^3 m4^3*. This word alignment model was, then, used to train a SMT system with the exact configuration of the rest of systems in this project. As expected, the BLEU score achieved by this system, 0.1681, was better that any of the other systems, but the improvement was not so significant as it could be expected according to the high degree of noised found during the construction of the gold standard corpus. Therefore, sentence misalignment does not explain the low BLEU score of the systems of this project. Given this, it would be, in my opinion, interesting to test to what extent the potentially divergent vocabulary used in the translation units extracted from the package *Compiz Fusion* to create two unseen data sets have also influenced those BLEU scores.

Table 6.1. Results using GIZA++ configuration *m1[5] hmm[5] m3[3] m4[3]* for the following versions of the parallel corpus: *Baseline, Tag-lemmas to Tag-lemmas* and *Lemmas to Lemmas*.

	m1[5] hmm[5] m3[3] m4[3]
Words to Words (Baseline)	0.1559
Tag-lemmas to Tag-lemmas	**0.1627**
Lemmas to Lemmas	0.1597

In order to determine the statistical significance of the differences in BLEU score observed in Table 6.1, an additional set of ten experiments —for SMT systems built using GIZA++ configuration *i)*— was run for each of the combinations included in this table. Thus, ten new testing data sets were randomly generated from the original testing data set, containing each of them only 475 randomly extracted sentences —avoiding repetition of sentences within each of data sets— out of the 950 sentences of the original testing data set. For each of the combinations of models in Table 5, the BLEU score was computed for each small data set; it was found the mean, the standard deviation and, finally, it was run a paired t-test to establish the statistical significance of the observed differences.

Table 6.2. Paired t-test for Baseline, Lemmas to Lemmas and Tag-lemmas to Tag-lemmas for GIZA++ configuration *m1[5] hmm[5] m3[3] m4[3]*.

	Words to Words (Baseline)	**Lemmas to Lemmas**	**Tag-lemmas to Tag-lemmas**
Mean	0.15786	0.16102	0.16455
StdDev	0.009824482	0.01065226	0.01092553
t	-	-2.2375	-5.2119
p-value	-	0.05205	0.0005552*

Table 6.2 shows the results obtained after running the t-test, in which each model is compared to Baseline. The order of the models included in this table is determined by the mean of BLEU score obtained for the ten small data sets. This table shows that for GIZA++ configuration *m1[5] hmm[5] m3[3] m4[3]* the difference between the Baseline model and *Tag-lemmas to Tag-lemmas* is also significant.

A second series of experiments was conducted using GIZA++ configuration *ii)*. In this series of different combinations of (non)morphologically transformed versions of the corpus were used to develop different SMT systems (see Table 6.3).

Table 6.3. Results using GIZA++ configuration *m1^5 hmm^5 m4^3 m6^3* for the following versions of the parallel corpus: *Baseline, Tag-lemmas to Tag-lemmas, Words to Tag-Lemmas, Lemmas to Words, Words to Lemmas, Tag-lemmas to Lemmas and Lemmas to Tag-lemmas*.

	m1^5 hmm^5 m4^{3}m6^3
Words to Words (baseline)	0.1573
Tag-lemmas to Tag-lemmas	0.1605
Lemmas to Lemmas	0.1392
Tag-lemmas to Words	0.1580
Words to Tag-lemmas	0.1606
Lemmas to Words	**0.1628**
Words to Lemmas	0.1528
Tag-lemmas to Lemmas	0.1574
Lemmas to Tag-lemmas	0.1613

Table 6.3, above, shows that all the combinations tried perform better than Baseline and that, although *Tag-lemmas to Tag-lemmas* still performs better than the Baseline model, the combination with the highest score is *Lemmas to Words*.

In order to determine the statistical results showed on Table 6.3, the same methodology as for GIZA++ configuration *i)* was followed.

Table 6.4. Paired t-test for the 10 small data sets for each combination of models. * marks a statistically significant difference.

	Words to Words	Lemmas to Lem-mas	Words to Lem-mas	Tag-lemmas to Lem-mas	Tag-lemmas to Words	Words to Tag-lemmas	Lemmas to Tag-lemmas	Lemmas to Words	Tag-lemmas to Tag-lemmas
Mean	0.157	0.140	0.153	0.1582	0.1585	0.160	0.1622	0.1625	0.1629
StdDev	0.009	0.012	0.0098	0.0097	0.0097	0.008	0.011	0.0081	0.0104
t	-	8.15	5.59	-0.389	-0.729	-1.57	-2.14	-2.26	-4.51
p-value	-	1.89e-05*	0.0003*	0.7063	0.4846	0.1497	0.0606	0.0502	**0.0014***

As can be observed in Table 6.4, a statistically significant difference was found for only three pairs: *Lemmas to Lemmas, Words to Lemmas and Tag-lemmas to Tag-lemmas*. In the case of the first two, as the positive sign of the *t* value indicates, the difference is in favor of Baseline. However, the case of *Tag-lemmas to Tag-lemmas*, the difference is in favor of the non-Baseline system.

It is also worth noting the case of the model *Lemmas to Lemmas*. Thus, although it had performed quite well for GIZA++ configuration *i)*, for configuration *ii)* it was the worst model of the series. In fact, there was an statistically significant difference between the BLEU score of this model and the Baseline, but the difference is favorable to Baseline, which is significantly better. The rest of the models, although some of them, such as *Lemmas to Tag-lemmas* and *Lemmas to Words*, have p-values close to 0.05, their score is not significantly better than Baseline's.

Finally, the last experiment conducted involved the comparison of both *Tag-lemmas to Tag-lemmas* models build using the two different GIZA++ configurations used in this research.

Table 6.5. Paired t-test for both Tag-lemmas to Tag-lemmas models built using GIZA++ configurations $m1^5$ hmm^5 $m3^3$ $m4^3$ —configuration i)— and $m1^5$ hmm^5 $m4^3$ $m6^3$ — configuration ii).

	Configuration i)	Configuration ii)
Mean	0.16455	0.16299
StdDev	0.01092553	0.01043024
t	-	1.2032
p-value	-	0.2596

As Table 6.5 shows there is no statistically significant difference between the two GIZA++ different for the Tag-lemmas to Tag-lemmas version of the parallel corpus.

6.3 A DIFFERENT INSIGHT IN WORD-TO-WORD ALIGNMENTS

Although, as stated in section 6.1, the main evaluation measure chosen for this research project was the NIST BLEU score, this measure does not directly evaluate the quality of word-to-word alignments. Instead, BLEU score is an indirect measure that gives an indication of word alignments quality depending on how good a MT systems performs when reusing them to translated unseen events.

In order to get a different and more direct insight into word alignments, the quality of all the word-to-word alignments produced by GIZA++ for each of the combinations of versions of the parallel corpus was also evaluated using a hand-aligned gold standard.

Thus, a small gold standard was built by randomly extracting pairs of source and target sentences out of the same parallel corpus used to train GIZA++ models. As pointed out by Och and Ney (2003), word alignments are "performed in a completely unsupervised way [...], [therefore] there is no need to have a test corpus separate from the training corpus" (34). Once a sufficient number of candidate pairs was extracted, each of these pairs, which were not allowed to be repeated along this small test set, were hand-aligned using Jean Mark Gawron's word aligner editor, *aligner-1.0* [6]. In order to maintain the coherence and consistency among word alignments produced by hand, the gold standard was annotated following the Blinker Project translational equivalence annotation guidelines, Melamed (1998a) and Melamed (1998b). As discussed in Section 3.1.2, during the process of building the gold standard a significant number of misaligned sentences were encountered. For the purpose of building a noise-free gold standard corpus, those misaligned pairs were discarded. The final size of the gold standard was 300 sentences.

Most of the word alignment research conducted in the last four years has used Alignment Error Rate (AER) as the measure of word-to-word alignment quality Och and Ney (2003). AER is based on the assumption that word alignment is an inherently ambiguous task (*vid.* Melamed (1998b)); therefore, word alignments in hand-aligned gold standards should be specified in two different ways: as Sure (S), that is, unambiguous alignments; or as Possible (P), that is, ambiguous alignments. Following this perspective, word alignments recall and precision are measured as follows:

$$Recall : \frac{|A \cap S|}{|S|} \tag{6.1}$$

$$Precision : \frac{|A \cap P|}{|A|} \tag{6.2}$$

where S and P stand for Sure and Possible alignments, respectively; and A stands for the Alignments produced by GIZA++ algorithms.

Combining recall and precision, as defined above, AER is calculated as follows:

$$AER(S, P; A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} \tag{6.3}$$

Although AER is the measure used by most of the research published since Och and Ney (2003) proposed it, the AER was not used in this project because, as Fraser and Marcu (2007)

[6] Word aligner editor *aligner-1.0* can be freely downloaded from http://www-rohan.sdsu.edu/~gawron/mt_plus/mt/course_core/code/aligner/index-1.0.html

point out, AER does not punish unbalanced precision and recall. Thus, it is possible to obtain equally good AER evaluations for balanced and unbalanced pairs of recall and precision. Following a Fraser and Marcu (2007, 297) example, consider two alignment systems A and B such that $|P \cap A| = 50$ and $|S \cap A| = 50$, for A, and $|P \cap A| = 75$ and $|S \cap A| = 25$, for B. The GIZA++ score for system A would be 0.5 for both recall and precision, while the GIZA++ score for system B would be 0.75 and 0.25, for precision and recall respectively. These results make the difference between systems A and B obvious. However, the AER metric cannot capture this difference because, according to equation 6.3, both models have an AER score of 0.5. In general, AER can be unreliably maximized by "favoring precision over recall, which can be done by simply guessing very few alignment links" Fraser and Marcu (2007, 297). This is a general problem for error-rate measures.

Given this mathematical shortcoming in AER formulation, Fraser and Marcu (2007) suggest a different word alignment quality measure, name F-Measure, which is formalized as follows:

$$\textit{F-Measure} = \frac{1}{\frac{\alpha}{Precision} + \frac{1-\alpha}{Recall}} \tag{6.4}$$

In this formulation the α-parameter allows models with unbalanced Precision and Recall to be penalized. Given the flexibility that this additional factor provides, F-Measure can be tuned to provide a task-specific measure of word alignment quality. For example, by finding the optimal value of α that gives the best correlation between BLEU score and F-Measure, a BLEU score specific balance between Precision and Recall can be found.

As for the distinction between Sure and Possible alignments, Fraser and Marcu (2007, 299) show that a better correlation between BLEU score and F-measure can be found when annotating the gold standard with Sure-only alignment links. In the end, one has to take into account the fact that the distinction between Sure and Possible alignments is also a highly subjective distinction, if hand annotation of word alignment is an ambiguous task *per se*. Thus, it is common practice in research projects to explicitly discuss and define which alignments should be annotated as Sure and which as Possible. From this point of view, how Sure and Possible alignments are defined in each particular research has an substantial influence on Precision and Recall. In order to avoid this source of subjectivity, it was decided, following Fraser and Marcu (2007) suggestion, to annotate the gold standard described above with Sure-only alignment links. Following this strategy,

Given Fraser and Marcu (2007) protocol of annotating Sure-only links, Recall and Precision equations are reformalized as follows:

$$Recall : \frac{|A \cap S|}{|S|} \tag{6.5}$$

$$Precision : \frac{|A \cap S|}{|A|} \tag{6.6}$$

As noted above, under the Fraser and Marcu (2007) approach to word alignment, measurement of word alignment quality requires finding the optimal value of α for correlating word alignment score with BLEU score. Following Fraser and Marcu (2007), the coefficient of determination (r^2) was used for this task. This is the square of the Pearson product moment correlation coefficient (r), whose equation is:

$$r = \frac{N \sum xy - (\sum x)(\sum y)}{\sqrt{N \sum x^2 - (\sum x)^2}\sqrt{N \sum y^2 - (\sum y)^2}} \tag{6.7}$$

, where *N* stands for the number of event and *x* and *y* for each of the distributions that are being correlated.

Thus, for each of the GIZA++ configurations used in the experiments above an r^2 correlation coefficient was calculated (see Figures 6.1 and 6.2).

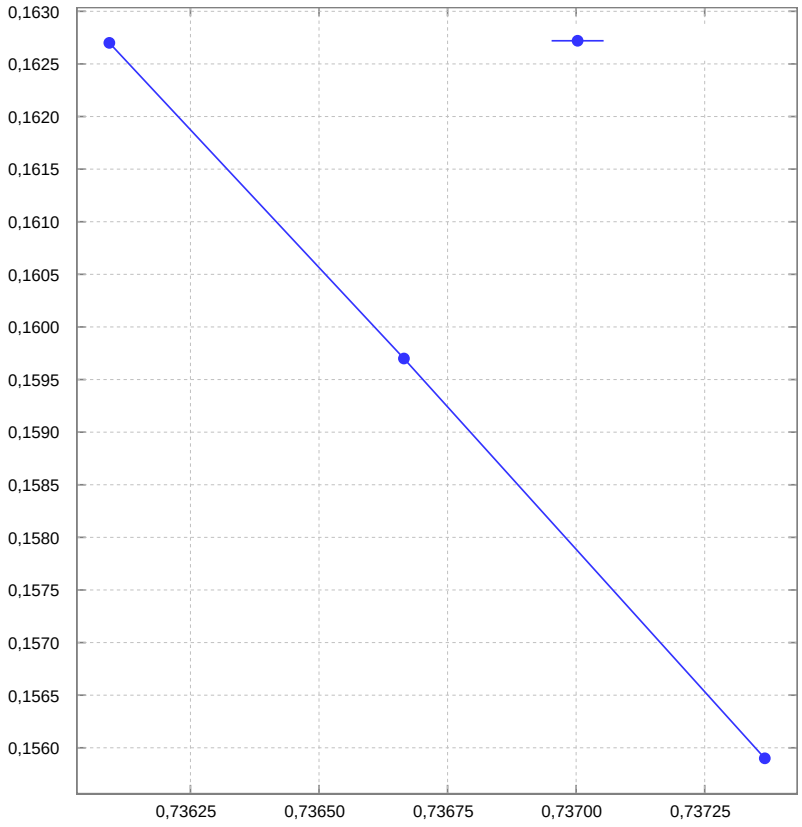

Figure 6.1. F-Measure (x-axis) α = 0.6 against BLEU score (y-axis) for m1[5] hmm[5] m3[3] m4[3], r^2 = 1.00.

As Figure 6.1 shows, for GIZA++ configuration *m1*^{5} *hmm*^{5} *m3*^{3} *m4*^{3} there is a perfect correlation between BLEU score and F-Measure when α is set to 0.6. Although the differences in F-Measure do not seem to be substantial, this correlation is, unexpectedly, inverse. In other words, the model with highest BLEU score, *Tag-lemma to Tag-lemma*, is the model with the lowest word alignment quality and the model with the lowest BLEU score, Baseline, is in contrast the model with the highest word alignment quality.

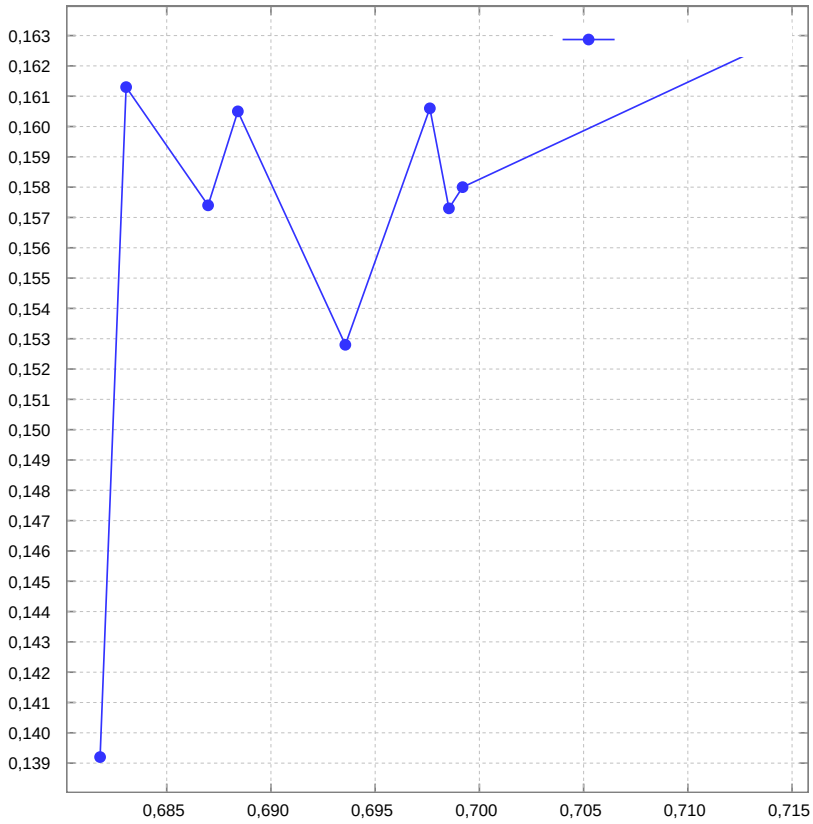

Figure 6.2. F-Measure (x-axis) α = 0.1 against BLEU score (y-axis), for m1^5 hmm^5 m4^3 m6^3, r^2 = 0.242.

In the case of GIZA++ configuration *m1*^{5} *hmm*^{5} *m4*^{3} *m6*^{3} (Figure 6.2), the best correlation that could be found between BLEU score and F-Measure is 0.242, when α is set to 0.1. Although, unlike the other GIZA++ configuration, the correlation is in this case positive, that is, models with higher F-Measures tend to correlate with higher BLEU scores; the correlation coefficient found is fairly small. Therefore, as for the experiments conducted in this research it is not true that the F-Measure works as a good predictor for BLEU scores.

It needs to be noted that Fraser and Marcu (2007) use F-Measure and r^2 correlation coefficient, as defined above, to capture the relation between differences in word alignment

quality and BLEU score achieved when using corpora of different sizes. In contrast, in this research project this correlation coefficient is used to explain the differences observed, not for corpora of different sizes, but for versions of the same corpus, that is, with an invariable size.

Since BLEU score and F-Measure have been proven, despite their drawbacks, to be fairly good evaluation measures, what is, in my opinion, most likely to be failing in the comparison carried out in this section is the type of alignment links annotated when building the gold standard corpus. Although an error analysis study would be necessary to directly explore how GIZA++ alignment models output relates to the hand-annotated gold standard used in this research, it does not seem, in my opinion, too adventurous to speculate that there must be some alignments that, depending on the properties of the versions of the corpus used to train GIZA++ algorithms, can be captured, whereas others just cannot be captured. It is necessary to say that Och and Ney (2003) and Fraser and Marcu (2007) are not the only proposals to create hand-annotated gold standards. Thus, Ahrenberg et al. (2000), who also do not distinguish between either Sure and Possible alignments, introduced, along with Sure-only alignments analogous to Fraser and Marcu (2007), another type of alignment called *partially correct alignment*. From their point of view, since word-to-word alignment algorithms can find one-to-many and many-to-one relations —and of course when symmetrized many-to-many relations—, it is necessary to beware that, given their formal properties, there is nothing that prevents word alignment algorithms from partially finding many-to-one and one-to-many relations. Hence, the pertinence of their proposal of including a partially correct alignments within the hand-annotated alignments paradigm. Seeking a better correlation between F-Measure and BLEU, this inclusion would need the redefinition of Precision and Recall so they take into account this type of alignments.

I will conclude with a discussion of Table 6.6, which summarizes the evaluation of all the GIZA++ configurations that were trained during this project. For the majority of GIZA++ configurations, there are only data for a few combinations of versions of the corpus. Due to time limitations it was only possible to reuse some of these GIZA++ configurations to train Phrase-based MT systems. Besides this, as can be observed in Figures 6.1 and 6.2, each of the configurations for which Moses models were developed have very different optimal α values in terms of F-Measure and BLEU score correlation. For this reason, all F-Measures in Table 6.6 have been calculated setting α to 0.5, which does not favor either Precision nor Recall.

Table 6.6. Table 10: Summary of evaluation data of all the GIZA++ configurations trained during this research project. In this table, P, R and F stand for Precision, Recall and F-Measure, respectively. Numbers in the first column stand for GIZA++ configurations as follows: 1- *m1*^{5} *hmm*^{5}, 2- *m1*^{5} *m2*^{5}, 3- *m1*^{5} *hmm*^{5} *m3*^{3}, 4- *m1*^{5} *m2*^{5} *m3*^{3}, 5- *m1*^{5} *hmm*^{5} *m3*^{3} *m4*^{3}, 6- *m1*^{5} *hmm*^{5} *m4*^{3}, 7- *m1*^{5} *m2*^{5} *m3*^{3} *m4*^{3}, 8- *m1*^{5} *hmm*^{5} *m3*^{3} *m4*^{3} *m5*^{3}, 9- *m1*^{5} *hmm*^{5} *m4*^{3} *m5*^{3}, 10- *m1*^{5} *hmm*^{5} *m3*^{3} *m4*^{3} *m6*^{3}, 11- *m1*^{5} *hmm*^{5} *m4*^{3} *m6*^{3}.

		Words to Words	Words to Lem-mas	Words to Tag-lemmas	Lemmas to Lem-mas	Lemmas to Tag-lemmas	Lemmas to Words	Tag-lemmas to Tag-lemmas	Tag-lemmas to Lem-mas	Tag-lemmas to Words
1	P	0.75	-	-	0.756	-	-	0.75	-	-
	R	0.86	-	-	0.848	-	-	0.851	-	-
	F	**0.801**	-	-	**0.799**	-	-	**0.798**	-	-
2	P	0.686	-	-	0.683	-	-	0.68	-	-
	R	0.829	-	-	0.814	-	-	0.818	-	-
	F	**0.751**	-	-	**0.742**	-	-	**0.743**	-	-
3	P	0.858	-	-	0.869	-	-	0.86	-	-
	R	0.688	-	-	0.675	-	-	0.688	-	-
	F	**0.764**	-	-	**0.76**	-	-	**0.764**	-	-
4	P	0.828	-	-	0.846	-	-	0.826	-	-
	R	0.61	-	-	0.597	-	-	0.595	-	-

(table continues)

Table 6.6 (continued)

		Words to Words	Words to Lem-mas	Words to Tag-lemmas	Lemmas to Lem-mas	Lemmas to Tag-lemmas	Lemmas to Words	Tag-lemmas to Tag-lemmas	Tag-lemmas to Lem-mas	Tag-lemmas to Words
	F	**0.703**	-	-	**0.7**	-	-	**0.692**	-	-
5	P	0.860	-	-	0.876	-	-	0.883	-	-
	R	0.607	-	-	0.595	-	-	0.603	-	-
	F	**0.712**	-	-	**0.709**	-	-	**0.71**	-	-
6	P	0.781	-	-	0.809	-	-	0.797	-	-
	R	0.689	-	-	0.669	-	-	0.677	-	-
	F	**0.732**	-	-	**0.732**	-	-	**0.732**	-	-
7	P	0.848	-	-	0.867	-	-	0.853	-	-
	R	0.578	-	-	0.571	-	-	0.563	-	-
	F	**0.688**	-	-	**0.689**	-	-	**0.679**	-	-
8	P	0.862	-	-	0.875	-	-	0.862	-	-
	R	0.604	-	-	0.6	-	-	0.604	-	-
	F	**0.71**	-	-	**0.712**	-	-	**0.71**	-	-
9	P	0.811	-	-	0.834	-	-	0.821	-	-
	R	0.641	-	-	0.619	-	-	0.633	-	-

(table continues)

Table 6.6 (continued)

		Words to Words	Words to Lem-mas	Words to Tag-lemmas	Lemmas to Lem-mas	Lemmas to Tag-lemmas	Lemmas to Words	Tag-lemmas to Tag-lemmas	Tag-lemmas to Lem-mas	Tag-lemmas to Words
	F	**0.716**	-	-	**0.711**	-	-	**0.715**	-	-
10	P	0.859	-	-	0.874	-	-	0.862	-	-
	R	0.61	-	-	0.597	-	-	0.605	-	-
	F	**0.714**	-	-	**0.709**	-	-	**0.711**	-	-
11	P	0.781	0.765	0.76	0.809	0.807	0.817	0.798	0.802	0.8
	R	0.69	0.686	0.691	0.67	0.672	0.704	0.678	0.676	0.69
	F	**0.733**	**0.724**	**0.724**	**0.733**	**0.733**	**0.757**	**0.733**	**0.734**	**0.741**

Since an F-Measure calculated using Sure-only alignment links is not a good BLEU score predictor, as the above discussion showed, only limited evaluation is possible for any of the GIZA++ configurations for which word alignment quality data are available but for which no Moses model was built, and no BLEU scores were computed.

It is interesting to note, though, that there is a divergence in performance between GIZA++ configurations incrementally trained on Model 3 (configuration 3 through 11) and above and those configurations trained only using Model 1 and either HMM model or Model 2 (configurations 1 and 2). Thus, whereas those configurations incrementally trained on Model 3 and above consistently have a bigger precision than recall; configurations trained on Model 1 and HMM or Model 2 have a much bigger recall than precision.

Since in “applications such as statistical machine translation [...] a higher recall is more important” Och and Ney (2003, 33) than a higher precision, it would be interesting to verify the BLEU scores achievable by GIZA++ configurations, such configuration 1 and 2, when their word alignments are used to build Phrase-based MT systems.

Chapter 7

Conclusions

The first conclusion that can be drawn from this research is that benefits of morphological transformation depend on the language-pair languages to which they are applied. The only transformation applied to this parallel corpus that coincided with a strategy previously used was replacing word forms by their lemma representation. This strategy did not achieve better results than Baseline, in contrast to what has been reported in Goldwater and McClosky (2005) and Nguyen and Shimazu (2006).

Another important conclusion is that differences in the complexity of morphological inflection of each of the languages do not determine *a priori* which language should be object of morphological transformations. Thus, English is clearly less morphologically complex than Galician, but the only morphological transformation that achieved a statistically significant improvement over Baseline involved the transformation of both sides of the parallel corpus.

In addition, it seems also fair to conclude that the benefits of morphological transformation of a parallel corpus depend on which word-to-word alignment algorithm is used. Thus, it seems clear that different GIZA++ configurations yielded to different MT results and that, in some cases, changes were quite dramatic.

Finally, given the low correlation between BLEU score and F-Measure found in this research, the last conclusion that can be drawn is that, at least when measuring word alignment quality of different versions of the same parallel corpus, a reformulation of word alignment links is necessary. Thus, in order to make F-Measure work as a good predictor of Phrase-based SMT systems behavior, it will be necessary to determine which properties of a particular version of a corpus favor different types of alignment links.

Chapter 8

Future Research

As a continuation of the research conducted during this Master's thesis, future work will be carried out from two different points of view.

From a more practical point of view, future work will prioritize the following two directions:

1. Corpus augmentation: It is well known that SMT system heavily rely on the size of the corpus used to train them to achieve competitive results. Since the software localization English-Galician parallel corpus was compiled new localization projects have been initiated and some of them have been completed already. For this reason, the parallel corpus will be augmented with new available data so that future experiments can benefit from a bigger collection of training data.
2. Minimization of noise sources: As noted during the description of all the preprocessing steps of the parallel corpus used in this research, sentence-to-sentence realignment was an obligatory step to be taken due to the fact that the POS-tagger used to morphologically annotate the corpus needed to have each token in a different line. Although no systematical evaluation of the noise this realignment introduced in the corpus was performed, it was found that a substantial proportion of sentences were misaligned. For this reason, the second direction to be taken in future work would necessarily pass by avoiding sentence-to-sentence realignment. In this sense, preliminary experiments are being conducted to develop a Maximum Entropy POS-tagger that will allow POS-tagging without making necessary to sentence realign the corpus.
3. Research expansion: As mentioned in chapter 2, this research project is situated within the context designed by the approval of the GPN, which intends the promotion of Galician within the IT sector. This objective can only be accomplished by localizing into Galician a vast inventory of computational applications and by developing, at the same time, a standardized technical apparatus to help maintain the consistency among all the localizations. Drawing, precisely, from this perspective, besides the development of a domain-specific English-Galician SMT system, this research project also intends the development of automatically extracted bilingual lexicons that, in conjunction with Translation Aid Tools, such as the above cited SMT system, can help translators in the localization of those computational applications. Since for the extraction of high quality bilingual lexicons, a high word alignment precision is required, it is worth noting that the version of the corpus that achieved the highest overall BLEU score —as seen in table 6.3—, *Lemmas to Words*, is also the version with the highest precision in table 6.6. This version of the corpus, as well as other versions that have achieved promising results during this research, namely *Lemmas to Tag-lemmas* and, of course, *Tag-lemmas*

to Tag-lemmas, will be explored in detail so that can be exploited in the development of high quality bilingual lexicons. Furthermore, new symmetrizing strategies, known to provide high precision word alignment quality —for example, the intersection of the bidirectional word alignments or a more conservative version of symmetrizing algorithm, such as *grow-diag*, described in Section 4.1.1— will be tried in further experiments to test their influence on the BLEU scores of all the different SMT systems trained during this project.

4. Error analysis study: In order to explore how GIZA++ alignment models output relates to the hand-annotated gold standard used in this research. Seeking to improve the correlation between F-Measure and BLEU score, this type of study will provide valuable information to determine what type of alignments can or cannot be capture by different GIZA++ alignment models. This information will be, in my opinion, of crucial importance to develop a new typology of alignment links to hand-annoted gold standards intended to be used to measure word alignment quality.

Among the numerous unanswered questions that could arise from this research, some of them, because they are, in fact, derived from the conclusions above, have special importance in that they will, from a more theoretical point of view, inspire future research:

- Would it be possible to establish a language typology to determine which type of morphological transformation is more beneficial for a particular pair of languages?
- Which properties of a pair of languages or a particular morphological transformation determine the behavior of a particular word alignment algorithm? Is there a systematical correlation among these factors?
- What type of alignment links, found by each of the word alignments algorithms when trained with different versions of the parallel corpus, increase or lower Phrase-based MT systems performance?
- Would it be possible to develop a new typology of alignment links to hand-annotate gold standards, so that the F-measure of different versions of the same parallel corpus can better correlate with the performance of Phrase-based SMT systems and, therefore, be used as a predictor of MT systems behavior?

8.0 ?

Abney, S. 2002. Bootstrapping. In *Procedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 360–367.

Ahrenberg, L., M. Andersson, and M. Merkel. 1998. A Simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Texts. In *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, 29–35.

Ahrenberg, L., M. Merkel, A. Hein, and J. Tiedemann. 2000. Evaluation of word alignment systems. In *Proceedings of 2nd International Conference on Language Resources & Evaluation (LREC 2000)*.

Al-Onaizan, .Y, J. Curin, M. Jahr, K. Knight, J. Lafferty, D. Melamed, F. J. Och, D. Purdy, N.A. Smith, and D. Yarowsky. 1999. Statistical machine translation. Technical report. JHU.

Ayan, N., B. Dorr, and N. Habash. 2004. Multi-Align: Combining linguistic and statistical techniques to improve alignments for adaptable MT. In *Proceedings of the 6th Conference of the AMTA (AMTA-2004)*, 17–26.

Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, and P. Roosin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics* 16(2):79–85.

Brown, P., S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, and R. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics* 19(2):263–311.

Brown, P., S. Della Pietra, V. Della Pietra, J. Lafferty, and R. Mercer. 1992. Analysis, statistical transfer, and synthesis in machine translation. In *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92), Empiricist vs. rationalist methods in MT*, 83–100.

Callison, C., and M. Osborne. 2003. Bootstrapping Parallel Corpora. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond*, 44–49.

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

Caseli, H., A. Silva, and M. Nunes. 2004. Evaluation of Methods for Sentence and Lexical Alignment of Brazilian Portuguese and English Parallel Texts. In *In Proceedings of the 7th SBIA, LNAI 3171*, 184–193.

Civera, J., and A. Juan. 2006. Mixtures of IBM Model 2. In *EAMT-2006: 11th Annual Conference of the European Association for Machine Translation*, 159–167.

Cooper, R. L. 1990. *Language Planning and Social Change*. Cambridge: Cambridge University Press.

Delisle, J., C. Lee-Jahnke, and M. Cormier. 1999. *Terminologie de la Traduction. Translation Terminology. Terminología de la Traducción. Terminologie der Übersetzung*. John Benjamins.

DeNeefe, S., K. Knight, W. Wang, and D. Marcu. 2007. What Can Syntax-based MT Learn from Phrase-based MT? In *Proceedings of EMNLP-CoNLL*.

Francis, W. N. 1964. A standard sample of present-day English for use with digital computers. Technical report. Report to the U.S Office of Education on Cooperative Research Project No. E-007. Providence: Brown University.

Francis, W. N, and H. Kučera. 1982. *Frequency analysis of English usage. Lexicon and Grammar.* Boston: Houghton Mifflin.

Fraser, A., and D. Marcu. 2007. Measuring Word Alignment Quality for Statistical Machine Translation. *Computational Linguistics* 33(3):293–303.

Fung, P. 1998. *Machine Translation and the Information Soup*. Chap. A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel Corpora, 1–17. Heidelberg: Springer Berlin.

Gale, W. A. 1991. Identifying Word Correspondences in Parallel Texts. In *Proceedings of the workshop on Speech and Natural Language*, 152–157.

Galley, M., M. Hopkins, K. Knight, and D. Marcu. 2004. What's in a translation rule? In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL-04)*.

Gamallo, P. 2005. Extraction of Translation Equivalents from Parallel Corpora Using Sense-Sensitive Contexts. In *Proceedings of Conference of the European Association for Machine Translation (EAMT'05)*.

Gamallo, P. 2006. *Computational Processing of the Portuguese Language*. Chap. Using Natural Alignment to Extract Translation Equivalents, 41–49. Heidelberg: Springer Berlin.

Gamallo, P., and S. S. Docío. 2005. El tratamiento de la polisemia en la extracción de léxicos bilingües a partir de corpora paralelos. *Procesamiento del Lenguaje Natural* 35:103–110.

Gamallo, P., and J. R. Pichel. 2005. An Approach to Acquire Word Translations from Non-Parallel Texts. In *12th Portuguese Conference on Artificial Intelligence (EPIA'05)*.

Gee, J. P. 1999. *An Introduction to Discourse Analysis: Theory and Method*. London: Routledge.

Goldwater, S., and D. McClosky. 2005. Improving statistical MT through morphological analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 676–683.

Guinovart, X. G., and E. F. Sacau. 2004. Métodos de optimización de la extracción de léxico bilingüe a partir de corpus paralelos. *Procesamiento del lenguaje natural* 33:133–140.

Huang, J-X., and K-S. Choi. 2000. Chinese-Korean Word Alignment Based on Linguistic Comparison. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 392–399.

Jurafsky, D., and H.J. Martin. To be published in spring 2008. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.

Knight, K. 1999a. A Statistical MT Tutorial Workbook. from http://www.isi.edu/natural-language/mt/wkbk.rtf, April 30.

Knight, K. 1999b. Decoding Complexity in Word-Replacement Translation Models. *Computational Linguistics* 25(4):607–615.

Koehn, P., A. Axelrod, A. Birch, C. Callison, M. Osborne, and D. Talbot. 2005. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. In *MT Eval Workshop*.

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

Koehn, P., F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, 127–133.

Lee, Y-S. 2004. Morphological analysis for statistical machine translation. In *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, 57–60.

Liu, Y., Q. Liu, and S. Lin. 2005. Log-linear Models for Word Alignment. In *Proceedings of the 43rd Annual Meeting of the ACL*, 459–466.

Mauser, A., E. Matusov, and H. Ney. 2006. Training a Statistical Machine Translation System Without GIZA++. In *International Conference on Language Resources and Evaluation*, 715–720.

May, J., and K. Knight. 2007. Syntactic Re-Alignment Models for Machine Translation. In *Proceedings of EMNLP-CoNLL*.

Melamed, I.D. 1997. A Word-to-Word Model of Translational Equivalence. In *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, 490–497.

Melamed, I.D. 1998a. Annotation Style Guide for the Blinker Project. Technical report. IRSC Technical Report No. IRCS-98-06.

Melamed, I.D. 1998b. Manual Annotation of Translational Equivalence: The Blinker Project. Technical report. IRCS Technical Report No. 98-07.

Monteagudo, H. 2002. A lingua galega na sociedade: descrición da situación actual e perspectivas de futuro. In *A Normalización Lingüística a debate*, ed. H. Monteagudo, S. García Conde, H. López, and X. Subiela. 7–46. Vigo: Edicións Xerais de Galicia.

Moore, C. R. 2004. Improving IBM Word-Alignment Model 1. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 519–526.

Nguyen, T.P., and A. Shimazu. 2006. Improving phrase-based statistical machine translation with morphosyntatic transformation. *Machine Translation* 20:147–166.

Nießen, S., and H. Ney. 2000. Improving SMT quality with morpho-syntactic analysis. In *Proceedings of COLING*.

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

Och, F. J. 1999. An Efficient Method for Determining Bilingual Word Classes. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics,* 71–76.

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

Och, F. J., and H. Ney. 2000a. A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th conference on Computational linguistics*, 1086–1090.

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

Och, F.J., and H. Ney. 2000b. Improved statistical alignment models. In *Proceedings of 38th Annual meeting of the ACL*, 440–447.

Och, F.J., and H. Ney. 2004. The alignment template approach to statistical machine translations. *Computational Linguistics* 30(4):417–449.

Papineni, K.A., S. Roukos, T. Ward, and W.J. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical report. IBM Research Division, Thomas J. Watson Research Center.

Schmid, H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the Conference on New Methods in Language Processing*.

Tiedemann, J. 1998. Extraction of translation equivalents from parallel corpora. In *Proceedings of the 11th Nordic Conference on Computational Linguistics NODALI98*.

Tiedemann, J. 2003. Combining Clues for Word Alignment. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, 339–346.

Toutanova, K., and C Tolga, H. Manning. 2002. Extensions to HMM-based Statistical Word Alignment Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 87–94.

Xunta de Galicia. 2005. *Plan Xeral de Normalización Lingüística*. Santiago de Compostela: Xunta de Galicia.

ABSTRACT OF THE THESIS

IMPROVING WORD-TO-WORD ALIGNMENTS USING MORPHOLOGICAL INFORMATION

by
Pablo Malvar Fernández
Master of Arts in Linguistics
San Diego State University, 2008

All current Statistical Machine Translation systems rely on an initial layer of word-to-word alignment; not surprisingly, alignment quality has been shown to be a key factor in performance at higher levels of abstraction. This thesis is an investigation into the effects of part-of-speech tagging and morphological transformations of a parallel corpus on aligment quality.