UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València

Interactive Post-Editing in Machine Translation

Master's Thesis

Master's Degree in Artificial Intelligence, Pattern Recognition and Digital Imaging

Author: Miguel Domingo Ballester

Supervisor: Francisco Casacuberta Nolla

September 11, 2015

Abstract

The current state of the art in Machine Translation (MT) is far from being good enough, with a post-process carried out by a human agent being necessary in many cases in order to correct translations.

Statistical post-editing of a MT system has been used in the past to improve the translation quality of that system. Additionally, research on interactive translation prediction has been done with the aim of reducing the human post-editing effort. In this thesis, a new methodology that combines both techniques is proposed in order to, given a MT system, increase the translation quality of that system and reduce the effort that the human agent needs to make in order to correct the translation of that system.

This methodology is tested on different scenarios (to connect with the output of a rule-based machine translation system, and as a method to adapt an statistical MT system from one domain to another) with different corpora, obtaining very encouraging results.

Keywords: statistical machine translation; statistical post-editing; rule-based machine translation; domain adaptation; interactive translation prediction

Resumen

El estado actual del arte en traducción automática (Machine Translation, MT) todavía no es lo suficientemente bueno, siendo en muchos casos necesario un post-proceso llevado a cabo por un agente humano a fin de corregir las traducciones.

La post-edición estadística de un sistema de MT se ha utilizado en el pasado para mejorar la calidad de traducción de dicho sistema. Además, se han llevado a cabo investigaciones en traducción mediante predicción interactiva con el objetivo de reducir el esfuerzo humano de post-edición. En esta tesis se propone una nueva metodología que combina ambas técnicas a fin de, dado un sistema de MT, incrementar la calidad de traducción de dicho sistema y reducir el esfuerzo que el agente humano ha de hacer a la hora de corregir las traducciones de dicho sistema.

Esta metodología ha sido probada en diferentes escenarios (para conectar la salida de un sistema de traducción basado en reglas, y como método para adaptar un sistema de MT estadístico de un dominio a otro) con diferentes córpora, obteniendo resultados muy esperanzadores.

Palabras clave: traducción automática estadística; post-edición estadística; traducción automática basada en reglas; adaptación al dominio; traducción mediante predicción interactiva

Resum

L'estat actual de l'art a traducció automàtica (Machine Translation, MT) encara no és prou bona, sent necessari en molts casos un postprocés realitzat per un agent humà a fi de corregir les traduccions.

La postedició estadística d'un sistema de MT s'ha utilitzat en el passat per a millorar la qualitat de traducció de dit sistema. A més, s'han dut a terme investigacions en traducció mitjançant predicció interactiva amb l'objectiu de reduir l'esforç humà de postedició. En aquesta tesi es proposa una nova metodologia que combina dues tècniques a fi de, donat un sistema de MT, incrementar la qualitat de traducció de dit sistema i reduir l'esforç que l'agent humà ha de realitzar a l'hora de corregir les traduccions de dit sistema.

Aquesta metodologia ha sigut provada a diferents escenaris (per a connectar l'eixida d'un sistema de traducció basat en regles, i com a mètode per adaptar un sistema de MT estadístic d'un domini a un altre) amb diferent còrpora, obtenint resultats molt esperançadors.

Paraules clau: traducció automàtica estadística; postedició estadística; traducció automàtica basada en regles; adaptació al domini; traducció mitjançant predicció interactiva

Acknowledgments

I would like to thank my family. Specially my mum, for all her unconditional love.

My most sincere gratitude to Dr. Francisco Casacuberta, for supervising this thesis and giving me the opportunity to collaborate with the PRHLT Research Center, introducing me to the world of research.

I'm also grateful to Dr. Germán Sanchis, whose help and advices were crucial for the development of this thesis.

Finally, I would like to thank my labmates. Specially Mara and Álvaro, whose advices have helped me a lot throughout this work.

The research leading to these results has received funding from the Ministerio de Economía y Sostenibilidad (MINECO) under grant RTC-2014-1466-4, and Generalitat Valenciana under grant PROMETEOII/2014/030.

Contents

List of Figures

List of Tables

List of Equations

Overview

This master's thesis focus on developing a methodology that combines statistical post-editing and interactive machine translation, in order to automatically increase the translation quality of a machine translation system, and reduce the effort that human translators need to make in order to correct the translations done by those systems. The structure of this thesis is as follows:

- *Chapter 1* introduces the main aspects of the disciplines in which this thesis is settled.
- In *Chapter 2*, a methodology for improving translation quality and reducing human post-editing effort is proposed.
- In *Chapter 3*, experiments are run in order to assess the proposed methodology.
- Finally, *Chapter 4* contains the conclusions of this work, and proposes possible improvements to carry out.

Chapter 1

Introduction

Language is an inherent characteristic of human beings, which enables them to communicate between them. However, language diversity supposes a great challenge to this communication.

Approaches to this challenge can be traced back to the 17th century, with the idea of creating a universal language. Unfortunately, proposals to this idea relied too much on philosophical concepts [1], and this language was never found.

More recently, at the end of World War II and the beginning of the Cold War, with the expertise in breaking enemy codes, the idea of using computers to translate between languages was born. However, although initial researches were perceived as a success and it was thought that Machine Translation (MT) would be a well-solved problem in a few years; the progress of MT evolved at a much slower pace than expected, and funding was severely cut [2].

Nowadays, there's an increasing need for translating from one language to another. Health services, political institutions, and education are just a few examples in which this necessity can be found. To cope with this need, research in MT has been increasing during the last sixty years. However, this discipline is still a growing baby, and the weight of this task still falls on the shoulders of human translators.

A good example to better illustrate this need is the European Union (EU). With 28 state members and 24 official languages, a multilingual organisation like the EU needs high quality translation and relies on professional linguists to keep it running smoothly [3]. According to [4], as of 2015, this supposes a staff of around 2500 people working full time on translating documents and on other language-related tasks.

1.1 Machine Translation

MT aims at translating from a source language into a target language by means of a computer. Although the use of mechanical dictionaries to overcome the barriers of language was first suggested in the 17th century [5], it wasn't until the 1950's that this research area arose. Its first proposals were based on information theory, taking the simplistic view that differences between languages laid in their vocabularies and the permitted word orders.

The main strategies that have been applied to MT can be classified as follows [6]:

- According to the **input type**: text or speech.
- According to the **type of application** which uses the translation: applications that translate the input into a database query; applications that produce an approximated translation of the input for its correction in a post-edition stage by the user; applications that interactively generate the output in collaboration with the user; or fully automated translation systems.
- According to the **translation technology**: rule-based systems or corpus based systems.

1.1.1 Rule-Based Systems

Rule-Based Machine Translation (RBMT) was one of the first approaches used in MT. RBMT systems are based on linguistic information extracted from dictionaries and grammars, and on a set of translations rules created by human translators. These rules are what determines how to translate from one language to another.

Usually, these systems are split in three stages: an **analysis step**, in which information is extracted from the source text; a **transference step**, in which the results of the analysis are transform into an abstract representation; and a **generation step**, in which the target text is generated.

1.1.2 Corpus-Based Systems

Corpus-based systems use translation examples (also known as corpora or parallel texts) from one language to another. These examples are used to infer the translation of the source text. Once a corpus-based system has been implemented, its software can be quickly adapted to be used in other domains and with other language pairs (as opposed to rule-based systems, which are specific for a given language pair).

Corpus-based systems can be classified as follows:

- **Example-Base Machine Translation (EBMT) systems**: these systems use a set of translations examples as its main knowledge base. Translation process is generated through two steps: first, a set of hypothesis similar to the source text are extracted from the corpus (*comparison*); and second, the hypothesis are recombined to generate the final translation of the source text (*recombination*).
- **Statistical Machine Translation (SMT) systems**: these systems base their translations on statistical models and other models from information theory. They require a great amount of parallel texts containing relevant information for the translation process. These texts are used to estimate the parameters of the models mentioned before, which are used to infer the translation of a new source text.
- **Other corpus-based systems**: there are other alternatives to implement corpus-based systems, such as the *finite state* approach, which applies the mathematical tools provided by the automata theory; or the *context-free grammar* approach, which applies context-free grammars to MT.

1.2 Statistical Machine Translation

SMT approaches the MT problem of generating translations, with an statistical point of view. Statistical models are involved in the translation process. These models estimate their parameters from the parallel texts of the available corpora. Therefore, as long as the required corpora is available, SMT systems can work with many different language pairs.

More formally, given a sentence $\mathbf{x}$ in a source language, the problem in MT is to find its corresponding translation $\mathbf{y}$ in a target language. This problem is formalized by SMT as follows [7]:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} Pr(\mathbf{y}|\mathbf{x}) \tag{1.1}$$

Applying Bayes' theorem, this can be seen as:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} Pr(\mathbf{y}) \cdot Pr(\mathbf{x}|\mathbf{y}) \tag{1.2}$$

This last equation is known as *fundamental equation of machine translation* [7]. The term $Pr(\mathbf{y})$ of this equation represents the well-formedness of $\mathbf{y}$, and is usually called the *language model* probability (*n-gram* models are usually

adopted [8]). Finally, the term $Pr(\mathbf{x}|\mathbf{y})$ corresponds to the *translation model*, which represents the relation between the source sentence and its translation.

In practice, all of these models (and possibly others) are often combined into a *log-linear model* for $Pr(\mathbf{y}|\mathbf{x})$ [9]:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \left\{ \sum_{n=1}^{N} \lambda_n \cdot log(f_n(\mathbf{y}, \mathbf{x})) \right\} \tag{1.3}$$

where $f_n(\mathbf{y}, \mathbf{x})$ can be any model that represents an important feature for the translation; N is the number of models (or features); and λ_n are the weights of the log-linear combination.

One of the most popular instantiations of log-linear models are those including phrase-based models [10, 11]. The basic idea behind phrase-based translation is the segmentation of the source sentence into phrases; the translation of those source phrases into target phrases; and the reordering of those translated phrases in order to compose the target sentence.

Therefore, there are three main computational challenges of SMT [7]:

1. Estimating the **language model** probability.
2. Estimating the **translation model** probability.
3. Finding an efficient and effective **global search** method.

These challenges come as modules in order to build a translation system based on Bayes' rule. A prepocess and a postprocess stages for the sentences are also included, in order to increase the performance of the system. Figure 1.1 shows an scheme of this process.

1.3 Domain Adaptation

Domain Adaptation (DA) is a topic of increasing interest over the last few years [12]. Often, there is a mismatch between the target domain of an SMT system (commonly known as the *in-domain*) and the domain from which training data are available (commonly known as the *out-of-domain*). This mismatch leads to a reduction of the translation quality [13]. Therefore, the aim of domain adaptation is to improve the performance of the systems trained with out-of-domain data.

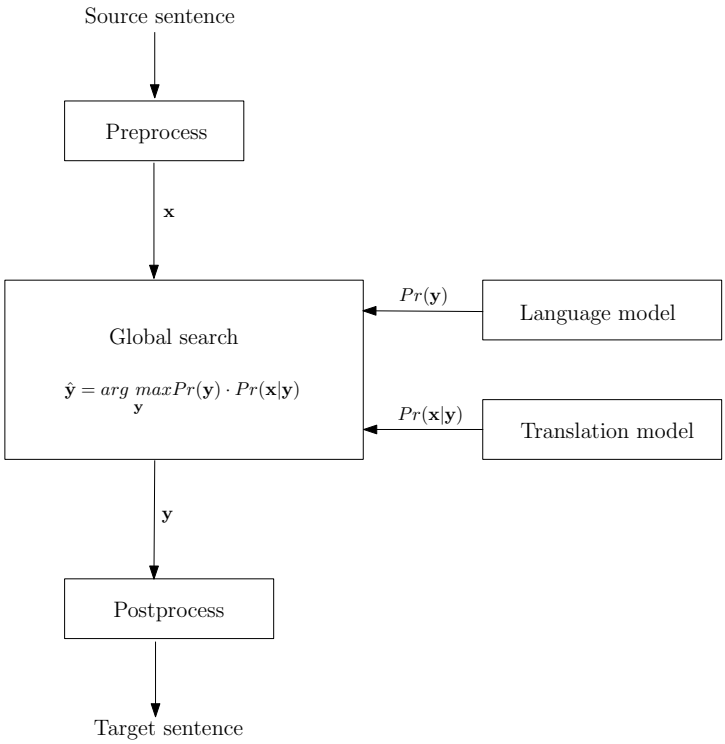

Figure 1.1: Architecture of the translation process based on Bayes' rule. Figure extracted from [11].

Different approaches have raised in this paradigm. Some of them reside in the selection of the most adequate sentences of the out-of-domain data with which to train the system [14, 15]; while others reside in modifying the probabilities of the existing model by using a mixture model that optimizes the coefficients to the adaptation domain. Among these last ones, mixtures models have been applied to *word alignment* [16]; *language modelling* [17, 18]; and *the translation model* [19, 20].

1.4 Post-Editing

The current state of the art in MT is far from being good enough, and translation quality is not as good as it should be. In many cases, a manual correction is needed in order to improve the quality of translations. This manual correction is known as *post-editing*.

More formally, post-editing is the process of improving a machine-generated translation with a minimum of manual labour [21]. It is possibly the oldest form of human-machine cooperation for translation [22].

1.5 Computer-Assisted Translation

Computer-Assisted Translation (CAT) seeks to provide human translators with as many tools as possible in order to facilitate their work through computer software.

CAT has evolved during the years and is a broad and imprecise term. The tools that can be included in a CAT environment range from very simple ones (like a spell checker), to more sophisticated ones (like a language search-engine). Some of the most notable tools are:

- **Translation memories**: Translation Memories (TM) are programs that store previously translated texts and their source texts in a database. These texts are divided into segments of different size and length. This is done so that, at the start of a new translation session, the program searches for segments already existing in the TM, and replaces them in the text to translate, reducing the length of the text the human translator has to translate.
- **Language search-engines**: language search-engines are similar to TM, but instead of storing the translated texts and dividing them in segments, they access an online repository that contains segments from lots of TMs.
- **Terminology management**: terminology management is a software that provides translators with means to search for a certain terminology appearing in the documents that are being translated.
- **Interactive machine translation**: Interactive Machine Translation (IMT) is a paradigm that combines SMT and human post-editing in an interactive process that obtains the final translation with the collaboration of human and machine. There are different approaches to IMT, being one of them *interactive translation prediction* (see Section 1.6).
- **Crowd-assisted translation**: crowd-assisted translation are online platforms in which a great number of translators (usually, bilingual people without a translation background) collaborate to translate documents.

1.6 Interactive Translation Prediction

Interactive Translation Prediction (ITP) was first introduced by Barrachina et al. [23] as a new methodology inside the CAT framework. This methodology proposes an alternative to the classical MT approach (MT plus manual correction), considering translation as an interactive process where human and computer collaborate to generate the final translation. Generally, this interactive process is a left-to-right process, although other types of interaction are possible.

Figure 1.2 shows an example of an ITP session. At this session, a source English sentence $\mathbf{x}$ = "The cough may last for 1-2 months or longer" is to be translated into a Spanish target sentence $\hat{\mathbf{y}}$. At the beginning, the system proposes a complete translation $\mathbf{s_h}$ = "La tos puede durar 1 ó 2 meses o más". Then, the user marks the prefix $\mathbf{p}$ = "La tos puede durar" as correct and types the next word w = "1-2". After that, the system suggests a new suffix $\mathbf{s_h}$ = "meses o más" that completes the validated prefix and the word the user has typed. This process continues, with a new prefix validation followed by new input from the user (if necessary), and so on; until the user considers the translation to be complete and satisfactory.

source (x): The cough may last for 1-2 months or longer
desired translation ($\hat{\mathbf{y}}$): La tos puede durar 1-2 meses o más tiempo

IT-0	$\mathbf{p}$	
	$\mathbf{s_h}$	La tos puede durar 1 ó 2 meses o más
IT-1	$\mathbf{p}$	La tos puede durar
	w	1-2
	$\mathbf{s_h}$	meses o más
IT-2	$\mathbf{p}$	La tos puede durar 1-2 meses o más
	w	tiempo
	$\mathbf{s_h}$	
END	$\mathbf{p}$	La tos puede durar 1-2 meses o más tiempo

Figure 1.2: ITP session to translate a sentence from English into Spanish. The desired translation is the translation the user wants to obtain. At IT-0, the system proposes a translation ($\mathbf{s_h}$). At IT-1, the user accepts the first four words ("La tos puede durar") by moving the mousse to the next position, and tells the system that the following word (w) should be "1-2". Then, the system suggests completing the sentence with "meses o más" (a new $\mathbf{s_h}$). Iteration 2 is similar to iteration 1. At the final iteration, the user accepts the current translation.

1.6.1 Statistical Framework

The crucial step of the ITP process is the production of the new suffix [23]. This suffix will be selected according to its probability, which should be maximized according to the available information. That is, the new suffix will be given by:

$$\hat{\mathbf{s}}_\mathbf{h} = \arg\max_{\mathbf{s_h}} Pr(\mathbf{s_h}|\mathbf{x}, \mathbf{p}) \tag{1.4}$$

Applying Bayes' theorem, this can be rewritten as:

$$\hat{\mathbf{s}}_\mathbf{h} = \arg\max_{\mathbf{s_h}} \frac{Pr(\mathbf{p}, \mathbf{s_h}|\mathbf{x})}{Pr(\mathbf{p}|\mathbf{x})} \tag{1.5}$$

Finally, taking into account that $Pr(\mathbf{p}|\mathbf{x})$ does not depend on $\mathbf{s_h}$, this can be seen as:

$$\hat{\mathbf{s}}_\mathbf{h} = \arg\max_{\mathbf{s_h}} Pr(\mathbf{p}, \mathbf{s_h}|\mathbf{x}) \tag{1.6}$$

This last equation is similar to Equation 1.1. The main difference is that, in this case, the search procedure is limited to those target sentences $\mathbf{y}$ whose prefix is equal to $\mathbf{p}$. Therefore, provided that the search procedures are adequately modified, we can use the same MT models [24].

1.6.2 Implementation Approach

Among the approaches for the implementation of the ITP methodology, common implementations rely on the *wordgraph* data structure [6, 23, 25].

A wordgraph is a weighted directed acyclic graph whose nodes represent a partial translation of a given sentence. Its edges are labelled with a word (or group of words) of the target sentence, and weighted according to the scores given by an SMT model (for more details, see [26]).

Figure 1.3 shows an example of a wordgraph generated for the translation of the source sentence *«You have a cold»*. This translation results in the target sentence *«tiene un resfriado»*. Note that the scores on the edges are not probabilities, since there isn't any normalization.

The main advantage of this implementation is that the wordgraph only needs to be generated at the beginning of the ITP process (of a given source sentence), and the ITP suffixes can be obtained by incrementally processing the wordgraph at each interaction. Thanks to this, the system is very efficient in terms of the time cost per interaction.

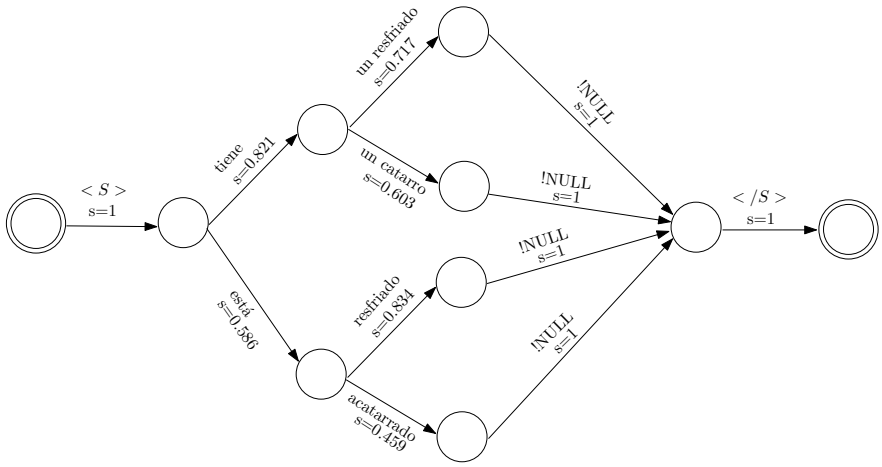

Figure 1.3: Example of a wordgraph generated during the translation of the source sentence *«You have a cold»*, which results in the target sentence *«tiene un resfriado»*.

A common problem with this approach rises when the user inserts a prefix that cannot be found in the wordgraph, since in such a situation the system is unable to find a path through the wordgraph and, therefore, cannot provide any suffix. To solve this problem, we have incorporated an stochastic error-correction to our implementation.

Chapter 2

Interactive Post-Editing

2.1 Introduction

The current state of the art in MT is far from being good enough. In many cases, a human post-editing process is necessary (see Section 1.4). In order to reduce human post-editing effort, research in CAT has increased during the years (see Section 1.5). Among this research, ITP is one of the most innovative directions (see Section 1.6).

On the other hand, *statistical post-editing* has proved to be a good technique to automatically improve the translation quality of a given MT system under certain circumstances (see Section 2.2).

In this chapter, we introduce a new methodology that combines both techniques. This methodology can be applied to any MT system in order to increase the quality of its translations, and to reduce the human post-editing effort needed to correct those translations.

2.2 Statistical Post-Editing

Statistical Post-Editing (SPE) is a method for automatically post-editing a MT system by training an SMT system that takes as an input the output of the MT system. Usually, this is accomplished by translating the data needed for training the SMT system with the original MT system (the one to post-edit), and then training the system with the output resulting from that translation. Figure 2.1 shows an illustration of this process.

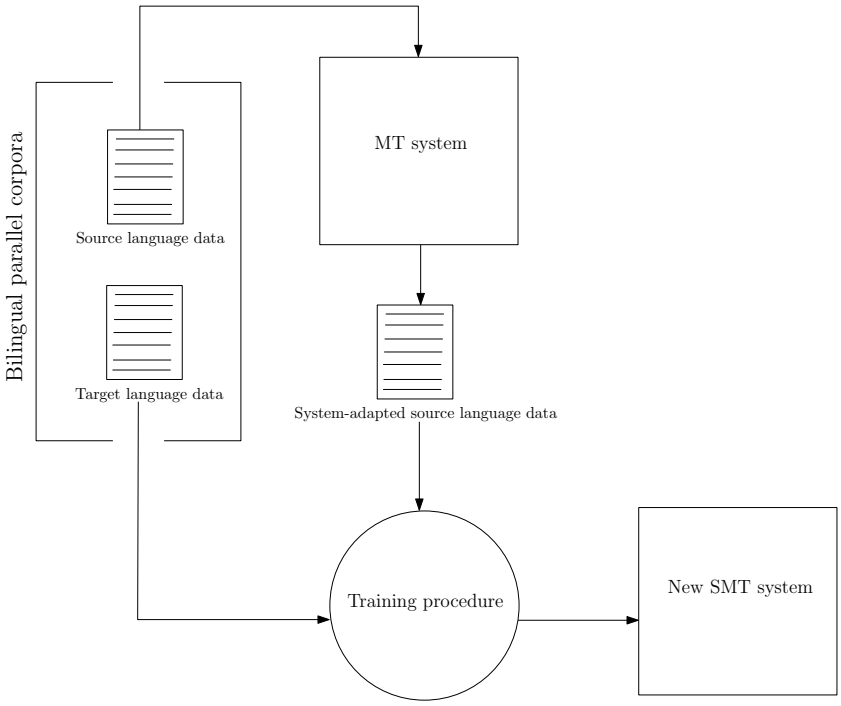

Figure 2.1: SPE SMT system training process. The source language data of the bilingual parallel corpora is first translated with the MT system to post-edit. Then, the resulting translation is used as the new source language data and, together with the target language, is used to train the new SMT system.

To the best of our knowledge, this method was first introduced by Simard et al. [27]. To the date, the main research areas regarding SPE have been:

- Post-editing a **RBMT** system.
- Using SPE as a **domain adaptation** technique.
- Post-editing an **SMT** system.

Among those areas, post-editing an SMT system has not had very satisfactory results [27, 28]. Most likely, this is due to the original system already extracting all the information contained in the data, not leaving any new information for the SPE system to extract (and hence, not leaving room for the SPE system to improve translation quality).

Therefore, in this thesis we shall focus only in the first two research areas (post-editing a RBMT system, and using SPE as a domain adaptation technique).

2.2.1 Rule-Based Machine Translation Scenario

In this scenario, SPE has been used to automatically post-edit a RBMT system in order to improve its translation quality. For this purpose, in-domain data is translated with the RBMT system, and the resulting translation is used for training a new SMT system. From this point forward, every text to translate will be first translated with the RBMT system, and then with the new SPE system. Figure 2.2 shows an illustration of this process.

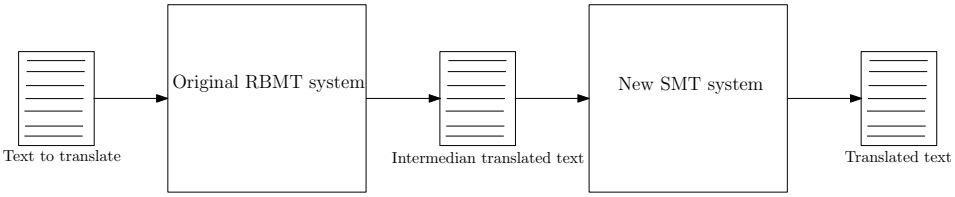

Figure 2.2: SPE process of a RBMT system. The text to translate is first translated with the original RBMT system. Then, the resulting translation is translated with the new SMT system, obtaining the final translation.

SPE has proved to be a reliable method to apply in this scenario [27, 29–31]. In most cases, the combination of RBMT and SPE yields better results than training a new SMT system with the in-domain data.

2.2.2 Domain Adaptation Scenario

In this scenario, a MT system has been trained with domain specific data, and the text to translate belongs to another different domain. The goal of SPE is to adapt the system into the domain of the text to translate. In order to do this, in-domain data is translated with the out-of-domain system (the original system), and the resulting translation is used for training a new in-domain SMT system. From this point forward, every text to translate (belonging to this new domain) will be first translated with the out-of-domain system, and then with the new SMT system. Figure 2.3 shows an illustration of this process.

SPE has proved to be a reliable method to apply in this scenario [27, 32]. In most cases, applying SPE to the domain adaptation scenario yields better results than training a new SMT system with the in-domain data.

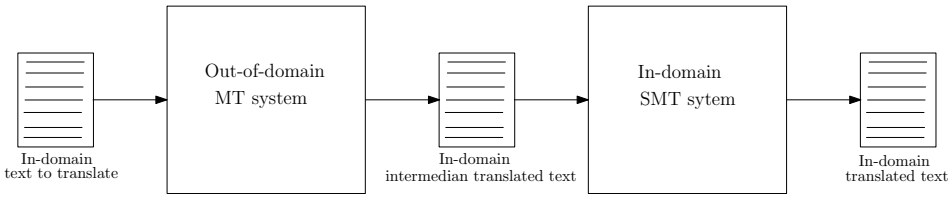

Figure 2.3: SPE as a domain adaptation technique process. The in-domain text to translate is first translated with the out-of-domain system. Then, the resulting translation is translated with the in-domain SMT system, obtaining the final translation.

2.3 Interactive Post-Editing

In this section, we introduce our new methodology that combines SPE and ITP techniques. Prior to the use of this methodology, it is necessary to train a new SMT system that takes as input the output of the original MT system. This is done by translating some corpora with the original MT system, and using the resulting translation as the new corpora (which is used to train the new SMT system).

The steps this methodology consists of are as follows:

- **SPE step**: in this step, we translate the output of the original MT system with the new SMT system.
- **ITP step**: in this step, we apply the ITP methodology to the SMT system used in the previous step.

This way, we are trying to benefit from both techniques to both increase translation quality and reduce human post-editing effort.

Figure 2.4 illustrates the full process of translating a given source sentence $\mathbf{x}$ with our methodology. First of all, this sentence is translated with the original MT system, which generates the new source sentence $\mathbf{z}$. Then, $\mathbf{z}$ is fed into our system which, interacting with the user, generates the final target sentence $\mathbf{y}$. In this interaction, the user receives the original source sentence $\mathbf{x}$ and the target sentence $\mathbf{y}$, and generates a new target sentence $\mathbf{y}'$, which is fed back into our system to produce a new target sentence $\mathbf{y}$. This interaction will have as many iterations as needed, until the user considers the translation to be complete and satisfactory.

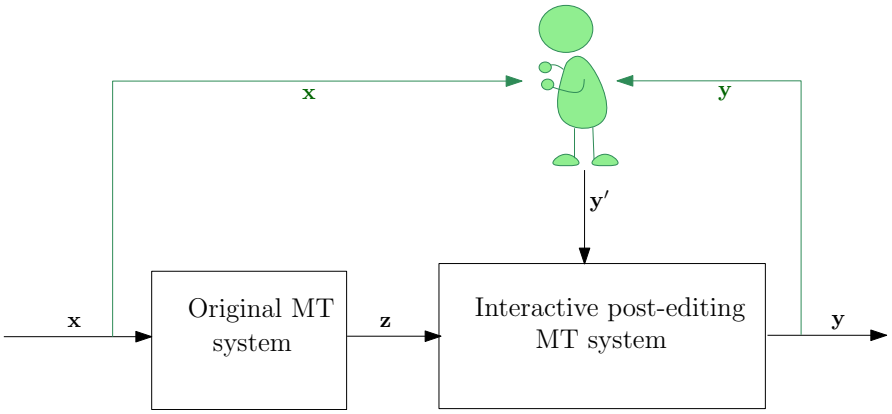

Figure 2.4: Interactive post-editing process. A source sentence $\mathbf{x}$ is fed into the system. This sentence is first translated with the original MT system, generating the new source sentence $\mathbf{z}$. Then, the new source sentence is fed into the interactive post-editing system which, with the interaction of the user, generates the final target sentence $\mathbf{y}$.

2.3.1 Statistical Framework

Our methodology takes as an input the output of the original MT system. More formally, given a sentence $\mathbf{x}$ in a source language and its corresponding translation $\mathbf{y}$ obtained with the original MT system, our system takes $\mathbf{y}$ as the new source sentence $\mathbf{z}$.

As in ITP, the crucial step of our methodology is the production of the new suffix, which will be given by:

$$\hat{\mathbf{s}}_\mathbf{h} = \arg\max_{\mathbf{s_h}} Pr(\mathbf{s_h}|\mathbf{x}, \mathbf{z}, \mathbf{p}) \tag{2.1}$$

Applying Bayes' theorem this can be rewritten as:

$$\hat{\mathbf{s}}_\mathbf{h} = \arg\max_{\mathbf{s_h}} \frac{Pr(\mathbf{p}, \mathbf{s_h}|\mathbf{x}, \mathbf{z})}{Pr(\mathbf{p}|\mathbf{x}, \mathbf{z})} \tag{2.2}$$

Finally, taking into account that $Pr(\mathbf{p}|\mathbf{x}, \mathbf{z})$ does not depend on $\mathbf{s_h}$, this can be seen as:

$$\hat{\mathbf{s}}_\mathbf{h} = \arg\max_{\mathbf{s_h}} Pr(\mathbf{p}, \mathbf{s_h}|\mathbf{x}, \mathbf{z}) \tag{2.3}$$

This last equation corresponds to the probability of the production of the new suffix in a general way. However, for simplicity, in this thesis we have removed the source sentence $\mathbf{x}$ from the suffix generation. Therefore, the production of the new suffix will be given by:

$$\hat{\mathbf{s}}_\mathbf{h} = \arg\max_{\mathbf{s_h}} Pr(\mathbf{p}, \mathbf{s_h}|\mathbf{z}) \tag{2.4}$$

This equation is closely related to Equation 1.6. Therefore, as with ITP, provided that the search procedures are adequately modified, we can use the same MT models [24].

Chapter 3

Experimental Framework

In this chapter, we test the proposed methodology on different scenarios (to connect with the output of a RBMT system, and as method to adapt an SMT system from one domain to another) using different corpora.

First, we introduce the software and corpora used in the experimentation; the metrics used to assess the results; and how the experiments are organized. Finally, we present and discuss the obtained results.

3.1 Software

In this section, we briefly describe the main software used in the experimental framework.

Apertium

Funded by the Spanish Government, Apertium [33] is an open-source toolkit for RBMT. It was originally design for the translation between similar language pairs, but it was expanded over the years to be able to deal with more divergent language pairs.

Moses

Moses [34] is an open-source toolkit which implements state of the art SMT techniques. Taking as an input a group of bilingual sentences, the toolkit trains a translation model for the given language pair. Moreover, it implements a decoder algorithm to obtain the most probable translation of a certain text in an efficient way.

SRILM

SRILM [35] is a toolkit for the efficient estimation and handle of large scale language models. It has became a classical tool for language modelling, applied not only in MT but in other research fields such as *speech recognition*, *handwriting text recognition* or *tagging and segmentation*.

MGIZA++

MGIZA++ [36] is a multi-threaded implementation of the GIZA++ [37] word alignment toolkit, which computes the word alignment of a bilingual sentence-aligned corpus by using language-independent statistical methods.

3.2 Evaluation Metrics

In this section, we describe all the evaluation metrics used to assess the quality of the results.

BLEU

BiLingual Evaluation Understudy (BLEU) [38], is a method for the automatic evaluation of machine translation. It computes the geometric average of the modified n-gram precision (p_n), multiplied by a factor BP that penalizes short sentences. Its equation can be seen at Equation 3.1.

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} \frac{log\ p_n}{N}\right) \tag{3.1}$$

WER

Word Error Rate (WER) [37], computes the minimum number of substitution (S), deletion (D) and insertion (I) operations needed to convert the word strings of the output sentences of the translation system, into the word strings of the reference sentences. This compute is normalized by the number of words in the reference sentences. Its equation can be seen at Equation 3.2.

$$WER = \frac{S + D + I}{N} \tag{3.2}$$

WSR

Word Stroke Ratio (WSR) [39], measures the number of word strokes a user would need to perform in order to obtain the translation they desires, normalized by the total number of words in that translation. Its equation can be seen at Equation 3.3.

$$WSR = \frac{number\ word\ strokes}{number\ reference\ words} \tag{3.3}$$

E-R

Estimated Effort-Reduction (E-R) [40], is the relative difference between WER and WSR. This metric gives an estimation of the reduction in human effort, in terms of words to be corrected. It's computation can be seen at Equation 3.4.

$$E\text{-}R = \frac{WER - WSR}{WER} \tag{3.4}$$

3.3 Corpora

In this section, we introduce the corpora used in the experimental framework. A total of three different corpora have been used in this thesis: EMEA corpus, EU corpus, and europarl corpus.

3.3.1 EMEA

EMEA corpus is formed by documents from the *European Medicines Agency* [41], and is publicly available on the Internet. This corpus has been used to train some of the in-domain SMT systems used in the experiments (see Section 3.4). The main features of this corpus are shown in Table 3.1.

To obtain these partitions, we have first removed every repeated line from both languages of the corpus. Then, we have shuffled the lines, making sure to maintain the alignment between the language pair. After that, we have lower-cased and tokenized each sentence by means of the scripts included with Moses. Finally, we have divided the corpus, taking the last 5000 sentences of each language as the test set; the next 10000 as the development set; and the rest as the training set. In the case of the training and development sets, each sentence containing more than 100 words has been removed[1] (again, using the scripts included with Moses).

[1] This is done to avoid conflicts with the MERT procedure (see Section 3.4).

Table 3.1: EMEA corpus statistics for the en-es language pair.

		English	Spanish
Training	Sentences	292K	
	Running words	5M	5.6M
	Vocabulary	67K	81K
Development	Sentences	9.8K	
	Running words	171K	192K
	Perplexity (3-grams)	30.43	30.76
Test	Sentences	5K	
	Running words	90K	101K
	Perplexity (3-grams)	31.72	31.61

3.3.2 EU

EU corpus was extracted from the *Bulletin of the European Union* [42], which exists in all official languages of the European Union, and is publicly available on the Internet. This corpus has been used to train some of the in-domain SMT systems used in the experiments (see Section 3.4). The main features of this corpus are shown in Table 3.2.

Table 3.2: EU corpus statistics for the en-es language pair.

		English	Spanish
Training	Sentences	212.4K	
	Running words	5.2M	5.8M
	Vocabulary	40.3K	53.8K
Development	Sentences	2K	
	Running words	49.8K	55.8K
	Perplexity (3-grams)	37.42	31.98
Test	Sentences	800	
	Running words	20K	22.8K
	Perplexity (3-grams)	43.82	36.46

In a similar way as we have done with EMEA corpus, we have shuffled, lower-cased and tokenized each sentence of the corpus. Then, we have removed all sentences containing more than 100 words, and taken the last 2000 sentences as the development set, and the rest as training set. In this case, a test set used in previous works was already defined.

3.3.3 Europarl

Europarl corpus was extracted from the *Proceedings of the European Parliament* [43], which are written in all official languages of the European Union, and is

publicly available on the Internet. This corpus has been used to train the out-of-domain SMT systems used in the experiments (see Section 3.4). The main features of this corpus are shown in Table 3.3.

Table 3.3: Europarl corpus statistics for the en-es language pair.

		English	Spanish
Training	Sentences	1.9M	
	Running words	53.1M	55.5M
	Vocabulary	130K	190K
Development	Sentences	10K	
	Running words	268K	279K
	Perplexity (3-grams)	58.07	60.01

Once more, similarly as we have done with the previous corpora, we have shuffled, lower-cased and tokenized each sentence of the corpus. Then, we have removed all sentences containing more than 100 words, and taken the last 2000 sentences as the development set, and the rest as training set. In this case, since we are using the corpus as an out-of-domain, we haven't defined any test set.

3.4 Experimental Set-Up

Since SPE works best when post-editing a RBMT system and in the domain adaptation scenario (see Section 2.2), we wanted to assess how our methodology worked in each of those situations.

As baseline, we have considered an SMT system trained with each one of the in-domain corpora (see Section 3.3). For training this system, and all the SMT systems used in this thesis, we have used the Moses toolkit (see Section 3.1). Phrase pairs were extracted from symmetrised word alignments generated by MGIZA++. The weights of the log-linear model were optimized by means of the Minimum Error Rate Training (MERT) procedure [44]. Finally, an n-gram word-based language model was estimated on the target side of the parallel corpora using the improved KneserNey smoothing [45], by means of the SRILM toolkit. The suitable level of n-grams to use was one of the parameters to estimate (see Section 3.5).

For the scenario in which the MT system to improve is a RBMT system, we have used Apertium as our system. With the output of this system (the input being the in-domain data), we have trained the SMT system of the SPE step. Finally, using the Moses decoder, we have obtained the resulting wordgraph of translating the test dataset with the SMT system of the SPE step, and fed it into a software that implements the ITP functionality used in this work. This functionality allowed us to simulate real users by using the references of the test dataset.

Lastly, regarding the scenario in which the domains of the MT system and the text to translate differ, we have trained an SMT system with the out-of-domain data (see Section 3.3). Then, like in the previous situation, we have trained the SMT system of the SPE step with the output of this system (the input being the in-domain data). After that, we have obtained the resulting wordgraph of translating the test dataset with the SMT system of the SPE step, and fed it into the software that implements the ITP functionality (once more, using Moses decoder to generate the wordgraph).

This procedure has been done several times, each time the in-domain data belonging to a different corpus (see Section 3.3).

Finally, since our aim was to both improve translation quality and reduce human post-editing effort, we have divided our experiments in two different categories:

- **Translation quality**: in this category, we apply only the SPE step of our methodology.
- **Human post-editing effort**: in this category, we focus in the ITP step of our methodology, using for the SPE step the SMT system that performed best in the previous category.

3.5 Results

3.5.1 Translation Quality

In this category, we narrow our experiments to the SPE step of our methodology.

Rule-Based Machine Translation

The first of our experiments consisted in applying the SPE step of our methodology to an existing RBMT system (see Section 3.4) using both in-domain corpora (see Section 3.3).

EMEA corpus

Table 3.4 shows the results obtained on this experiment with EMEA corpus. It must be noted that, since we are using the vanilla version of Apertium, the translation quality has a great disadvantage in comparison to an SMT system.

Besides applying the SPE step of our methodology, we have also tried different n-gram levels for each of the language models used for training the systems. In this case, the baseline obtains its best result with the use of 3-grams, and the system of the SPE step does it with 6-grams.

Table 3.4: System performance on EMEA test (English to Spanish) for the RBMT translation quality experiment. The baseline is the system tagged as *Moses*.

System	**en-es**	
	n-grams	**BLEU**
Moses	2	56.6
	3	58.2
	4	54.9
	5	55.6
Apertium	–	17.6
Apertium + SPE	2	50.6
	3	58.0
	4	58.7
	5	59.3
	6	**59.5**
	7	57.8

As it was mentioned before, the Apertium system performs far worse than the baseline (around 40 points of difference on BLEU). However, when applying our methodology, the SPE step improves greatly the translation quality (nearly 42 points of improvement on BLEU). Furthermore, the translation quality is better than the baseline (around 1 point of improvement on BLEU).

EU corpus

Table 3.5 shows the results obtained on this experiment with EU corpus. Once more, it must be noted that, since we are using the vanilla version of Apertium, the translation quality has a great disadvantage in comparison to an SMT system.

As with EMEA corpus, besides applying the SPE step of our methodology, we have also tried different n-gram levels for each of the language models used for training the systems. In this case, the baseline obtains its best result with the use of 5-grams, and the system of the SPE step does it with 6-grams.

Once more, we can see that the Apertium system performs far worse than the baseline (around 34 points of difference on BLEU). However, when applying our methodology, the SPE step improves greatly the translation quality (around 33 points of improvement on BLEU). Despite this improvement, in this case the translation quality is not better than the baseline (around 0.6 points of difference on BLEU).

Table 3.5: System performance on EU test (English to Spanish) for the RBMT translation quality experiment. The baseline is the system tagged as *Moses*.

System	**en-es**	
	n-grams	**BLEU**
Moses	2	45.4
	3	47.4
	4	48.5
	5	**48.6**
	6	43.2
Apertium	–	14.8
Apertium + SPE	2	44.5
	3	46.4
	4	47.2
	5	47.6
	6	48.0
	7	45.0

Domain Adaptation

The next experiment we run consisted in applying the SPE step of our methodology to an SMT system trained with out-of-domain data (see Section 3.4), using both in-domain corpus (see Section 3.3). Additionally, we have also tried different n-gram levels for each of the language models used for training the systems.

EMEA corpus

For EMEA corpus, the baseline obtains its best result with the use of 3-grams, the out-of-domain system does it with 5-grams, and the system of the SPE step does it with 6-grams. Table 3.6 shows the results obtained with this corpus.

As expected, due to the difference between domains, the out-of-domain system performs far worse than the baseline (around 35 points of difference on BLEU). However, when applying our methodology, the SPE step improves greatly the translation quality (around 33 points of improvement on BLEU). Nonetheless, despite this improvement, the translation quality is not better than the baseline (around 1.5 points of difference on BLEU).

Table 3.6: System performance on EMEA test (English to Spanish) for the DA translation quality experiment. The baseline is the system tagged as *Moses (EMEA)*.

System	**en-es**	
	n-grams	**BLEU**
Moses (EMEA)	2	56.6
	3	**58.2**
	4	54.9
	5	55.6
Moses (europarl)	2	21.6
	3	22.3
	4	23.2
	5	23.9
	6	23.4
Moses (europarl) + SPE (EMEA)	2	53.8
	3	55.3
	4	56.1
	5	56.5
	6	56.6
	7	54.2

EU corpus

Table 3.7 shows the results obtained on this experiment with EU corpus. As with EMEA corpus, besides applying the SPE step of our methodology, we have also tried different n-gram levels for each of the language models used for training the systems. In this case, the baseline obtains its best result with the use of 5-grams, the out-of-domain system does it with 4-grams, and the system of the SPE step does it with 6-grams.

In this case, corpora domains are closer (the out-of-domain belonging to the proceedings of the European Parliament, and the in-domain belonging to the bulletins of the European Union). Nonetheless, the out-of-domain system still performs far worse than the baseline (around 14 points of difference on BLEU). However, when applying our methodology, the SPE step improves greatly the translation quality (around 12.5 points of improvement on BLEU). Despite this improvement, translation quality is not better than the baseline (around 1.5 points of difference on BLEU).

Table 3.7: System performance on EU test (English to Spanish) for the DA translation quality experiment. The baseline is the system tagged as *Moses (EU)*.

System	**en-es**	
	n-grams	**BLEU**
Moses (EU)	2	45.4
	3	47.4
	4	48.5
	5	**48.6**
	6	43.2
Moses (europarl)	2	31.7
	3	33.4
	4	34.2
	5	34.1
	6	34.0
Moses (europarl) + SPE (EU)	2	43.8
	3	45.1
	4	46.1
	5	46.7
	6	46.9
	7	44.0

Discussion of the results

Results show that, in all cases, our methodology increments significantly the translation quality of the original MT system (from 12 to 42 points of improvement on BLEU). However, the final quality is not always better than the one obtained by the baseline system (an SMT system trained with the in-domain corpus).

For the experiment in which the original MT system was a RBMT system and the in-domain corpus was EMEA corpus, our methodology increased translation quality from 17.6 to 59.5 points of BLEU. In this case, the final quality was better than the baseline (whose translation quality was of 58.2 points of BLEU).

When the original MT system was a RBMT system and the in-domain corpus was EU corpus, our methodology increased translation quality from 14.8 to 48.0 points of BLEU. However, the translation quality of the baseline was of 48.6 points of BLEU.

For the experiment in which the original MT system was an SMT system trained with the out-of-domain corpus and the in-domain corpus was EMEA corpus, our methodology increased translation quality from 23.9 to 56.6 points of BLEU. However, the translation quality of the baseline was of 58.2 points of BLEU.

Finally, when the original MT system was an SMT system trained with the out-of-domain corpus and the in-domain corpus was EU corpus, our methodology increased translation quality from 34.2 to 46.9 points of BLEU. However, the translation quality of the baseline was of 48.6 points of BLEU.

3.5.2 Human Post-Editing Effort

The experiments in this category focus on the ITP step of our methodology. In order to be able to estimate the reduction in human post-editing effort, we represent translation quality in terms of WER. This way, we can make a comparison with the WSR used on the ITP step. Moreover, WER and WSR are the classic metrics used in ITP.

Rule Based Machine Translation

This experiment consisted in applying the ITP step of our methodology to the systems that gave the best results for the previous RBMT experiment (see Section 3.4), using both in-domain corpora (see Section 3.3).

EMEA corpus

Table 3.8 shows the results obtained on this experiment with EMEA corpus. As it happened when computing the results of the SPE step in terms of BLEU, our systems improves significantly the translation quality of Apertium (around 31 points of improvement on WER), with the resulting translation quality being better than the baseline (around 2 points of improvement on WER).

Table 3.8: System performance on EMEA test (English to Spanish) for the RBMT human post-editing effort experiment. The baseline is the system tagged as *Moses*.

System	**en-es**			
	n-grams	**WER**	**WSR**	**E-R**
Moses	3	42.3	36.4	14.1
Apertium	–	71.6	–[2]	–
Apertium + SPE	2	47.8	45.8	4.1
	3	41.5	37.2	10.3
	4	40.8	36.3	11.0
	5	40.3	36.5	9.5
	6	**40.3**	**36.2**	**10.2**
	7	41.6	37.5	10.0

[2]This measure cannot be computed since it's not possible to compute the word graph with apertium.

When applying the ITP step, we obtain an estimated effort-reduction of around 10 percent. Moreover, this final effort is slightly better than the baseline (with an estimated effort-reduction of 0.6 percent).

EU corpus

Table 3.9 shows the results obtained for this experiment with EU corpus. As it happened when computing the results of the SPE step in terms of BLEU, our systems improves significantly the translation quality of Apertium (around 26 points of improvement on WER), but this quality is not better than the baseline.

Table 3.9: System performance on EU test (English to Spanish) for the RBMT human post-editing effort experiment. The baseline is the system tagged as *Moses*.

System	**en-es**			
	n-grams	**WER**	**WSR**	**E-R**
Moses	5	**44.6**	**44.0**	**1.3**
Apertium	–	70.5	–[3]	–
Apertium + SPE	2	47.7	48.4	-1.7
	3	46.0	46.0	0.1
	4	45.6	44.8	1.8
	5	45.1	44.3	1.8
	6	44.7	44.5	0.3
	7	47.6	47.0	1.3

In this case, when applying the ITP step, the estimated effort-reduction is not significantly better (0.3 percent) for the system whose translation quality was the best. However, the system trained with 5-grams obtains better results, with an estimated effort-reduction of near 2 percent. Nonetheless, human post-editing effort is not better than the baseline (WSR is greater).

Domain Adaptation

This experiment consisted in applying the ITP step of our methodology to the systems that gave the best results for the previous domain adaptation experiment (see Section 3.4), using both in-domain corpora (see Section 3.3).

EMEA corpus

Table 3.10 shows the results obtained for this experiment with EMEA corpus. As it happened when computing the results of the SPE step in terms of BLEU, our systems improves significantly the translation quality of the out-of-domain

[3]This measure cannot be computed since it's not possible to compute the word graph with apertium.

system (around 25 points of improvement on WER), but this quality is not better than the baseline.

Table 3.10: System performance on EMEA test (English to Spanish) for the DA human post-editing effort experiment. The baseline is the system tagged as *Moses (EMEA)*.

System	**en-es**			
	n-grams	**WER**	**WSR**	**E-R**
Moses (EMEA)	3	**42.3**	**36.4**	**14.1**
Moses (europarl)	5	67.0	71.9	-7.3
Moses (europarl) + SPE (EMEA)	2	45.1	42.2	6.5
	3	43.8	40.0	8.7
	4	43.1	39.2	9.0
	5	42.8	39.2	8.5
	6	42.6	39.0	8.5
	7	44.9	41.8	6.9

When applying the ITP step, we obtain an estimated effort-reduction of around 8.5 percent. However, this effort is not better than the baseline (WSR is greater).

It must be noted that, in this case, applying the ITP methodology to the out-of-domain system yields a negative effort-reduction. This is probably due to the difference between domains (the system coming from a parliamentary domain and the text to translate coming from a medical domain). Since the difference is great, the new predictions of the ITP process aren't very accurate, which results in a greater number of interactions needed in order for the user to achieve their desired translation.

EU corpus

Table 3.11 shows the results obtained for this experiment with EU corpus. As it happened when computing the results of the SPE step in terms of BLEU, our systems improves significantly the translation quality of the out-of-domain system (around 8 points of improvement on WER), but this quality is not better than the baseline.

When applying the ITP step, we obtain an estimated effort-reduction of around 0.2 percent. However, once more, this effort is not better than the baseline (WSR is greater).

It must be noted that, in some cases, the ITP methodology is yielding a negative effort-reduction. This could be due to all the possible translations contained in the wordgraph being too alike, which makes each new prediction very similar to the previous one and, therefore, more interactions are needed in order to obtain the user's desired translation.

Table 3.11: System performance on EU test (English to Spanish) for the DA human post-editing effort experiment. The baseline is the system tagged as *Moses (EU)*.

System	en-es			
	n-grams	**WER**	**WSR**	**E-R**
Moses (EU)	5	**44.6**	**44.0**	**1.3**
Moses (europarl)	4	54.6	54.7	-0.2
Moses (europarl) + SPE (EU)	2	48.9	50.6	-3.5
	3	48.0	47.5	1.0
	4	46.7	46.3	0.7
	5	46.1	46.4	-0.5
	6	46.2	46.1	0.2
	7	48.9	49.8	-1.8

Discussion of the results

Results show that, in all cases, our methodology reduces the estimated human post-editing effort with respect of the SMT system of the SPE step (the best SMT system obtained in the experiments of the previous category). However, the final human post-editing effort (estimated by means of WSR) is not always better than the one obtained by the baseline system (an SMT system trained with the in-domain corpus).

For the experiment in which the original MT system was a RBMT system and the in-domain corpus was EMEA corpus, our methodology obtained an estimated effort-reduction of 10.2 percent (from a WER of 40.3 points into a WSR of 36.2 points). In this case, the final effort was slightly better than the one estimated for the baseline (36.4 points of WSR).

When the original MT system was a RBMT system and the in-domain corpus was EU corpus, our methodology obtained an estimated effort-reduction of 0.3 percent (from a WER of 44.7 points into a WSR of 44.5 points). However, the estimated effort of the baseline was of 44.0 points of WSR.

For the experiment in which the original MT system was an SMT system trained with the out-of-domain corpus and the in-domain corpus was EMEA corpus, our methodology obtained an estimated effort-reduction of 8.5 percent (from a WER of 42.6 points into a WSR of 39.0 points). However, the estimated effort of the baseline was of 36.4 points of WSR.

Finally, when the original MT system was an SMT system trained with the out-of-domain corpus and the in-domain corpus was EU corpus, our methodology obtained an estimated effort-reduction of 0.2 percent (from a WER of 46.2 points into a WSR of 46.1 points). However, the estimated effort of the baseline was of 44.0 points of WSR.

Chapter 4

Conclusions

4.1 Conclusions

In this thesis, we have proposed a new methodology that combines SPE and ITP in order to increase the translation quality of an existing MT system, and reduce the effort a human agent needs to make in order to correct the translations of that system.

This methodology has been tested by means of different MT systems and with the use of different corpora, obtaining very encouraging results. These results show that, in all cases, our methodology improved the translation quality of the original system and reduced the human post-editing effort.

However, these results aren't better than those obtained with the system proposed as a baseline (with an exception). This means that, although we have succeed in improving the original system, we could have trained a new SMT system with the in-domain corpora, applied the ITP methodology to it, and we would have obtained better results.

An exception to this is the experiment in which we applied our methodology to a RBMT system using EMEA corpus. In this experiment, the results obtained by our methodology were better than the baseline. Moreover, in those cases in which results were worse than the baseline, the difference was of no more than two points.

Overall, results show that we are in the right path to develop a methodology that improves the translation quality and human post-editing effort of an existing MT system (yielding better results that using a new SMT system), but we still have some more work to do.

4.2 Future Work

Among the future work, it would be interesting to experiment with more MT systems and more diverse corpora. Specially, it would be interesting to try a tuned version of a RBMT system, in stead of just using a vanilla version.

Another thing to do is to incorporate a new module into our methodology that enables the possibility to have more than one MT system. This way, in parallel with the SPE step proposed in this work, we could train an SMT system (with the in-domain corpora) that also translates the source sentence. Then, we would incorporate this new module after the SPE step. This module would receive as input the two translations of the source sentence (the one obtained in the SPE step, and the one obtained by the new SPE system), select the best of them, and feed it to the ITP step. This would combine the benefits of the methodology proposed in this thesis with the strength of an SMT system trained with the in-domain corpora.

Finally, in this thesis we have used the whole in-domain corpus (both for training the baseline and for training the new systems). However, it would be interesting to test how would our methodology work using only a part of the corpus (starting with a small part, and incrementally increasing it). This would reflect those cases in which there is a small quantity of in-domain data, and would enable the opportunity of analysing the relation between the computational cost in improving the system (by either using our methodology or training a new system with the in-domain data) and the improvement in quality and human post-editing effort.

Bibliography

[1] R. Descartes. *Descartes to Mersenne, 20 November 1629*. Descartes: Philosophical Letters. Oxford Clarendon Press, 1970. Translated and edited by Anthony Kenny.

[2] Y. Bar-Hillel. *The Present Status of Automatic Translation of Languages*, volume 1 of *Advances in Computers*. Academic Press, 1960.

[3] Translation and the european union. `http://ec.europa.eu/dgs/translation/translating/index_en.htm`, 2014.

[4] The european commission's in-house translation service. `http://ec.europa.eu/dgs/translation/whoweare/index_en.htm`, 2015.

[5] W. J. Hutchins. *Machine Translation: A Brief History*. Concise history of the language sciences: from Sumerians to the cognitivists. Pergamon Press, 1995.

[6] D. Ortiz-Martínez. *Advances in Fully-Automatic and Interactive Phrase-Based Statistical Machine Translation*. PhD thesis, Universidad Politécnica de Valencia, 2011. Advisors: Ismael García Varea and Francisco Casacuberta.

[7] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[8] F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA, USA, 1997.

[9] F. J. Och and H. Ney. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 295–302, 2002.

[10] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proceedings of the Human Language Technology and North American*

Association for Computational Linguistics Conference, volume 1, pages 48–54, 2003.

[11] R. Zens, F. J. Och, and H. Ney. Phrase-based statistical machine translation. In *Advances in artificial intelligence. 25. Annual German Conference*, volume 2479 of *LNCS*, pages 18–32, 2002.

[12] K. Shah, L. Barrault, and H. Schwenk. Translation Model Adaptation by Resampling. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 392–399. Association for Computational Linguistics, 2010.

[13] C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158. Association for Computational Linguistics, 2007.

[14] A. Sethy, P. Georgiou, and S. Narayanan. Selecting relevant text subsets from web-data for building topic specific language models. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 145–148. Association for Computational Linguistics, 2006.

[15] R. C. Moore and W. Lewis. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224. Association for Computational Linguistics, 2010.

[16] J. Civera and A. Juan. Domain adaptation in statistical machine translation with mixture modelling. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 177–180. Association for Computational Linguistics, 2007.

[17] B. Zhao, M. Eck, and S. Vogel. Language model adaptation for statistical machine translation with structured query models. In *Proceedings of the 20th International Conference on Computational Linguistics*. Association for Computational Linguistics, 2004.

[18] P. Koehn and J. Schroeder. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227. Association for Computational Linguistics, 2007.

[19] G. Foster and R. Kuhn. Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 128–135. Association for Computational Linguistics, 2007.

[20] B. Chen, M. Zhang, A. Aw, and H. Li. Exploiting n-best hypotheses for smt self-enhancement. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies:*

Short Papers, pages 157–160. Association for Computational Linguistics, 2008.

[21] TAUS. Postediting in practice. TAUS report `https://www.taus.net/reports/postediting-in-practice`, 2010.

[22] S. O'brien, L. Winther Balling, M. Carl, M. Simard, and L. Specia. *Post-editing of machine translation: Processes and Applications*. Cambridge Scholars Publishing, 2014.

[23] S. Barrachina, O. Bender, F. Casacuberta, J. Civera, E. Cubel, S. Khadivi, A. L. Lagarda, H. Ney, J. Tomás, and E. Vidal. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28, 2009.

[24] F. J. Och, R. Zens, and H. Ney. Efficient search for interactive statistical machine translation. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics*, volume 1, pages 387–393, 2003.

[25] J. González-Rubio, D. Ortiz-Martínez, J. M. Benedí, and F. Casacuberta. Interactive machine translation using hierarchical translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 244–254, 2013.

[26] N. Ueffing, F. J. Och, and H. Ney. Generation of word graphs in statistical machine translation. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 156–163, 2002.

[27] M. Simard, C. Goutte, and P. Isabelle. Statistical phrase-based post-editing. In *Proceedings of NAACL*, 2007.

[28] H. Béchara, R. Rubino, Y. He, Y. Ma, and J. Genabith. An evaluation of statistical post-editing systems applied to rbmt and smt systems. In *Proceedings of COLING 2012*, pages 215–230. The COLING 2012 Organizing Committee, 2012.

[29] P. Isabelle, C. Goutte, and M. Simard. Domain adaptation of mt systems through automatic post-editing. In *Proceedings of the MT Summit X*, pages 10–14, 2007.

[30] M. Simard, N. Ueffing, and P. Isabelle. Rule-based translation with statistical phrase-based post-editing. In *Proceedings of the ACL-2007 Workshop on Statistical Machine Translation (WMT-07)*, pages 203–206, 2007.

[31] A. L. Lagarda, V. Alabau, F. Casacuberta, R. Silva, and E. Díaz-de Liaño. Statistical post-editing of a rule-based machine translation system. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT) 2009*, pages 217–220, 2009.

[32] R. Rubino, S. Huet, F. Lefèvre, and G. Lenarés. Statistical post-editing of machine translation for domain adaptation. In *Proceedings of the European Association for Machine Translation (EAMT)*, pages 221–228, 2012.

[33] M. Forcada, M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. Tyers. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144, 2011.

[34] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180. Association for Computational Linguistics, 2007.

[35] A. Stolcke. Srilm-an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 257–286, 2002.

[36] Q. Gao and S. Vogel. Parallel implementations of word alignment tool. In *Proceedings of the ACL 2008 Software Engineering, Testing, and Quality Assurance Workshop*, pages 49–57, 2008.

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

[38] K. Papineni, S. Roukos, T. Ward, and W. Jing-Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[39] J. Tomás and F. Casacuberta. Statistical phrase-based models for interactive computer-assisted translation. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 835–841. Association for Computational Linguistics, 2006.

[40] V. Romero, A. H. Toselli, and E. Vidal. Using mouse feedback in computer assisted transcription of handwritten text images. In *Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR)*, pages 96–100, 2009.

[41] J. Tiedemann. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume V, pages 237–248. Borovets, Bulgaria, 2009.

[42] S. Khadivi and C. Goutte. Tools for corpus alignment and evaluation of the alignments (deliverable d4. 9). Technical report, Technical report, TransType2 (IST-2001-32091), 2003.

[43] P. Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the Machine Translation Summit*, pages 79–86, 2005.

[44] F. J. Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, 2003.

[45] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 310–318, 1996.