

Difficulty Estimation of Machine Translation

Ana Sofia Vieira de Jesus Almeida

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisor(s): Professor Isabel Maria Martins Trancoso

Doctor João de Almeida Varelas Graça

Examination Committee

Chairperson: Professor João Fernando Cardoso Silva Sequeira

Supervisor: Professor Isabel Maria Martins Trancoso

Member of the Committee: Professor Nuno João Neves Mamede

November 2017

Dedicated to my parents Carlos and Ana Paula,
my twin sister Ana Filipa and
my brother João.

Acknowledgments

I would like to begin by thanking my supervisors, Isabel Trancoso and João Graça for the opportunity of developing this research work in the area of machine translation. I would also like to thank Rúben Solera Ureña, Helena Moniz, Adel Abugren and Ramón Fernandez for all the support provided during this journey. I would like to extend my thanks to all the people of INESC-ID and Unbabel who contributed directly or indirectly to this thesis.

I would like to thank my friends who accompanied me in these harsh academic years. Finally, I would like to express my gratitude to my parents, my brother and my sister for supporting me throughout my academic career and through the process of researching and writing this thesis. This achievement would not have been possible without them. Thank you.

Lisbon, 15 October 2017

Ana Sofia Vieira de Jesus Almeida

Resumo

A estimação da dificuldade de tradução de um texto ou de uma frase é uma tarefa complexa e pouco explorada. Este trabalho propõe sistemas automáticos capazes de prever a dificuldade de tradução de um texto e de uma só frase. O tópico está relacionado com a tarefa de estimação da qualidade. Por isso, os nossos métodos de classificação tiveram em consideração as características adotadas nesta tarefa e os valores do *Multidimensional Quality Metrics* (MQM). O método proposto teve também em consideração as anotações feitas por um experiente anotador, e as características sugeridas por este. A principal contribuição deste trabalho reside no estudo da dificuldade de tradução ao nível do texto e da frase, e na construção de dois classificadores que atingem os 75.50 % de exatidão ao nível do texto e 77.67 % ao nível da frase. Os resultados sugerem que o MQM pode avaliar a dificuldade de tradução e que as características de qualidade se correlacionam com a dificuldade. Foi possível verificar que existe uma relação da dificuldade de tradução com a legibilidade do texto e com a *Human-Targeted Translation Edit Rate* (HTER) da frase. Adicionalmente, confirmou-se que o processo de pós-edição de um texto é afetado pela dificuldade de tradução do mesmo, implicando um texto difícil um maior número de edições e um maior tempo de edição. Por último, verificou-se que basta que uma frase seja difícil de traduzir para que o texto a que pertence também o seja. O sistema agora desenvolvido irá ser utilizado para selecionar editores de forma mais eficiente, e consequentemente melhorar a qualidade final da tradução. Este trabalho apresenta alguns resultados promissores e irá ajudar projetos futuros na área da dificuldade de tradução.

Palavras-chave: aprendizagem de máquina supervisionada, dificuldade de tradução ao nível do texto e frase, métricas de avaliação de legibilidade, tradução automática.

Abstract

Estimating the translation difficulty is a complex and little explored task. This thesis proposes automatic systems capable of predicting the translation difficulty of both texts and isolated sentences. The topic is closely related to the quality estimation task. Hence, our classification method took into account the values of the Multidimensional Quality Metrics (MQM) and the features adopted in the quality estimation task. The proposed method also took into account the annotations made by an expert annotator, and the features suggested by him. The main contribution of this work lies in the study of translation difficulty at both text and sentence levels, and in the development of two classifiers that reach 75.50 % accuracy at the text level and 77.67 % at the sentence level. The results suggest that MQM can assess the sentence translation difficulty and that the quality features correlate with difficulty. There is a relationship of the translation difficulty with the text readability and the Human-Targeted Translation Edit Rate (HTER) of a sentence. In addition, the post-editing process of a text is affected by the difficulty of translating it. Therefore, a difficult text implies a greater number of edits and a longer editing time. Finally, it was found that a sentence that is difficult to translate is enough for the text to which it belongs to be so. The developed system will be used to select more efficiently the editors, and thus improve the final quality of the translation. This dissertation presents promising results and may help future projects in the area of translation difficulty.

Keywords: machine translation, readability assessment metrics, supervised machine learning, translation difficulty at text and sentence levels.

Contents

Acknowledgments	v
Resumo	vii
Abstract	ix
List of Tables	xiv
List of Figures	xv
Nomenclature	xvii
1 Introduction	1
1.1 Motivation	1
1.2 Goals	3
1.3 Dissertation Structure	4
2 State of the Art	5
2.1 Machine Translation	5
2.2 Post-editing effort	8
2.3 Readability	9
2.4 Text Difficulty and Machine Translation	12
3 The Unbabel Annotation tool	15
3.1 Error typology	15
3.2 Error Taxonomy	16
3.3 Error penalty	19
3.4 Difficulties associated with the annotation tool	20
4 Corpora Description	21
4.1 The WMT-News corpus	21
4.2 The MCS corpus	21
4.2.1 Data collection	22
4.2.2 Data analysis	22
4.2.3 Data partition	25
4.2.4 Balance of the data set	26
4.3 The AMTA corpus	28

4.3.1	Data collection	28
4.3.2	Data analysis	28
4.3.3	Data partition	30
4.3.4	Balance of the data set	32
5	Preliminary Experiments	35
5.1	Most common errors made by Machine Translation	35
5.2	Impact of human edition in machine translation errors	36
6	Feature extraction and classifiers	41
6.1	Text and sentence classification module	42
6.2	Text analysis module	43
6.3	Sentence analysis module	46
6.4	Weka and Algorithms	47
6.4.1	ZeroR	47
6.4.2	Naive Bayes	47
6.4.3	Support vector Machine	48
6.4.4	Multilayer Perceptron	48
6.4.5	REPTree	48
6.4.6	K-Nearest Neighbor	48
7	Experimental results	51
7.1	Evaluation parameters	51
7.2	Baseline experiments	53
7.3	Relevant features experiments	55
7.4	Cross-corpora experiments	60
7.5	Discussion	62
8	Conclusions	65
8.1	Summary	65
8.2	Contributions	67
8.3	Future Work	67
	Bibliography	71
A	WMT-News Corpus	A.1
B	Corrected WMT-News Corpus	B.5

List of Tables

4.1	Characteristics of the MCS corpus.	22
4.2	Distribution of the MCS texts according to their length.	23
4.3	Distribution of the number of edits according to category.	24
4.4	Distribution of the time of edition according to category.	24
4.5	Examples of texts for each class considered.	26
4.6	Average values for parameters of the MCS corpus.	27
4.7	Distribution of sentences according to editor and MT system.	28
4.8	Example of AMTA sentences per topic.	29
4.9	Characteristics of the AMTA corpus.	29
4.10	Average MQM score per editor.	30
4.11	AMTA average classification method illustration.	31
4.12	Sentences distribution according to translation difficulty rating and topic.	31
5.1	Accuracy error types frequency according to severity.	37
5.2	Fluency error types frequency according to severity.	37
5.3	Style error types frequency according to severity.	38
5.4	Wrong language variety error types frequency according to severity.	38
5.5	Named entities error types frequency according to severity.	38
5.6	Total number of errors per category and severity.	39
6.1	Distribution of classes for each corpora type.	42
6.2	Machine Learning Algorithms used in Weka.	47
7.1	Confusion matrix.	53
7.2	Classifier results for each feature set and corpus.	54
7.3	Evaluation parameters values for the best classifier of the baseline experiments.	55
7.4	Confusion matrix for the best classifier of the baseline experiments.	55
7.5	Classifier results when using the 15 most relevant features of the MCS corpus.	56
7.6	Most relevant features for the MCS corpus and information gain rank.	57
7.7	Final classifier for the MCS corpus.	58
7.8	Final classifier for the AMTA threshold corpus.	58
7.9	Most relevant features for the AMTA threshold corpus and information gain rank.	59

7.10 Cross-corpora experiments.	60
7.11 Cross-corpora results.	60
7.12 Characteristics of text 1 and predictions on different scenarios.	61
7.13 Characteristics of text 2 and predictions on different scenarios.	61
7.14 Characteristics of text 3 and predictions on different scenarios.	62

List of Figures

1.1	Languages used on the Internet (2015).	1
1.2	Unbabel workflow.	2
4.1	Number of words per text vs time per word.	25
4.2	Distribution of manually classified texts.	26
4.3	Final distribution of classified texts according to the annotator.	27
4.4	Distribution of the four topics available in the AMTA corpus.	29
4.5	AMTA average classification method distribution.	32
4.6	AMTA threshold classification method distribution.	32
4.7	Final distribution of the AMTA average classification method.	33
4.8	Final distribution of the AMTA threshold classification method.	33
5.1	Most common errors made by MT [13].	36
5.2	Most common errors made by MT in WMT-News corpus.	36
6.1	Training module.	41
6.2	Prediction module.	42
7.1	Kappa statistic scale.	53

Nomenclature

Acronyms

AMTA Association for Machine Translation in the Americas

BLEU Bilingual Evaluation Understudy

HTER Human-Targeted Translation Edit Rate

INESC-ID Instituto de Engenharia de Sistemas e Computadores - Investigação e Desenvolvimento

MQM Multidimensional Quality Metrics

MT Machine Translation

MWE Multi-word expression

NMT Neural Machine Translation

RBMT Rule-Based Machine Translation

SMT Statistical Machine Translation

SVM Support Vector Machine

WEKA Waikato Environment for Knowledge Analysis

WMT Workshop on Machine Translation

Chapter 1

Introduction

This document explores the prediction of the translation difficulty at text and sentence levels. This is a complex problem that has not been explored by many researchers. As a proof of concept, we have chosen a single language pair (English-Spanish). The motivation for the development of this thesis is described in Section 1.1. The goals are enumerated in Section 1.2 and finally, the outline of the remain document is presented in Section 1.3.

1.1 Motivation

The enormous potential of machine translation (MT) has barely been explored. In a world flooded with information, Internet has the power to hold enormous quantities of data within different languages. According to Figure 1.1, only 55.7 % of Internet content is in English, leaving more than 40 % of the Internet content distributed by other languages.¹

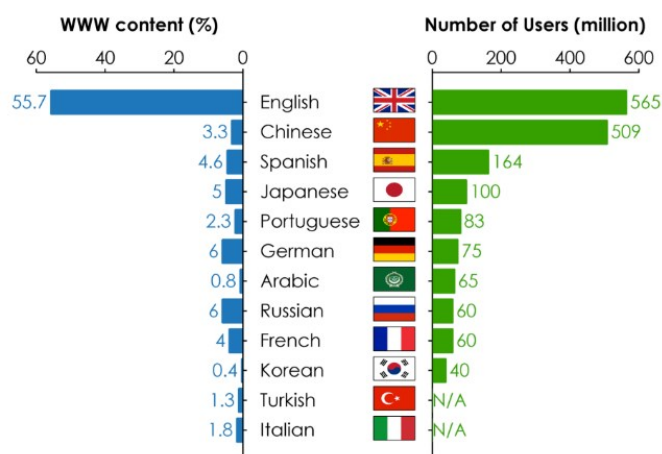


Figure 1.1: Languages used on the Internet (2015).

¹<https://unbabel.com/blog/top-languages-of-the-internet/> (visited on 16/02/2017)

Over the years, millions of people have joined the Internet community. Most of these people prefer reading and writing in their own native languages. Every year, the number of languages used in the Internet increases, thus it is critical to simplify the communication between people around the world through machine translation. Nowadays, it is common to use an MT system like Google translate ² to obtain a fast translation. However, a fast translation does not ensure a high quality. The quality of the translation is relative and depends on its final application. The economic potential of MT is also interesting and for that reason, the companies are investing on studying and improving their MT systems. The idea is to provide better service at lower price. Consequently, companies could captivate more users to their services and have more profit.

The topic of text difficulty has been studied over the years, which is proved by the development of over 200 readability algorithms. On the other hand, little attention has been paid to the estimation of translation difficulty. This task can be very attractive to companies that use MT, as this can help to distribute the post-editing work efficiently, saving both time and money through an appropriate editor selection.

This work was developed with the aim of developing automatic classification systems capable of classifying texts and sentences concerning their translation difficulty. Through these systems companies like Unbabel could select the editors efficiently. Unbabel is a Portuguese startup that combines machine translation and crowdsourced translation to provide a faster service with higher quality. By using an editor community that works online at the company platform, Unbabel ensures a faster service. The workflow of this Portuguese startup is present in Figure 1.2.³

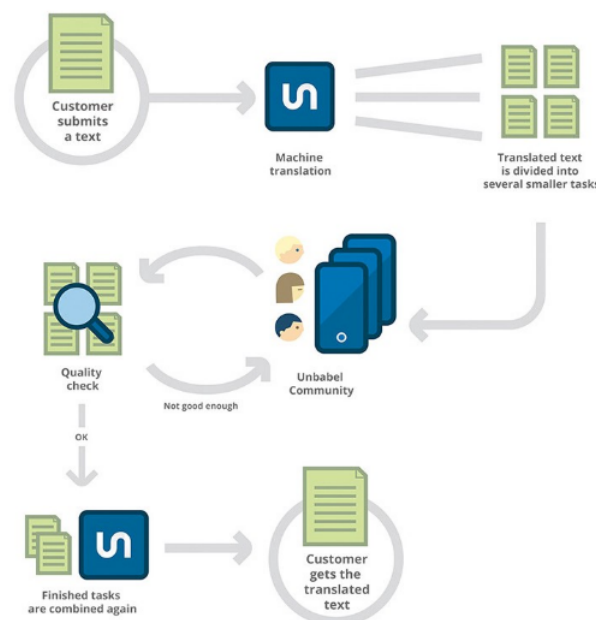


Figure 1.2: Unbabel workflow.

²<https://translate.google.pt/> (visited on 17/06/2016)

³<https://help.unbabel.com/hc/en-us/articles/205864998-Introduction> (visited on 09/06/2017)

When a text arrives at Unbabel, it is divided into small segments and later distributed to the editor community. The editors are bilingual (speak two languages) but not necessarily professionals. Thus, the number of editors available is much larger than the number of professional editors, and the costs per edition are dramatically reduced. On the online platform, the editors have access to the source and target text. In this way, the editors can ensure that the meaning of the source is present in the target, and that it is well written.

After editing the segment, the quality is automatically checked to determine if it is required to go through another edition. If the segment needs more editing, it will be put back on the platform and another editor will later edit the segment and ideally will increase its quality. If the quality is met, the segments of the source are gathered and delivered to the customer. In some cases, the full text is also reviewed by a senior editor in order to ensure consistency and fluidity.

The human post-editing work allows increasing the quality of the final translations. Because texts are divided into segments, the time of edition is relatively low and the use of non-professional editors reduces the involved costs.

1.2 Goals

Our main goal was to develop automatic classification systems capable of classifying a text and a sentence concerning their translation difficulty. These systems will ideally classify the texts and sentences that arrive at Unbabel and give important information to support better selection of the most suitable editors for the document. Some of the research questions we posed ourselves during the development of the two classifiers were:

- How can we measure translation difficulty?
- Is it possible to estimate translation difficulty at text and sentence levels? Which one has the best results?
- What are the most relevant features for each corpus level?
- Does translation difficulty correlates with readability and HTER?
- Is there any relationship between the translation difficulty and the efficiency of the post-editing task?
- It is enough to have a difficult to translate sentence for the text to which it belongs to be so?

1.3 Dissertation Structure

The remainder of this document is organized as the following: Chapter 2 gives an overview of the background and fundamental concepts concerning machine translation and difficulty. Section 2.1 introduces the machine translation definition and describes the types of available MT software. Section 2.2 presents the relationship between text difficulty and post-editing effort, and the post-editing effort measurement methods are referred. Section 2.3 discusses the features used by well-known readability algorithms that estimate text difficulty through semantic/syntactic features and statistical language modeling techniques. Finally, Section 2.4 addresses the text difficulty and MT.

In Chapter 3, the Unbabel annotation tool is presented. Section 3.1 explains the error typology and Section 3.2 describes the error taxonomy by enumerating all error types that can be selected by the annotators. The error penalty system is described in Section 3.3 and the difficulties associated with the annotation tool are mentioned in Section 3.4.

Chapter 4 presents the corpora used in this work. The corpus used in the preliminary experiments is addressed in Section 4.1. A text based corpus is described in Section 4.2 and Section 4.3 characterizes a sentence based corpus. In the last two corpora, four topics are approached: the data collection, the data analysis, the data partition and the balance of the data set.

In Chapter 5 the results of two preliminary experiments are discussed. The first experiment concerning the most common errors made by machine translation is presented in Section 5.1. Section 5.2 mentions the experiment that tries to understand the impact of human edition in machine translation errors, by checking what are the errors that the human editor can and can not eliminate through the edition process.

Chapter 6 focuses on the features extracted from the corpora and describes the machine learning algorithms used to train the classifiers. Section 6.2 presents the features extracted from the text based corpus and Section 6.3 explores the features extracted from the sentence based corpus. The algorithms used are described in Section 6.4.

The parameters adopted to evaluate the classifiers and the obtained results are presented in Chapter 7. Finally, conclusions drawn from this work and references for possible future contributions are presented in Chapter 8.

Chapter 2

State of the Art

The primary goal of this project was the estimation of text and sentence translation difficulty, and the exploration on how the translation difficulty can be measured.

A definition of machine translation (MT) and a description of the types of MT software available are presented in Section 2.1. Furthermore, the evaluation of MT systems is also approached by pointing out and describing the usual methods used for this purpose. The relationship between text difficulty and post-editing effort is mentioned in Section 2.2. In the same section, the post-editing effort measurement methods are also referred. Section 2.3 overviews the background on the features used by well-known readability algorithms that estimate text difficulty through semantic/syntactic sources and statistical language modeling techniques. Finally, Section 2.4 addresses text difficulty and MT.

2.1 Machine Translation

MT involves the translation from a source to a target language through software. The goal is to produce a translation with high quality, where the meaning of the text in both target and source languages must be the same. However, the output is often post-edited by human editors, in order to achieve a higher quality score. At first, the translation task may seem easy, but it is not solely about replacing words and therefore, it is mandatory to analyze the text in order to understand the links between its elements and consequently achieve a better translation. One of the biggest challenges in MT is to achieve high quality, given that many characteristics can make this task more complex such as ambiguity (words/expressions that have several meanings), variability (expressing the same thing in many different ways) and idiomatic expressions, characteristics of each natural language. The text content can give a hint for the selection of the most indicated vocabulary. However, this way might not be enough to overcome bilingual lexical ambiguities.

MT faces several challenges like:

- The bilingual lexical ambiguities, variability and idiomatic expressions;
- Scarcity of resource language pairs;
- Lack of robustness across multiple domains;
- The structural and lexical differences between languages.

There are different types of MT methods [11] like Rule-Based Machine Translation (RBMT) and Statistical Machine Translation (SMT). Rule-Based Machine Translation adopts countless rules and dictionaries for each pair of languages. This type of method relies on human effort (linguists) to produce the applicable rules. The method parses the text and uses the rules and dictionaries to get a translation. Then, the grammar structure of the source language is transferred to the target language. Statistical Machine Translation depends on having statistical models trained with a parallel corpus at the sentence level and can achieve high-quality translations when large corpora are available. The model parameters result from the analysis of the corpora and although it does not use any knowledge of grammar, it can achieve a good fluency when compared to RBMT. Note that SMT can often find exceptions to the rule-based and reduce the human resources cost.

The most recent approach to teach a machine how to translate is based on deep neural networks and it is called Neural Machine Translation (NMT) [24]. It is based on how the human brain works, and consists of a set of nodes that relate to each other and can represent single words, sentences or another segment. NMT requires access to both source and target sentences as training data, and the relationship between the nodes is created through bilingual texts with which the system is trained. The NMT seems to be a promising approach and the community will continue to explore its contribution to the machine translation area. This approach was not used in this work.

In order to know which MT systems perform better, several metrics that try to quantify the translation quality have been proposed over the years. The most reliable method includes human evaluation, which is a subjective and very time-consuming type of evaluation. Moreover, there are different translation types. Human translation can obtain very different translations while using the same pair of source-target languages and source text. This happens due to the existence of more than one way of correctly translating a text. On the other hand, MT has the challenge of understanding how software can produce high-quality translations. Since evaluating MT output using human judgments is slow, some automatic measures were created. They judge the quality of MT output by comparing the system output (candidates) against a reference or a set of reference translations. Automatic evaluation methods include Bilingual Evaluation Understudy (BLEU), Metric for Evaluation of Translation with Explicit Ordering (METEOR), Translation Edit Rate (TER) and TER-Plus.

BLEU [29] is an algorithm proposed by IBM based on the principle that the MT output and the professional human translation should be as similar as possible towards having a higher translation quality. The BLEU score is calculated by counting the number of n-grams in the system output that occur in the set of reference translations. It is precision-oriented by measuring how much the system output is correct instead of seeing if the reference is totally reproduced in the system output. This metric has values between zero and one, where one indicates a perfect translation, and all tokens are equally weighted. The result is an average of the scores of the n-grams. Longer n-gram matches in equal order with the references achieve stronger BLEU credits. A brevity penalty is used to penalize MT sentences that are shorter than reference. The number of words of the candidate (w_c) and the reference (w_r) sentences are compared and the penalty value is calculated through the equation 2.1.

$$BP = \begin{cases} 1 & , \text{if } w_c \geq w_r \\ e^{(1-w_r/w_c)} & , \text{otherwise} \end{cases} \quad (2.1)$$

The evaluation method adopted by NIST (National Institute of Standards and Technology) is based on BLEU but the evaluation includes adjustments by calculating the degree of information that a particular n-gram gives.

METEOR [1] appeared to improve BLEU. It adopts word-to-word matches between the reference and the translation to compute the score. Words that do not match are stemmed and then, matched. When having more than one reference, a score for each reference is calculated and then the highest score is selected. When comparing METEOR with BLEU, the first gets a better correlation with human judgments.

TER [37] measures the number of edits (insertions, substitutions or deletions of single words and shifts of word sequences) which are required to match the MT system output with the reference, and it is in line with human effort. This measure only takes into account exact matches and can only handle one reference at the time. When more than one reference translation is available, TER works like METEOR, by comparing scores of hypothesis against each one of the references translations available. A TER score of zero represents a perfect match between raw translation (machine translation output) and reference, and a higher score implies a worse translation quality. Hence, the score must be as low as possible and can be calculated through equation 2.2.

$$TER = \frac{\text{number of edits}}{\text{number of reference words}} \quad (2.2)$$

It is possible to estimate the post-editing effort by analyzing a large number of sample translations. A higher TER score also means the translation requires more post-editing due to its lack of quality. The Human-targeted TER (HTER) considers the TER between the MT and its post-edited version.

TER-plus [38] emerged as an upgrade version of TER by implementing paraphrase, stemming, synonyms and optimizing costs to improve the principal weaknesses of TER, while increasing correlation with human judgments. Contrary to TER, TER-plus takes stemming and synonyms into account for word matches instead of exact matches, providing a way of dealing with limited reference translations. The method identifies stems using the Porter stemming algorithm [31] and synonyms through WordNet [15] (a lexical database for English). While in TER the edit costs are all equal, in TER-plus the costs can be learned automatically through a hill-climbing search meaning, an iterative algorithm that starts with an arbitrary solution and then tries to find an optimal solution by incrementing the parameters.

The existence of many different automatic evaluation methods demonstrates the difficulty of this multifaceted problem.

2.2 Post-editing effort

The post-editing task refers to the correction of raw machine translated output by human editors according to specific guidelines [4]. It is an expensive, time-consuming task, and is extremely difficult to predict the duration required to complete a post-editing task. This section focuses on understanding how post-editing effort and text difficulty are related.

The post-editing effort is highly correlated with TER reflecting that complex texts require higher post-edit effort than the easy texts. There are three dimensions of effort: temporal, cognitive and technical [3]. Generally, productivity is measured through temporal effort, in the form of processing speed, and cognitive effort in the form of eye-tracking [4].

Sharon O'Brien studied the relationship between controlled language (CL) and post-editing effort. CL refers to a set of rules that a writer should follow when constructing sentences, and helps to increase text readability. They explored the principle that the application of CL rules would improve MT output and consequently decrease the post-editing effort required [3]. In this case the Choice Network Analysis (CNA) that focuses on the changes made by post-editors, was applied as a measure for cognitive effort and the number of deletions, editions and insertions made by a post-editor was computed as a way of estimating the technical effort. CNA is a method for creating models of how translation is made and by taking into account the complexity of the choices available to the translator can estimate the difficulty of parts of source texts [28]. The sentences that were subsequently translated and post-edited showed an increased processing speed when CL rules were applied. The fact that source text Part-of-Speech (PoS) may have an impact on the time and the quality of translation [18], makes easier to construct rules that try to take an advantage of this knowledge.

Other methods to measure post-editing effort include think-aloud protocol (TAP) and Translog [2]. TAP involves participants thinking aloud while post-editing, which brings a slow-down effect in the process. Therefore, it is not the best alternative to measure cognitive post-editing effort. Translog acquires user activity during post-editing task recording key strokes and mouse movements. Typed words, num-

ber of deleted characters and time required to complete the post-editing task is the type of information that can be extracted from this tool. The problem with Translog lies on not knowing what happens during the pauses made by the translator or post-editor. The long pauses occur before the post-editor makes changes in the hardest parts of the sentence identified by the CNA [2]. There are some features that affect MT and they are known as Negative Translatability Indicators (NTI). Using controlled language in such a way that NTI can be eliminated would improve MT output and decrease post-editing effort.

2.3 Readability

There are many definitions of Readability. McLaughlin (1969) defined readability as: "the degree to which a given class of people finds certain reading matter compelling and comprehensible" [25], while Edgar Dale and Jeanne Chall (1949) had another interpretation: "The total sum (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers has with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting" [14]. Readability has been measured through readability algorithms over the years. One of the research questions of this work is precisely about the usability of readability in the translation difficulty estimation.

Some say the precursor of readability algorithms was Thorndike, an American psychologist who devoted his career to the Teachers college, Columbia University and wrote "The Teacher's work book" in 1921 [40]. The main idea was to focus on vocabulary, particularly on word frequency. It was assumed that the words that appeared less often were less known and therefore, increased text difficulty.

Over the years, more than 200 readability algorithms were created. However, most of them are no longer used. There are readability algorithms using the same features but adopting different weights, in order to give different degrees of importance to each feature, making a direct comparison between them, impossible. The most common features are the average sentence length, the average word length (using the number of syllables or the number of characters), the number of words per sentence, the average word frequency and the percentage of hard words. These types of features are used in the most known readability algorithms like Flesch Reading Ease Formula [16], Dale-Chall [5], Gunning's "Fog-Index" [19], Lexile [36] and Automated readability index [35].

The readability formulas:

- get the grade level that a person needs to be able to read a text;
- are text-based;
- are easy to use;
- can identify if a text is or not too complex for a reader.

On the other hand, they can not tell if a person will be capable of interpreting a text and are just quantitative measures.

The Flesch Reading Ease Formula is one of the oldest readability formulas to assess the grade-level of the reader that uses the average sentence length and the average number of syllables per word as features. According to this formula, the best text should contain short sentences and words. Flesch-Kincaid emphasizes sentence length over word length while using the same features as the Flesch Reading Ease Formula.

Dale-Chall [5], Gunning's "Fog-Index" [19] and Automated readability index [35] use the percentage of hard words and the average sentence length as lexical and grammatical features. The definition of hard words is also different depending on the author: generally hard words are words that do not appear on the known word lists but, in the Gunning's "Fog Index" [19] hard words are words with more than 3 syllables.

Unlike other readability measures that use the information about the syllables per word, the Automated readability index (ARI) and the Coleman-Liau from 1967 relied on characters per word. These algorithms do not require an analysis of the characters that create the words but take into account the word length in characters. According to Coleman, "There is no need to estimate syllables since word length in letters is a better predictor of readability than the word length in syllables" [9].

The Lexile Framework [36] measures the syntactic and semantic complexity using the sentence length and the word familiarity, respectively. It is based on the syntactic axiom: "the shorter the sentences, the easier the passage is to read" and the semantic axiom: "the more familiar the words, the easier the passage is to read." Generally, longer sentences and words with lower frequency lead to higher Lexile measures and shorter sentences and words with higher frequency lead to lower Lexile measures.

The developers of Accelerated Readers reading software have also released ATOS [26], a readability formula that estimates the difficulty of books and other texts. It takes into account some of the most important factors such as average sentence length, average word length, vocabulary grade level and number of words of a book, as well as less common aspects such as verb forms, pronoun use, length of clauses, paragraph length and ratio of unique to total words. One of the most important aspects is the use of the book length, which most readability measures do not avail.

Then there are other readability algorithms using statistical language modeling techniques such as Collins-Thompson & Callan [10], Schwarm & Ostendorf [32] or Heilman, Collins-Thompson & Eskenazi [21].

Collins-Thompson & Callan have put together a readability measure [10] based on an extension of the multinomial Naive Bayes classification algorithm that combines multiple language models (one for each of the 12 American grade levels) to determine the most likely grade language model given the text T , or equivalently, the model G_i that maximizes $L(G_i|T)$.

Schwarm & Ostendorf [32] use Support Vector Machines (SVMs) while Heilman, Collins-Thompson & Eskenazi readability algorithm [21] is based on a linear function of grammatical and lexical features (relative frequencies of word unigrams). Normally, the adoption of sentence length, as a measure of grammatical complexity, assumes that a longer sentence is more grammatically complex than a shorter one, which is often but not always the case. Parameters are estimated using a Gaussian distribution with zero mean and a user-specified variance over the parameters.

Lastly, Edward Fry presented the Fry readability formula [17] that estimate the grade reading level through a graph where the average number of syllables per 100 words is on the x-axis and the average number of sentences per 100 words is on the y-axis. The intersection of these averages determines the reading level of the text. This method of 1968 is a more intuitive in the way of estimating the grade reading level. However, it is not always able to obtain a conclusive result.

A study [30] analyzed different readability factors in order to link discourse structure with text quality and identify the individual factors: vocabulary, discourse relations, average number of verb phrases and length of the text as the strongest predictors of readability. On the other hand, the use of rare words, technical terminology and complex syntax decreases readability, and the average number of words per sentence or the average number of characters per word are bad predictors.

Some experiments determined that the rates of occurrence of certain parse tree constructions become more frequent as ESL levels (low-intermediate, high-intermediate, advanced) increases [22]. The average sentence length of texts increase with the age and reading level of the audience. Also, grammatical features may be especially important to second language readability (students of second language often struggle with grammatical structures).

Readability classifiers have been addressed in previous Master theses at IST, namely by Pedro Curto [12], whose thesis aimed at the selection of adequate materials for teaching European Portuguese as a second language, for different proficiency levels. The system extracts 52 features grouped in seven groups: PoS, syllables, words, chunks and phrases, averages and frequencies, and some extra features. Two experiments were made concerning the evaluation of the classification task: the first one based on a five-level scale (A1, A2, B1, B2, C1) and the last based on a simplified three-level scale (A, B, C). The corpus used to train the classifier consists of 212 previously classified texts, provided by Instituto Camões. The lowest precision was obtained for the PoS feature group. In both cases (five-level and three-level scale), the most influential features were the number of words, the number of different words, the number of dependencies, the number of tree nodes, and the number of sentences. The classifier achieved an accuracy of 79.25 % for the first experiment and 86.32 % for the second, having one level distance for most of the errors.

2.4 Text Difficulty and Machine Translation

MT evaluation usually works by comparing the MT system output with reference translations to provide a quality score that correlates with the degree of resemblance between two translations. There is a relationship between text difficulty and the quality of MT output. Intuitively, difficult texts would provide a poor quality in machine translation.

A pilot study [8] with ILR - Interagency Language Roundtable reckons texts from 5 different languages (Spanish, Farsi, Arabic, Russian and Korean) and their correspondent reference translations. There are 4 levels in the ILR scale of language proficiency/difficulty. Level 1 is the basic that requires elementary reading skills and level 4 is the top level which requires advanced professional proficiency. By comparing the NIST/BLEU scores of MT output with input text difficulty, it was possible to find a relationship between the input text difficulty and the machine translation performance. The main conclusion reveals that increasing ILR levels causes lower performance from a MT system. Furthermore, having a perfect word translation does not imply intelligibility to the human readers because the word order has a particular influence.

In 2010, a thesis entitled Locating and Reducing Translation difficulty [27] focused on locating difficult-to-translate phrases (DTP) and reducing their translation difficulty. In order to classify a sentence as easy or difficult, the process begins by translating the sentence and measuring its quality. Considering a threshold, a decision is made and two classes are created. Translation quality is measured through BLEU by comparing the MT system output with the human references. A DTP is a phrase that is weakly translated using a specific MT system and as consequence has very low BLEU value. In order to do some word aligning between source and reference sentences, some sentences were added to increase the size of the corpus and consequently increase word alignment quality. Through GIZA++ software the words were aligned and a corpus of parallel phrases was build. The most frequent reasons behind phrase difficulty include:

- Unknown source language word;
- Lexical ambiguity;
- Articles/Punctuation/Numbers Deletion/Insertion;
- Cross lingual subject verb object order differences;
- Word form error (plural, gerund);
- Translation divergence (concept expression differences across two languages).

A study about the relationship between text difficulty and translation accuracy [20] states that text difficulty with the purpose of translation is a function of cognitive effort, meaning that when many choices exist, the translator has to make a significant cognitive effort to select the right one, and the other way around is also true, a few choices makes the translation process much easier. Reiss (1982) proposed that text difficulty for the purpose of translation depends on 5 aspects:

- The subject matter (semantic aspect);
- The register (material aspect);
- The type of language used (functional aspect);
- The pragmatics of the reader (pragmatic aspect);
- The historical-cultural context (temporal, local or cultural aspect).

An experiment with 20 students of Western Sydney University was performed. Eleven of those 20 students spoke Spanish, and during one semester they had to translate texts of approximate 260 words, with a maximum time of 1 hour and 10 minutes. Here the dictionaries were allowed. Their findings suggest that text difficulty is related to:

- The text itself;
- The Translator's competence;
- The mode of translation.

With the goal of being able to automatically rate Machine Translatability [41], a method that, if applied to every component of a source text, can automatically identify the machine-translatable and non-machine-translatable parts to a specific MT system, was developed. The non-machine-translatable parts provide important information to a revision and posterior improvement of the final translation. Instead of requiring reference sentences, this method requires an MT system which is capable of doing bidirectional translation, this is, capable of translating to source and target languages. C-measure measures the similarity between source sentence and its back translation (translates from source to target and then translate the target into the source). The chosen languages were Japanese and English. This similarity was based on BLEU. Without any reference translation, it is still possible to find sentences that are difficult to translate by MT systems, by choosing sentences with low C-measure values.

Sanjun Sun was able to develop a formula to predict a text's translation difficulty level for a translator through translator's pre-translation rating [39]. There was a statistically significant relationship between the translation difficulty score and the self-predicted level of translation difficulty at a 95.0 % confidence level. The adjusted R-squared statistic was 0.462. The author also included a study about the influence of readability in translation difficulty and founded that a text's readability only partially accounts for its translation difficulty. The translation difficulty was assessed by the NASA Task Load Index, a multidimensional scale for measuring mental workload, and 15 short passages to be translated from English to Chinese were used. The participants included third-year undergraduate and first-year graduate students in translation from a university in Shandong Province and Beijing. The study concluded that translation quality was an unreliable indicator of translation difficulty, while the time required to perform a human translation was weakly correlated to the translation difficulty.

The authors asked about the most challenging part of the process on the post-translation survey and gave three options:

- Source text comprehension;
- Reverbalization in the target language;
- The two were equally difficult.

The answers indicate that the hardest part of the translation process is the process of reverbalization (77 %), the writing part. However, the source text comprehension (13 %) may also give some problems to the translator. Only 10 % of the participants responded that both tasks were equally difficult.

Chapter 3

The Unbabel Annotation tool

Error annotation is a very important task, allowing the evaluation of a translation service. Annotations can be automated or made by humans. Human annotation is expensive and time consuming, although it is more accurate. It is, however, more difficult to achieve consistency and objectivity. Annotations can be performed by one annotator or by multiple annotators, being important in this last case to verify the agreement among all the annotators, in order to provide information on the reliability of the results. Unbabel is implementing an inter-annotator agreement structure and the preliminary results for the translation between English and Italian demonstrate an 83 % inter-annotator agreement.

The Unbabel annotation tool used to know the average quality of a group of translations in an objective way is described in this chapter. The error typology used by Unbabel is mentioned in Section 3.1 and Section 3.2 describes the error taxonomy by identifying all types of errors that can be chosen in the tool. The error penalty system is described in Section 3.3 and Section 3.4 enumerates difficulties associated with the Unbabel annotation tool.

3.1 Error typology

The guidelines used by Unbabel to create an error typology were based on the Multidimensional Quality Metrics (MQM) framework [23], a template developed in the Quality Translation 21 project (QT21). The QT21 project was created with the aim of breaking down barriers, specifically language barriers, and is funded by the European Union Horizon 2020 research and innovation program. QT21 intends to improve statistical and machine-learning based translation models, enhance evaluation and learn from mistakes, by systematically analyzing quality barriers informed by human translators. In addition, scalability is also important in order to ensure that learning and decoding with these models is efficient and that reliance on data is minimized. The MQM framework defines metrics and scorecards to estimate the quality of translated texts. The framework evaluates different aspects of the quality of translation and is multidimensional, which means that it is used to measure quality in different aspects such as accuracy, fluency, style and terminology.

A numeric score for the translation quality that takes into account different translation quality issues and their severity (minor, major and critical) is calculated after the error annotation. The Localization Industry Standards Association quality assurance (LISA QA) model was developed to be used by human reviewers when assessing a text or product. Most of its categories are not suitable for automatic assessment as they rely on human judgment. The scores are presented in percentage. The MQM of a sentence can be calculated with standard LISA severity weights through equation 3.1.

$$MQM\ score[\%] = 100 - \left(\frac{Issues\ minor + 5 \times Issues\ major + 10 \times Issues\ critical}{Word\ count} \right) \times 100 \quad (3.1)$$

Each severity is associated with a weight reflecting the difference between having several minor issues that may not influence the quality of the translation, and a critical issue that can drastically alter the meaning of the translation. An error-free translation scores 100 %.

3.2 Error Taxonomy

The error taxonomy defined at Unbabel is divided into seven categories: accuracy, fluency, style, terminology, wrong language variety, named entities, formatting and encoding. The annotator can choose from 41 error types available in the annotation tool. The error taxonomy is shown below and the errors available are underlined.

(1) **Accuracy** (Errors in the translation of meaning)

• **Mistranslation** (errors in substitutions)

- Overly Literal

- False Friend (when a word has been translated in a different meaning because it looks and sounds similar)

- Should not have been translated (expressions that do not have equivalents in the target language and that do not have to be translated)

- Lexical Selection (the terms selected are not correct for the context or are not accurate to convey the meaning of the original text)

- Omission

- Untranslated

- Addition

(2) **Fluency** (problems that affect the reading and the good comprehension of the text)

- **Inconsistency**

- **Word Selection** (the same term has been translated differently throughout the text)

- **Tense Selection** (temporal cohesion is not correct)

- **Coherence** (meaning is not complete or is hard to understand)

- **Duplication**

- **Spelling** (misspelled word)

- **Orthography** (orthographic errors)

- **Capitalization** (wrong use or absence of capital letters)

- **Diacritics** (symbols substituted or missing)

- **Typography**

- **Punctuation**

- **Unpaired Quote Marks and Brackets**

- **Whitespace** (there is an extra or missing whitespace)

- **Inconsistency in Character use** (specially added for Chinese, to mark the inconsistency in the use of traditional and simplified characters)

- **Grammar**

- **Function Words**

- **Prepositions**

- **Conjunctions**

- **Determiners**

- **Word Form**

- **Part-of-Speech** (the grammatical category is not correct or is not natural)

- **Agreement** (problems with number and person between words)

- **Tense/Mood/Aspect**

- **Word Order**

- **Sentence Structure** (the structure of the sentence is not correct or is too close to the original language)

(3) **Style** (issues concerning register and fluency)

- **Register** (the register used is not correct for the context)
- **Inconsistent Register**
- **Repetitive Style** (unnecessary repetitions of expressions or words)
- **Awkward Style** (style used is not accurate)

(4) **Terminology**

- **Noncompliance with client or company style guide** (translation does not follow the client's instructions)
- **Noncompliance with the glossary and vocabulary** (word translated is not the same that is shown in the glossary)

(5) **Wrong language variety**

(6) **Named entities**

- **Person** (name of the person has/has not been translated when it should/should not have been)
- **Organization**
- **Location**
- **Function**
- **Product**
- **Amount** (errors involving the substitution of unit, conversions)
- **Time** (date and time mistranslated due to different formats)

(7) **Formatting and encoding** (errors that were provoked by the platform layout)

3.3 Error penalty

The errors may have different impacts on the translation quality. There are minor errors that can be ignored while others can completely change the meaning of the text, leading to conclusions that are very different from those expected. Thus, errors can be cataloged in three different categories:

(1) **MINOR:** Errors that do not lead to a loss of meaning and would not confuse the user but would be noticed. This type of errors decreases fluency or makes the content less appealing. These errors include:

- Double spaces;
- Use of the decimal point instead of a comma;
- Misplaced commas;
- Missing hyphens;
- Repetitions of the same term in the same sentence.

(2) **MAJOR:** Errors that may mislead the user or hinder proper use of the product/service due to a significant change in the meaning or because errors appear in a visible or important part of the content. Examples of major errors are:

- Lack of agreement;
- Wrong grammatical subject;
- Tense/Mood/Aspect issues;
- Coherence issues;
- Wrong word order;
- Wrong function word.

(3) **CRITICAL:** Errors that may carry health, safety, legal or financial implications, damage the company's reputation, cause the application to crash or negatively modify/misrepresent the functionality of a product or service, or which could be seen as offensive. These critical errors include:

- Translation does not make sense;
- Word selection that may have a negative influence on the reader towards a certain product;
- Different meaning from the source text that may lead to legal, healthy or economic repercussions;
- Whenever the meaning of the source text is changed.

Each error category is associated with a different weight penalty that is used when calculating the MQM score (equation 3.1).

The Unbabel annotation tool shows to the annotator both source and target texts. The errors found in the target text can be annotated and later classified according to the error taxonomy and penalty. The annotator indicates also the level of fluency, how natural the target text sounds on a scale of zero to five, where five represents a higher fluency.

3.4 Difficulties associated with the annotation tool

There are situations that can confuse the annotators when using the annotation tool. For instance, an error can be categorized differently and it is up to the annotator to decide the most appropriate error type. It can be difficult to decide which error type should be selected when a word is associated with more than one error type. According to the instructions given by Unbabel in the annotation tool, only one category can be selected. Thus, when a word has more than one type of error associated with, the annotator should select the category with higher severity, in order to be always considered the worst case scenario.

The taxonomy presented was developed to annotate the final translations, when the document is ready to be delivered to the client. However, the taxonomy is also used when machine translation texts are annotated, which may not be the most appropriate given that these texts are associated with different types of errors from those found in a post-edited text.

Chapter 4

Corpora Description

This chapter presents the corpora used to develop all classifiers. Section 4.1 mentions the corpus used in the preliminary experiments (WMT-News). In Section 4.2 a text based corpus is described (MCS), and Section 4.3 presents a sentence based corpus (AMTA). The last two corpora include different topics ranging from Mails and Customer Support texts to sentences regarding climate, Mexican, Norway and software.

4.1 The WMT-News corpus

This section addresses the WMT-News corpus that included the first 15 sentences from the annual Workshop on Machine Translation news corpus.¹ This corpus contained the source (in English) and the target (in Spanish) sentences. Although the corpus was larger, only a small part was used in the preliminary experiments. A pre-processing at the source and target sentences that included the elimination of spaces before commas and apostrophes was required. The hypotheses sentences were attained through the translation into Spanish by Google translate of the English source sentences. The source, hypothesis and target sentences are presented in Appendix A.1. Next, a Spanish researcher corrected the Google translate output (hypothesis) from the WMT-News corpus. The corrected sentences are shown in Appendix B.1.

4.2 The MCS corpus

This section addresses a text based corpus concerning Mails and Customer Support texts. The data collection is explained in Section 4.2.1 and the data analysis is made in Section 4.2.2. The data partition is explored in Section 4.2.3 and Section 4.2.4 approaches the balance of the data set.

¹http://opus.lingfil.uu.se/WMT-News/En-es_sample.html (visited on 17/06/2016)

4.2.1 Data collection

Data collection is not always trivial, especially when confidentiality is required, adding issues to the task. Although this data can not be disseminated for confidentiality reasons, some properties may be shared.

Unbabel collected 200 uncategorized texts concerning Mails and Customer Support. Due to the "abundance" of data concerning English to Spanish translations, the language pair English-Spanish was selected. Hence, the 200 source texts were in English. The source texts were translated by the Unbabel's MT system and the output, 200 Spanish texts were edited by several non-professional human editors. Both machine translated and the post-edit texts were included in the corpus. Information concerning the number of edits performed by the editors, as well as the time of edition for each text, was provided by the company.

4.2.2 Data analysis

After the data collection, a data analysis is presented in order to get a better understanding of the MCS corpus. The MCS texts had a typical structure which is as follows:

- Greetings;
- Mail body;
- Closing line.

The characterization of the corpus is shown in Table 4.1. Lexical diversity was calculated by dividing the number of distinctive words by the text length. The number of adjectives and prepositions was calculated using the Part-of-Speech (PoS) tag and the Stanford Named Entity recognizer was used to compute the number of named entities present in the texts. The multi-word expressions were found through a protocol that is explained in Section 6.2.

Table 4.1: Characteristics of the MCS corpus.

Feature	Average Value per text
Number of sentences	5.80
Number of words	71.59
Number of syllables per word	1.40
Lexical diversity	0.82
Number of adjectives	4.59
Number of prepositions	7.90
Number of named entities	1.42
Number of Multi-word expressions	1.88
Number of edits	26.99
BLEU	0.45
Automated Readability Index	10.80
Time of edition (in seconds)	44.68
HTER	0.38

Mail and Customer Support texts are generally short, having on average six sentences per text. Text length is related with the lexical diversity. A smaller text tends to have a higher lexical diversity and a bigger text, due to the repetition of the words, is associated with lower lexical diversity values. In this particular case, the texts were small and the lexical diversity was on average, greater than 0.8. The reduced text length (about 70 words) justified the weak presence of multi-word expressions and named entities. On the other hand, the presence of adjectives and prepositions was more accentuated than other features, which did not cause any surprise.

The Human-Targeted Translation Edit Rate (HTER) that can give an idea of the human effort required to produce a good translation had a low value, around 0.38. This lower value is justified by the text type and length, as Mails and Customer Support texts tend to be easier to translate and post-edit than other types of texts like articles or even books.

BLEU reached 0.45 on average. This was not a particularly good value since the maximum and desirable value is one, indicating a perfect translation. The score was calculated through a public code.²

A readability formula, Automated Readability Index (ARI) was also included so that we could had an idea of the average comprehension difficulty of the texts. The result (10.8) indicated that a person with 14 to 15 years old or within the ninth grade would be capable of understanding the texts of the MCS corpus.

Table 4.1 gives only a partial idea of the corpus. Next, some parameters are better explored so that a general idea can be formed. Table 4.2 exhibits how the 200 texts are distributed according to their length. The corpus was divided into three categories: small (S), medium (M) and large (L). The first category represented the smaller texts with less than 100 words. The M category represented all texts with 100 to 200 words and finally, category L comprehended all larger texts with more than 200 words. 76 % of the texts from the MCS corpus belonged to S category, followed by 21 % from category M. As the average length of a mail text was approximately 70 words, only 3 % of the corpus belonged to the larger category. Texts with the highest number of words tended to be customer support texts as they contain a different purpose, which can include explaining a procedure to the user.

Table 4.2: Distribution of the MCS texts according to their length.

Category	Size of text	Percentage of texts
S	<100 words	76 %
M	100 to 200 words	21 %
L	>200 words	3 %

For each category, an analysis concerning the distribution of the number of edits made by human editors and the time of edition (in seconds) was performed. Table 4.3 presents the distribution of the number of edits according to each category. In category S, 40 % of the texts had up to 10 edits and only 3 % required more than 50 edits. Moreover, 37 % of texts required more than 20 edits. The number of edits was approximately 10 to 20 % of the text length. In the case of category M, only 9 % of the texts had less than 20 edits and the majority of texts, 47 % required 40 to 60 edits. It turns out that the text

²<https://github.com/rohinisb/BLEUcalculator> (visited on 02/05/2017)

length had, naturally, influence on the number of edits made by the editors. Finally, for texts of category L, the number of edits was once again higher than the previous cases. For larger texts the number of edits reached 50 % of the text length.

Table 4.3: Distribution of the number of edits according to category.

Category S		Category M		Category L	
Interval	Percentage of texts	Interval	Percentage of texts	Interval	Percentage of texts
0 edits	7 %	<20 edits	9 %	<100 edits	50 %
0 - 10 edits	33 %	20 - 40 edits	30 %	100 - 130 edits	33 %
10 - 20 edits	23 %	40 - 60 edits	47 %	>130 edits	17 %
20 - 30 edits	15 %	60 - 80 edits	5 %		
30 - 40 edits	9 %	>80 edits	9 %		
40 - 50 edits	10 %				
>50 edits	3 %				

Table 4.4 represents the distribution of the edit time by category. Only 5 % of the smaller texts took more than 2 minutes to edit, while 81 % took up to 1 minute to edit. Practically all texts that belonged to category M took less than 60 seconds to edit, which indicated that a longer text may require shorter editing time, per word. For the last category considered, which contemplated texts with more than 200 words, the editing time was also reduced. The texts of the MCS corpus apparently do not need more than 2 minutes to be edited and have a good translation quality. Additionally, bigger texts seemed to be related with lower word edition time.

Table 4.4: Distribution of the time of edition according to category.

Category S		Category M		Category L	
Interval	Percentage of texts	Interval	Percentage of texts	Interval	Percentage of texts
<30 sec	47 %	<30 sec	35 %	<30 sec	0 %
30 - 60 sec	34 %	30 - 60 sec	50 %	30 - 60 sec	50 %
60 - 120 sec	14 %	60 - 120 sec	7 %	60 - 120 sec	33 %
>120 sec	5 %	>120 sec	8 %	>120 sec	17 %

The time of edition is given by

$$Time\ of\ edition = time\ of\ reading + actual\ time\ of\ edition, \quad (4.1)$$

meaning that a higher number of words would lead to an increase on the time of edition, not necessarily because the text is harder but due to the extra time spent on reading a larger text. We studied the behavior of the time of edition per word and the result suggested that the time spent to get familiar with the content of smaller texts was higher than for bigger texts. Thus, bigger texts were associated with higher processing speed (less time per word) as can be verified by Figure 4.1.

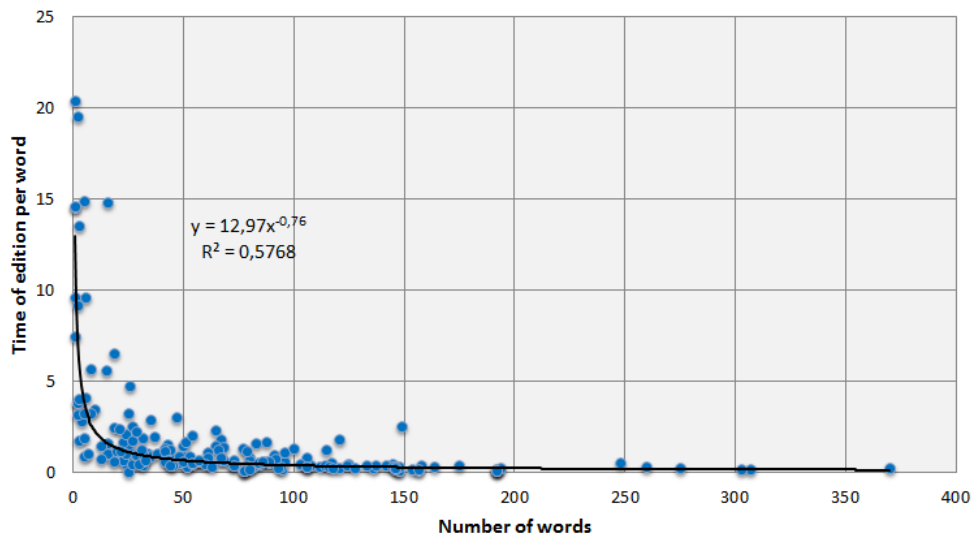


Figure 4.1: Number of words per text vs time per word.

The correlation between the number of words and the time of edition per word showed a R^2 value of 0.58 which indicates that the model accounted for 58 % of the data. This value confirmed the intuition that in a larger text, the time of editing per word is smaller.

4.2.3 Data partition

An expert Spanish annotator has manually classified all 200 source texts concerning their translation difficulty, and has provided a justification for their classification. Because a manual classification is an expensive and time-consuming task, only one annotator has manually classified the texts. The expert Spanish annotator is involved since the beginning in all the annotation process at Unbabel, and is responsible for the development of all the batches in the annotation tool. The goal of this classification was to make every text correspond to a single class, according to single-label classification.

Three classes concerning text translation difficulty were originally considered: Easy, Medium and Difficult:

Easy - Texts that had simple sentences and little content.

Medium - Texts that had some complex structures and a few errors, but did not have a technical context. They could be also texts with fewer words that, due to lack of context, would be quite difficult to translate.

Difficult - Texts with a technical scope and complicated structures that included the presence of many subordinate clauses (also called dependent clauses).

Table 4.5 shows an example for each class considered.

Table 4.5: Examples of texts for each class considered.

Class	Text
Easy	Hi there I have now heard back from xxx to say that they have contacted you with your booking confirmation. Can you confirm this? Thanks, xxx
Medium	Hello xxx, Thanks for your purchase. You have been refunded \$242, and the subscription for your xxx database has been appropriately reverted to 2016-07-30. You will not be charged for updates going forward. Please let me know if I can be of any further assistance. Kind regards,
Difficult	Hello xxx, I am unfortunately not seeing a correction request that covers 190.60.113.0. Could you please submit a request for this network at http://xxxx.com ? Several of your other requests have, however, been accepted. If they are not live after tomorrow's scheduled update, they should be live after next Tuesday. Kind regards,

4.2.4 Balance of the data set

Many "real-world" problems are based on unbalanced class data [6]. An unbalanced data set exists when the number of samples in each class is poorly distributed, leaving one or two classes with the majority of the samples and the rest of the classes with fewer samples.

As shown in Figure 4.2, when considering three classes, a very unbalanced corpus was obtained, with only three texts on the Difficult class. The majority of the texts were assigned to the Easy or Medium classes.

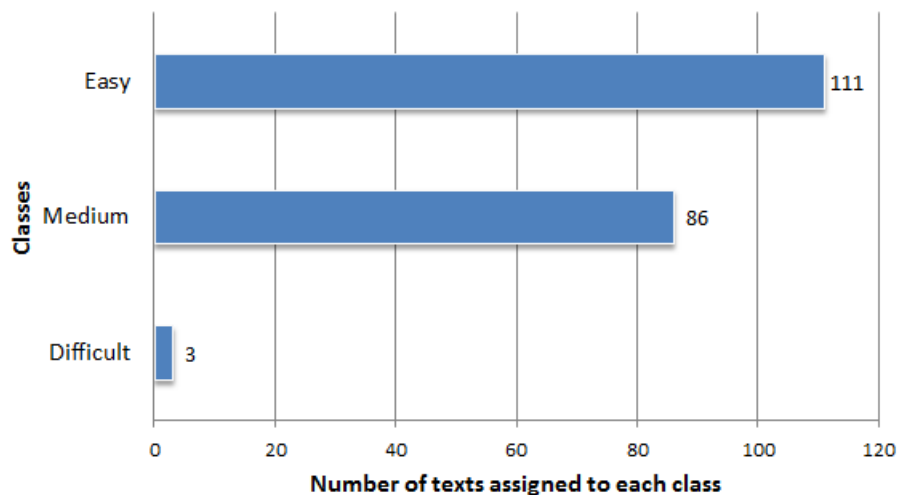


Figure 4.2: Distribution of manually classified texts.

The existence of only three difficult texts made the distribution of the corpus unbalanced, and made it impossible to create a classifier capable of identifying this class (Difficult). Since no more training data was available, we choose to neglect the minority class and reduce the number of available classes from three to two: Easy and Difficult. The distribution according to each class available is exhibited in Figure 4.3.

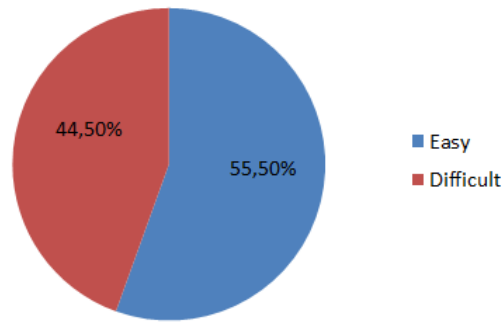


Figure 4.3: Final distribution of classified texts according to the annotator.

After reducing the number of available classes, the distribution of the classified texts became more balanced with 111 easy texts and 89 difficult texts. The 200 texts and their classification were used to train the classifiers and obtain the classification models.

Table 4.6 displays the average values per class of several parameters available on the MCS corpus.

Table 4.6: Average values for parameters of the MCS corpus.

Class	Automated Readability Index	Flesch Reading Ease	BLEU	HTER	Time of edition (sec)	Number of edits
Easy	9.90	66.01	0.46	0.37	33.33	16.69
Difficult	11.92	57.87	0.45	0.39	58.83	39.82

Readability scores of a text as measured by the Automated Readability Index and the Flesch Reading Ease were found to be correlated with text translation difficulty. However, they can not predict the translation difficulty level, as they involve only reading comprehension. Additionally, the time of edition and the number of edits which are commonly used performance measures were found to be related to the level of translation difficulty. It made sense that easy to translate texts were associated with a shorter editing time and a smaller number of edits. Independently of the class, the values of BLEU and HTER were too close and therefore no conclusive conclusions could be drawn, regarding these parameters.

4.3 The AMTA corpus

This section addresses a sentence based corpus concerning topics like climate, Norway, Mexican and software. The data collection is explained in Section 4.3.1 and the data analysis is made in Section 4.3.2. The data partition is explored in Section 4.3.3 and Section 4.3.4 approaches the balance of the data set.

4.3.1 Data collection

The data used in this corpus is from the Association for Machine Translation in the Americas (AMTA) and consisted of 785 sentences that were previously used in other work [33]. The authors explored the effect of different MT systems in the post-edit process, discovering that an MT system with a lower BLEU score implies a higher PE effort and produces a bad quality post-edit output.

The source, machine-translated and post-edited sentences were provided in the corpus. Additionally, the information concerning the editor, the machine translation system, the sentence topic, and the MQM score was also made available. Another MQM score was calculated based on the annotations made by an expert Spanish annotator. This corpus will be named AMTA.

4.3.2 Data analysis

Information regarding the editor, the MT system, the topic and the MQM score was given, for each one of the 785 sentences available. In total, there were nine editors and nine MT systems. All editors had at least 2 years of translation experience and all MT systems had similar BLEU scores. Four different topics namely climate, Mexican, Norway and software were addressed, and there was a single topic per sentence. Table 4.7 shows the sentences distribution by editor and MT system.

Table 4.7: Distribution of sentences according to editor and MT system.

Editor	Percentage of sentences	MT system	Percentage of sentences
TR1	9.2 %	MT1	11.6 %
TR2	10.8 %	MT2	10.6 %
TR3	12.6 %	MT3	9.4 %
TR4	11.0 %	MT4	10.7 %
TR5	11.0 %	MT5	12.0 %
TR6	10.6 %	MT6	10.8 %
TR7	11.0 %	MT7	12.2 %
TR8	12.4 %	MT8	10.6 %
TR9	11.6 %	MT9	12.1 %

The nine editors had about the same amount of sentences to edit, with editor TR3 and TR8 editing the largest number of sentences with 99 and 97 sentences, respectively. MT3 was the least used MT system with only 9.4 % of sentences. On the other hand, MT7 and MT9 were the most used systems having translated respectively, 96 and 95 sentences. In general, there was a balance between all MT systems. As for topics, the corpus had only four and its distribution within the 785 sentences is presented in Figure 4.4.

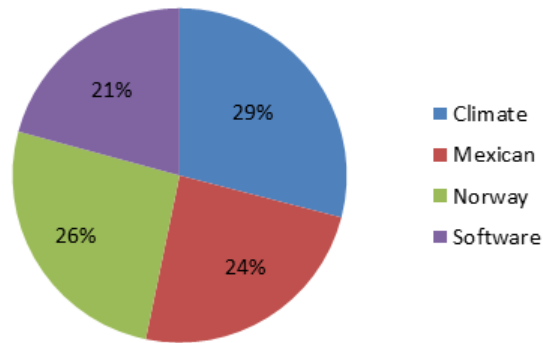


Figure 4.4: Distribution of the four topics available in the AMTA corpus.

The topic of climate was the most discussed in the corpus. Nonetheless, the distribution of all sentences was practically uniform. The topics were unrelated to each other and one example per topic is shown in Table 4.8.

Table 4.8: Example of AMTA sentences per topic.

Topic	Sentence
Climate	Rising temperatures are warming the world's oceans, which provide energy for hurricanes to grow.
Mexican	Besides, we're aware that many of the elite Mexicans in the ruling class don't like us.
Norway	Norway has handled its oil wealth very carefully - all but a small percentage of money from the industry is invested in a special fund for the benefit of future generations.
Software	Hard robots require a sophisticated feedback mechanism to help them determine how much force to apply during surgery so they do not damage our delicate tissues and organs.

From the sentences, another type of information was retrieved. Table 4.9 presents some of the data collected from the AMTA corpus.

Table 4.9: Characteristics of the AMTA corpus.

Feature	Average Value per sentence
Number of words	25.96
Number of syllables per word	1.43
Lexical diversity	0.94
Number of adjectives	2.30
Number of prepositions	3.46
Number of named entities	2.15
Number of Multi-word expressions	0.47

On average, the sentences had 26 words. Sentence length is related to lexical diversity which in this case was relatively high, being on average greater than 0.9.

4.3.3 Data partition

The basic idea for this corpus was to use the MQM scores as a proxy for the translation difficulty. The original MQM scores provided proved to be useless for the classification method we wanted to use. About 90 % of the sentences had an MQM score greater than 90 %, making it impossible to perform a classification based on these scores. Thus, a new annotation of the corpus was made using the Unbabel annotation tool.

Table 4.10 indicates the average MQM scores per editor. We show both the original MQM scores and the ones obtained through the new annotation. In the AMTA paper, all editors had a higher average MQM score than the scores obtained through the new annotation. In the latter case, the sentences were first annotated by an expert Spanish annotator through the Unbabel annotation tool, and later the MQM score was calculated. The annotator was the same that has manually classified the MCS corpus. The average MQM score in the AMTA paper was higher than the average MQM score obtained through the new annotation (85 %).

Table 4.10: Average MQM score per editor.

Editor	Average MQM score AMTA paper	Average MQM score expert annotator
TR1	98.10 %	87.18 %
TR2	96.12 %	83.30 %
TR3	95.96 %	80.59 %
TR4	97.50 %	86.97 %
TR5	96.83 %	84.39 %
TR6	96.55 %	86.79 %
TR7	97.12 %	87.78 %
TR8	97.37 %	86.25 %
TR9	95.01 %	82.84 %
All	96.69 %	85.00 %

This set of 785 sentences was rated for its translation difficulty based on its MQM score. This classification method is a proxy for translation difficulty. It assumes that a lower translation quality is associated with greater translation difficulty. Two classification methods were considered: a method based on the average MQM score of each editor (called AMTA average), and another method based on a given threshold (named AMTA threshold). Next, both methods are described.

As in this particular case there were nine different editors, the sentence classification was performed independently for each editor, so that the quality of the editor did not interfere with the classification.

Based on the average MQM score per editor, two classes were created:

Easy - When the sentence MQM score was greater than the average MQM score of the editor.

Difficult - When the sentence MQM score was lower than the average MQM score of the editor.

Table 4.11 illustrates the average classification method considering two editors.

Table 4.11: AMTA average classification method illustration.

Editor	Average MQM score	Sentence MQM score	Sentence	Class
TR1	87.18 %	91.30 %	Grey swans are just as bad as black swans, but they can be predicted to some degree thanks to new modeling techniques.	Easy
		56.52 %	It's possible that the storm surges from a grey swan could reach up to 23 feet in Dubai and 36 feet in Tampa.	Difficult
TR2	83.30 %	94.74 %	And Pena Nieto vowed to continue Mexico's war against the drug cartels, even though he offered no specifics.	Easy
		77.42 %	People here are still loath to mention by name Anders Breivik, the right-wing, racist extremist who gunned down and killed 77 men, women and children last year.	Difficult

In the case of the threshold method, a threshold of 85 % was considered. If the sentence had an MQM score greater than 85 %, it was considered easy, and if the score was lower than 85 %, it was classified as difficult. This threshold value was chosen for being the average MQM score obtained through the new annotation. Other threshold values were explored, but the selected value was what made the most sense for our analysis, considering that an MQM score greater than 95 % equals professional translation quality. The aim was that a sentence that did not had professional translation quality could still be considered easy. Thus, a threshold lower than 95 % was required.

Table 4.12 gives the distribution of the sentences according to the translation difficulty rating and the sentence topic. The topic was not used as a feature due to its scarcity.

Table 4.12: Sentences distribution according to translation difficulty rating and topic.

Class	Climate	Mexican	Norway	Software
Easy	72 %	75 %	69 %	71 %
Difficult	28 %	25 %	31 %	29 %

It did not appear to be a direct relationship between the topic of the sentence and its translation difficulty. Most of the sentences belonged to the class - Easy, and the distribution was very similar for all topics. The HTER mean values for the two classes were also calculated. The corpus presented an average value of 0.42 for the class - Easy and 0.50 for the class - Difficult. Thus, there was a relationship between HTER and translation difficulty rating, which associates less effort with easy to translate sentences.

4.3.4 Balance of the data set

The balance of the corpus can affect the classification results. When a machine learning algorithm is applied to an unbalanced corpus, higher accuracy values tend to be obtained. However, this does not mean that the model is good, since the algorithms tend to choose the majority class. In these cases, is appropriate to check other evaluation parameters such as precision, recall, F-measure and the confusion matrix. These parameters give more information about the model's performance and allows you to check for adjustments to be made.

Thus, Figure 4.5 shows the distribution of the 785 classified sentences according to the AMTA average classification method, and Figure 4.6 presents the distribution according to the AMTA threshold classification method.



Figure 4.5: AMTA average classification method distribution.

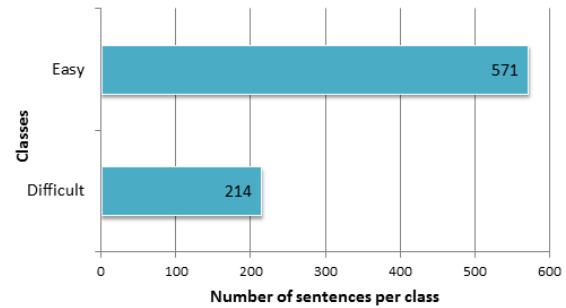


Figure 4.6: AMTA threshold classification method distribution.

Through both classification methods the vast majority of the sentences were classified as easy to translate and only a small portion was classified as difficult, raising the problem of the unbalanced class data. There are several techniques to outpace this problem, namely under-sampling and oversampling. Under-sampling is a technique that randomly selects a sample from the majority class and eliminates it, until the number of instances in both classes be the same. Thus, this technique has an associated problem which is, the loss of potentially relevant information from the left out samples. In the case of oversampling, the instances of the minority class are duplicated until the number of samples of the largest class is reached. Alternatively, you can create additional instances based on the data, so that they are not copies of the existing ones. In this case, the problem is the risk of overfitting.

In this work, an over-sampling technique, the Synthetic Minority Oversampling technique (SMOTE) was applied [7]. SMOTE is an oversampling approach in which the minority class is over-sampled by creating "synthetic" examples rather than oversampling with replacement. The synthetic samples are created as follows: take the difference between the sample under consideration and its nearest neighbor, multiply this difference by a random number between zero and one, and add it to the sample under consideration. No sentences are generated, only synthetic feature values.

The SMOTE was applied to both AMTA average and AMTA threshold classification methods output. This filter created "synthetic" examples, increasing the corpus length. The AMTA corpus, classified according to the average MQM score per editor increased from 785 to 1116 sentences. 331 sentences that belonged to the minority class: Difficult were created. Figure 4.7 presents the distribution of both classes for the AMTA average case. The AMTA corpus classified according to the 85 % MQM score threshold have also increased. After applying the SMOTE, its length achieved the 1106 sentences. This technique was applied using WEKA,³ and later the randomize filter was also applied. This filter changes the order of the samples so that cross-validation results are not affected. The final distribution of AMTA threshold is shown in Figure 4.8.

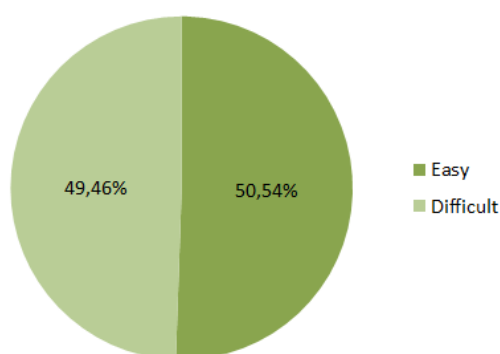


Figure 4.7: Final distribution of the AMTA average classification method.

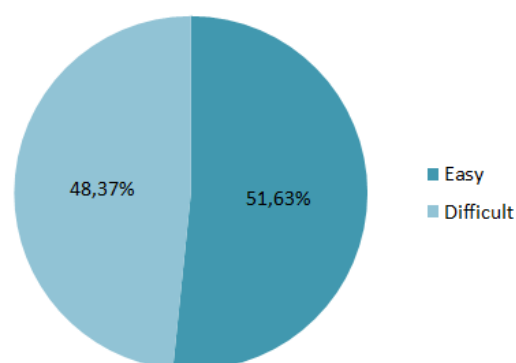


Figure 4.8: Final distribution of the AMTA threshold classification method.

After applying the SMOTE filter, the AMTA corpus achieved a higher level of equilibrium, regardless of the classification method used. Both classification methods were taken into account in the experiments and are addressed as AMTA average and AMTA threshold, for a better understanding.

³<http://www.cs.waikato.ac.nz/ml/weka> (visited on 20/02/2017)

Chapter 5

Preliminary Experiments

In order to discover what are the linguistic parameters responsible for making a text and a sentence harder to translate, two preliminary experiments were conducted. The idea was to acquire the best linguistic parameters for later application in the development of the classifiers, and understand the most common type of errors obtained at the machine translation system output.

The first experiment concerning the most common errors made by machine translation is presented in Section 5.1. The other experiment that tries to understand the impact of human edition in machine translation errors is mentioned in Section 5.2.

5.1 Most common errors made by Machine Translation

In order to understand what parameters may influence the difficulty of translating a sentence from English to Spanish, an investigation on the most common errors in MT system Google translate (free, easy to use) was carried out.

The WMT-News corpus described in Chapter 4 was used in this first experiment. The corrected WMT-News corpus (Appendix B.1) was analyzed and the most common errors made by Google translate included: no subject-verb agreement, lack of articles before the name of political parties, adoption of wrong verb form, usage of wrong prepositions and cases of literal translation. These findings were in agreement with the research from 2014 [13] which showed that the most common errors made by MT belonged to grammar and syntax categories. This research used a corpus of four newspaper articles. Figure 5.1 presents the most common errors in MT according to [13], and Figure 5.2 shows the most common errors in MT obtained with the WMT-News corpus.

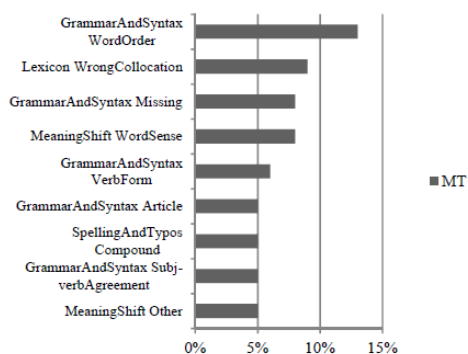


Figure 5.1: Most common errors made by MT [13].

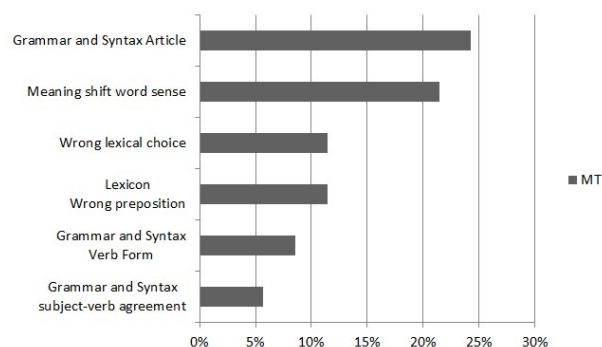


Figure 5.2: Most common errors made by MT in WMT-News corpus.

According to the authors [13], grammar and syntax have a huge impact on MT quality, combining the most common errors in the MT process. The most frequent error is the word order that achieves almost 15 % of all existing errors, while the nine most common errors combined cover almost 70 % of total MT errors. In the case of the WMT-News corpus, grammar and syntax errors were largely superior to any other error types. When comparing the results, one may observe that in the case of Figure 5.2, the articles and the meaning shift word sense represented almost 50 % of the total number of errors, whereas in Figure 5.1 the same errors added up to only 15 %.

It is important to note that the research referred to dates from 2014. Since the last years have brought a huge advance in the field of machine translation, the most common errors made by MT identified above may no longer make sense, when using the most modern systems such as DeepL translator. It innovates by combining a supercomputer with a breakthrough in neural machine translation. DeepL's neural networks train on billions of high-quality translated sentences provided by the search engine Linguee. "The translation quality is unmatched, with a huge drop in translation errors compared to the competition".¹

5.2 Impact of human edition in machine translation errors

The data annotated through the Unbabel annotation tool described in Chapter 3, refers to 30 jobs with a total of 2800 words. The jobs corresponded to translations from English to Spanish within the period between 04-04-2016 and 01-05-2016, regarding Mails and Customer Support texts. This data was a subset of texts, which we did not have access to. The data contained information about the errors, namely their type and distribution according to the severity scale described in Section 3.3. Additionally, the frequency of each error was provided. The annotations were made at two different points of the translation process: at the output of the machine translation system and after the first human edition.

The machine translated texts had a fluency of 1.32 while the post-edit texts had a 2.97 value of fluency on a 0-5 scale. As expected, the human edition was able to increase text fluency. However, the value can still be increased through another human edition.

¹<https://www.deepl.com/press.html> (visited on 14/09/2017)

Although the annotation tool described in Section 3.2 had seven categories, this data only had errors from five of those categories. For each of these categories, a table including the information regarding the error frequency and its severity for both scenarios considered: machine translated texts and the post-edited texts is showed. For each category the information about the total number of minor, major and critical errors is also provided. The information concerning accuracy errors is shown in Table 5.1.

Table 5.1: Accuracy error types frequency according to severity.

Accuracy error types	Machine translated texts			Post-edited texts		
Mistranslation	Minor	Major	Critical	Minor	Major	Critical
Overly Literal	70	17	4	40	2	0
False Friend	0	14	4	0	2	0
Lexical Selection	32	24	2	47	5	0
Omission	0	2	0	1	0	0
Untranslated	0	3	0	0	2	0
Addition	0	2	0	0	0	0
Total	102	60	10	88	11	0

Accuracy errors have to do with the translation of the meaning. The machine translated texts had 172 accuracy errors where the overly literal and the lexical selection were the most common error types. Although most of the errors had minor severity, there were still 60 major and 10 critical errors. In the post-edited texts, the number of errors dropped drastically to 99 and the critical errors were completely eliminated. The addition error seemed to be the easiest to deal with, since it was the only error type totally removed by the editor. The data concerning fluency errors is presented in Table 5.2.

Table 5.2: Fluency error types frequency according to severity.

Fluency error types	Machine translated texts			Post-edited texts		
Inconsistency	Minor	Major	Critical	Minor	Major	Critical
Word Selection	0	2	0	0	0	0
Tense Selection	0	1	0	0	0	0
Coherence	0	11	17	2	6	0
Spelling						
Orthography	6	0	0	5	1	0
Capitalization	24	0	0	4	0	0
Typography						
Punctuation	8	0	0	8	0	0
Unpaired Quote Marks and Brackets	0	7	0	0	2	0
Whitespaces	0	0	0	3	0	0
Grammar						
Function Words						
Prepositions	11	7	0	4	0	0
Conjunctions	4	6	0	5	0	0
Determiners	3	2	0	1	0	0
Word Form						
Part-of-Speech	5	2	0	1	1	0
Agreement	0	17	1	0	11	0
Tense/Mood/Aspect	18	26	1	19	3	0
Word order	2	6	1	1	1	0
Sentence Structure	8	4	0	17	1	0
Total	89	91	20	70	26	0

Fluency errors affect the reading and the good comprehension of the text. The most common critical errors were coherence errors, that were easily eliminated by a single human editor. The inconsistency errors (word and tense selection) were completely eliminated and the number of the others error types was drastically reduced through a single human edition. The machine translated texts had strong problems of coherence, capitalization, agreement and tense/mood/aspect. Although most of those errors had a minor severity, the agreement and tense/mood/aspect error types were the most problematic, and the edition by the human editor was not totally capable of eliminating these errors. From the 200 initial fluency errors, 104 were eliminated but numerous errors have still to be managed. Additionally, three whitespaces errors were added by the editor. This type of error presented a minor severity but was still noticed in the post-edited text. Table 5.3 exhibits the collected information about style errors.

Table 5.3: Style error types frequency according to severity.

Style error types	Machine translated texts			Post-edited texts		
	Minor	Major	Critical	Minor	Major	Critical
Inconsistent Register	0	11	0	0	5	0
Repetitive Style	2	1	0	0	0	0
Total	2	12	0	0	5	0

Style errors concern mainly register problems. This type of errors was not common on this data and there was no registry of a single critical style error. The repetitive style error was easily sorted out by the editor and from the 14 initial errors, only five inconsistency register errors were not eliminated through the human edition. The wrong language variety errors are presented in Table 5.4.

Table 5.4: Wrong language variety error types frequency according to severity.

Wrong language variety error types	Machine translated texts			Post-edited texts		
	Minor	Major	Critical	Minor	Major	Critical
Wrong language, variety	0	0	0	0	2	0
Total	0	0	0	0	2	0

Wrong language variety includes for instance, the usage of Latin American Spanish terms and expressions instead of European Spanish ones. Curiously, this type of error did not occur in the machine translated texts but it was created by the human editor. This can be related with the editor origin and the better solution would be to get another editor, and perform a new edition. The named entities errors are indicated in Table 5.5.

Table 5.5: Named entities error types frequency according to severity.

Named Entities error types	Machine translated texts			Post-edited texts		
	Minor	Major	Critical	Minor	Major	Critical
Person	0	0	2	0	0	0
Total	0	0	2	0	0	0

Named entities errors concern seven types: person, organization, location, function, product, amount and time. This specific data presented only person errors: when the name of the person has/has not been translated when it should/should not have been or is incorrect. The presence of only one named entities errors type can be related to the text type included in the 30 jobs considered, and that may also explain why a critical severity had been chosen.

After all error types from five different categories have been shown, Table 5.6 presents a resume with the total number of errors per category and severity, in order to understand which categories have a higher number of errors and verify the impact of the human edition on the total number of errors.

Table 5.6: Total number of errors per category and severity.

Categories of errors	Machine translated texts			Post-edited texts		
	Minor	Major	Critical	Minor	Major	Critical
Accuracy errors	102	60	10	88	11	0
Fluency errors	89	91	20	70	26	0
Style errors	2	12	0	0	5	0
Wrong language variety errors	0	0	0	0	2	0
Named Entities errors	0	0	2	0	0	0
Errors per severity	193	163	32	158	44	0
Total number of errors	388			202		

The number of annotated errors was high and not evenly distributed along the five different error categories. Accuracy and fluency categories had the highest number of errors and included almost all critical ones. The other three categories had fewer errors and most of them were corrected by the editor. Through the human edition process the critical errors were totally eliminated and the number of errors decreased in almost every category, with the exception of the wrong language variety category that had no errors in the machine translated texts and two major errors were added through the edition process. The total number of errors decreased, going down from 388 to 202, corresponding to a 48 % error reduction after the first human edition. This number was not good enough and therefore, it would be useful to do further edition by another human editor in order to decrease the number of errors and consequently, increase the translation final quality.

Chapter 6

Feature extraction and classifiers

This chapter presents the architecture of a supervised machine learning problem, explores the features extracted from both corpora and describes the machine learning algorithms used. In Section 6.1 the text and sentence classification modules are indicated. The different features extracted from the MCS corpus are presented in Section 6.2 and Section 6.3 explores the features extracted from the AMTA corpus. Finally, Section 6.4 describes the machine learning algorithms used.

In supervised machine learning problems the architecture is divided in two modules: training and prediction module. A classification model is generated by using the pre-classified input (text or sentence) and the information provided by the extracted features. Then, when a new input is provided, the same features are extracted and the classification model generated in the training module is applied, producing a prediction label for the input text or sentence. The training and prediction modules schemes are exhibited in Figure 6.1 and Figure 6.2, respectively.

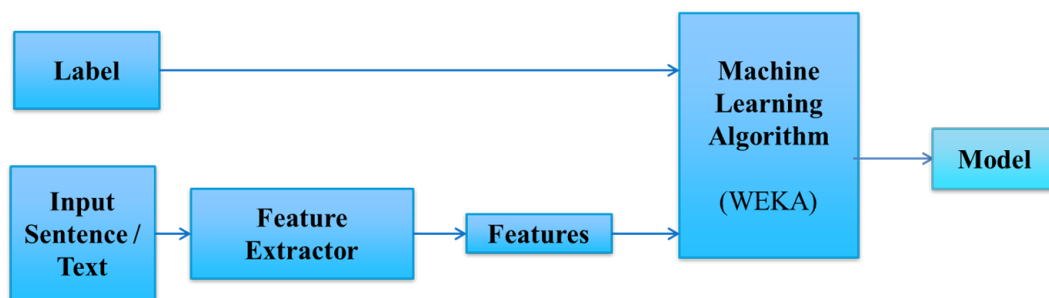


Figure 6.1: Training module.

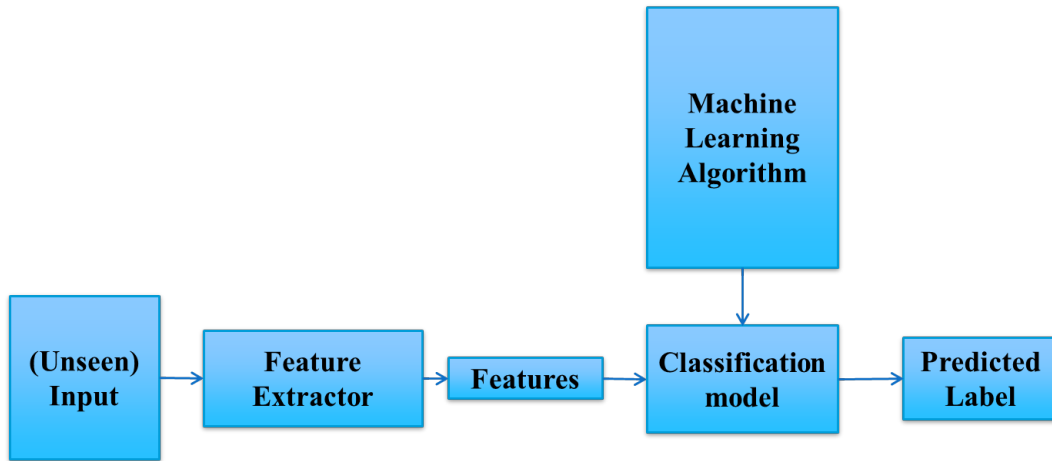


Figure 6.2: Prediction module.

In order to visualize the study concerning the translation difficulty as a supervised machine learning problem, some data need to be gathered. The required data included the corpora, the labels and the features. The corpora were presented in Chapter 4. The way in which the labels were obtained and the features were selected, is indicated in the following sections.

6.1 Text and sentence classification module

The problem of classifying a text or a sentence as easy or difficult to translate to another language was considered as a supervised machine learning problem, where the correct label was required for all the input data.

The corpora described in Chapter 4 were classified differently. The MCS corpus, composed of 200 texts about Mails and Customer Support was manually classified as to their translation difficulty, by an expert Spanish annotator. On the other hand, the AMTA corpus focused on four different topics: climate, Mexican, Norway and software, was classified depending on the sentence MQM score. The AMTA corpus was divided in two: AMTA average and AMTA threshold, depending on the classification method used. Recalling, Table 6.1 indicates the distribution of the classes considered, according to the corpora type. All corpora showed a similar balance level.

Table 6.1: Distribution of classes for each corpora type.

Corpus Type	Corpus name	Easy	Difficult
Text based	MCS	55.50 %	44.50 %
Sentence based	AMTA average	50.54 %	49.46 %
Sentence based	AMTA threshold	51.63 %	48.37 %

6.2 Text analysis module

This section explores the features extracted from the MCS corpus. Two feature modules were extracted: the features suggested and indicated by the annotator (annotator benchmark) and the QuEst++ baseline.

Annotator benchmark features

The annotator benchmark features were selected based on the suggestions of the annotator who manually classified the MCS corpus. This feature set included the following 19 features:

- Average word length (syllables) in the source text;
- Average word length (syllables) in the target text;
- Average word length (characters) in the target text;
- Percentage of adjectives in source text;
- Percentage of adjectives in target text;
- Percentage of named entities in source text;
- Percentage of multi-word expressions in source text;
- Percentage of prepositions in source text;
- Percentage of prepositions in target text;
- Lexical diversity in source text;
- Lexical diversity in target text;
- Percentage of verbs in source text;
- Percentage of verbs in target text;
- Number of sentences in source text;
- Number of sentences in target text;
- Automated Readability Index of source text;
- Flesch Reading Ease of source text;
- Average number of dependencies in the source text;
- Source and target word count ratio.

The average word length was measured both in terms of characters and syllables. This distinction was made in order to understand if the type of measure had some influence on the results. Using Part-of-Speech (PoS) tags, it was possible to calculate the percentage of adjectives, prepositions and verbs of every text.

In order to detect named entities, the Stanford Named Entity Recognizer that considers seven classes: location, person, organization, money, percent, date and time was used. Afterwards, the named entities founded were manually verified.

Multi-word expressions (MWE) are very important in the translation process. An MWE can be considered and translated as an unique "word", instead of being divided in several words with no relation attached. As the presence of MWE can affect the text/sentence translation difficulty, it is important to identify them. An MWE in a source language may not have a direct lexical equivalence in the target language. The best approaches to translate MWE involve example based MT, because in this case each MWE can be listed as an example with its translation in the target language. An estimation of the number of MWE was calculated through a protocol based on [34], which is the following: First get the top 10,000 most popular words/phrases searched for by Linguee (world's largest translation search engine) users, then queries wordnik (online English dictionary and language resource that provides dictionary and thesaurus content) for the definition of each one. Admit that any phrase without a definition is a Multi-word expression. Finally, search MWE in source text or sentence with this MWE database.

Lexical diversity gives information about the variety of words that exists in the corpus and it was calculated by dividing the number of distinctive words by the text or sentence length. A higher diversity may increase the text and sentence difficulty.

The version 0.3.1 of *textstat* package was used to extract the Automated Readability Index and the Flesch Reading Ease values. Finally, the number of dependencies was calculated through the Stanford Dependency Parser.

QuEst++ baseline features

In addition to the annotator benchmark features, the QuEst++ baseline features were also extracted. QuEst++ is an open source software developed by professor Lucia Specia's team at the University of Sheffield and contributions from a number of researchers. It has two main modules: a *Java* module to extract a number of word-, sentence-, and document-level features, and a *Python* module that interacts with the scikit-learn toolkit for machine learning.

In this work the *Java* module was adopted to extract the baseline features. Although QuEst++ is an open source tool for translation quality estimation, some of their features can be interesting for the translation difficulty estimation problem. The features were extracted through the extractor provided.¹

¹<https://github.com/ghpaetzold/questplusplus> (visited on 03/07/2017)

The 17 QuEst++ baseline features included:

- Number of tokens in the source text;
- Number of tokens in the target text;
- Average source token length;
- LM log probability of source text;
- LM log probability of target text;
- Number of occurrences of the target word within the target hypothesis (averaged for all words in the hypothesis - type/token ratio);
- Average number of translations per source word in the text (as given by IBM 1 table thresholded such that $\text{prob}(t|s) > 0.2$);
- Average number of translations per source word in the text (as given by IBM 1 table thresholded such that $\text{prob}(t|s) > 0.01$) weighted by the inverse frequency of each word in the source corpus;
- Percentage of unigrams in quartile 1 of frequency (lower frequency words) in a corpus of the source language (SMT training corpus);
- Percentage of unigrams in quartile 4 of frequency (higher frequency words) in a corpus of the source language;
- Percentage of bigrams in quartile 1 of frequency of source words in a corpus of the source language;
- Percentage of bigrams in quartile 4 of frequency of source words in a corpus of the source language;
- Percentage of trigrams in quartile 1 of frequency of source words in a corpus of the source language;
- Percentage of trigrams in quartile 4 of frequency of source words in a corpus of the source language;
- Percentage of unigrams in the source text seen in a corpus (SMT training corpus);
- Number of punctuation marks in the source text;
- Number of punctuation marks in the target text.

6.3 Sentence analysis module

This section refers to the features extracted from the 785 sentences of the AMTA corpus. The remaining sentences that were created synthetically through SMOTE had their own feature values, created synthetically. We extracted two feature modules: the features indicated by the annotator (annotator benchmark) and the QuEst++ baseline.

Annotator benchmark features

The annotator benchmark features extracted from the AMTA corpus were not exactly the same as those extracted from the MCS corpus. Due to the existence of some features that did not make sense to be applied to the sentence level, only 15 features were extracted. The number of sentences and the features related with the readability metrics were not extracted from the AMTA corpus. All features were extracted using code developed in *Python* or the public resources already mentioned. These features were:

- Average word length (syllables) in the source sentence;
- Average word length (syllables) in the target sentence;
- Average word length (characters) in the target sentence;
- Percentage of adjectives in source sentence;
- Percentage of adjectives in target sentence;
- Percentage of named entities in source sentence;
- Percentage of multi-word expressions in source sentence;
- Percentage of prepositions in source sentence;
- Percentage of prepositions in target sentence;
- Lexical diversity in source sentence;
- Lexical diversity in target sentence;
- Percentage of verbs in source sentence;
- Percentage of verbs in target sentence;
- Number of dependencies in source sentence;
- Source and target word count ratio.

QuEst++ baseline features

The feature extractor module provided by QuEst++ is capable of extracting sentence-level features. Thus, the QuEst++ baseline features extracted from the MCS corpus were also extracted from the AMTA corpus.

6.4 Weka and Algorithms

The goal of our classifiers was to classify a text and a sentence as being easy or difficult to translate from English to Spanish. In order to test several machine learning algorithms without wasting too much time, Weka was used. Weka is a collection of machine learning algorithms for data mining tasks. The machine learning algorithms applied in the experiments are shown in Table 6.2. Next, the six algorithms are briefly explained.

Table 6.2: Machine Learning Algorithms used in Weka.

Learning Method	Algorithm
Bayes	Naive Bayes
Linear	Support Vector Machine (SVM) Multilayer Perceptron
Lazy	K-Nearest Neighbor (KNN)
Rules	ZeroR
Decision tree	REPTree

6.4.1 ZeroR

ZeroR is the simplest algorithm that relies on target and ignores all predictors. It stands for zero rules because in fact, no rule is applied. It does not have any predictability power but it is commonly used to create a baseline performance as a benchmark for other classification methods. It always predicts the majority class through the development of a frequency table for class target. In this work, the ZeroR algorithm was used as a baseline.

6.4.2 Naive Bayes

Naive Bayes is an algorithm based on the Bayes theorem. The Naive term comes from the assumption made by the algorithm that there are no dependencies between the features. It assumes that each feature is independent of another for a given class variable, even when the dependency is clear. Because of this assumption this algorithm requires fewer data than other algorithms.

It is an easy and fast algorithm for predicting the class of a test data set, it works well on multi-class prediction and when the assumption of independence holds, generally yields a better performance than most of the algorithms.

6.4.3 Support vector Machine

Support vector Machine (SVM) is a linear algorithm that tries to find the best separation line between different classes. In some cases, a line is not enough and the kernel trick is used to take low dimensional input space into a higher dimensional space where a hyper-plane can be found. When classifying, the idea is to assure the best classification possible by ensuring, at the same time, the largest margin feasible.

For multi-class classification, the problem is divided into several binary classification problems so that standard SVM algorithm can be applied.

SVM is an effective algorithm in high dimensional spaces and memory efficient by using a subset of training points into the decision function (support vectors). Nonetheless, it is quite slow for large data sets.

6.4.4 Multilayer Perceptron

The Multilayer Perceptron (MLP) is a feedforward artificial neural network model. It is constituted by an input layer, an output layer and a hidden layer that can contain several layers itself. Each node of the hidden layer is a function of nodes from the input layer and the output is a linear function of nodes from the hidden layer. Each layer is fully connected to the next one because MLP constructs a directed graph, where all nodes need to be connected.

All nodes, with the exception of the input ones, are neurons with nonlinear activation functions. One of the most common activation functions includes the sigmoid and the hyperbolic tangent function. The backpropagation learning algorithm is needed to train the MLP and the weights are learned from the training set. They are calculated in order to minimize the error using the Gradient Descent method. MLP is capable of solving problems which are not linearly separable.

6.4.5 REPTree

REPTree stands for Reduced Error Pruning Tree and is a fast decision tree learner. It splits data into a growing set and a pruning set, builds a decision tree and prunes it using the mean squared error (MSE) on the predictions made. A tree is created in every iteration and then, the best one is selected.

6.4.6 K-Nearest Neighbor

The model representation for K-Nearest Neighbor (KNN) is the entire training data set and the K value is determinant in the classification. The K value is only determined by trial and error and depends on the problem. If K is too small, the method is susceptible to noise in the data and the model is much more dependent on the particular training examples chosen. A larger K yields smoother decision regions and provides probabilistic information. The selection of the K parameter is not easy. It should be large enough to avoid overfitting but small enough to avoid oversimplifying the distribution.

The distance weighting method allows weighting the contributions of each of the K neighbors according to their distance to the query instance x_q , giving greater weight to closer neighbors. Weight by $1/\text{distance}$ weight the vote of each neighbor according to the inverse of its distance from x_q . The weight by $1/\text{distance}$ method weight the vote of the neighbor according to its similarity with x_q , where similarity is defined as

$$w_i = 1 - d(x_q, x_i). \quad (6.1)$$

KNN is included in the lazy learning method because no learning model is required and all work happens when a prediction is requested. It is a non-parametric machine learning algorithm as it makes no assumption on the functional form of the problem. The computational complexity increases with the size of the data set.

Chapter 7

Experimental results

This chapter intends to introduce the evaluation parameters used in the classifiers evaluation and to explore the results obtained for all the experiments made. The selected parameters to evaluate all the classifiers performance are introduced in Section 7.1. In Section 7.2, the results obtained regarding the baseline experiments are shown. The results concerning the most relevant features to be used are presented in Section 7.3. Section 7.4 exhibits the results of the cross-corpora experiments. Lastly, the results are discussed in Section 7.5.

7.1 Evaluation parameters

To evaluate the performance of the classifiers the accuracy, precision, recall, F-measure, Root Mean Square Error (RMSE), Area Under the Receiver Operator Curve (AUC) and Kappa statistic were selected. Additionally, the confusion matrix was also used.

Accuracy gives the percentage of instances correctly classified by dividing the number of instances correctly classified by the total number of instances. The accuracy can be calculated through

$$Accuracy = \frac{tp + tn}{tp + fn + fp + tn} \quad (7.1)$$

where

tp = true positive

fn = false negative

fp = false positive

tn = true negative.

The value of the accuracy alone is not enough to evaluate the performance of the classifier, especially when there are unbalanced class data. In this case, a classifier can achieve 90 % accuracy simply by always selecting the majority class. Thus, although it is set in most cases, leading to a high value of accuracy, the model found can not correctly classify the instances of the minority class. This is called the accuracy paradox. Thus, it is important to look at other parameters such as precision and recall.

Precision and recall enables the evaluation of the model performance in the presence of unbalance class data. Precision measures the certainty of a positive test result and can be seen as a measure of classifier exactness. It can be calculated through

$$Precision = \frac{tp}{tp + fp}. \quad (7.2)$$

A low precision value can indicate a large number of false positives.

Recall measures the amount of true positives that are captured by the model and can be thought as a measure of classifiers completeness. The formula of the recall is similar to the precision but takes into account the false negatives instead of the false positives.

$$Recall = \frac{tp}{tp + fn} \quad (7.3)$$

A low recall indicates many false negatives. As accuracy and recall are both important parameters for disambiguate accuracy, a parameter has been created that allows the balance between the precision and the recall to be conveyed. That parameter is called F-measure and is calculated through

$$F_measure = \frac{2 \times Precision \times Recall}{Precision + Recall}. \quad (7.4)$$

RMSE measures the difference between the observed values and the expected values. It penalizes the most serious errors, that is, the errors that are further from the expected value. Equation 7.5 presents the formula for calculating the RMSE where $e = (\text{observed value} - \text{expected value})$.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=0}^n e_i^2} \quad (7.5)$$

AUC gives an estimated probability that evaluates the performance of the classification based on the area below the ROC curve. The ROC curve represents the relationship between true positives and false positives and it can be interpreted as the probability of a classifier hit (true positive) instead of failing the classification (false positive) considering a random example. In order to compare different classifiers, the Area Under the Receiver Operator Curve is used since it summarizes the results of the ROC curve through a single value. The AUC value is related to the classifier performance and therefore should be as large as possible.

Kappa statistic evaluates the agreement between the classifier results and the expected ones. A common cited scale [42] for kappa statistic is presented in Figure 7.1.

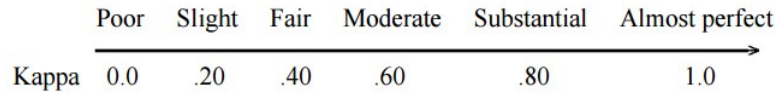


Figure 7.1: Kappa statistic scale.

The confusion matrix manages to summarize the performance of a classifier by giving an idea of the errors of which the classification model is making in making predictions. From this matrix, it is possible to check the most common error types and adjust the parameters used during the training. In this way, the application of this representation can help to consider other hypothesis and to improve the final performance of the classifier. Table 7.1 presents the confusion matrix representation.

Table 7.1: Confusion matrix.

Confusion Matrix		Gold standard	
		Easy	Difficult
Outcome	Easy	TP	FP
	Difficult	FN	TN

Taking into account the goal of this work, building a classifier capable of classifying a text and a sentence with respect to its translation difficulty, different experiments were conducted. The experiments were all performed on Weka and used cross-validation. The K-fold cross-validation splits the observations into K sets with the same length. Then, it uses (K-1) sets to train and one set to test. The process is repeated K times. One of the biggest advantages of cross-validation is that all observations are used in training and validation, and each observation is used for validation exactly once. The k-fold cross-validation estimator has a lower variance than a single hold-out set estimator, which is important when the data is limited. In a single hold out set, where 90 % of the data would be used for training and the remaining 10 % for testing, the test set would have very small dimensions, so there would be a greater variance in the performance estimated for different samples of data.

7.2 Baseline experiments

The goal of the baseline experiments was to determine which machine learning algorithm and feature set enabled the best results, for each corpus type: text and sentence based. Thereby, all corpora described in Chapter 4 and different classifiers were tested considering three possible feature sets: a set with the annotator benchmark features, a set with the QuEst++ baseline features, and a set with all features. A 5-fold and a 10-fold cross-validation were selected as test options for the classifiers. As the AMTA corpus was divided in two: AMTA average and AMTA threshold, the choice of the best corpus at the sentence level was made from the results of these baseline experiments. Table 7.2 presents the best classifiers results considering each feature set and corpus.

Table 7.2: Classifier results for each feature set and corpus.

Feature set	Annotator benchmark features	QuEst ++ baseline features	All features
MCS			
Accuracy	69.50 %	74.50 %	72.00 %
Classifier	SVM	SVM	SVM
Folds on cross-validation	10	10	10
AMTA average			
Accuracy	75.72 %	75.72 %	75.99 %
Classifier	KNN	KNN	KNN
Folds on cross-validation	10	10	10
AMTA threshold			
Accuracy	76.31 %	76.04 %	77.12 %
Classifier	KNN	KNN	KNN
Folds on cross-validation	10	10	10

In the case of the MCS corpus, the best machine learning algorithm was always the Support Vector Machine. The best accuracy score, 74.50 % was achieved applying only the QuEst++ baseline feature set and a 10-fold cross-validation. We concluded that the annotator benchmark features only made the classifier performance worse, and that the features used in the quality estimation task were in line with the text translation difficulty.

For the AMTA corpus, the results did not change significantly depending on the feature set considered, always obtaining accuracy values around 76 %. Nonetheless, the best outcome was found when considering all features. For both corpora, the cross-validation may be tested considering $K = 10$, and the best algorithm was the KNN. Particularly, for the AMTA average corpus the accuracy was 75.99 %. The AMTA threshold corpus can achieve a higher value, reaching 77.12 % of accuracy.

The results of both sentence level corpus were superior to the results of text level corpus, suggesting that MQM scores, used in the sentence classification method, can be used to assess translation difficulty. In all cases, the best classifier exceeded the reference results obtained through the ZeroR algorithm. The reference accuracy was in the order of 50 %, and this was largely surpassed by the results indicated in Table 7.2.

Table 7.3 indicates the values of the evaluation parameters described in Section 7.1, for the best classifier per corpus presented in Table 7.2. For better judgment of the results, Table 7.4 shows the data of the confusion matrices.

Table 7.3: Evaluation parameters values for the best classifier of the baseline experiments.

Corpus	Accuracy	Precision	Recall	F-measure	RMSE	AUC	Kappa statistic
MCS	74.50%	0.75	0.75	0.74	0.51	0.73	0.47
AMTA average	75.99%	0.76	0.76	0.76	0.48	0.76	0.52
AMTA threshold	77.12%	0.77	0.77	0.77	0.47	0.78	0.54

Table 7.4: Confusion matrix for the best classifier of the baseline experiments.

Corpus	TP	FN	FP	TN
MCS	95	16	35	54
AMTA average	438	126	142	410
AMTA threshold	447	124	129	406

The text based corpus had more difficulties in correctly classifying difficult texts, presenting 35 false positives and a few false negatives. It had a fair agreement between the classifier results and the expected ones by showing a 0.47 value for the Kappa statistic. The F-measure value was relatively low and the RMSE was greater than 0.5. The difficulty of predicting the difficult texts can be related to the distribution of the data, as there were more easy than difficult to translate texts in the corpus. Note that only three texts were manually annotated as difficult to translate.

Regarding sentence based corpora, the best result appeared in the AMTA threshold corpus. Corpora differ only in the adopted classification method. Both corpora showed similar Kappa statistic values and they indicated also a fair agreement. Since the AMTA threshold presented the best results at the sentence level, from now on only the results of the MCS and AMTA threshold corpus, the best corpus at the text and sentence levels respectively, will be stated.

As already mentioned, the sentence based corpus yielded the best results. This could be due to having more data in the AMTA corpus, or simply because the classification method was more controlled. This method was based on an already existing and proven framework (MQM), whereas the MCS was classified by a human.

7.3 Relevant features experiments

These experiments intended to discover the most relevant features and fine-tune the parameters of the algorithms. Feature selection is important because when redundant attributes are present, the algorithms can be misleading and the maintenance of irrelevant features can result in overfitting. The advantages of feature selection include decreasing overfitting, improving accuracy and minimizing training time. The feature set that produced the best classifier in the baseline experiments was considered. Two Weka attribute evaluators: InfoGain and GainRatio were tested and different numbers of attributes were retained. The InfoGain and GainRatio evaluate the worth of an attribute by measuring respectively,

the information gain and the gain ratio with respect to the class. These two attribute ranking methods are based on entropy. Entropy is commonly used in information theory and is considered as a measure of unpredictability. Equation 7.6 indicates its formula.

$$H(X) = - \sum p(x) \log(p(x)), \quad (7.6)$$

where $p(x)$ is the marginal probability density function for the random variable x . Based on entropy, equation 7.7 shows the formula of information gain.

$$InfoGain(Class, Attribute) = H(Class) - H(Class|Attribute) \quad (7.7)$$

The change of entropy is the information that is gained by the attribute. A notable problem occurs when information gain is applied to attributes that can take on a large number of distinct values. The gain ratio attempts to overcome this bias by penalizing the multiple-valued attributes.

$$GainRatio(Class, Attribute) = (H(Class) - H(Class|Attribute)) / H(Attribute) \quad (7.8)$$

The MCS corpus had as the best result an accuracy of 74.50 % through SVM and 10 folds on cross-validation. The feature set included only QuEst++ baseline features. In this case, the attribute evaluator that showed the best outcome was InfoGain. After several experiments where we changed the number of considered attributes, we concluded that of the 17 initial features only 15 were necessary. Table 7.5 presents the classifier results obtained when using the 15 most relevant features for the MCS corpus.

Table 7.5: Classifier results when using the 15 most relevant features of the MCS corpus.

Accuracy	Precision	Recall	F-measure	RMSE	AUC	Kappa statistic	Confusion matrix			
							TP	FN	FP	TN
75.00 %	0.75	0.75	0.75	0.50	0.74	0.48	95	16	34	55

There was an increase of 0.5 % in the accuracy achieved by the classifier. The values of F-measure and AUC increased slightly while RMSE decreased. As far as the confusion matrix is concerned, there was only one change that consisted in decreasing the number of false positives in a unit. Improved results indicate that the discarded features were irrelevant. Table 7.6 indicates the 15 most relevant features and their information gain rank.

Table 7.6: Most relevant features for the MCS corpus and information gain rank.

Ranked	Features
0.24	Number of tokens in the target text
0.23	Number of tokens in the source text
0.21	Number of occurrences of the target word within the target hypothesis (averaged for all words in the hypothesis - type/token ratio)
0.14	Number of punctuation marks in the target text
0.14	LM log probability of source text
0.12	LM log probability of target text
0.11	Number of punctuation marks in the source text
0.07	Percentage of trigrams in quartile 1 of frequency of source words in a corpus of the source language
0.06	Average source token length
0.05	Percentage of trigrams in quartile 4 of frequency of source words in a corpus of the source language
0	Percentage of bigrams in quartile 4 of frequency of source words in a corpus of the source language
0	Percentage of unigrams in the source text seen in a corpus (SMT training corpus)
0	Percentage of bigrams in quartile 1 of frequency of source words in a corpus of the source language
0	Average number of translations per source word in the text (as given by IBM 1 table thresholded so that $\text{prob}(t s) > 0.2$)
0	Percentage of unigrams in quartile 4 of frequency (higher frequency words) in a corpus of the source text

The number of tokens and the language model log probability of the text were relevant features for the translation difficulty estimation task. The features related to the percentage of unigrams, bigrams and trigrams were also important. There were some features that had rank equal to zero. This happened because when InfoGain is used, the features are considered individually, hence the information gain is zero. However, in certain cases one feature may need another feature to boost accuracy and hence, when considered together it produces predictive value.

For MCS, the best algorithm was the SVM. This algorithm has two parameters that can be changed: the C and the Kernel. C is essentially a regularization parameter, which controls the trade-off between achieving a low error on the training data and minimizing the norm of the weights. With a larger C, the optimization will choose a smaller margin hyper-plane if that hyper-plane does a better job on getting all the training points classified correctly. A smaller C ensures a larger margin even if the hyper-plane misclassifies more points. The typical values of C are exponents of 10 like 0.01, 0.1, 1, 10 and 100.

A Kernel is a similarity function and is at the basis of the kernel trick, which allows many learning algorithms based on linear models to build nonlinear models easily. A kernel function takes two input vectors as arguments and returns a real value corresponding to the inner product of the images of these vectors in some feature space without having to actually map the points to the feature space.

The SVM standard parameters include $C = 1.0$ and a Polynomial Kernel with the exponent equal to one. Equation 7.9 shows the formula of the polynomial kernel.

$$K(x, y) = \langle x, y \rangle^p \quad (7.9)$$

We tested several values of parameter C and tried different kernels. The best result was achieved using $C = 3.0$, that is, considering a lower margin. The best Kernel was the Normalized Polynomial Kernel with the exponent equals to two. Equation 7.10 presents the formula of the Normalized Polynomial Kernel.

$$K(x, y) = \frac{\langle x, y \rangle^2}{\sqrt{\langle x, x \rangle \langle y, y \rangle}} \quad (7.10)$$

where $\langle x, y \rangle$ is the Polynomial Kernel.

The final classifier results for the MCS corpus are shown in Table 7.7.

Table 7.7: Final classifier for the MCS corpus.

Accuracy	Precision	Recall	F-measure	RMSE	AUC	Kappa statistic	Confusion matrix			
							TP	FN	FP	TN
75.50 %	0.76	0.76	0.75	0.50	0.74	0.50	95	16	33	56

The result was not much higher than previously achieved. There was a gain of 0.5 % in relation to the accuracy and the gain of one more difficult text correctly classified. In conclusion, the corpus consisting of 200 Mails and Customer Support texts achieved an accuracy of 75.50 %. Cross-validation and the SVM algorithm were used. The most relevant features at the text level appeared to be the features used in the quality estimation task.

The AMTA threshold corpus reached the best result, 77.12 % of accuracy when using KNN and 10 folds on cross-validation. The best feature set included all features. The best attribute evaluator was InfoGain. After several experiments, we concluded that only 20 features would be necessary to achieve a better score. The biggest advantage for this corpus is the necessity of extracting fewer features, decreasing the processing speed. Table 7.8 indicates the final classifier results for the AMTA threshold corpus.

Table 7.8: Final classifier for the AMTA threshold corpus.

Accuracy	Precision	Recall	F-measure	RMSE	AUC	Kappa statistic	Confusion matrix			
							TP	FN	FP	TN
77.67 %	0.78	0.78	0.78	0.47	0.78	0.55	452	119	128	407

Once again, the improvement was small but still important. Moreover, the number of features to be considered dropped drastically from 32 to 20. The number of false negatives and false positives dropped and consequently, the number of true positives and true negatives increased, improving the final result. Table 7.9 exhibits the 20 most relevant features and their info gain rank. The features from the QuEst++ baseline are identified by an asterisk (*).

Table 7.9: Most relevant features for the AMTA threshold corpus and information gain rank.

Ranked	Features
0.06	* Percentage of trigrams in quartile 1 of frequency of source words in a corpus of the source language
0.06	* Number of punctuation marks in the target sentence
0.05	* Number of punctuation marks in the source sentence
0.05	Number of dependencies in source sentence
0.05	* Number of tokens in the target sentence
0.05	* Average source token length
0.04	* Number of tokens in the source sentence
0.04	* Percentage of bigrams in quartile 4 of frequency of source words in a corpus of the source language
0.04	Average word length (syllables) in the source sentence
0.04	Source and target word count ratio
0.04	* LM log probability of target sentence
0.03	* LM log probability of source sentence
0.03	* Percentage of unigrams in quartile 4 of frequency (higher frequency words) in a corpus of the source document
0.02	Lexical diversity in source sentence
0.02	Percentage of verbs in source sentence
0.02	Percentage of prepositions in target sentence
0.02	Lexical diversity in target sentence
0.02	Percentage of verbs in target sentence
0.02	Percentage of prepositions in source sentence
0.01	* Number of occurrences of the target word within the target hypothesis (averaged for all words in the hypothesis - type/token ratio)

Again, the most informational features from QuEst++ baseline included the number of tokens and punctuation marks. In this corpus, the most relevant features from the annotator benchmark were the number of dependencies followed by the average word length and the source and target word count ratio. There was a balance in the contribution of each feature set for the most relevant features to be considered, meaning that each feature set was equally important for this corpus type. In conclusion, the AMTA threshold corpus, classified according to a 85 % MQM threshold, obtained the maximum accuracy of 77.67 % through the KNN algorithm and assuming 10 folds in cross-validation.

Considering the two corpora types, we conclude that the sentence based corpus achieved always better results. The difference of variability between the MCS and the AMTA corpus may help to explain these results. The AMTA corpus used more editors and MT systems while covering more topics. Additionally, more data was available and the classification method used was different. The most relevant QuEst ++ baseline features included the number of punctuation marks, the number of tokens, the average source token length and the LM log probability. These were the most important features for both corpora, which means that the features appeared to be transverse to the corpus level. On the other hand, the sentence based corpus also required features such as the number of dependencies or the target word count ratio, which belonged to the annotator benchmark feature set.

7.4 Cross-corpora experiments

These experiments were made in order to verify if a classifier trained with a corpus level and tested on another one presented good results. On these experiments only the best classifier for each corpus was applied. The four additional features extracted from the MCS corpus were not considered in the model used, enabling the experiments of Table 7.10 to be performed.

Table 7.10: Cross-corpora experiments.

Train			Test
Corpus	Classifier	Cross-validation	Corpus
MCS	SVM	K=10	AMTA threshold
AMTA threshold	KNN	K=10	MCS

The experiments were made on Weka and the results are shown in Table 7.11.

Table 7.11: Cross-corpora results.

Accuracy	Precision	Recall	F-measure	RMSE	AUC	Kappa statistic	Confusion matrix			
							TP	FN	FP	TN
51.63%	0.51	0.52	0.36	0.70	0.50	0	569	2	533	2
39.50%	0.26	0.40	0.27	0.78	0.44	-0.11	2	109	12	77

The cross-corpora results were poor, especially for a binary classifier. Considering that the corpora were classified according to different classification criteria, and the most relevant features were not the same for both corpora levels, the results did not surprise. Apparently, when the classifier is trained with a text level corpus and is tested on a sentence level corpus, the results are superior, achieving a maximum of 51.63 % accuracy.

Additionally, experiments were made in order to understand if the presence of a single difficult to translate sentence was enough for the text which it belongs to be also considered difficult. In order to carry out this preliminary study, the three texts that were considered by the expert Spanish annotator as the most difficult to translate among the 200 texts of the MCS corpus were used. The sentence classification method used the number of edits made between the machine translation output and its post-edit version. The method of sorting sentences was based on the number of edits and not on the MQM score, because these data were not available for the MCS corpus. Given these values, the sentences were ordered from the most difficult (higher number of edits) to the easiest. Several new texts were created, resulting from the gradual removal of the most difficult sentences. The features used by the model built for the text level were then extracted from the new texts, and predictions were made. The aim was to check whether texts containing at least one difficult sentence, in the sense of the number of edits, were classified as difficult.

Table 7.12 presents the characteristics of the first text considered, and the predictions obtained taking into account several scenarios, where different texts made from text 1 are contemplated. The predictions were obtained using the best model obtained for the text level.

Table 7.12: Characteristics of text 1 and predictions on different scenarios.

Number of sentence	Number of edits	Sentence order by difficulty	Document	Prediction
1	0	5	Text 1	Difficult
2	8	2	Text 1 without sentence 5	Easy
3	0	4	Text 1 without sentence 2	Difficult
4	1	1, 3, 6		
5	12			
6	0			

Of the six sentences that make up text 1, two had a greater number of edits and were consequently considered the most difficult to translate sentences. The text was initially considered difficult, and when the most difficult sentence in this text was removed, the text was classified as Easy. In the other two cases presented, sentence 5 was part of the text analyzed and was therefore always obtained a Difficult classification. This specific case confirms our intuition that the presence of a single difficult to translate sentence is enough for the text which it belongs, to be also considered difficult. Table 7.13 presents the characteristics of the second text and the predictions obtained for several scenarios considered. Each scenario contemplates different texts made from text 2.

Table 7.13: Characteristics of text 2 and predictions on different scenarios.

Number of sentence	Number of edits	Sentence order by difficulty	Document	Prediction
1	23	2	Text 2	Difficult
2	34	5	Text 2 without sentence 2	Difficult
3	6	1	Text 2 without sentence 5	Difficult
4	4	3	Text 2 without sentence 1	Difficult
5	33	4	Text 2 without sentence 2 and 5	Difficult
			Text 2 without sentence 2, 5 and 1	Easy

In the case of text 2, more than half of the text had a high number of edits and was therefore difficult to reach an easy text. It was once again concluded that it is enough to have a difficult to translate sentence so that the text is also classified as difficult. A prediction of an Easy text appeared only when all difficult sentences were removed. In this case, the number of edits of sentences was very different, ranging from 4 to 34 edits. This difference facilitates the identification of the most difficult sentences and demonstrates again that our intuition is probably correct. The characteristics of the third text and the predictions resulting from different scenarios are shown in Table 7.14.

Table 7.14: Characteristics of text 3 and predictions on different scenarios.

Number of sentence	Number of edits	Sentence order by difficulty	Document	Prediction
1	5	2, 6	Text 3	Difficult
2	11	3	Text 3 without sentence 3	Difficult
3	8	1	Text 3 without sentence 2 and 6	Difficult
4	1	4	Text 3 without sentence 2, 6 and 3	Difficult
5	0	5	Text 3 without sentence 2, 6, 3 and 1	Easy
6	11			

In the case of the last text under consideration, the same conclusion was again reached. It is enough that a sentence is difficult to translate so that the text to which it belongs to be so. Thus, the text was only classified by the previously developed model as easy when all difficult sentences were removed.

7.5 Discussion

The results obtained in both text and sentence levels are promising, and demonstrated that translation difficulty can be assessed using not only the opinions of an expert annotator but also the MQM scores. We developed classifiers that predict translation difficulty using features suggested by an expert Spanish annotator and features from the quality estimation task.

The text based corpus composed of 200 texts was manually classified by an expert Spanish annotator as to its translation difficulty. The best classifier found for this corpus reached 75.50 % accuracy and was obtained through SVM. Two sets of features were considered: annotator benchmark and QuEst++ baseline. The first one was purely based on the opinions and ideas of the expert Spanish annotator, who manually annotated the texts. The other consisted of features that were typically used in the quality estimation task. The best result was obtained using only the second feature set. Through attribute selection, we saw that only 15 features were needed for the classifier to reach the peak of its performance.

In the case of sentence level, two different classification methods were considered. The AMTA average corpus was rated through the average MQM score, and the AMTA threshold corpus was rated considering a threshold of 85 %, for each sentence. The results obtained were better in the AMTA threshold corpus achieving a 77.67 % accuracy and using the KNN algorithm. Regarding the features, this corpus needed both feature sets while there were only 20 most relevant features. There was a balance in the contribution of each feature set for the most relevant features to be considered.

Comparing the results of the text and sentence level corpus we concluded that the sentence based corpus presented better results. Although the results were not very different from those obtained in the text based corpus, the difference may have an explanation. The explanation may lie in the variability of the corpus. The AMTA corpus used more editors and covered more topics. In contrast to the MCS corpus, the AMTA corpus used nine MT systems. Moreover, the sentence based corpus had more available data and the classification method was more controlled. It was based on an already existing and proven framework (MQM), whereas the MCS was classified by a human.

The cross-corpora experiments showed that a classifier trained with a text level corpus and tested on a sentence level corpus achieved a better result than the opposite scenario. However, the result (51.63 % accuracy) was relatively low, especially for a binary classifier. Additionally, it seems that a sentence that is classified as difficult to translate is enough for the text to which it belongs to be so.

Chapter 8

Conclusions

This chapter is divided into three sections. A summary of the work developed is presented in section 8.1. In section 8.2 the main contributions of this study are listed and finally, section 8.3 introduces ideas for future work.

8.1 Summary

At the beginning of this work two preliminary experiments were developed concerning the most common errors made by MT and the impact of human editions in MT errors. In the first experiment, 15 sentences of the WMT-News corpus were selected, and in the second, data from the Unbabel annotation tool was used. Through these experiments, it could be seen that the most common errors made by MT belonged to grammar and syntax categories. In the case of the studied sentences, the most frequent errors included: no subject-verb agreement, lack of articles, wrong verb form and literal translations. As for the impact of human edition in MT errors, the edition process drastically reduced the number of errors, and critical errors were completely eliminated. The accuracy and fluency categories had the highest number of errors, and there were small wrong language variety errors added by some editors. In addition, one edition may be insufficient to achieve the desired quality.

We proposed different classifiers to predict the translation difficulty of texts and sentences. Two different corpora were used. The MCS corpus that consisted of 200 Mails and Customer Support texts, and the AMTA corpus that had 785 sentences regarding four different topics: climate, Mexican, Norway and software. The texts were manually annotated by an expert Spanish annotator concerning their translation difficulty. The sentences were classified according to their MQM score, a threshold of 85 % having been adopted to separate between easy and difficult sentences.

The feature extraction module extracts the features of the texts and sentences. In this work two feature sets were considered: the annotator benchmark and the QuEst++ baseline. The first set included the features suggested by the expert Spanish annotator and the second feature set included features commonly used in the quality estimation task. The quality features showed promising results in trans-

lation difficulty estimation, which reinforces the idea that the quality and the difficulty of translation are closely related.

Finally, six machine learning algorithms were used for training and testing the classifiers. Several experiments were performed, considering different feature sets and machine learning methods. In addition, experiments to understand the most relevant features of each corpus were also conducted. The best results were obtained at the sentence level, resulting in an accuracy of 77.67 %. Although the results were not much higher than those obtained at the text level (75.50 %), these can be justified by the variability of the corpus. The sentence level corpus had greater variability at the level of topics, editors and MT systems used. On the contrary, the text corpus used only one editor, a text type and a single MT system. Additionally, the AMTA had more data, and the classification method was more controlled. That is, it was based on an already existing and proven framework (MQM), whereas the MCS was classified by a human. In the case of MCS, only QuEst++ baseline features were relevant, while in the sentence-based corpus, both feature sets were used.

Two classifiers were built by being trained with a corpus level and tested with another one. The best results were obtained for the classifier trained with the text based corpus and tested on the sentence based corpus. The result of 51.63 % accuracy was relatively low, especially for a binary classifier. However, a promising outcome was not expected given the differences in the classification criteria used and the most relevant features discovered. Additionally, experiments have been made to test an intuitive relationship between text and sentence difficulty. The results obtained showed that it was enough for a text to include a single difficult to translate sentence for the text to be considered difficult. This relationship can not be generalized, since it was obtained through a small exploratory analysis of three texts.

Readability scores of a text as measured by the Automated Readability Index and the Flesch reading were found to be correlated with text translation difficulty. However, they can not by themselves predict translation difficulty level, as they involve only reading comprehension. The time of edition and the number of edits were also correlated with text translation difficulty, as difficult to translate texts tended to require longer editing time and a greater number of edits. With respect to the HTER, it was only possible to draw conclusions in the case of the sentence based corpus. The results indicate that there is a greater human effort in sentences with greater translation difficulty.

Sanjun Sun found that a text's readability only partially accounts for its translation difficulty level and that the time-on-task was significantly, but weakly, related to translation difficulty level. [39] Our study presents conclusions similar to the Sanjun Sun, regarding the relationship of readability with translation difficulty level. Moreover, it demonstrates promising results that will facilitate future studies on developing a translation difficulty formula.

8.2 Contributions

The main contributions of this work are:

- Development of a classifier with a 75.50 % accuracy that predicts text translation difficulty based on quality estimation features and a classifier with a 77.67 % accuracy that predicts sentence translation difficulty. These systems can be used to select editors and increase the quality of the final translation.
- Verification of the ability of MQM scores for assess translation difficulty. Study on the most relevant features towards predicting text and sentence translation difficulty. The quality estimation features presented promising results in translation difficulty estimation.
- Confirmation of the relationship between readability and text translation difficulty. A difficult text presented higher translation difficulty. However, readability alone is not able to predict text translation difficulty.
- Verification of the relationship between the post-editing process and the text translation difficulty. The harder to translate texts imply more edits and take longer to be edited. The HTER was higher in the more difficult to translate sentences.
- Preliminary exploration of the hypothesis that a sentence that is difficult to translate suffices for the text to which it belongs to be so. The results obtained support this hypothesis.

Additionally, a corpus of 200 texts annotated by an expert Spanish annotator concerning their translation difficulty is available.

8.3 Future Work

There is a wide set of research lines that can be done following this work, namely a more detailed study of the features that influence both sentence and text translation difficulty. It would be interesting to include features about lexical rarity, the rarity of PoS n-grams, or certain unusual verbal forms that might indicate greater translation difficulty. Additionally, a couple of suggestions are given below:

- Considering the reduced size of the corpus manually annotated regarding translation difficulty, unsupervised learning techniques that do not depend on a previously classified corpus could be investigated.
- Study different text types as novels or scientific texts and explore different sentences types like interrogative, exclamatory, declarative, and imperative.
- Try to increase the dimension of the corpora and consider other language pairs. The increase in the number of texts and sentences will decrease the sparsity of many features extracted in this work.

Bibliography

- [1] Banerjee, S. and Lavie, A. (2007). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Association of Computational Linguistics*, pages 65–72.
- [2] Brien, S. O. (2005). Methodologies for Measuring the Correlations between Post-Editing Effort and Machine Translatability. *Mach Translat*, pages 37–58.
- [3] Brien, S. O. (2006). Controlled language and post-editing. *Machine Translation*, November:197 – 215.
- [4] Brien, S. O. (2011). Towards predicting post-editing productivity. *Springer*, July:197–215.
- [5] Chall, J. S., & Dale, E. (1995). Readability revisited, the new Dale-Chall readability formula. *Cambridge, MA: Bookline Books*.
- [6] Chawla, N. V. (2005). Chapter 40 DATA MINING FOR IMBALANCED DATASETS : AN OVERVIEW. *Department of Computer Science and Engineering, University of Notre Dame, USA*.
- [7] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE : Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence reasearch* 16, pages 321–357.
- [8] Clifford, R., Granoien, N., Jones, D., Shen, W., and Weinstein, C. (2003). The Effect of Text Difficulty on Machine Translation Performance – A Pilot Study with ILR-Rated texts in Spanish, Farsi, Arabic, Russian and Korean. *LREC 2004 FOURTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION*, pages 343–346.
- [9] Coleman, Meri; Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.
- [10] Collins-Thompson, K. and Callan, J. (2004). A Language Modeling Approach to Predicting Reading Difficulty. *Hlt-Naacl*.
- [11] Costa-Jussà, M. R., Farrús, M., Marino, J. B., and Fonollosa, J. A. R. (2012). Study and comparison of rule-based and statistical catalan-spanish machine translation systems. *Computing and Informatics*, 31:245–270.
- [12] Curto, P., Mamede, N., and Baptista, J. (2014). Automatic readability classifier for European Portuguese. *Conference INFORUM 2014*, pages 309–324.

- [13] Daems, J., Macken, L., and Vandepitte, S. (2014). On the origin of errors: A fine-grained analysis of MT and PE errors and their relationship. *Lrec*, pages 62–66.
- [14] Dale, E. & Chall, J. S. (1949). The concept of readability. *National Council of Teachers of English, Elementary*.
- [15] Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. *MIT Press*.
- [16] Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32:221–233.
- [17] Fry, E. (1968). A Readability Formula That Saves Time. *Journal of Reading*, 11:513–516, 575–578.
- [18] Green, S., Heer, J., and Manning, C. D. (2013). The Efficacy of Human Post-Editing. *Proceeding CHI 13 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.
- [19] Gunning, R. (1968). The Fog Index After Twenty Years. *Journal of Business Communication*, 6:3–13.
- [20] Hale, S. B., Hale, S., and Campbell, S. (2002). The Interaction Between Text Difficulty and Translation Accuracy and Translation Accuracy. *John Benjamins Publishing Company*.
- [21] Heilman, M., Collins-Thompson, K., and Eskenazi, M. (2008). An analysis of statistical models and features for reading difficulty prediction. *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications - EANL 08*, pages 71–79.
- [22] Heilman, M. J., Collins-Thompson, K., Callan, J., and Eskenazi, M. (2007). Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. *Conference: Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 460–467.
- [23] Lommel, A. R. and Uszkoreit, H. (2006). Multidimensional Quality Metrics : A Flexible System for Assessing Translation Quality. *Lisa*.
- [24] Luong, M.-t. and Manning, C. D. (2015). Stanford Neural Machine Translation Systems for Spoken Language Domains. *International Workshop on Spoken Language Translation*, pages 2–5.
- [25] McLaughlin, G. (1969). SMOG grading: A new readability formula. *Journal of reading*, 12:639–646.
- [26] Milone, M. and Biemiller, A. (2014). The development of ATOS: The Renaissance readability formula. *Renaissance Learning*.
- [27] Mohit, B. (2010). *LOCATING AND REDUCING TRANSLATION* by Behrang Mohit Bachelor of Computer Science, Carnegie Mellon University, 2000 Masters of Information Management and Systems, University of California at Berkeley, 2003 Masters of Intelligent Systems. PhD thesis.
- [28] O'Brien, S. (2006). *Machine-Translatability and Post-Editing Effort: An Empirical Study using Translog and Choice Network Analysis*. PhD thesis.

- [29] Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. *40Th Annual Meeting of the Association for Computational Linguistics (ACL)*, (July):311–318.
- [30] Pitler, E. (2008). Revisiting Readability : A Unified Framework for Predicting Text Quality. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- [31] Porter, M. F. (1980). An algorithm for suffic stripping. *Emerald Group Publishing Limited*, 14 no. 3:130–137.
- [32] S. Schwarm and M. Ostendorf (2005). Reading level assessment using support vector machines and statistical language models. *In Proceedings of the 43rd Annual Meeting on the Association for Computational Linguistics, Ann Arbor, USA*.
- [33] Sanchez-torron, M. (2016). Machine Translation Quality and Post-Editor Productivity. *Proceedings of AMTA 2016*.
- [34] Schneider, N., Smith, N. A., Schneider, N., and Smith, N. A. (2015). A Corpus and Model Integrating Multiword Expressions and Supersenses. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1537–1547.
- [35] Senter, R.J.; Smith, E. (1967). Automated Readability Index. *Wright-Paterson AFB, Ohio: Aerospace Medical Division*.
- [36] Series, W. P. (2001). Assessing the Lexile Framework : Results of a Panel Meeting. *U.S. Department of Education, National Center for Education Statistics*, 08.
- [37] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of Association for Machine Translation in the Americas*, August:223–231.
- [38] Snover, M. G., Madnani, N., Dorr, B., and Schwartz, R. (2009). TER-Plus: Paraphrase, semantic, and alignment enhancements to translation Edit Rate. *Machine Translation*, 23:117–127.
- [39] Sun, S. and Shreve, G. M. (2014). Measuring translation difficulty An empirical study. *John Bne-jamins Publishing company*, pages 98–127.
- [40] Thorndike, E. L. (1921). The Teacher's Word Book. *Teachers College, Columbia University*.
- [41] Uchimoto, K., Hayashida, N., Ishida, T., and Isahara, H. (2005). Automatic Rating of Machine Translatability. *MT Summit X*, pages 235–242.
- [42] Viera, A. J. and Garrett, J. M. (2005). Understanding Interobserver Agreement : The Kappa Statistic. *Family Medicine*, May:360–363.

Appendix A

WMT-News Corpus

1	Source	Right-wing populists triumph in Austria, have total of 29 percent.
	Hypothesis	Los populistas de derecha triunfo en Austria, tiene total de 29 %.
	Target	Triunfo de la derecha populista en Austria, juntos forman el 29 %.
2	Source	According to the first preliminary results of the early parliament elections in Austria have brought about a perceptible weakening of both parties in the present large coalition and a significant boost for the right-wing populist parties.
	Hypothesis	De acuerdo con los primeros resultados preliminares de las elecciones al Parlamento primeros en Austria han dado lugar a un debilitamiento perceptible de ambas partes en el presente coalición grande y un impulso significativo para los partidos populistas de derecha.
	Target	Las elecciones anticipadas al parlamento austriaco han traído según los primeros resultados aproximados, un apreciable debilitamiento de ambos partidos de la gran coalición actual y un expresivo fortalecimiento del partido de la derecha populista.
3	Source	The Austrian People's Party (OVP), where the position of the current head, Wilhelm Molterer, is being severely jolted, suffered particularly great losses.
	Hypothesis	El Partido Popular de Austria (OVP), donde la posición del actual jefe, Wilhelm Molterer, está siendo severamente sacudido, sufrió particularmente grandes pérdidas.
	Target	Unas fuertes y especiales pérdidas ha sufrido el Partido Popular Austríaco (OVP), donde se tambalea seriamente la posición del actual presidente, Wilhelm Molterer.
4	Source	Conversely, the campaign leader of the Alliance for the Future of Austria (BZO), Carinthian Governor Jorg Haider, is preparing a triumphant return to national politics.
	Hypothesis	Por el contrario, el líder de la campaña de la Alianza para el Futuro de Austria (BZO), Carinthian gobernador Jorg Haider, está preparando un regreso triunfal a la política nacional.
	Target	Por el contrario, el líder de la Alianza para el futuro de Austria (BZO), el jefe del gobierno del estado de Carintia, Jorg Haider, prepara su regreso triunfal a la política nacional.

Continued on next page

5	Source	According to the preliminary results, the Social Democratic Party (SPO) remains the strongest in the country with 29.8 percent of votes, however, it has lost 5.5 percent of votes since the last elections in 2006.
	Hypothesis	Según los resultados preliminares, el Partido Socialdemócrata (SPO) sigue siendo el más fuerte en el país con un 29.8 por ciento de los votos, sin embargo, se ha perdido un 5.5 por ciento de los votos desde las últimas elecciones de 2006.
	Target	Según los resultados aproximados, la socialdemocracia (SPO) permanece como el partido más fuerte en el país con un 29.8 por ciento de votos, por lo que pierde un 5.5 por ciento de voto en comparación con las últimas elecciones del año 2006.
6	Source	OVP, which with its 25.6 percent loses nearly nine percent of votes, fared even worse.
	Hypothesis	OVP, que con su 25.6 por ciento pierde casi el nueve por ciento de los votos, fue aún peor.
	Target	Peor todavía ha terminado el OVP que pierde un 25.6 por ciento, casi un nueve por ciento de votos.
7	Source	These are the worst results of both large parties in Austrian post-war history, and particularly for the People's Party, who urged the early elections, it is literally a catastrophic result.
	Hypothesis	Estos son los peores resultados de los dos grandes partidos en la historia de la posguerra de Austria, y en particular para el Partido Popular, quien instó a las elecciones anticipadas, que es, literalmente, un resultado catastrófico.
	Target	Son los peores resultados, literalmente catastróficos, en la historia de estos dos grandes partidos austríacos desde la posguerra y especialmente para los populares que convocaron estas elecciones anticipadas.
8	Source	At the beginning of July, when OVP left the coalition, the People's Party still had a significant lead on SPO in the polls.
	Hypothesis	A principios de julio, cuando OVP abandonó la coalición, el Partido Popular todavía tenía una ventaja significativa sobre SPO en las encuestas.
	Target	Todavía a principios de julio, cuando el OVP salió de la coalición, los populares tenían en lo que se refiere a preferencia ante el SPO, una ventaja considerable.
9	Source	Voters, however, apparently punished them for letting the government flounder.
	Hypothesis	Los votantes, sin embargo, al parecer, los castigó por dejar que la platija gobierno.
	Target	Está claro, sin embargo, que los votantes los castigaron por dejar que el gobierno naufragara.
10	Source	In the first reactions to the results, there were already speculations about the possible resignation of the party head and current vice-chancellor Wilhelm Molterer.
	Hypothesis	En las primeras reacciones a los resultados, ya había especulaciones sobre la posible renuncia del jefe del partido y actual vicescanciller, Wilhelm Molterer.
	Target	Ya en las primeras reacciones a los resultados se dejaron oír especulaciones sobre la dimisión del presidente del partido y, hasta hoy, vicescanciller, Wilhelm Molterer.

Continued on next page

11	Source	Observers anticipate that this could take place as early as at the extraordinary meeting of the party leadership on Monday.
	Hypothesis	Los observadores anticipan que esto podría tener lugar tan pronto como en la reunión extraordinaria de la dirección del partido el lunes.
	Target	Los observadores creen que ya este lunes se podría llegar a ello en la reunión excepcional de la presidencia del partido.
12	Source	Such a development would certainly simplify the journey toward the increasingly most likely recourse from the election results, that is, the renewal of the large coalition of SPO and OVP.
	Hypothesis	Tal desarrollo, sin duda, simplificar el viaje hacia la cada vez más probable es que el recurso a partir de los resultados de las elecciones, es decir, la renovación de la gran coalición de SPO y OVP.
	Target	Tal evolución evidentemente facilitaría el camino del constante y más aproximado punto de partida de los resultados electorales, es decir, de la renovación de la gran coalición entre SPO y OVP.
13	Source	Given the strengthening of both right-wing populist parties - the Freedom Party (FPO) gained a preliminary 18 percent and BZO eleven percent of the votes - however, at the same time, the Social Democrats expressed fear of a repetition of the year 1999, when the People's Party agreed with the populists (FPO was still united at that time, it broke away from BZO in 2005) on a common government, which eventually provoked sanctions from the European Union.
	Hypothesis	Teniendo en cuenta el fortalecimiento de ambos partidos populistas de derecha - el Partido de la Libertad (FPO) ganó una preliminar 18 por ciento y BZO once por ciento de los votos - sin embargo, al mismo tiempo, los socialdemócratas expresaron el temor de que se repitiera el año 1999, cuando el Partido Popular estaba de acuerdo con los populistas (FPO fue todavía unido en ese momento, se separó de BZO en 2005) en un gobierno común, lo que finalmente provocó sanciones de la Unión Europea.
	Target	En vista de que el refuerzo de ambos partidos de la derecha populista - los liberales (FPO) han ganado aproximadamente un 18 por ciento y el BZO, un once por ciento de los votos - y que los socialdemócratas al mismo tiempo expresaron sus temores de que se repitiera lo mismo del año 1999 cuando los populares negociaron con los populistas (entonces FPO era una unidad, pero en el año 2005 se desmembró de BZO) para un gobierno de coalición que llegó a generar una sanción por parte de la Unión Europea.

Continued on next page

14	Source	Haider, who has already announced that he is prepared to work together with any party and presumes he will return to Vienna to national politics, is evidently banking on this development.
	Hypothesis	Haider, que ya ha anunciado que está dispuesto a trabajar junto con cualquiera de las partes y presume que regresará a Viena a la política nacional, es evidente que la banca en este desarrollo.
	Target	Por esta evolución apuesta claramente Haider quien ya anunció que está preparado para colaborar con cualquier partido y que ya asume su vuelta a Viena, a la política nacional.
15	Source	The Green Party also got into the parliament, but suffered a slight loss and fell from third to fifth place among Austrian political parties.
	Hypothesis	El Partido Verde también se metió en el parlamento, pero sufrió una ligera pérdida y cayó del tercer al quinto lugar entre los partidos políticos austriacos.
	Target	También el Partido Verde ha conseguido llegar al parlamento aunque ha perdido ligeramente y descendió del tercer al quinto puesto entre los partidos políticos austriacos.

Appendix B

Corrected WMT-News Corpus

1	Hypothesis	Los populistas de derecha triunfo en Austria, tiene total de 29%.
	Corrected Hypothesis	Los populistas de derecha triunfan en Austria, y consiguen en total el 29%.
2	Hypothesis	De acuerdo con los primeros resultados preliminares de las elecciones al Parlamento primeros en Austria han dado lugar a un debilitamiento perceptible de ambas partes en el presente coalición grande y un impulso significativo para los partidos populistas de derecha.
	Corrected Hypothesis	Los primeros resultados provisionales de las elecciones anticipadas al Parlamento en Austria muestran un debilitamiento dos partidos de la actual gran coalición y un ascenso significativo de los partidos populistas de derecha.
3	Hypothesis	El Partido Popular de Austria (OVP), donde la posición del actual jefe, Wilhelm Molterer, está siendo severamente sacudido, sufrió particularmente grandes pérdidas.
	Corrected Hypothesis	El Partido Popular de Austria (OVP), en el que la posición del actual líder, Wilhelm Molterer, se tambalea seriamente, sufrió particularmente grandes pérdidas.
4	Hypothesis	Por el contrario, el líder de la campaña de la Alianza para el Futuro de Austria (BZO), Carinthian gobernador Jorg Haider, está preparando un regreso triunfal a la política nacional.
	Corrected Hypothesis	Por el contrario, el jefe de campaña de la Alianza para el Futuro de Austria (BZO), (del estado) de Carintia Jorg Haider, prepara su regreso triunfal a la política nacional.

Continued on next page

5	Hypothesis	Según los resultados preliminares, el Partido Socialdemócrata (SPO) sigue siendo el más fuerte en el país con un 29.8 por ciento de los votos, sin embargo, se ha perdido un 5.5 por ciento de los votos desde las últimas elecciones de 2006.
	Corrected Hypothesis	Según los resultados provisionales, el Partido Socialdemócrata (SPO) sigue siendo el más fuerte en el país con un 29.8 por ciento de los votos; sin embargo, ha perdido un 5.5 por ciento de los votos desde las últimas elecciones en 2006.
6	Hypothesis	OVP, que con su 25.6 por ciento pierde casi el nueve por ciento de los votos, fue aún peor.
	Corrected Hypothesis	Al OVP, que con un 25.6 por ciento pierde casi el nueve por ciento de sus votos, le fue aún peor.
7	Hypothesis	Estos son los peores resultados de los dos grandes partidos en la historia de la posguerra de Austria, y en particular para el Partido Popular, quien instó a las elecciones anticipadas, que es, literalmente, un resultado catastrófico.
	Corrected Hypothesis	Estos son los peores resultados de los dos grandes partidos en Austria desde de la guerra, y en particular para el Partido Popular, que instó a forzó por las elecciones anticipadas, es realmente un resultado catastrófico.
8	Hypothesis	A principios de julio, cuando OVP abandonó la coalición, el Partido Popular todavía tenía una ventaja significativa sobre SPO en las encuestas.
	Corrected Hypothesis	A principios de julio, cuando el OVP abandonó la coalición, el Partido Popular todavía tenía una ventaja considerable sobre el SPO en las encuestas.
9	Hypothesis	Los votantes, sin embargo, al parecer, los castigó por dejar que la platija gobierno.
	Corrected Hypothesis	Sin embargo, al parecer, los votantes les castigaron por dejar que el gobierno naufraguara.
10	Hypothesis	En las primeras reacciones a los resultados, ya había especulaciones sobre la posible renuncia del jefe del partido y actual vicescanciller, Wilhelm Molterer.
	Corrected Hypothesis	En las primeras reacciones a los resultados, ya hubo especulaciones sobre la posible dimisión del líder del partido y actual vicescanciller, Wilhelm Molterer.
11	Hypothesis	Los observadores anticipan que esto podría tener lugar tan pronto como en la reunión, extraordinaria de la dirección del partido el lunes.
	Corrected Hypothesis	Algunos comentaristas pronostican que esto podría tener lugar ya en la reunión extraordinaria de la dirección del partido del lunes.

Continued on next page

12	Hypothesis	Tal desarrollo, sin duda, simplificar el viaje hacia la cada vez más probable es que el recurso a partir de los resultados de las elecciones, es decir, la renovación de la gran coalición de SPO y OVP.
	Corrected Hypothesis	Estos acontecimientos, sin duda simplificarían el camino hacia la cada vez más probable solución tras los resultados electorales, es decir, la renovación de la gran coalición entre el SPO y el OVP.
13	Hypothesis	Teniendo en cuenta el fortalecimiento de ambos partidos populistas de derecha - el Partido de la Libertad (FPO) ganó una preliminar 18 por ciento y BZO once por ciento de los votos - sin embargo, al mismo tiempo, los socialdemócratas expresaron el temor de que se repitiera el año 1999, cuando el Partido Popular estaba de acuerdo con los populistas (FPO fue todavía unido en ese momento, se separó de BZO en 2005) en un gobierno común, lo que finalmente provocó sanciones de la Unión Europea.
	Corrected Hypothesis	Teniendo en cuenta el ascenso de los dos partidos populistas de derecha -el Partido de la Libertad (FPO) ha conseguido provisionalmente un 18 por ciento y el BZO un once por ciento de los votos - sin embargo, al mismo tiempo, los socialdemócratas expresaron su temor a que se repitiera (la situación) del año 1999, cuando el Partido Popular acordó un gobierno de coalición con los populistas (el FPO todavía estaba unido en ese momento, se segregó del BZO en 2005), lo que finalmente provocó sanciones por parte de la Unión Europea.
14	Hypothesis	Haider, que ya ha anunciado que está dispuesto a trabajar junto con cualquiera de las partes y presume que regresará a Viena a la política nacional, es evidente que la banca en este desarrollo.
	Corrected Hypothesis	Haider, que ya ha anunciado que está dispuesto a colaborar con cualquier partido y asume que regresará a Viena a la política nacional, cuenta, evidentemente, con que esto ocurra.
15	Hypothesis	El Partido Verde también se metió en el parlamento, pero sufrió una ligera pérdida y cayó del tercer al quinto lugar entre los partidos políticos austriacos.
	Corrected Hypothesis	El Partido Verde también entró en el parlamento, pero sufrió un ligero descenso y cayó del tercer al quinto puesto entre los partidos políticos austriacos.