Difficulty Estimation of Machine Translation

Ana Sofia Almeida
Electrical and Computer Engineering, Instituto Superior Técnico, Lisbon, Portugal
Email: ana.jesus.almeida@ist.utl.pt

***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.

I. Introduction

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. Only 55.7 % of Internet content is in English, leaving more than 40 % of the Internet content distributed by other languages.[1]

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 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.

In this work, our main goal was to develop automatic classification systems capable of classifying texts and sentences concerning their translation difficulty. As a proof of concept, we have chosen a single language pair (English-Spanish).

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?

The remainder of this paper is organized as follows: Section II overviews the background on MT and readability. Section III describes the text and sentence based corpora used in this work. The features extracted from both levels and the machine learning algorithms used are mentioned on Section IV. Section V mentions the experiments performed and the results obtained. We draw our conclusions in Section VI.

II. Background

This work focused on MT, readability and on how text and sentence translation difficulty can be measured. As such, in the remainder of this section we describe the fundamental concepts and the related work on these areas.

A. 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

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

[2] https://translate.google.pt/ (visited on 17/06/2016)

is often post-edited by human editors, in order to achieve a higher quality score. MT faces several challenges like:

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

There are different types of MT methods [1] such as Rule-Based Machine Translation (RBMT) and Statistical Machine Translation (SMT). RBMT 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 models are trained with a parallel corpus at the sentence level, and can achieve high-quality translations when large corpora are available.

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) [2]. 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. This promising 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. 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.

Two of the most popular automatic evaluation methods are the Bilingual Evaluation Understudy (BLEU) and the Translation Edit Rate (TER).

BLEU [3] is 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 and has values between zero and one, where one indicates a perfect translation. 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 the 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 1.

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

TER [4] 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. A TER score of zero represents a perfect match between the raw translation (machine translation output) and the reference, and a higher score implies a worse translation quality, as well as a requirement for more post-editing. This metric is calculated through equation 2.

$$TER = \frac{number\ of\ edits}{number\ of\ reference\ words} \quad (2)$$

B. Readability

Readability has been measured over the years through more than 200 readability algorithms. One of the research questions concerns the usability of readability in the translation difficulty estimation. The readability formulas get the grade level that a person needs to be able to read a text, are text-based, and 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 quantitative measures.

The Flesch Reading Ease Formula [5] 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. The Automated Readability Index uses the percentage of hard words and the average sentence length as lexical and grammatical features.

A study [6] 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.

Readability classifiers have been addressed in previous Master theses at IST, namely by Pedro Curto [7], 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 lowest precision was obtained for the PoS feature group. In both cases, 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.

C. Text Difficulty and Machine Translation

In 2010, a thesis entitled Locating and Reducing Translation difficulty [8] focused on locating difficult-to-translate phrases (DTP), and reducing their translation difficulty. A DTP is a phrase that is weakly translated using a specific MT system, and as consequence has very low BLEU value. The most frequent reasons behind phrase difficulty include:

- Unknown source language word;
- Lexical ambiguity;
- Articles and punctuation;
- 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 [9] 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).

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 [10]. 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 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.

III. Corpora Description

In this section, we mention the corpora used. How the corpora were collected and the classification method adopted are explained in detail. A discussion on the balance of the data sets is also made.

A. The MCS corpus

The MCS corpus included 200 uncategorized Mails and Customer Support texts. The source texts were translated by the Unbabel's MT system and the output was edited by several non-professional human editors. Both machine translated and 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 also provided by the company.

The characterization of the corpus is shown in Table I. 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.

TABLE I: 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

The texts were generally short, due to their type. As a consequence, they presented a high lexical diversity value and a lower HTER value, not requiring much edition. The Automated Readability Index value of 10.8 indicated that a person with 14 to 15 years old, or within the ninth grade would be capable of comprehend the MCS corpus texts. Table II exhibits how the texts were distributed according to their length.

TABLE II: Distribution of 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 %

The corpus was divided into three categories: small (S), medium (M) and large (L). Most of the texts were small, having less than 100 words, and just 3 % contain more than 200 words. The largest texts tended to be customer support texts as they contain a different purpose, which can include explaining a procedure to the user.

We studied the relationship between the text length and the time of edition per word, and the result suggests that bigger texts are associated with higher processing speed (less time per word).

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.

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).

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. When considering three classes, a very unbalanced corpus was obtained, with only three texts on the Difficult class. Thus, we choose to neglect the minority class and reduce the number of available classes from three to two: Easy and Difficult. After reducing the number of available classes, the distribution of the classified texts became more balanced with 55.50 % of easy texts and 44.50 % of difficult texts.

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

TABLE III: Average values for parameters of the MCS corpus.

Class	ARI	Flesch Reading Ease	Time of edition (sec)	Number of edits
Easy	9.90	66.01	33.33	16.69
Difficult	11.92	57.87	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.

B. The AMTA corpus

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 [11]. 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.

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. The sentences were equally distributed across the editors, MT system and topic. Table IV presents some of the data collected from the AMTA corpus.

TABLE IV: 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 MWE	0.47

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. This happened because about 90 % of the sentences had an MQM score greater than 90 %, making impossible to perform a classification based on these MQM scores. Thus, a new annotation of the corpus was made by an expert Spanish annotator, and later the MQM score was calculated. The annotator was the same that has manually classified the MCS corpus. The annotator used the Unbabel annotation tool that has an error taxonomy divided into seven categories: accuracy, fluency, style, terminology, wrong language variety, named entities, formatting and encoding. There were 41 error types available. One of three categories (Minor, Major, Critical) was assigned to each error. Each category had an associated weight, that was taken into account when calculating the MQM score. Table V indicates the average MQM scores per editor. We show both the original MQM scores and the ones obtained through the new annotation.

TABLE V: 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. The average MQM score obtained through the new annotation was 85 %. 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 each 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.

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.

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. In the case of oversampling, the instances of the minority class are duplicated until the number of samples of the largest class is reached. In this case, the problem is the risk of overfitting.

In this work, an oversampling technique, the Synthetic Minority Oversampling technique (SMOTE) was applied [12]. SMOTE is an oversampling approach in which the minority class is over-sampled by creating ”synthetic” examples rather than oversampling with replacement. The SMOTE was applied to both AMTA average and AMTA threshold classification methods output. This filter creates ”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. 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. After applying the SMOTE filter, the AMTA corpus achieved a higher level of equilibrium, regardless of the classification method used.

Recalling, Table VI indicates the distribution of the classes considered, according to the corpora type.

TABLE VI: 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 %

IV. Feature extraction and classifiers

This section addresses the different features extracted from both corpora and the selected machine learning algorithms.

A. Text analysis module

Two sets of features were extracted from the MCS corpus: the annotator benchmark and the QuEst++ baseline features.

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.

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

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. 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 17 features were extracted through the extractor provided.[4] The 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 prob (t|s) >0.2);
- Average number of translations per source word in the text (as given by IBM 1 table thresholded such that 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.

B. 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. Again, two sets of features were extracted.

[4]https://github.com/ghpaetzold/questplusplus (visited on 03/07/2017)

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 made 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. The QuEst++ baseline features extracted from the MCS corpus were also extracted from the AMTA corpus, as the QuEst++ has a module to extract features at sentence level.

The goal of our classifiers was to classify a text and a sentence as being easy or difficult to translate from English to Spanish. Several machine learning algorithms were tested using Weka. Weka is a collection of machine learning algorithms for data mining tasks. The machine learning algorithms applied in the experiments included the ZeroR, the Naive Bayes, the Support Vector Machine (SVM), the Multilayer Perceptron, the K-Nearest Neighbor (KNN) and the REPTree algorithms. The ZeroR method was used as benchmark.

V. Experimental results

This section mentions the evaluation parameters and describes the results obtained in the experiments performed.

A. Evaluation parameters

In order to evaluate the performance of the classifiers some evaluation parameters were selected, namely the accuracy, precision, recall, F-measure, Root Mean Square Error (RMSE) and Area Under the Receiver Operator Curve (AUC). Additionally, the confusion matrix was also used.

B. Baseline experiments

The baseline experiments aimed to identify the best machine learning algorithms and the best feature set, for each corpus considered. For each feature set, six algorithms were tested and for each of them the Weka's standard parameters were adopted. Table VII presents the best classifiers results considering each feature set and corpus.

TABLE VII: Classifiers 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 cross-validation	10	10	10
AMTA average			
Accuracy	75.72 %	75.72 %	75.99 %
Classifier	KNN	KNN	KNN
Folds cross-validation	10	10	10
AMTA threshold			
Accuracy	76.31 %	76.04 %	77.12 %
Classifier	KNN	KNN	KNN
Folds cross-validation	10	10	10

In the case of the MCS corpus, the best machine learning algorithm was always the SVM. 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.

Table VIII indicates the values of the evaluation parameters for the best classifier of MCS and AMTA threshold corpora, presented in Table VII. For better judgment of the results, Table IX shows the corresponding confusion matrices.

TABLE VIII: Evaluation parameters for the best classifier of the baseline experiments.

Accuracy	Precision	Recall	F-measure	RMSE	AUC
74.50%	0.75	0.75	0.74	0.51	0.73
77.12%	0.77	0.77	0.77	0.47	0.78

TABLE IX: Confusion matrix for the best classifier of the baseline experiments.

Corpus	TP	FN	FP	TN
MCS	95	16	35	54
AMTA threshold	447	124	129	406

The text based corpus had more difficulties in correctly classifying difficult texts. The difficulty of predicting the difficult texts could be related to the distribution of the data, as there were more easy than difficult to translate texts in the corpus. According to the manual annotation, there were only 3 difficult to translate texts, which explains the confusion of this classifier.

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. That is, it was based on an already existing and proven framework (MQM), whereas the MCS was classified by a human.

C. Relevant features experiments

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 improving accuracy and minimizing training time. These experiments intended to discover the most relevant features and fine-tune the parameters of the algorithms. 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. Table X indicates the 15 most relevant features for the MCS corpus and their information gain rank.

TABLE X: 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 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.

The final classifier results for the MCS corpus are shown in Table XI. As a consequence of using only the most relevant features, the results have slightly improved.

TABLE XI: Final classifier for the MCS corpus.

Accuracy	Precision	Recall	F-measure	RMSE	AUC
75.50 %	0.76	0.76	0.75	0.50	0.74

TABLE XII: Confusion matrix for the final classifier of the MCS corpus.

Corpus	TP	FN	FP	TN
MCS	95	16	33	56

The results were not much higher than previously achieved. There was a gain of 1 % in terms of accuracy, and two additional difficult texts were 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 in the baseline experiments when using the KNN algorithm and 10 folds on cross-validation. The best feature set included the two modules considered. After several experiments, we concluded that only 20 features would be necessary to achieve a better score. Table XIII exhibits the 20 most relevant features and their info gain rank. The features from the QuEst++ baseline are identified by an asterisk (*).

TABLE XIII: 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. These features seem to be transversal across different corpora levels. 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. It follows that, to predict the sentence translation difficulty both source and target sentences were useful. 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. Table XIV indicates the final classifier results for the AMTA threshold corpus, and the corresponding confusion matrix is shown in Table XV.

TABLE XIV: Final classifier for the AMTA threshold corpus.

Accuracy	Precision	Recall	F-measure	RMSE	AUC
77.67 %	0.78	0.78	0.78	0.47	0.78

TABLE XV: Confusion matrix for the final classifier of the AMTA threshold corpus.

Corpus	TP	FN	FP	TN
AMTA threshold	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. 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.

D. 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 XVI to be performed.

TABLE XVI: 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 XVII.

TABLE XVII: Cross-corpora results.

Accuracy	TP	FN	FP	TN
51.63%	569	2	533	2
39.50%	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 XVIII presents the characteristics of one of the texts considered, and the predictions obtained for each scenario are shown in Table XIX.

TABLE XVIII: Characteristics of the text.

Number of sentence	Number of edits	Sentence order by difficulty
1	23	2
2	34	5
3	6	1
4	4	3
5	33	4

TABLE XIX: Predictions on different scenarios.

Document	Prediction
Text	Difficult
Text without sentence 2	Difficult
Text without sentence 5	Difficult
Text without sentence 1	Difficult
Text without sentence 2 and 5	Difficult
Text without sentence 2, 5 and 1	Easy

More than half of the sentences had a high number of edits. A prediction of an Easy text appeared only when all difficult sentences were removed, indicating that when a text contemplates a single difficult sentence, the text is also considered difficult. The analysis performed to the other two texts provided similar results.

VI. Conclusions and future work

In this thesis 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 translation 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. [10] 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.

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. It would be also interesting to consider different language pairs and study different text types as novels or scientific texts, and explore different sentences types like interrogative, exclamatory, declarative, and imperative. Additionally, it would be good to increase the dimension of the corpora, decreasing the sparsity of many features extracted in this work.

Acknowledgements

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.

References

[1] M. R. Costa-Jussà, M. Farrús, J. B. Marino, and J. A. R. Fonollosa, "Study and comparison of rule-based and statistical catalan-spanish machine translation systems," *Computing and Informatics*, vol. 31, pp. 245–270, 2012.

[2] M.-t. Luong and C. D. Manning, "Stanford Neural Machine Translation Systems for Spoken Language Domains," *International Workshop on Spoken Language Translation*, pp. 2–5, 2015.

[3] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a method for automatic evaluation of machine translation," *40Th Annual Meeting of the Association for Computational Linguistics (ACL)*, no. July, pp. 311–318, 2002.

[4] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," *Proceedings of Association for Machine Translation in the Americas*, vol. August, pp. 223–231, 2006.

[5] R Flesch, "A new readability yardstick," *Journal of Applied Psychology*, vol. 32, pp. 221–233, 1948.

[6] E. Pitler, "Revisiting Readability : A Unified Framework for Predicting Text Quality," *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008.

[7] P. Curto, N. Mamede, and J. Baptista, "Automatic readability classifier for European Portuguese," *Conference INFORUM 2014*, pp. 309–324, 2014.

[8] B. Mohit, "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, 2010.

[9] S. B. Hale, S. Hale, and S. Campbell, "The Interaction Between Text Difficulty and Translation Accuracy and Translation Accuracy," *John Benjamins Publishing Company*, 2002.

[10] S. Sun and G. M. Shreve, "Measuring translation difficulty An empirical study," *John Bnejamins Publishing company*, pp. 98–127, 2014.

[11] M. Sanchez-torron, "Machine Translation Quality and Post-Editor Productivity," *Proceedings of AMTA 2016*, 2016.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE : Synthetic Minority Oversampling Technique," *Journal of Artificial Intelligence reasearch 16*, pp. 321–357, 2002.