Linguistic Indicators for Quality Estimation of Machine Translations

Mariano Felice

Main advisor Dr. Lucia Specia

Co-advisors Dr. Xavier Blanco
Dr. Constantin Orăsan

A project submitted as part of a programme of study for the award of *Erasmus Mundus International Masters/MA in Natural Language Processing & Human Language Technology.*

May 2012

To my family

Abstract

This work presents a study of linguistically-informed features for the automatic quality estimation of machine translations. In particular, we address the problem of estimating quality when no reference translations are available, as this is the most common case in real world situations. Unlike previous attempts that make use of internal information from translation systems or rely on purely shallow aspects, our approach uses features derived from the source and target text as well as additional linguistic resources, such as parsers and monolingual corpora.

We built several models using a supervised regression algorithm and different combinations of features, contrasting purely shallow, linguistic and hybrid sets. Evaluation of our linguistically-enriched models yields mixed results. On the one hand, all our hybrid sets beat a shallow baseline in terms of Mean Average Error but on the other hand, purely linguistic feature sets are unable to outperform shallow features.

However, a detailed analysis of individual feature performance and optimal sets obtained from feature selection reveals that shallow and linguistic features are in fact complementary and must be carefully combined to achieve optimal results. In effect, we demonstrate that the best performing models are actually based on hybrid sets having a significant proportion of linguistic features. Furthermore, we show that linguistic information can produce consistently better quality estimates for specific score intervals.

Finally, we analyse many factors that may have an impact on the performance of linguistic features and suggest new directions to mitigate them in the future.

Acknowledgements

First and foremost, I *need* to thank my family for their unflagging support over the years (especially these two unusual ones) as much as for everything they have done to make me become who I am today. Anything I do will always be for them.

That being said, I would like to express my gratitude to my supervisor, Dr. Lucia Specia, for introducing me to the challenge of estimating machine translation quality and believing in me from the very beginning. Her constant guidance, encouragement and constructive criticism have been invaluable for me.

I am also indebted to my co-supervisors, Dr. Xavier Blanco and Dr. Constantin Orăsan, who gave me full support and positive feedback on many aspects of my work.

In addition, I would like to thank all my classmates and friends from the masters, who shared their life, knowledge and culture with me in the last two years. This experience would not have been the same without their company.

My gratitude goes also to Brian Davies for his thorough revision of this work and for being always willing to help me improve my English.

Last, but by no means least, I will always be grateful to the European Union for making my master's studies possible. This project was supported by the European Commission, Education & Training, Erasmus Mundus: EMMC 2008-0083, Erasmus Mundus Masters in NLP & HLT programme.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This introduction describes the main motivations behind this research (section 1.1), our objectives and research questions (section 1.2) and the main contributions of our work with regard to existing approaches (section 1.3). Finally, an outline of the structure of this dissertation is given (section 1.4).

1.1 Motivation

The decision to pursue this work on the automatic estimation of machine translation quality was due to a number of reasons. First and foremost, Machine Translation (MT) has become one of the most popular and necessary applications derived from the natural language processing field. From home users to big companies, it has been adopted as a convenient way to translate content automatically, either to understand or produce text in a foreign language. However, automatic translations can often be erratic and useless, which is why quality assessment becomes necessary to improve systems and use them wisely.

There are clearly two different perspectives from which the assessment of translation quality can be viewed: system development and system use. For system development, translation is generally *evaluated*, which implies the availability of gold standard translations. Thus, the output MT systems can be directly compared against human references and assigned a quality score by measuring their similarity. In fact, most MT evaluation metrics rely on lexical similarity measures, although some of them are more flexible than others.

From an end user's perspective, however, the lack of human references renders any evaluation irrelevant so an *estimation* must be done instead. Such estimation of quality is often based on characteristics derived from the source and target texts as well as information from additional resources. The work we pursue here follows this direction.

Our strongest motivation is actually studying how MT quality estimation may benefit from the use of linguistic information, following encouraging results in recent research.

Although shallow features have been extensively used in previous approaches, they are limited in scope. They convey no notion of meaning, grammar or content so they could be very biased towards describing only superficial aspects of translations. As a result, work on the automatic assessment of machine translations, either with or without references, has gradually moved from using only shallow features to include linguistic information, showing that it helps produce better estimations. In fact, linguistic features account for richer aspects of translations and are closer to the way humans make their judgements. Unlike most approaches which focus only on a few specific indicators, our work explores a rich variety of linguistic features that are derived from theoretical foundations of translation quality, most of which have not been used before.

In addition, we restrict the Quality Estimation (QE) approach to incorporate only MT system-independent features. This brings the advantage of enabling estimations even when the internal information of a translation system cannot be accessed, for example when using online translation systems.

Finally, assessing the quality of machine translations is essential in a variety of scenarios where reference translations are not available. Home users who translate content using online translation services are a typical example. In those cases, when a user has limited or no knowledge of the source language or cannot read the whole source text, they need to know in advance whether they can trust the translation, especially if it is used to make decisions. On the other hand, large corporations translating high volumes of text automatically need to know which specific portions of text need post-editing by human translators, and the same applies to freelance translators using MT systems to streamline their work. Provided an accurate quality metric could be used, the time translators spend assessing machine translations would be dramatically reduced and could be allotted to effective human translation instead.

1.2 Objectives

The aim of this work is to explore the effect of linguistic information in the automatic estimation of machine translation quality. In order to do this, we use a machine learning algorithm to build a regression model that scores translations on a scale from 1 (lowest quality) to 5 (highest quality). These assessments can also be seen as a measure of how much post-editing is needed to make them suitable for publication, ranging from 'requiring complete re-translation' to 'requiring little to no editing'.

Our specific objectives are summarised in the following research questions:

1. *How does the performance of QE models integrating linguistic features compare to models using only shallow features?*

   To this end, we compare models using only linguistic or shallow features with other hybrid sets. Model performance is measured in terms of prediction error and correlation with gold standard scores and significance tests are also carried out to determine whether any difference found is statistically significant.

2. *What are the best and worst performing linguistic features and why?*

   By applying different training schemes, we set out to discover which linguistic features are the best and worst in our models so that we can draw conclusions on the type of information they handle and how this correlates with quality. In addition, learning about the performance of individual features allows us to identify possible causes for performance.

3. *What is the best performing combination of features?*

   For this purpose, we apply a feature selection algorithm to find an optimal set of features that would maximise performance on our datasets. Further analysis of the resulting set will indicate the proportion of shallow and linguistic features and hence provide a valuable insight into the role of linguistic information.

Our experiments are restricted to translation from English into Spanish and features from the source and target text only, using a standard dataset produced for a shared task on quality estimation so that our results can be later compared to well-known baselines and state-of-the-art systems.

Additionally, we examine theoretical principles and criteria from translation studies from which we derive most of our linguistic features.

1.3 Main Contributions

There are basically two main contributions of our work to the task of QE for machine translations. The first of them is the comparative study of models containing shallow versus linguistic features, in addition to a few hybrid sets. Although other approaches have proposed the use of shallow and linguistic information, none of them was targeted at contrasting models using only one type or the other, let alone comparing them to hybrid sets. Moreover, we present a detailed study of pros and cons of a much wider range of linguistic features and how they can complement shallow features in optimal hybrid sets.

Our second significant contribution is the introduction of novel linguistic features that have not been explored before, such as information about subject-verb agreement

or lexicon estimation using a spell checker. Furthermore, most of our linguistic features are derived from theoretical criteria, which also constitutes an innovative approach.

Part of this work was also submitted to the Quality Estimation Shared Task of the Seventh Workshop on Statistical Machine Translation (WMT 2012) and accepted for publication (Felice and Specia, 2012). However, the analysis and results presented in this work vary from those in the aforementioned publication because of differences in the implementation of some features (see section 4.2.5).

1.4 Dissertation Outline

The remainder of this work is structured in four chapters. Chapter 2 introduces machine translation assessment (section 2.1) and describes the difference between reference-based (section 2.2) and reference-free (section 2.3) approaches. This last section also makes a distinction between Confidence Estimation (section 2.3.1) and Quality Estimation (section 2.3.2). In Chapter 3, we summarise guidelines and criteria for producing and assessing translations as suggested in specialised literature (section 3.1) and go on to derive related computational features that are later used in our models (section 3.2). Chapter 4 describes our experimental setup (section 4.1), including datasets, resources used to extract the proposed features, evaluation metrics and training of our models. Section 4.2 provides the results of our experiments together with a detailed analysis of performance. Finally, Chapter 5 presents conclusions and directions for future work.

Chapter 2

Related Work

This chapter gives an overview of the main approaches to the assessment of machine translations. First, a brief introduction is given (section 2.1), followed by a quick review of reference-based evaluation and its most significant metrics (section 2.2). After that, reference-free estimation approaches are described (section 2.3), defining the key concepts of Confidence Estimation (section 2.3.1) and Quality Estimation (section 2.3.2). Finally, a short summary of potential applications of reference-free approaches is given (section 2.3.3).

2.1 MT Quality Assessment

Machine translation has always been one of the most ambitious tasks within natural language processing. While many commercial systems in use today employ a classic rule-based approach (Way, 2010), newer systems make use of Statistical Machine Translation (SMT), where correspondences between the source and target language are learnt automatically from parallel corpora (Koehn, 2010). Specifically, phrase-based SMT (Koehn, Och and Marcu, 2003) is the most successful approach today, showing consistent improvement on general domains (Callison-Burch, Koehn, Monz, Peterson et al., 2010; Callison-Burch, Koehn, Monz and Zaidan, 2011). However, achieving high translation quality is still an open issue for MT systems, especially for long and complex sentences.

The design of automatic MT assessment metrics is a challenging task. Firstly, they may use reference translations (section 2.2) or not (section 2.3), which largely determines their purpose and application. Secondly, they need to be versatile enough so as to allow variations in translation but at the same time penalise oddities and deviations from the original. Lastly, they should exhibit good correlation with human judgements in order to be reliable. Given that formalising the factors that determine the quality of translations is a hard task (see section 3.1), it is not surprising that defining successful metrics is also difficult.

		WER	TER	BLEU-2
REF.	Israeli officials are responsible for airport security			
TRANS. A	Israeli officials responsibility of airport safety	57%	57%	9%
TRANS. B	airport security Israeli officials are responsible	71%	28%	61%

Table 2.1: Example of WER, TER and BLEU-2 scores for two translation hypotheses and one reference, based on examples by Koehn (2010).

Most approaches make use of 'shallow' information but there has also been a noticeable move towards the inclusion of linguistic information in the latest research. In fact, much of the previous work is actually devoted to exploring what *features* are the most significant and where they should be extracted from, be it the source text, output translation, references, MT system parameters or a combination of those.

The following sections describe in more detail the two principal approaches and metrics proposed so far.

2.2 Reference-based Evaluation

Evaluating machine translations is essential for the development, improvement and fine-tuning of MT systems and is typically addressed by computing similarity metrics between system output and human references. Quality is thus measured in terms of 'human-likeness', under the assumption that the more a machine translation resembles a human translation, the better it is, although this is not universally accepted (Albrecht and Hwa, 2007a).

While some of these metrics are based solely on edit distance, like PER (Tillmann, Vogel, Ney, Zubiaga et al., 1997), WER (Nießen, Och, Leusch and Ney, 2000) and TER (Snover, Dorr, Schwartz, Micciulla et al., 2006a), others concentrate on more lexical aspects. The popular BLEU metric (Papineni, Roukos, Ward and Zhu, 2002), for example, relies on n-gram overlapping and has been specially designed to approximate human judgements at the corpus level, thus performing poorly on individual sentences. Other metrics built on similar ideas include NIST (Doddington, 2002), ROUGE (Lin and Och, 2004), GTM (Melamed, Green and Turian, 2003; Turian, Shen and Melamed, 2003) and METEOR (Banerjee and Lavie, 2005). Table 2.1 shows an example evaluation using three of these metrics: WER, TER and BLEU-2 (a version of BLEU based on unigrams and bigrams).

These n-gram lexical metrics have been the subject of strong criticism, especially BLEU (Callison-Burch, Osborne and Koehn, 2006), which has been found to contradict human judgements in many cases. In addition to this fact, reference-based evaluation suffers from two major limitations. Firstly, there can be many different good translations for a given source text but a reference translation represents only one of them. As a result, if an automatic translation fails to match its reference,

it will be regarded as bad although this is not necessarily the case. Although most metrics try to minimise this problem by allowing the use of multiple references, collecting multiple translations for a single input text is difficult and expensive.

Secondly, given that references are produced by human translators, they are not only limited but also inconvenient and expensive in practice. As a result, evaluation metrics can only be computed on texts with existing human translations, which greatly restricts their application. A common workaround, however, is to use the output from auxiliary MT systems as *pseudo-references* (Albrecht and Hwa, 2007b, 2008) but this is far from an optimal solution. In consequence, this type of metric is acceptable for system evaluation but clearly unsuitable for evaluating ad-hoc translations.

In order to overcome these limitations, a range of reference-independent methods have been proposed, which are described in section 2.3.

2.2.1 Lexical Features vs. Linguistic Features

Most recent works put heavy emphasis on linguistic information in clear contrast to earlier approaches. In fact, they are proposed as alternatives for standard metrics which are often criticised for taking account of lexical aspects only. On the contrary, these new efforts integrate information from different linguistic aspects like syntax and semantics to provide better approximations of overall translation quality. This has led to a classification of approaches in two different groups: 'lexical' and 'linguistic' metrics.

Work by Amigó, Giménez and Verdejo (2009), for instance, describes a combination of standard lexical metrics with syntactic and semantic information to generate a new metric called ULC. Experiments carried out to analyse the correlation of this metric with human judgements at the system and phrase level revealed that in both cases the new metric achieved better results than individual lexical metrics. It was thus concluded that additional linguistic information contributed to overcoming the intrinsic limitation of lexical metrics. A further description and exploration of this approach is given by Giménez and Màrquez (2010).

Other approaches employ machine learning techniques to integrate n-gram statistics with simple linguistic features. One such example is the ROSE metric (Song and Cohn, 2011), which combines precision and recall from lexical n-gram matches with information on punctuation, content and function words, and even n-grams over part-of-speech (PoS) tags. Scoring performance using Support Vector Machines (SVM) regression with a linear kernel function fell slightly below BLEU while ranking results were mixed: for translation from foreign languages into English, ROSE was significantly better than BLEU but the opposite was found for translation in the opposite direction.

Yang, Sun, Zhu, Li et al. (2011), on the other hand, experiment with regression and ranking algorithms using as few as six linguistic features, such as information on content words, noun phrases and a parsing score. Results show that regression achieves better generalization performance than previous attempts and correlation with human judgements is higher than that of non-linguistic metrics. Ranking performance, however, was not found to be as good as initially presumed. In all, the most interesting contribution of this work is the empirical demonstration that very few linguistic features may suffice for improving the performance of existing lexical metrics.

2.3 Reference-free Assessment

Reference-free translation assessment approaches emerged to overcome the limitation of evaluation metrics that required human references for their computation. Although this approach enables much wider application, the lack of the expected outputs (references) make it impossible to measure quality in terms of segment comparison so new information has to be exploited as a result. The following sections give a chronological overview of the most significant attempts and introduces key concepts that are fundamental to the methodology used in this work.

2.3.1 Confidence Estimation

The first attempts at assessing MT quality without references were conceived as *Confidence Estimation* (CE) problems. Under this methodology, MT assessment is seen as the problem of estimating how confident systems are in their output rather than how good translations are objectively. Instead of using human references, CE metrics rely mainly on parameters and information of the MT system and the translation process, in addition to a few complementary features from the source and target text. Given that system features are the primary focus of this approach, they are also referred to as *confidence features* in the latest research.

In their seminal work, Blatz, Fitzgerald, Foster, Gandrabur et al. (2004) set the objectives of CE for the assessment of machine translations and provided the first experimental results. The aim of their work was to provide a numerical quantity which would serve as a direct indicator of translation quality, making a distinction between *weak* and *strong* CE: while the former yields a free numerical score which can later be mapped onto a quality scale or even a probability of correctness, the latter is directly aimed at estimating such probabilities.

Experiments were carried out for the purpose of evaluating quality at both sentence level and sub-sentence level (i.e. words and n-grams) using Chinese to English datasets. In both cases, the aim was to train a system from samples

Group	**Description**
Word alignment information	Maximum and average alignment distance between words in the source and target texts.
Average target word statistics	Average word frequency in the output translation and n-best hypotheses (*n-best lists*) plus a few variations.
Selected parameters of the translation model	Log probability of the translation hypothesis based on a trigram model, length penalty, cost of using auxiliary translation rules, etc.
Basic syntactic information	Number of mismatched parentheses and quotation marks.
Centre hypothesis features	Edit distance from the n-best hypotheses to the central hypothesis.
IBM Model 1[1] features	Conditional log probability of the source given the target sentence and vice versa.
N-best lists	Information on hypothesis ranking, average length, list density and scores.
Phrase-based language models	1, 2 and 3-gram probability and perplexity of the target sentence plus type-token ratios.
Search based features	Scores and pruning information from the decoding process.
Semantic similarity	Statistics on strongly related words within the target sentence.
Sentence length features	Number and ratios of tokens in the source and target segments.
Source language model	Log probability and perplexities of the source sentence.
Source n-gram frequency statistics	Percentage of 1, 2 and 3-grams in different frequency quartiles of an auxiliary corpus for the source sentence.
Target language model	Percentage of 1, 2 and 3-grams in different frequency quartiles of an auxiliary corpus for the target sentence plus other length-based features.
Translation consistency features	Average number of times a source word or phrase is translated identically in the n-best list.

Table 2.2: Summary of features proposed by Blatz, Fitzgerald, Foster, Gandrabur et al. (2004).

associating a set of features to a quality label by using a variety of machine learning techniques. Other proposals have also been made for word-level error detection (Xiong, Zhang and Li, 2010; Bach, Huang and Al-Onaizan, 2011). However, given the aim of this work, we only focus on sentence level experiments.

A total of 91 features accounting for different aspects of translation were proposed for the sentence-level task. Following the CE principle, references were not used for the extraction of features but were used instead for computing automatic quality scores for each segment as a replacement for manual annotation. Table 2.2 summarises the types of features used in these experiments.

[1]The simplest of IBM translation models that estimates lexical translations and their probabilities from a source and target text (Brown, Della Pietra, Della Pietra and Mercer, 1993).

All the sentences in their datasets were assigned automatic quality scores based on NIST or a modified version of WER using auxiliary reference translations. Different machine learning algorithms were subsequently tested with the resulting dataset in order to learn a quality function. Results showed that evaluation was harder at the sentence level than at subsentence level, that multilayer perceptrons outperformed Bayes models and that features based on the target text and n-best lists delivered better results than the rest. Nevertheless, the power to discriminate good from bad translations achieved in the experiments was found to be poor, apparently due to the use of automatic metrics that correlate only moderately with human judgements.

For this reason, an additional experiment with human annotators was carried out so as to analyse the correlation of six popular automatic metrics with human judgements at the sentence level. This revealed that there exists both low inter-annotator agreement and low metric correlation.

Quirk (2004) carried out similar CE experiments using a considerably smaller dataset tagged with human scores instead. Again, a series of machine learning alternatives were evaluated among which simple linear regression was found to be the best. His results also served to prove that a relatively small set of translation samples with human quality annotations achieves better performance than a large set of samples with automatic quality scores.

Translation quality has also been strongly associated with fluency. Work by Gamon, Aue and Smets (2005) describes the combination of an SVM classifier and language models in order to study the correlation with human-rated fluency and overall quality at the sentence level. Results show that such estimated scores fall slightly behind those for BLEU but were still considered successful given the fact that reference translations were not required for their computation. Other attempts targeted at fluency and adequacy (Albrecht and Hwa, 2007a,b, 2008) have introduced the use of *pseudo-references* instead, which are nothing but translations of the same sentence produced by other MT systems. Vectors were constructed for each of the sentences in the chosen dataset comprising the scores of several standard metrics computed from the pseudo-references (like BLEU, WER, PER, METEOR and ROUGE) together with fluency scores derived from target-language treebanks. Features were later paired with human quality judgements and used to build an SVM regression model in order to produce an automatic quality estimator. The study concludes that although the best correlation rates are achieved with human references, pseudo-references are still a viable option since they produce comparable results. In fact, the regression metric trained on machine references outperformed human-based BLEU and METEOR scores, with regression estimates performing consistently better than any of the standard reference-based metrics alone.

Black-box features	**Glass-box features**
Sentence length	SMT model score
Type-token ratio	Phrase and word probabilities
N-gram log probability and perplexity	Information from n-best hypotheses
Mismatches in punctuation	Aborted nodes in the decoder's search graph
PoS tagging on source and target texts	Untranslated words

Table 2.3: Examples of black-box and glass-box features.

The choice of features for machine learning techniques is undoubtedly a key issue for the accurate estimation of translation quality. Specia, Turchi, Cancedda, Dymetman et al. (2009) draw a clear distinction between *black-box* and *glass-box* features. The first group refers to features that are essentially extracted from the source and target text, including both individual and contrastive features (e.g. individual lengths and their ratio). In addition, they can also include features extracted with additional resources as long as they are related to the source and target text only, such as n-gram statistics or part-of-speech tags for a sentence. On the other hand, glass-box features use information from the MT system that produced the translation and the decoding process, such as hypothesis scores, n-best lists or phrase probabilities. Features computed from resources used to train an SMT system, such as its parallel corpus or derived translation tables, could also be considered within this group since they use privileged information that is not normally available to end users (see section 3.2.1.2) but this distinction was not originally made by the authors. Given the clear focus of these two groups, it is then natural to consider glass-box features as *CE features* and black-box features as *non-CE features*. A few examples are included in Table 2.3.

Experiments were conducted using Partial Least Squares regression on four different datasets containing: 1) automatically annotated NIST scores, 2) 1-to-4 human scores according to post-editing needs 3) 1-to-5 human adequacy scores and 4) word-based post-editing time. Learnt NIST scores deviated an average of 1.5 points on a scale from 1 to 18, which was considered acceptable, while the learnt 1-4 and 1-5 human scores deviated on average 1.1 and 0.6-0.7 points respectively. Conclusions regarding post-editing time, however, were less homogeneous among systems and require further study.

One revealing finding of this work is that black-box features are generally more discriminating, with very little gain from glass-box features for manually annotated datasets. Although this seems to contradict the observations by Blatz, Fitzgerald, Foster, Gandrabur et al. (2004), their work uses only automatic scores, which is why their conclusions cannot be directly compared.

Many of the latest developments in confidence estimation have been specifically targeted at aiding post-editing. Based on previous work (Specia, Turchi, Cancedda,

Dymetman et al., 2009; Specia, Turchi, Wang, Shawe-Taylor et al., 2009), He, Ma, van Genabith and Way (2010) proposed the integration of an SMT system and a Translation Memory (TM) in order to accelerate a post-editor's job. Using the proposed solution, whenever a new sentence in a source language needs to be translated, two candidate translations are obtained: one from the TM and another from the SMT system. Later, an auxiliary system which has been trained to recognise which of the two translations requires less post-editing, decides on the best translation and proposes that one to the user. Hence, post-editing times should be reduced given the 'intelligent' decision of the recommendation system.

Experiments were conducted by using part of an English-French translation memory, the phrase-based SMT system Giza++ (Och and Ney, 2003) and a binary SVM classifier. Training samples were obtained by using a separate subset of source-language sentences from the same TM. For each of these sentences, the two candidate translations from the SMT system and the TM were generated and automatically annotated with TER scores as a way of representing post-editing effort.

The set of features proposed in that work can be classified into three categories: SMT features (such as phrase-translation model scores, language model probability and word penalty), TM features (namely the *fuzzy match* cost of a suggestion) and system-independent features (like source and target language model perplexities and a pseudo-source fuzzy match score using a re-translation of the candidate into the source language). Following our earlier classification, the first and second groups constitute pure CE features while the third comprises non-CE features.

Despite training on an indirect measure of post-editing effort, the results obtained in such experiments were encouraging. In particular, 0.85 precision and 0.89 recall were registered while a considerable reduction of post-editing effort was said to be achieved by using the recommendation system. A later study with human post-editors presents even better results (He, Ma, Roturier, Way et al., 2010).

In a similar fashion, Specia and Farzindar (2010) proposed new models for predicting translation quality within a post-editing environment. Experiments were conducted with two different datasets. The first of them comprised French-English translations from the legal domain evaluated using Human-targeted Translation Edit Rate (HTER), a variation of TER that uses a post-edited hypothesis as reference (Snover, Dorr, Schwartz, Micciulla et al., 2006b). The second dataset included the same Spanish-English sentences used in previous work (Specia, Turchi, Cancedda, Dymetman et al., 2009) but enriched with TER scores. Each of these datasets was then used to build SVM regression models using a total of 86 shallow features from the source and target texts as well as auxiliary corpora. While results revealed some difficulty predicting scores accurately, correlation was higher for HTER, probably owing to the use of post-edited versions of the sentences rather than free references.

Overall, the approach was found to be a computationally cheap alternative for estimating post-editing effort.

From a different perspective, González-Rubio, Ortiz-Martínez and Casacuberta (2010) considered the post-editing effort estimation problem within an Interactive-predictive Machine Translation framework (i.e. interacting with an SMT system during the translation process). Confidence measures were proposed for filtering the initial translations generated by the SMT system before actually presenting them to the user, in an attempt to save them time editing translations that were judged to be already acceptable. Using a single word-level feature derived from an IBM Model 1 system, the authors computed two different sentence-level measures that were used to evaluate translation candidates produced by the SMT system. Since these scores are estimated, it is possible that some translations considered acceptable do have errors. However, such a loss in quality was not found to be extremely significant and was hence tolerated in order to favour a reduction in post-editing effort.

2.3.1.1 Combination of Approaches

As described earlier, reference-based and confidence estimation metrics tackle the MT quality problem from two clearly distinct perspectives. While the former are more suitable for comparing different MT systems and have fairly good correlation with human judgements at the corpus level, the latter are ideally suited for analysing only one MT system and correlate better with human judgements at the segment level (usually sentences). For this reason, studying how they could be used in conjunction to provide better estimates of MT quality is highly desirable.

To the best of our knowledge, the first experimental approach to the subject is the work by Specia and Giménez (2010). Their report describes the use of both classic reference-based metrics (such as BLEU, METEOR, TER or WER) and CE scores as features for SVM regression. Experiments were conducted using different datasets containing human scores for post-editing effort and the results were evaluated in terms of correlation.

The individual performance of reference-based and CE metrics was then compared to that of a few proposed combinations, including a simple averaged scheme and SVM regression. Results show that hybrid metrics yield better correlation with human judgements at sentence level than each individual metric, with statistically significant improvements in most cases. Nevertheless, it has also been observed that some datasets had greater affinity with either the reference-based or CE metrics, mainly due to their composition and the type of score used in each case. In spite of this, the hypothesis that a combination of metrics would yield better results than metrics on their own was actually confirmed. The highest correlation rate reported was 0.608 over the English-Spanish dataset for a hybrid metric using SVM regression.

Although their work provides an alternative method for MT evaluation that is robust and covers many more aspects than standard metrics, its application is limited by a number of factors, such as the need for linguistic resources and the availability of human references. As a result, the methodology is mainly suited for testing and tuning MT systems.

2.3.2 Quality Estimation

Quality Estimation (QE) has emerged as a broader term to refer to reference-free quality assessment (Specia, Raj and Turchi, 2010). Although initially used to describe reference-free approaches that did not include system-dependent features, it is now used to refer to any quality prediction system in general, regardless of whether they use glass-box (CE) or black-box (non-CE) features.

This general conception has allowed researchers to abandon the idea of quality assessment as a system-dependent task and experiment with features from different sources. As a result, the focus is no longer on how confident a particular system is about its output but rather how good a translation is on its own. Given this new scope, many approaches seem to have moved from using CE features to using non-CE features.

Estimating MT quality using only system-independent features may be desirable for a number of reasons. First of all, it enables the assessment of translations without requiring access to the internal components of an MT system, which is ideal for situations where commercial systems are used. Secondly, system-independent approaches allow the assessment of translations produced by any MT system regardless of their paradigm (rule-based, statistical, hybrid, interactive, etc.). Thirdly, it would also be possible to assess human translations, for example to find the most appropriate one out of a pool in collaborative environments. Lastly, ignoring system features makes an approach less computationally costly. However, relying only on black-box features could also be disadvantageous. Specifically, it can be very challenging to produce a set of features that could be as discriminative as CE features, specially when this type of feature has been found to be very informative in previous work (Blatz, Fitzgerald, Foster, Gandrabur et al., 2004).

The first experiments using only system-independent features were proposed by Specia, Turchi, Wang, Shawe-Taylor et al. (2009), although their approach is further developed in a later report (Specia, Raj and Turchi, 2010). Their proposal makes use of 74 features, of which 39 relate to the source text and the remaining 35 to the evaluated translation. A model was built using SVM epsilon regression and was later evaluated in terms of correlation with human judgements. In all cases, the correlation coefficients for the QE approach were found to be significantly higher than those for common MT evaluation metrics, with gains of over 50% in many cases.

Soricut and Echihabi (2010), on the other hand, proposed a model for document-level ranking that includes a variety of system-independent features from the source and target texts as well as additional corpora and pseudo-references.

2.3.2.1 Introduction of Linguistic Features

Most of the approaches for QE described so far have made use of similar shallow features, such as n-gram statistics, token counts, aspects of punctuation, etc. Given that the information provided by these features can be limited, new approaches have started to exploit linguistic information in an attempt to account for richer aspects of translation that could contribute to a more accurate assessment of quality.

Xiong, Zhang and Li (2010), for instance, combined word posterior probabilities with a few linguistic indicators in order to train a binary classifier for word-level error detection. Their linguistic features were grouped into two categories: lexical information (the surface form of a word and its PoS tag) and syntactic information (whether a word is linked to other words in a sentence according to a parser). Experiments were conducted using different combinations of features with a maximum entropy classifier and automatic binary classes derived from WER scores over reference translations. A comparison of their results with previous work revealed that linguistic features were able to produce better estimations than word posterior probabilities, especially when more than one feature was used. In addition, it was observed that combining all the features achieved even better results, supporting the hypothesis that linguistic indicators are complementary to shallow information.

Specia, Hajlaoui, Hallett and Aziz (2011) used linguistic features for estimating the *adequacy* of translations from Arabic into English. A total of 122 system-independent features were used to build an SVM regression model to predict METEOR scores and three different classifiers using 1-4 human annotations. Apart from many common features from the source and target texts, a set of contrastive features was introduced to capture the accuracy of translations, such as ratios between source and target length, content and function words, PoS tags and differences in constituents, dependency relations and named entities. Although the proposed classification models always outperformed a majority class classifier, the contribution of adequacy features was not found to be consistently beneficial. On the other hand, regression results were more optimistic, deviating as little as 10% from the expected scores.

In another approach, Hardmeier (2011) proposed the exploitation of constituency and dependency relations from the source and target texts independently to improve estimates. Experiments were conducted using the same English-Spanish human-rated datasets as in other previous approaches (Specia, Cancedda and Dymetman, 2010) plus another English-Swedish subtitle dataset also manually annotated with scores from 1 to 4.

Constituency and dependency trees were produced for the source and target sentences and integrated into a binary SVM classifier along with a subset of features previously used by Specia, Turchi, Wang, Shawe-Taylor et al. (2009). A comparison of different classifiers revealed that the models using only syntactic tree kernels perform slightly worse than those trained on explicit features but the best performance is actually achieved in models that integrate both types of information. In consequence, this demonstrated that the addition of syntactic information helped improve performance of shallow models.

In a more complex scenario, Bach, Huang and Al-Onaizan (2011) addressed word-level QE by using linguistic features derived automatically from annotated parallel corpora. Their approach defines three basic types of features: 1) source side features, which associate the occurrence of source PoS tags or tokens to target words 2) alignment context features, which associate neighbouring words in the source text with target words, and 3) dependency structure information, which encodes similarities between dependency relations in the source and target texts. Different 'instances' of these features are then learnt from a very large parallel corpus and subsequently used to build a weighted vector model that predicts a binary class for each target word ('good' or 'bad'). In another variation, a classifier was trained to output a label related to a post-edited version ('good', 'insertion', 'substitution', 'shift'). These word-level scores were then used for estimating sentence-level scores using the Viterbi algorithm.

Experiments show that the addition of the proposed linguistic features to existing datasets produces consistently better results than models without these features. Moreover, the authors report a Pearson correlation coefficient of 0.6 between their sentence-based predictions and HTER scores.

Finally, work by Pighin and Màrquez (2011) describes a methodology for projecting source annotations onto target translations so that they can be exploited for QE. As an example, the authors projected predicate-argument labels from English into Spanish and used these estimates for pairwise ranking quite successfully.

It should also be noted that all these approaches use different representations of linguistic information. Xiong, Zhang and Li (2010), for example, use strings for lexical features and a binary indicator for syntactic information, Hardmeier (2011) employs tree kernels that are particularly suitable for SVM algorithms, Bach, Huang and Al-Onaizan (2011) create a large number of binary features from word and PoS patterns and Pighin and Màrquez (2011) generate projected predicate-argument structures. In our work, however, we follow the approach by Specia, Hajlaoui, Hallett and Aziz (2011) and use attribute/value pairs as a way of avoiding complex and costly representations.

2.3.3 Applications

Reference-free MT assessment has been suggested for a wide range of applications:

- *Hypothesis Re-ranking.* Confidence estimation measures have been proposed and tested for re-ranking the n-best list of candidates handled by SMT systems, instead of relying on their internal score (Blatz, Fitzgerald, Foster, Gandrabur et al., 2004; Bach, Huang and Al-Onaizan, 2011).

- *Aiding translation and post-editing.* Experiments using CE scores and SMT systems interactively have proved to reduce translation effort in exchange for a slight decrease in quality (González-Rubio, Ortiz-Martínez and Casacuberta, 2010). Similar aims are pursued by using translation memories (He, Ma, van Genabith and Way, 2010; He, Ma, Roturier, Way et al., 2010), visual post-editing environments (Bach, Huang and Al-Onaizan, 2011) or prediction models based on HTER, post-editing effort and time (Specia, 2011a).

- *Warning users about unreliable translations.* The reliability of a translation could be a determining factor in deciding whether it can be trusted to make a decision or should only be considered for gist. In this regard, works on the estimation of adequacy (Specia, Hajlaoui, Hallett and Aziz, 2011) and word-level errors (Bach, Huang and Al-Onaizan, 2011) may be particularly helpful.

- *Document ranking.* Soricut and Echihabi (2010) have employed a QE approach for producing a ranking of translated documents according to their quality. Although their system is aimed at ranking translations of different source documents rather than alternative hypothesis for a single source text, they show QE can also be useful for document selection.

- *Filtering.* As in binary classification, many authors suggest the use of quality scores for filtering bad translations in post-editing environments (Gamon, Aue and Smets, 2005; Specia and Farzindar, 2010; Specia, Raj and Turchi, 2010). This would allow translators to edit sentences which are only worth editing and retranslate bad ones completely, for example.

- *System combination.* Many authors have explored the possibility of using CE scores to compare and select the best translations produced by different SMT systems (Blatz, Fitzgerald, Foster, Gandrabur et al., 2004; Quirk, 2004; Specia, Raj and Turchi, 2010). A combination of the best translations would thus yield better overall results than using only one system.

Despite the number of described proposals for QE, this task is still relatively new and gradually arousing interest within the MT research community. In particular,

the use of linguistic features is an open problem and requires further research in order to develop more mature models. It is then expected that specialised meetings and competitions in this area, such as the recent WMT 2012 Quality Estimation Shared Task, would help develop this line of research as much as they have helped others in the past.

Chapter 3

Estimating MT Quality

Quality is never an accident. It is always the result of intelligent effort.

—John Ruskin

The first part of this chapter provides a summary of key theoretical principles and criteria used to assess translations (section 3.1), which we use to derive our computational features. The rest of the chapter presents a classification scheme (section 3.2.1) and a detailed description of each implemented feature (section 3.2.2).

3.1 Translation Quality

Having a clear definition of what translation quality means and how it can be measured is essential to learn how to assess it properly. However, there is no straightforward answer to this question. Many translation experts have tried to define the principles of good translation over the years but there are, in fact, no fixed rules that guarantee a good result. The way a text is translated depends on a number of factors such as the nature of the message, its purpose and audience (Weissbort and Eysteinsson, 2006) so it is difficult to formulate a generalisation.

Nevertheless, there are some very general aspects that translations are expected to exhibit. One of the most classic quotations from the field of translation studies is from Tytler (1791), who postulated the following three 'General Rules' for translation:

1. A translation should give a complete transcript of the idea of the original work.
2. The style and manner of writing should be of the same character as that of the original.
3. The translation should have all the ease of the original composition.

Renowned translator Yan Fu (1973) also stated his three translation principles as **fidelity**, **fluency** and **elegance**, greatly influencing Chinese translation during

the 20th century. Another notable contribution is that of Nida, who introduced the concepts of **functional equivalence** to describe fidelity to form and content (Nida, 1964) and **dynamic equivalence** (Nida, 1964; Nida and Taber, 1969) to characterise a natural-sounding translation that produces the same effect as the original.

Theoretical discussions on the desirable aspects of translation go far beyond this summary, although most of them revolve around the same matters. Munday (2008) gives an extensive chronological description of such theories.

In addition, many authors have concentrated on criteria for assessing rather than producing translations. House (1977, 1997), for instance, proposes a 'model' for Translation Quality Assessment (TQA) that compares a source and target text on dimensions such as function, genre, register and language use. The following list is a summary of the most significant parameters proposed for QTA by different authors (al Qinai, 2000):

- Textual typology and tenor.
- Formal correspondence.
- Coherence of thematic structure
- Cohesion.
- Text-pragmatic (dynamic) equivalence
- Lexical properties (register).
- Grammatical/syntactic equivalence.

Since our interest in translation quality is eminently practical, we also considered hands-on guidelines for translators from which we could derive other indicators of quality. In this regard, it is worth mentioning the criteria used by the Institute of Linguists in the United Kingdom to assess translations for their Diploma in Translation (IoL, 2011):

1. Comprehension, accuracy and register.
2. Grammar (morphology, syntax, etc.), cohesion, coherence and organisation of work.
3. Technical aspects: punctuation, spelling, accentuation, transfer of names, dates, figures, etc.

Additional detailed criteria for examination purposes are proposed by the Association of American Translators (Doyle, 2003).

3.2 Features

In order to build a computational model for quality estimation, we had to encode the formal aspects of translation quality described in section 3.1 into representations that could be stored as feature-value pairs and processed by machine learning algorithms. A classification and detailed description of such representations (*features*) is provided in the following sections.

3.2.1 Classification Criteria

By classifying our features, it is possible to understand their differences and similarities as well as make some comparisons. The classification criteria we have put forward are explained below.

3.2.1.1 Linguistic Knowledge

Since the purpose of our work is to study the contribution of linguistic information to QE, we must classify our features into two distinct groups: **linguistic** and **shallow** features.

We consider linguistic features those that require at least some minimal knowledge of general linguistics or the language they operate on, such as the number of nouns in a sentence, information on phrase structure or the lexicon of a certain language. By contrast, shallow or non-linguistic features can be extracted without requiring proper linguistic knowledge and could range from the number of tokens[1] or punctuation marks[2] in a sentence to statistical data such as n-gram probabilities.

3.2.1.2 Origin

In reference-free MT assessment, features can be extracted from three main sources: the source text (original), the target text (translation) or the system that generated the translation.

Features that exploit information from the MT system or the process used to generate the translation (such as phrase table probabilities or an SMT decoder score) are often called **glass-box** features, since they can 'see' how the translation was produced. On the other hand, features extracted from the source and target text as well as some relationships between them are called **black-box** features, since they are totally uninformed about the translation process. A third category of **grey-box**

[1] Note that we have used the term *token* and not *word* here. While a token can be defined as a string of consecutive non-blank characters that conveys no a priori meaning, a word is a linguistic concept that carries semantic content and is part of the lexicon of a language.

[2] Although the notion of punctuation is essentially linguistic, we consider it very elementary and shallow for the purpose of assessing translation quality.

features can be attributed to those that are not strictly glass-box but somehow integrate information related to an MT system, such as statistics derived from the parallel corpus used for training.

It is usually stated that the origin of features is also a source of specific indicators which are useful for QE. Thus, features from the source text, the MT system and the target text are regarded as complexity, confidence and fluency indicators respectively (Specia, Hajlaoui, Hallett and Aziz, 2011).

3.2.1.3 Language Dependence

Many of our linguistic features are built on aspects that are common to many languages while others are specially tailored, and tied, to a particular language. We call the first group **language-independent** features because they apply to many languages and can often be extracted by using general-purpose resources, such as PoS taggers or corpora. Example features in this group include PoS counts, content and function words and named entities.

On the other hand, **language-dependent** features are specific to a language and rarely applicable to others, mainly because the observed phenomena does not exist in another language or they require substantial change. Examples of these include the identification of zero subjects and missing contractions in Spanish (like *al* or *del*).

A notable difference between these two categories lies in implementation. While language-independent features can be ported to other languages by using different underlying language-dependent resources (Specia, 2011b), this is not the case for language-dependent features since the linguistic phenomena they capture is generally not cross-lingual.

3.2.1.4 Resource Dependence

The difference between **resource-dependent** and **resource-independent** features lies in the fact that the former require external resources for their extraction (such as linguistic processors or corpora) while the latter can be computed using only the text that is being analysed, with no additional information. N-gram probabilities and counts of PoS tags are common examples of resource-dependent features whereas Type/Token Ratio and its variations are resource-independent.

3.2.1.5 Aspect

Features can also be classified according to the criteria that have been described in section 3.1. However, we will restrict our categories to the following aspects:

Complexity: degree of difficulty of the source text, a dimension that is often considered in many reference-free automatic approaches (Specia, Turchi, Wang, Shawe-Taylor et al., 2009; Hardmeier, 2011; Specia, Hajlaoui, Hallett and Aziz, 2011)

Fidelity (also **accuracy/adequacy**): dynamic equivalence (Nida, 1964; Nida and Taber, 1969), faithfulness to the meaning of the original text (Munday, 2008).

Grammaticality: conformity to the grammatical rules of a language.

Fluency: smooth, effortless and natural use of language (Brumfit, 1984; Crystal, 2010); closely related to grammaticality (Giannakopoulos, Karkaletsis and Vouros, 2012).

Coherence and cohesion: logical and meaningful structure of a text that makes it easy to read and interpret (Halliday and Hasan, 1976; Louwerse and Graesser, 2005).

Structure: surface technical aspects, as described by the Institute of Linguists (IoL, 2011).

It should be noted that these dimensions are not mutually exclusive which is why a single feature might be classified into more than one of these categories. One such example is subject-verb agreement, which is a sign of both grammaticality and fluency.

3.2.2 Proposed Features

The models we present in this work are based on a set of 147 features that attempt to identify many of the quality indicators that are described at the beginning of this chapter. Although a few more features were initially proposed, their implementation was thwarted by the lack of suitable resources, especially for Spanish. As a result, only those features that we were able to compute automatically are described. In addition, we use features extracted from the source and target texts only, which is why glass-box features are not included.

With regard to feature representation, we found that features that contrast information from the source and target text could be expressed in a variety of ways but this has rarely been explored in previous research. In order to find a suitable representation for this type of feature, we tested different relational schemes as part of our feature engineering process, including ordinary subtraction, division, trigonometric ratios and modified versions of precision and recall. Results showed that subtraction was the most appropriate combination scheme although keeping the

original information separate seemed to be even more effective than contrasting it in a single feature. All feature values in our experiments were treated as real numbers.

The following lists provide a description and classification of each feature together with a unique identifier that is used to refer to them throughout the rest of this work. Related features are grouped for the sake of clarity and examples are given in either the source or target language and in good or bad grammatical form to show how this affects estimations. Novel features are marked with * .

3.2.2.1 Linguistic Features

1-5 *Content words.* These refer to the words in a language that carry the meaning of a sentence, such as nouns, full verbs and adjectives (van Gelderen, 2005). They are also generally called **open-class** words, since they allow the addition of new items (a new noun, for example). With the help of a PoS tagger, we extract the number and proportion of content words in the source (`source-cont-words`, `source-cont-words-pcent`) and target sentences (`target-cont-words`, `target-cont-words-pcent`) and compare them (`s-t-ratio-cont-words-pcent`) as a way of measuring semantic content.

Identifiers:

`source-cont-words` `target-cont-words`
`source-cont-words-pcent` `target-cont-words-pcent`
`s-t-ratio-cont-words-pcent`

Example:

Cleft lips and palates affect around one in 700 babies born in the UK.
`source-cont-words` = 7

6-9 *Function words.* These include words such as determiners, pronouns, prepositions, conjunctions, auxiliary verbs and adverbs of degree (van Gelderen, 2005) that help combine content words to form sentences. They are also called **closed-class** words because additions to this set are not normally permitted. These words are usually considered **stopwords** in automatic text processing.

Absolute numbers and proportions are also extracted from the source (`source-func-words`, `source-func-words-pcent`) and target sentence (`target-func-words`, `target-func-words-pcent`). An infrequent proportion of content and function words is expected to indicate a poorly grammatical translation.

Identifiers:

`source-func-words` `target-func-words`
`source-func-words-pcent` `target-func-words-pcent`

Example:

They got down and looked at him. `source-func-words` = 4

10-12 *Nouns.* Proportion of nouns in the source (`source-n-pcent`) and target sentence (`target-n-pcent`), including both common and proper nouns. A ratio between these proportions is also computed (source/target: `s-t-ratio-n-pcent`).

Identifiers:

`source-n-pcent` `target-n-pcent` `s-t-ratio-n-pcent`

Example:

Ford retira más de 140.000 autos en EEUU por limpiaparabrisas defectuosos. `s-t-ratio-n-pcent` = 4/12 = 0.33

13-15 *Verbs.* Proportion of verbs in the source (`source-v-pcent`) and target sentence (`target-v-pcent`), including both full and auxiliary verbs. A ratio between these proportions is also computed (source/target: `s-t-ratio-v-pcent`).

Identifiers:

`source-v-pcent` `target-v-pcent` `s-t-ratio-v-pcent`

Example:

Fans would have been likely to die of boredom.
`source-v-pcent` = 4/10 = 0.4

16 *Pronouns.* We only compute the ratio of the proportions of pronouns in the source and target sentences.

Identifiers:

`s-t-ratio-pron-pcent`

Example:

As far as they were concerned, it was just between them and their friends.

En lo que a ellos respecta, fue sólo entre ellos y sus amigos.
`s-t-ratio-pron-pcent` = (2/14)/(2/13) = 0.93

17-19 *Noun phrases (NP).* We use automatic chunking for identifying the phrase structure of a sentence and count the number of noun phrases (`source-np`, `target-np`). In our implementation, we count only the outermost NPs in a sentence, which means we do not check for embedded NPs. Our intuition, as in many other cases, is that a comparison of such phrases between the source and target text (`t-s-diff-np`) might serve as a shallow approach to quantifying differences in content.

Identifiers:

`source-np` `target-np` `t-s-diff-np`

Example:

> [$_{\text{NP}}$ *You*] *need to make sure* [$_{\text{NP}}$ *you*] *speak to* [$_{\text{NP}}$ *someone with the authority to do a deal*]. `source-np` = 3

20-22 *Verb phrases (VP).* Extraction of these phrases is similar to noun phrases. Again, VP embedding is not taken into account.

Identifiers:

`source-vp` `target-vp` `t-s-diff-vp`

Example:

> *Grecia* [$_{\text{VP}}$ *compra tiempo*]*, pero los problemas de fondo* [$_{\text{VP}}$ *siguen*]. `target-vp` = 2

23-25 *Prepositional phrases (PP).* Extraction of these phrases is similar to noun phrases. PP embedding is also disregarded.

Identifiers:

`source-pp` `target-pp` `t-s-diff-pp`

Example:

> [$_{\text{PP}}$ *En la reunión de Rajoy con sus ministros*] *también se han abordado reformas* [$_{\text{PP}}$ *en el ámbito financiero*]. `target-pp` = 2

26 *Explicit subjects.* In languages such as Spanish (our studied target language), the subject of a sentence can be stated explicitly or implicitly (Real Academia Española, 2009). In particular, we believe that studying explicit subject-predicate relations may help detect lack of fluency (for example, by the addition of extraneous information in subject position or the excessive repetition of an existent subject) and even fidelity to the original (by the missing elements of the structure, as shown in the example below).

The number of explicit subjects in a sentence is estimated by counting the number of cases where an NP is immediately followed by a VP. Although this approximation yields correct results in most cases, it is not foolproof and may produce inaccurate estimations in cases with complex grammatical structures, a fact that is further aggravated by the imperfections of automatic parsing.

Identifiers:

`target-exp-subj`*

Example:

> [$_{\text{NP}}$ *The kidnapping*] [$_{\text{VP}}$ *happened in the province of Agusan del Sur.*]
>
> [$_{\text{NP}}$ *El secuestro ocurrido en la provincia de Agusan del Sur.*]
>
> `target-exp-subj` $= 0$

27-28 *Pronominal subjects.* Within explicit subjects, we also count the number of cases where the subject NP is a pronoun (`target-pron-subj`) and estimate their proportion within the sentence (`target-pron-subj-pcent`). This feature is specially targeted at discovering superfluous, excessive or confusing pronouns that may sound unnatural in Spanish.

Identifiers:

`target-pron-subj`* `target-pron-subj-pcent`*

Example:

> * [$_{\text{PRON}}$ *Ella*] [$_{\text{VP}}$ *se cree que han matado a ella mediante asfixia utilizando una bolsa de plástico.*] `target-pron-subj` $= 1$

29 *Non-pronominal subjects.* Number of explicit subject cases where the subject NP is not a pronoun.

Identifiers:

`target-non-pron-subj`*

Example:

> [$_{\text{NP}}$ *El Gobierno*] [$_{\text{VP}}$ *le ofreció dádivas si desistía*] *pero* [$_{\text{NP}}$ *ella*] [$_{\text{VP}}$ *se negó*]. `target-non-pron-subj` $= 1$

30-31 *Zero subjects.* Since Spanish is a pro-drop language (Chomsky, 1981), explicit subjects are often omitted in fluent speech to avoid unnecessary repetition. In consequence, we use this feature as an indicator of fluency.

The number of zero subject cases (`target-zero-subj`) and their proportion within the sentence (`target-zero-subj-pcent`) are estimated by simply

counting the number of VPs within a sentence that are not immediately preceded by an NP. As in other mentioned cases, such a simplification in the interpretation of sentences may result in inaccurate estimations.

Identifiers:

`target-zero-subj`* `target-zero-subj-pcent`*

Example:

> [$_{\text{NP}}$ *Hijo de diputado*] [$_{\text{VP}}$ *atropelló*], [$_{\text{VP}}$ *mató*], [$_{\text{VP}}$ *huyó*] *y* [$_{\text{VP}}$ *está en libertad*]. `target-zero-subj` = 3

32 *Subject-verb agreement.* This feature, only implemented for the target language, gives an estimation of the number and proportion of cases where the subject of a verb phrase agrees grammatically with its main verb.

The number and proportion of agreements in a sentence (`target-s-v-agree`, `target-s-v-agree-pcent`) are computed only for explicit subject-verb cases and, in the case of Spanish, is based on the matching of three required aspects: person, number and gender (Real Academia Española, 2009). The way we conceived automatic checking of agreement is as follows.

First, we extract all the explicit subject-verb cases in the sentence. Second, we search the VP for the most specific verb in terms of person, number and gender (which is only applicable to participles). Finally, we check that the aspects of the main verb agree with those of specific types of words in the NP, namely determiners, pronouns, nouns and adjectives. Only when all these parameters match is a case counted.

Identifiers:

`target-s-v-agree`* `target-s-v-agree-pcent`*

Example:

> * [$_{\text{NP}}$ *Los noruegos*] [$_{\text{VP}}$ *nunca ha tenido dudas de que se trataba de un cohete ruso*]. `target-s-v-agree` = 0

34-36 *Deictics.* Deictic expressions (Romero, 2005) are considered one of the essential indicators of cohesion both in English (Halliday and Hasan, 1976) and Spanish (Mederos Martín, 1985; Uribe Mallarino, 2002). For this reason, we attempt to estimate whether this aspect is preserved in the translation by computing the ratio of deictic words (`t-s-diff-deixis`) found in the source (`source-deixis`) and target sentences (`target-deixis`).

Our approach is very simple and considers only one-word deictics, such as pronouns (personal deixis) and relative references to time[3] (temporal deixis) and space[4] (spatial deixis). Checking is done using language-specific lists that were manually compiled from different specialised sources.

Identifiers:

`source-deixis`* `target-deixis`* `t-s-diff-deixis`*

Example:

Now, he claims, there is no good solution to this. `source-deixis` $= 3$

Según él hoy ya no tienen una solución buena. `target-deixis` $= 3$

37-42 *Phrase structure.* We apply constituency parsing (Chomsky, 1957) to the source and target text in order to extract two shallow indicators: width (`source-ptree-width`, `target-ptree-width`) and depth (`source-ptree-depth`, `target-ptree-depth`) of their parse trees. Although languages structure content differently within a sentence, we believe that some underlying correspondences between the source and target trees (`t-s-diff-ptree-width`, `t-s-diff-ptree-depth`) could help assess the probability of the translation structure.

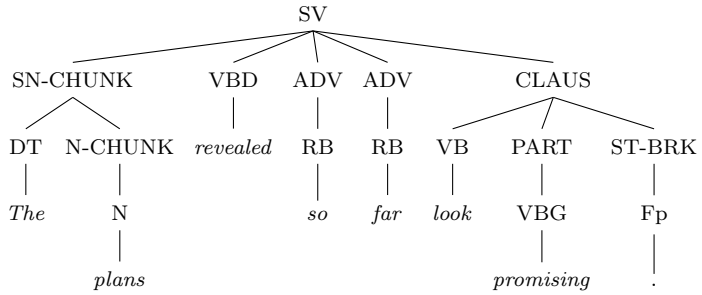
Ideally, trees should be exploited to retrieve more valuable information, as proposed by Hardmeier (2011). However, languages express things differently so a mismatch of structures between a source and target text is not expected to discriminate good from bad translations. In fact, even in the same language sentences could express equal information using quite different grammatical structures. For this reason, we have chosen to exploit aspects that somehow abstract from detailed structure, namely tree width (computed as the number of root node children) and depth (longest path from the root node to the leaves). By using these shallow parameters, we also help minimise the effect of inaccurate parsing of ungrammatical sentences.

Identifiers:

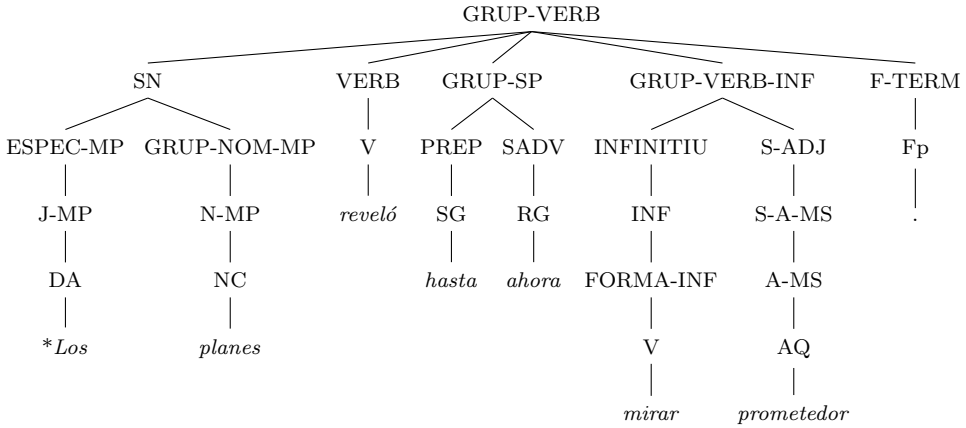
`source-ptree-depth` `target-ptree-depth` `t-s-diff-ptree-depth`
`source-ptree-width`* `target-ptree-width`* `t-s-diff-ptree-width`*

[3]E.g. *now*, *then*, *tomorrow* (English); *ahora*, *después*, *mañana* (Spanish).
[4]E.g. *here*, *there*, *these* (English); *aquí*, *allí*, *éstos* (Spanish).

Example output from FreeLing (Padró, Collado, Reese, Lloberes et al., 2010):



`source-ptree-width` = 5, `source-ptree-depth` = 4



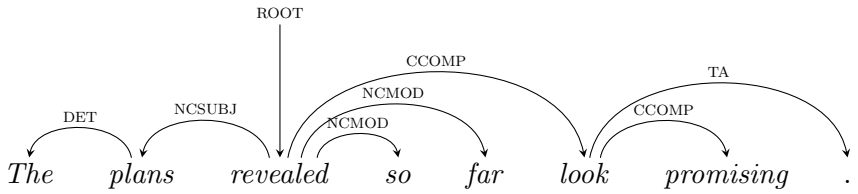
`target-ptree-width` = 5, `target-ptree-depth` = 6

43-48 *Dependency structure.* Following the same intuition as for phrase structure analysis, we also applied dependency parsing (Kubler, McDonald, Nivre and Hirst, 2009) to the source and target text in order to compute high-level width (`source-dtree-width`, `target-dtree-width`) and maximum depth (`source-dtree-depth`, `target-dtree-depth`), as well as their differences (`t-s-diff-dtree-depth`, `t-s-diff-dtree-width`).

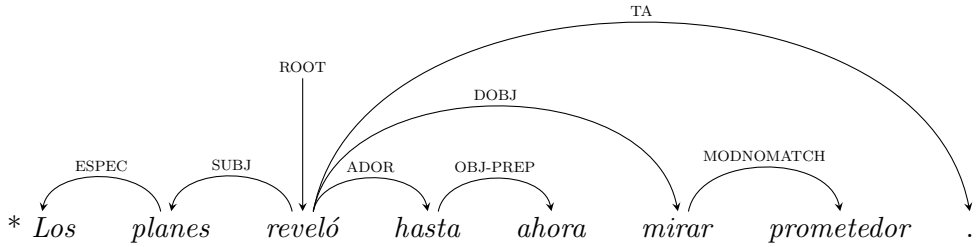
Identifiers:

`source-dtree-depth`* `target-dtree-depth`* `t-s-diff-dtree-depth`*
`source-dtree-width`* `target-dtree-width`* `t-s-diff-dtree-width`*

Example output from FreeLing:



`source-dtree-width` = 4, `source-dtree-depth` = 2



`target-dtree-width` = 4, `target-dtree-depth` = 2

49-63 *Named entities.* Named entity recognition (Jurafsky and Martin, 2009, chap. 22) is applied to the source and target sentences as a way of identifying important concepts that must be preserved in the translation. By means of an automatic entity recognition module, we are not only able to identify proper nouns but also classify them into standard categories (person, location, organisation and 'other').

Unless recognisers are very robust, identification relies mainly on capitalisation, which is why they might not perform well on lowercase text.

The number and differences of entities within a sentence pair are computed for each category (`source-ne-*`, `target-ne-*`, `t-s-diff-ne-*`) as well as in total (`source-ne`, `target-ne`, `t-s-diff-ne`).

Identifiers:

`source-ne`	`target-ne`	`t-s-diff-ne`
`source-ne-per`	`target-ne-per`	`t-s-diff-ne-per`
`source-ne-loc`	`target-ne-loc`	`t-s-diff-ne-loc`
`source-ne-org`	`target-ne-org`	`t-s-diff-ne-org`
`source-ne-other`*	`target-ne-other`*	`t-s-diff-ne-other`*

Example:

> [$_{\text{PER}}$ *Nabeel Shaath*] *said after* [$_{\text{PER}}$ *Mr Obama*]*'s speech to the* [$_{\text{OTHER}}$ *General Assembly*] *in* [$_{\text{LOC}}$ *New York*] *that going to the* [$_{\text{ORG}}$ *U.N.*] *is the 'only alternative to violence'.* `source-ne` = 5

64-65 *Language model of PoS tags: log-probability.* A common method for estimating whether a particular sequence of words conforms to a given language is the use of **n-gram language models**. N-grams are sequences of one or more tokens (unigrams, bigrams, trigrams, etc.) extracted from text that are used as a unit of analysis. An n-gram language model is basically a table containing n-grams extracted from corpora and their associated frequencies, so that they can be used to estimate probabilities for new sequences (Jurafsky and Martin,

2009, chap. 4). These language models usually combine more than one type of n-gram in order to achieve more robust results.

Formally, the probability of a sequence of tokens in terms of n-grams is defined as:

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k | w_{k-N+1}^{k-1})$$

where w_1^n is a sequence of n words and $P(w_k | w_{k-N+1}^{k-1})$ is the conditional probability of word w_k given the sequence of previous words in a model of N grams.

An interesting fact about this kind of modelling is that it can be applied to estimate any type of sequence, not necessarily words. In particular, we built a trigram language model to predict part-of-speech tags using a tagged version of a monolingual corpus for the target language. The rationale behind this decision is that a PoS sequence should serve as an abstraction from the lexicon of a language so it should provide better generalisations about likely sequences (i.e. permitted word combinations).

Since the probabilities of observed n-grams are usually very low and inconvenient for further calculations, a common practice is to represent them as **log-probabilities** (i.e. $P(w_1^n)$ is redefined as $\log P(a)$). The higher these probabilities are, the more likely the sequence is in a language, giving us an idea of its grammaticality and fluency.

For this feature, we implemented two variations: one using simple PoS tags such as noun, verb, adjective, etc. (`target-pos-logprob-bl`) and another using detailed morphosyntactic tags (`target-pos-logprob`) that encode information about many aspects of the words, such as number, gender, mood, tense, etc. In both cases we considered end-of-sentence markers.

Identifiers:

`target-pos-logprob`* `target-pos-logprob-bl`

Example:

* *Al final , un oficial y un protester fueron hospitalizados .*
PAL NC CM ART ADJ CC ART NC VSfin VLadj FS

`target-pos-logprob-bl` $= -129.30$

* *A el final , un oficial y un protester fueron*
SPS00 DA0MS0 NCFS000 Fc DI0MS0 AQ0CS0 CC DI0MS0 NCMS000 VSIS3P0
hospitalizados .
VMP00PM Fp

`target-pos-logprob` $= -179.09$

66-67 *Language model of PoS tags: perplexity.* Unlike log-probability, the **perplexity** of a sequence of words measures how likely it is that they branch out and combine with other words as well, so it can be viewed as an inverse version of probabilities. In fact, minimising perplexity is equivalent to maximising probability (Jurafsky and Martin, 2009, chap. 4). Formally:

$$PP(W) = \sqrt[n]{\frac{1}{P(w_1 w_2 ... w_n)}}$$

where W is a sequence of n words and $P(w_1 w_2 ... w_n)$ is the probability of all the words in the sequence estimated using an n-gram language model.

Again, two versions of this feature are computed using simple (`target-pos-ppl-bl`) and morphosyntactic PoS tags (`target-pos-ppl`), both including end-of-sentence markers.

Identifiers:

`target-pos-ppl`* `target-pos-ppl-bl`

Example:

* *Al final , un oficial y un protester fueron hospitalizados .*
PAL NC CM ART ADJ CC ART NC VSfin VLadj FS

`target-pos-ppl-bl` = 119.53, `target-pos-ppl` = 238.56

68 *Missing contractions.* This language-specific feature calculates the number of times that the target text failed to produce the Spanish contractions *del* (*de+el*) and *al* (*a+el*) in cases where they were required. The purpose of this feature is to check for appropriate use of language.

Identifiers:

`target-miss-contr`*

Example:

* *La policía pudo agarrar a el sospechoso gracias a un testigo.*

`target-miss-contr` = 1

69 *Dangling determiners.* We count the number of cases in the target text where determiners are not linked to any noun, as this is a very frequent grammatical mistake in machine translations. Our implementation makes use of a PoS tagger and considers a maximum context window of five words between determiners and nouns.

Identifiers:

`target-dang-det`*

Example:

> * *Brighton es una vibrante y fantásticamente creativa y estamos tratando de intentar representar que a través de la espiritualidad.*
>
> `target-dang-det` = 1

70 *Unknown words.* This feature estimates the number of words that appear in the target text and are not considered to be part of the target language lexicon. Failure to render words into the target language indicates poor translation ability and introduces noise, with a significant impact on quality. This is often the case for SMT systems that have been trained on one domain and tested on another: words that have not been seen in the training corpus remain untranslated in subsequent translations. In some cases, however, replication of these out-of-vocabulary words is a wise decision, as in the case of names (Habash, 2008), but our approach does not distinguish between such cases.

The number of unknown words is estimated by using an auxiliary spell checker for Spanish from which we make assumptions about valid words in the language.

Identifiers:

`target-unknown`*

Example:

> *La virtualisation de smartphones no es ciencia ficción.*
>
> `target-unknown` = 2

3.2.2.2 Shallow Features

71-75 *Tokens.* We count the number of all tokens in the source (`source-num-tokens`) and target text (`target-num-tokens`) and compute their proportions in both directions (`s-t-token-ratio`, `t-s-token-ratio`). Tokens are defined as strings of consecutive printable characters separated by blanks or punctuation. By this definition, a token does not necessarily constitute a word and could therefore be a number, symbol or punctuation mark.

This basic feature not only gives an estimation of difficulty, under the assumption that longer sentences are more difficult to produce and understand than shorter ones (Dubay, 2004), but we can also use it to compute other metrics that we describe below.

In order to explore how the length of the target text compares to the source, we also compute the difference in tokens between them and normalise it by the number of tokens in the source (`s-t-diff-tok-norm`).

Identifiers:

`source-num-tokens` `s-t-token-ratio` `s-t-diff-tok-norm`
`target-num-tokens` `t-s-token-ratio`

Example:

Informe: Más de 150.000 languidecen en prisiones en Norcorea
`target-num-tokens` = 10

76-77 *Types.* These refer to the number of unique tokens (Wetzel, 2009) in the source (`source-num-types`) and target text (`target-num-types`), which may serve as an indication of vocabulary size.

Identifiers:

`source-num-types` `target-num-types`

Example:

El año fiscal estadounidense arranca el 1 de octubre y finaliza el 30 de septiembre del año siguiente. `target-num-types` = 15

78-79 *Type/Token Ratio (TTR).* This ratio between the number of types and tokens in a text is often used as an indicator of lexical diversity (Johnson, 1944). While low TTR values reflect a poor vocabulary, high values indicate richer and less repetitive sentences. In turn, a high TTR may also suggest greater conceptual complexity. Formally, the relation is defined as:

$$\text{TTR} = \frac{\text{number of types}}{\text{number of tokens}}$$

Computing this ratio and some of its variations on the target text may help determine lexical quality and repetition, such as the overuse of function words by many MT systems. Although TTR can also be computed on the source text as a proxy for complexity, we computed it only on the target text. Two different versions of the metric were used, including and excluding punctuation respectively (`target-ttr`, `target-ttr-bl`).

Identifiers:

`target-ttr`* `target-ttr-bl`

Example:

* *Trabajadores había sido obligados a hacer dos horas de las horas extraordinarias por encima de su normal turnos de ocho horas.*

`target-ttr` = 18/22 = 0.82

80 *Corrected TTR.* A variation of TTR (Carroll, 1964) defined as:

$$\text{CTTR} = \frac{k}{\sqrt{2n}}$$

where k is the number of types and n is the number of tokens.

Identifiers:

`target-corr-ttr`*

Example:

* *Trabajadores había sido obligados a hacer dos horas de las horas extraordinarias por encima de su normal turnos de ocho horas.*

`target-corr-ttr` = 2.71

81 *Log TTR.* Another variation of TTR (Herdan, 1960), defined as:

$$\text{Log TTR} = \frac{\log k}{\log n}$$

where k is the number of types and n is the number of tokens.

Identifiers:

`target-log-ttr`*

Example:

* *Trabajadores había sido obligados a hacer dos horas de las horas extraordinarias por encima de su normal turnos de ocho horas.*

`target-log-ttr` = 0.94

82 *Guiraud's Index.* This is another variation of TTR, also known as Root TTR (Guiraud, 1954) and defined by:

$$\text{GI} = \frac{k}{\sqrt{n}}$$

where k is the number of types and n is the number of tokens.

Identifiers:

`target-gi`*

Example:

> * *Trabajadores había sido obligados a hacer dos horas de las horas extraordinarias por encima de su normal turnos de ocho horas.*
>
> `target-gi` = 3.84

83 *Uber Index.* Another variant of TTR (Dugast, 1980), defined as:

$$\text{UI} = \frac{\log^2 n}{\log n - \log k}$$

where k is the number of types and n is the number of tokens.

Identifiers:

`target-ui`*

Example:

> * *Trabajadores había sido obligados a hacer dos horas de las horas extraordinarias por encima de su normal turnos de ocho horas.*
>
> `target-ui` = 20.68

84 *Jarvis TTR.* This metric is based on the Uber Index and defined as:

$$\text{Jarvis TTR} = n^{\frac{-1}{\text{UI}} \log n}$$

where n is the number of tokens and UI is the value of the Uber Index for the sentence (Jarvis, 2002).

Identifiers:

`target-jarvis-ttr`*

Example:

> * *Trabajadores había sido obligados a hacer dos horas de las horas extraordinarias por encima de su normal turnos de ocho horas.*
>
> `target-jarvis-ttr` = 0.82

85-86 *Language model of words: log-probability.* Following the same concept that we used for computing PoS log-probabilities, we built a trigram language model from monolingual corpora to estimate token log-probabilities. This feature is applied on the source text in an attempt to measure its difficulty (`source-logprob`) and on the target text to estimate grammaticality (`target-logprob`).

The reason why this feature is considered shallow is that no linguistic information is involved in its computation, since tokens are treated as mere strings of characters.

Identifiers:

`source-logprob` `target-logprob`

Example:

* *El rastro Giffords especial de la recuperación desde la Jan. 8 disparos.*
`target-logprob` = −39.329

87-90 *Language model of words: perplexity.* We also use a trigram language model to estimate perplexities for the source and target texts, with (`source-ppl`, `target-ppl`) and without end-of-sentence markers (`source-ppl-no-smark`, `target-ppl-no-smark`).

Identifiers:

`source-ppl` `target-ppl`
`source-ppl-no-smark` `target-ppl-no-smark`

Example:

* *El rastro Giffords especial de la recuperación desde la Jan. 8 disparos.*
`target-ppl` = 227.045, `target-ppl-no-smark` = 334.507

91-95 *Source unigrams.* As in previous features, we use n-gram statistics to estimate the probability of the source text and get an idea of its difficulty. In this specific case, we compute the proportion of unique unigrams in the source text that are part of our language model (`source-uni`). Using this value, we can quantify how much the sentence deviates from the general lexicon of its language.

The same proportions are also computed using four separate subsets of the language model, which go from the 25% least frequent unigrams (q1) up to the 25% most frequent (q4).

Identifiers:

`source-uni` `source-uni-q3`
`source-uni-q1` `source-uni-q4`
`source-uni-q2`

Example:

A siviglian dancer pays tribute to another siviglian dancer.
`source-uni` = 0.86

96 *Source average unigram frequency.* This is an average of the frequency of each source type in a large auxiliary monolingual corpus, aimed at estimating natural usage.

Identifiers:

`source-avg-uni`

Example:

Negotiations concerning a ransom are reportedly already underway.
`source-avg-uni` $= 821,107.1$

97-101 *Source bigrams.* This is the proportion of unique bigrams from the source sentence that belong to our language model (`source-bi`). The intuition is that whenever the number of recognised bigrams increases, so does the probability that the sentence is well formed.

This feature also includes estimations using four frequency subsets of our language model (q1 to q4).

Identifiers:

`source-bi`
`source-bi-q1`
`source-bi-q2`
`source-bi-q3`
`source-bi-q4`

Example:

The bond market has climbed sharply. `source-bi` $= 0.83$

102-106 *Source trigrams.* This is the proportion of unique source trigrams that are also part of our language model (`source-tri`), which is expected to give more accurate results than unigrams and bigrams. Again, additional estimations were computed using four splits of our language model according to ascending frequency (q1 to q4).

Identifiers:

`source-tri`
`source-tri-q1`
`source-tri-q2`
`source-tri-q3`
`source-tri-q4`

Example:

He is now under arrest and faces up to three years in prison.
`source-tri` $= 0.75$

107 *Average token length.* This is only computed on the source text as another measure of difficulty, under the assumption that longer words are generally more complex and difficult to understand than shorter ones (Dubay, 2004).

Since tokens also include non-alphanumeric strings (such as numbers, punctuation or symbols), they are also taken into account for the average.

Identifiers:

`source-avg-tok-len`

Example:

> *Parasympathetic postganglionic neurons release acetylcholine while sympathetic postganglionic neurons release norepinephrine.*
>
> `source-avg-tok-len` = 9.5

108-110 *Non-alphabetic tokens.* Tokens that are not wholly composed of letters (such as numerical expressions) are counted in the source (`source-tok-not-az`) and target text (`target-tok-not-az`) to estimate how they are preserved. A ratio between these values is also computed (`s-t-ratio-tok-not-az`).

Identifiers:

`source-tok-not-az` `target-tok-not-az` `s-t-ratio-tok-not-az`

Example:

> *El G8 acuerda dar 38.000 millones dólares para las reformas en cuatro países árabes.*
>
> `target-tok-not-az` = 2

111-113 *Numbers.* The proportions of numbers (integers and decimals) in the source (`source-num-pcent`) and target text (`target-num-pcent`) are also computed in order to check how they are preserved. A difference between these values which is normalised by the target proportion is also calculated (`s-t-diff-num-norm`).

Identifiers:

`source-num-pcent` `target-num-pcent` `s-t-diff-num-norm`

Example:

> *En 2012 se venderán 118,9 millones de tablets.*
>
> `target-num-pcent` = 2/9 = 0.22

114-118 *Punctuation marks.* We compute the number and proportion of punctuation marks in the source (`source-punc`, `source-punc-pcent`) and target text (`target-punc`, `target-punc-pcent`) plus a difference normalised by the number of tokens in the target (`s-t-diff-punc-norm`). This and the following indicators are expected to capture the relationship between the use of punctuation in the source and target language.

Identifiers:

`source-punc` `target-punc` `s-t-diff-punc-norm`
`source-punc-pcent` `target-punc-pcent`

Example:

Tras bin Laden, ¿Quién es el más buscado del FBI?

`target-punc` = 3

119-120 *Colons.* We compute the absolute difference in the number of colons between the source and target segments (`s-t-diff-colons`) in addition to a variation which is normalised by the number of tokens in the target text (`s-t-diff-colons-norm`).

Identifiers:

`s-t-diff-colons` `s-t-diff-colons-norm`

Example:

UEFA President Platini: Reprieve for Poland and Ukraine

El presidente de la UEFA, Platini: Respiro para Polonia y Ucrania

`s-t-diff-colons` = 1 − 1 = 0

121-122 *Semicolons.* Absolute difference of semicolons between the source and target segments (`s-t-diff-semic`) plus a variation normalised by the number of tokens in the target text (`s-t-diff-semic-norm`).

Identifiers:

`s-t-diff-semic` `s-t-diff-semic-norm`

Example:

The Traffic Law only knows the term vehicle conductor; it does not define the rights and responsibilities of that person.

* *La Ley de Tráfico sólo conoce el plazo vehículo director; no define los derechos y responsabilidades de esa persona.*

`s-t-diff-semic` = 1 − 1 = 0

123-124 *Commas.* Absolute and normalised differences of commas between the source and target text.

Identifiers:

`s-t-diff-commas` `s-t-diff-commas-norm`

Example:

This year, interest in winter breaks is low due to the fact that there is still no snow on the Krkonoše.

* *Este año, el interés en invierno rompe es bajo, debido al hecho de que todavía no hay sobre la nieve Krkonoše.*

`s-t-diff-commas` $= 1 - 2 = -1$

125-126 *Full stops.* Absolute and normalised differences of full stops between the source and target text.

Identifiers:

`s-t-diff-periods` `s-t-diff-periods-norm`

Example:

Obama has always been reticent in regards to his prize.

Obama ha sido siempre reticente en cuanto a su premio.

`s-t-diff-periods` $= 1 - 1 = 0$

127-128 *Exclamation marks.* Absolute and normalised differences of exclamation marks between the source and target text. It must be borne in mind that, unlike English, Spanish requires two different exclamation marks: one at the beginning (*¡*) and another at the end of the sentence (*!*).

Identifiers:

`s-t-diff-excm` `s-t-diff-excm-norm`

Example:

How can any deny themselves the pleasure of my company!

* *¿Cómo puede negar cualquier por sí mismos el placer de mi empresa!*

`s-t-diff-excm` $= 1 - 1 = 0$

129-130 *Question marks.* Absolute and normalised differences of question marks between the source and target text. It must be borne in mind that, unlike English, Spanish requires two different question marks: one at the beginning (*¿*) and another at the end of the sentence (*?*).

Identifiers:

`s-t-diff-questm` `s-t-diff-questm-norm`

Example:

How do you tell if someone is lying?

¿Cómo saber si alguien está mintiendo?

`s-t-diff-questm` $= 1 - 2 = -1$

131 *Mismatched quotation marks.* This feature checks for mismatches of double or single quotation marks in the target text.

Identifiers:

`target-mis-qmarks`

Example:

* *Los periódicos rusos están describiendo el "Bulava" como "el cohete flightless".* `target-mis-qmarks` $= 0$

132 *Mismatched brackets.* This feature checks for mismatches of round, square and curly brackets in the target text.

Identifiers:

`target-mis-brackets`

Example:

Según AFP, los pistoleros forman parte del Nuevo Ejército Popular (NPA), que es la facción armada del Partido Comunista de Filipinas (CPP). `target-mis-brackets` $= 0$

133-147 *Average translations per token.* This feature estimates the average number of translations of each token in the source sentence, providing a measure of polysemic content in the source text. In order to compute this feature, correspondences between a source word and its possible translations are derived from the alignment of an auxiliary parallel corpus. The result of this alignment is a table where each source token is paired with tokens in the target language with a given probability (derived from the co-occurrences in the copora). Estimations of the number of translations can then be extracted directly from this table. In fact, we define two criteria to obtain different indicators.

The first criterion is a probability threshold, by which we only count a target token towards the number of possible translations of a source token if it has a probability greater than 0.01, 0.05, 0.1, 0.2 or 0.5 respectively.

The second criterion is a normalisation factor, that is used to mitigate the effect of token frequency on the estimations. We used three variations for this purpose: 1) no normalisation (`source-avg-trans-*`), 2) normalisation by source token frequency (`source-avg-trans-*-freq`) and 3) normalisation by inverse source token frequency (`source-avg-trans-005-*`).

By combining these two criteria, we obtained 15 different estimations that are then divided by the number of tokens in the source text to generate the averages.

The reason why these estimates are considered shallow is that they are computed using purely statistical methods on the surface tokens, without the help of linguistic information. A downside to this method is the artificial increase in translation probabilities caused by the unrestricted alignment of all source tokens with all target tokens in each training pair of the corpus (Saers and Wu, 2009).

Identifiers:

`source-avg-trans-001`
`source-avg-trans-001-freq`
`source-avg-trans-001-inv`
`source-avg-trans-005`
`source-avg-trans-005-freq`
`source-avg-trans-005-inv`
`source-avg-trans-01`
`source-avg-trans-01-freq`
`source-avg-trans-01-inv`
`source-avg-trans-02`
`source-avg-trans-02-freq`
`source-avg-trans-02-inv`
`source-avg-trans-05`
`source-avg-trans-05-freq`
`source-avg-trans-05-inv`

Example:

City officials have categorically denied that accusation.

`source-avg-trans-005` = 6.44

Many of the features described in this section have been proposed in previous QE approaches and were included in our models to evaluate how they interact with our new indicators. On the other hand, our work introduces 34 novel features (6 shallow and 28 linguistic), which have been marked with * throughout the section. Although a few of them are variations of existing indicators (e.g. `target-ttr`), others are original features for which specific extractors had to be implemented (e.g. `target-s-v-agree`, `target-unknown`).

Table 3.1 shows a classification of all our 147 features according to the criteria explained in section 3.2.1.

Numbers	Description	Ling. know.		Origin			Lang. dep.		Res. dep.		Aspect					
		L	S	S	T	GREY-B	LD	LI	RD	RI	COMP	FID	GRAM	FLU	CC	STR
1-5	Content words	X		X	X			X	X		X	X				
6-9	Function words	X		X	X			X	X					X	X	
10-12	Nouns	X		X	X			X	X		X	X				
13-15	Verbs	X		X	X			X	X		X	X				
16	Pronouns	X		X	X			X	X					X	X	
17-19	Noun phrases	X		X	X			X	X		X	X	X			
20-22	Verb phrases	X		X	X			X	X		X	X	X			
23-25	Prepositional phrases	X		X	X			X	X		X	X	X			
26	Explicit subjects	X			X		X		X					X		
27-28	Pronominal subjects	X			X		X		X					X		
29	Non-pronominal subjects	X			X		X		X					X		
30-31	Zero subjects	X			X		X		X					X		
32	Subject-verb agreement	X			X		X		X				X	X	X	
34-36	Deictics	X		X	X		X		X			X		X	X	
37-42	Phrase structure	X		X	X			X	X		X	X	X	X		
43-48	Dependency structure	X		X	X			X	X		X	X	X	X		
49-63	Named entities	X		X	X			X	X			X				X
64-65	Language model of PoS: log-probability	X			X			X	X		X		X	X		
66-67	Language model of PoS: perplexity	X			X			X	X		X		X	X		
68	Missing contractions	X			X		X		X				X	X	X	
69	Dangling determiners	X			X			X	X				X	X	X	
70	Unknown words	X			X			X	X			X	X	X	X	
71-75	Tokens		X	X	X			X		X	X	X				
76-77	Types		X	X	X			X		X	X	X				
78-79	Type/Token Ratio (TTR)		X		X			X		X	X			X		
80	Corrected TTR		X		X			X		X	X			X		
81	Log TTR		X		X			X		X	X			X		
82	Guiraud's Index		X		X			X		X	X			X		
83	Uber Index		X		X			X		X	X			X		
84	Jarvis TTR		X		X			X		X	X			X		
85-86	Language model of words: log-probability		X	X	X	X		X	X		X		X	X		
87-90	Language model of words: perplexity		X	X	X	X		X	X		X		X	X		
91-95	Source unigrams		X	X		X		X	X		X		X	X		
96	Source average unigram frequency		X	X		X		X	X		X		X	X		
97-101	Source bigrams		X	X		X		X	X		X		X	X		
102-106	Source trigrams		X	X		X		X	X		X		X	X		
107	Average token length		X	X				X		X	X					
108-110	Non-alphabetic tokens		X	X	X			X		X		X				X
111-113	Numbers		X	X	X			X		X		X				X
114-118	Punctuation marks		X	X	X			X		X		X			X	X
119-120	Colons		X	X	X			X		X		X			X	X
121-122	Semicolons		X	X	X			X		X		X			X	X
123-124	Commas		X	X	X			X		X		X			X	X
125-126	Full stops		X	X	X			X		X		X			X	X
127-128	Exclamation marks		X	X	X			X		X		X				X
129-130	Question marks		X	X	X			X		X		X				X
131	Mismatched quotation marks		X	X	X			X		X			X			X
132	Mismatched brackets		X	X	X			X		X			X			X
133-147	Average translations per token		X	X		X		X	X		X					

Table 3.1: Feature classification. Only black-box and grey-box features are used in this work.

Chapter 4

Evaluation

There is no difference between theory and practice... in theory.
But in practice, there is.
—Anonymous

The first part of this chapter focuses on our experimental setup (section 4.1), describing our dataset, tools, evaluation metrics and training algorithm. The second part (section 4.2) provides the results of our experiments (sections 4.2.1 and 4.2.2) and a detailed analysis of performance (sections 4.2.3 and 4.2.4).

4.1 Experimental Setup

The following sections describe the experimental setup we used to test our quality estimation proposal. We describe the datasets (section 4.1.1), tools for feature extraction (section 4.1.2), the baseline system used for comparison (section 4.1.3), evaluation metrics (section 4.1.4) and feature sets (section 4.1.5). Finally, we give a brief description of Support Vector Machines and the training phase (section 4.1.6).

4.1.1 Datasets

For our experiments, we used the official training and test sets provided for the WMT 2012 Quality Estimation Shared Task (Callison-Burch, Koehn, Monz, Post et al., 2012). The training data comprised 1,832 English sentences extracted from news texts and their translations into Spanish, produced by the SMT system Moses (Koehn, Hoang, Birch, Callison-Burch et al., 2007) trained on a reduced version of the English-Spanish Europarl parallel corpus (Koehn, 2005) provided for the WMT 2010 (Callison-Burch, Koehn, Monz, Peterson et al., 2010). Each of these pairs also included an average quality score computed from three human judgements of post-editing effort using the following scale:

1: The MT output is incomprehensible, with little or no information transferred accurately. It cannot be edited, needs to be translated from scratch.

2: About 50-70% of the MT output needs to be edited. It requires a significant editing effort in order to reach publishable level.

3: About 25-50% of the MT output needs to be edited. It contains different errors and mistranslations that need to be corrected.

4: About 10-25% of the MT output needs to be edited. It is generally clear and intelligible.

5: The MT output is perfectly clear and intelligible. It is not necessarily a perfect translation, but requires little to no editing.

The test set, on the other hand, consisted of another 422 instances produced in the same fashion.

4.1.2 Resources

Features described in section 3.2 were extracted using a variety of resources. We used TreeTagger (Schmid, 1995) for PoS tagging in English and FreeLing (Padró, Collado, Reese, Lloberes et al., 2010) for Spanish, relying also on their tokenisation and chunking for other features. Freeling was also used to obtain constituency and dependency trees and perform named entity recognition in both languages.

For our language models, we used the SRILM toolkit (Stolcke, 2002) and two different corpora: AnCora (Taulé, Martí and Recasens, 2008) for Spanish and the English section of the WMT 2010 Europarl-based corpus for English. The PoS language models were built using PoS-tagged versions of the same corpora.

Unknown words in the target language were estimated using the JMySpell[1] spell checker and the publicly available general Spanish (es_ES) dictionary from the OpenOffice suite[2]. In order to avoid wrong estimations, all named entities were filtered out before the spell-checking phase, since they are not expected to be part of the language.

The lists of deictic expressions we used were restricted to one-token words only and compiled manually from different sources.

4.1.3 Baseline System

We used the official baseline system provided for the WMT 2012 QE Shared Task to evaluate the performance of our models. The system was trained using the same

[1]Available at `http://kenai.com/projects/jmyspell`
[2]Available at `http://www.openoffice.org/`

learning algorithm as our models but uses only 17 shallow features that were found to be the best in previous work (listed in Table 4.1).

Besides using standard error metrics (section 4.1.4), direct comparison between the baseline and our models is also possible because they use the same machine learning setup.

4.1.4 Evaluation Metrics

Performance of regression models is often measured in terms of prediction error. For this reason, we adopted the following three standard error metrics to evaluate our models (Witten, Frank and Hall, 2011, chap. 5):

Mean Absolute Error (MAE): This is a measure of absolute error expressed in the same units as the predictions, which is why it can be easily interpreted. Formally, it is defined as:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i|$$

where n is the total number of evaluated instances, $\widehat{y}_i$ is the value predicted by the estimator and y_i is the expected true value.

Mean Squared Error (MSE): This is a common error metric that resembles variance and is therefore expressed in quadratic units. Its definition is given by:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

Although it is the most widely used evaluation metric, it is known to magnify the effect of outliers.

Root Mean Squared Error (RMSE): This metric is the square root of the MSE and can be seen as analogous to standard deviation:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}$$

Since it is expressed in the same units as the predicted values, its interpretation is also straightforward. Although RMSE values are similar in magnitude to MAE, they are slightly greater.

In addition to these metrics, we also studied the correlation between expected values and their predictions. To this end, we computed Pearson's correlation

coefficient, which is defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where x_i represents each expected value, y_i its prediction and $\bar{x}$ and $\bar{y}$ represent the means of their corresponding distributions.

Values of r obtained from this formula range from -1 to 1. A value of 1 indicates a perfect linear relationship between the variables which can be decreasing or increasing depending on whether r is -1 or 1 respectively. On the other hand, a value of 0 denotes no relationship between the variables whereas any other intermediate value is subject to interpretation. For the specific case of translation quality prediction, the closer the Pearson correlation coefficient is to 1 or -1, the more reliable the estimator.

4.1.5 Feature Sets

In order to evaluate the impact of our proposed linguistic features, we built different models using the following initial feature sets:

Baseline set: Includes 17 shallow features set as the official baseline in the WMT 2012 QE Shared Task. The model built from these features is known to be very strong and hard to beat, since it includes some of the best features from previous work in an attempt to push the state of the art within the shared task.

Linguistic set: Comprises all the 70 linguistic features described in section 3.2.2.1.

Shallow set: Comprises all the 77 shallow features described in section 3.2.2.2, which include the ones in the baseline set.

Baseline+linguistic set: Includes all the features from the baseline set plus the ones in the linguistic set, amounting to 87 mixed features.

Full set: Comprises the sum of all feature sets, yielding a total of 147 mixed features.

The list of features included in each set is shown in Table 4.1.

4.1.6 Training

We built our models using Support Vector Machines regression, since it has been proved to be successful for QE in previous work (see section 2.3.2).

SVMs are efficient supervised machine learning algorithms that can be used for both classification and regression problems. When applied to binary classification, SVMs find the optimal the optimal hyperplane that separates the classes and

Name	**Features**	**Total**
Baseline	source-avg-tok-len, source-avg-trans-001-inv, source-avg-trans-02, source-bi-q1, source-bi-q4, source-logprob, source-num-tokens, source-punc, source-tri-q1, source-tri-q4, source-uni, source-uni-q1, source-uni-q4, target-logprob, target-num-tokens, target-punc, target-ttr-bl	17
Linguistic	s-t-ratio-cont-words-pcent, s-t-ratio-n-pcent, s-t-ratio-pron-pcent, s-t-ratio-v-pcent, source-cont-words, source-cont-words-pcent, source-deixis, source-dtree-depth, source-dtree-width, source-func-words, source-func-words-pcent, source-n-pcent, source-ne, source-ne-loc, source-ne-org, source-ne-other, source-ne-per, source-np, source-pp, source-ptree-depth, source-ptree-width, source-v-pcent, source-vp, t-s-diff-deixis, t-s-diff-dtree-depth, t-s-diff-dtree-width, t-s-diff-ne, t-s-diff-ne-loc, t-s-diff-ne-org, t-s-diff-ne-other, t-s-diff-ne-per, t-s-diff-np, t-s-diff-pp, t-s-diff-ptree-depth, t-s-diff-ptree-width, t-s-diff-vp, target-cont-words, target-cont-words-pcent, target-dang-det, target-deixis, target-dtree-depth, target-dtree-width, target-exp-subj, target-func-words, target-func-words-pcent, target-miss-contr, target-n-pcent, target-ne, target-ne-loc, target-ne-org, target-ne-other, target-ne-per, target-non-pron-subj, target-np, target-pos-logprob, target-pos-logprob-bl, target-pos-ppl, target-pos-ppl-bl, target-pp, target-pron-subj, target-pron-subj-pcent, target-ptree-depth, target-ptree-width, target-s-v-agree, target-s-v-agree-pcent, target-unknown, target-v-pcent, target-vp, target-zero-subj, target-zero-subj-pcent	70
Shallow	s-t-diff-colons, s-t-diff-colons-norm, s-t-diff-commas, s-t-diff-commas-norm, s-t-diff-excm, s-t-diff-excm-norm, s-t-diff-num-norm, s-t-diff-periods, s-t-diff-periods-norm, s-t-diff-punc-norm, s-t-diff-questm, s-t-diff-questm-norm, s-t-diff-semic, s-t-diff-semic-norm, s-t-diff-tok-norm, s-t-ratio-tok-not-az, s-t-token-ratio, source-avg-tok-len, source-avg-trans-001, source-avg-trans-001-freq, source-avg-trans-001-inv, source-avg-trans-005, source-avg-trans-005-freq, source-avg-trans-005-inv, source-avg-trans-01, source-avg-trans-01-freq, source-avg-trans-01-inv, source-avg-trans-02, source-avg-trans-02-freq, source-avg-trans-02-inv, source-avg-trans-05, source-avg-trans-05-freq, source-avg-trans-05-inv, source-avg-uni, source-bi, source-bi-q1, source-bi-q2, source-bi-q3, source-bi-q4, source-logprob, source-num-pcent, source-num-tokens, source-num-types, source-ppl, source-ppl-no-smark, source-punc, source-punc-pcent, source-tok-not-az, source-tri, source-tri-q1, source-tri-q2, source-tri-q3, source-tri-q4, source-uni, source-uni-q1, source-uni-q2, source-uni-q3, source-uni-q4, t-s-token-ratio, target-corr-ttr, target-gi, target-jarvis-ttr, target-log-ttr, target-logprob, target-mis-brackets, target-mis-qmarks, target-num-pcent, target-num-tokens, target-num-types, target-ppl, target-ppl-no-smark, target-punc, target-punc-pcent, target-tok-not-az, target-ttr, target-ttr-bl, target-ui	77
Baseline+ linguistic	*All features from the baseline set plus the linguistic set.*	87
Full set	*All features from the linguistic set plus the shallow set.*	147

Table 4.1: Initial feature sets.

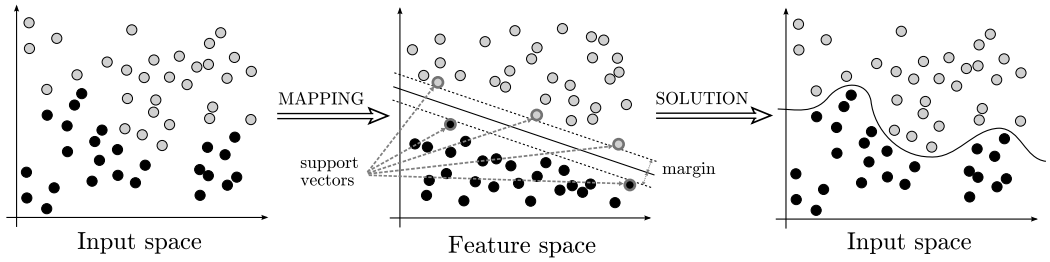


Figure 4.1: Two-dimensional classification using Support Vector Machines.

maximises the *margin* or distance between this hyperplane and its closest instances (the *support vectors*). If a problem is not linearly separable, a *kernel function* is used to map the data into a higher dimensional space where it can be linearly separated, providing a solution to the original problem. Figure 4.1 illustrates this strategy.

SVMs also implement a *cost* parameter (C) which controls the trade-off between allowing training errors and setting rigid margins. Whenever we increase this value, we are increasing the cost of misclassifications, forcing the creation of more accurate models that may result in overfitting. For this reason, it is important to find an optimal value for this parameter.

Support Vector Regression (SVR) problems are modelled by minimising an error function. Depending on the type of penalty that is imposed on misclassifications, we can distinguish two versions: epsilon-SVR and nu-SVR. Epsilon-SVR, the one adopted for our experiments, uses a parameter ϵ to define an acceptable distance (margin) between the inferred function and training instances, ignoring all cases beyond these limits (Figure 4.2). Besides C and ϵ, there is a third parameter, γ, that is used in radial basis functions (a type of kernel function) to control the non-linearity of the produced model.

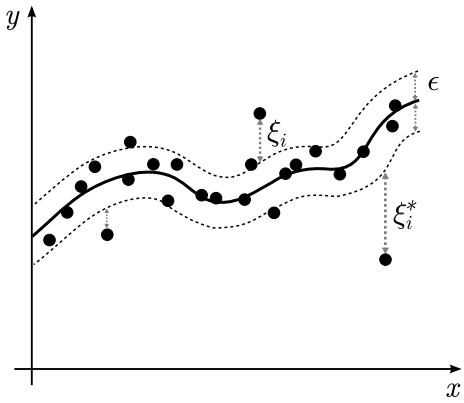


Figure 4.2: SVM epsilon regression. Variable ϵ represents margin size while ξ_i and ξ_i^* are slack variables specifying the upper and lower training errors respectively.

Feature set	MAE ↓	MSE	RMSE
Baseline+linguistic	**0.674**	0.681	0.825
Full set	0.681	0.710	0.842
Baseline	0.687	**0.672**	**0.820**
Shallow	0.691	0.713	0.844
Linguistic	0.716	0.791	0.889

Table 4.2: Error performance.

For our experiments, we used the LibSVM toolkit (Chang and Lin, 2011) to build epsilon regression models for each of our feature sets. Training was performed using a radial basis function kernel and optimal values for ϵ, C and γ estimated from prior five-fold cross validation on the training set. We also used the default tolerance value as the stopping criterion (=0.001).

Given that learning algorithms can be sensitive to high variance in the feature space, our training and test data were scaled to a range from -1 to 1.

4.2 Results

In the following sections we present the performance of our models in terms of error (section 4.2.1) and correlation (section 4.2.2). Further analysis of the efficacy of linguistic features is presented in section 4.2.3, where we also describe the application of a feature selection method to derive optimal feature sets. Finally, section 4.2.4 provides an interpretation of the results and a summary of the most challenging aspects of the task.

4.2.1 Overall Prediction Performance

We evaluated each of our models on the test set and measured their performance against gold standard annotations. The results of evaluation are shown in Table 4.2.

MAE values show that the two mixed sets of features (*baseline+linguistic* and *full set*) beat the baseline features, although these differences are slight and not statistically significant in any of the cases (paired t-test). On the other hand, the 'pure' feature sets (*shallow* and *linguistic*) fall slightly below the baseline, with the drop in the linguistic set being the only statistically significant at 94% confidence level (paired t-test, $p <= 0.06$). This would indicate that using a mix of linguistic and shallow features is more effective than using only one type or the other, as is further described in section 4.2.4. However, the rest of metrics give a different picture. If we consider MSE, the *baseline* set outperforms all the others although their differences are also very small. In any case, the three best performing sets are still the *baseline*, *baseline+linguistic* and the *full set*.

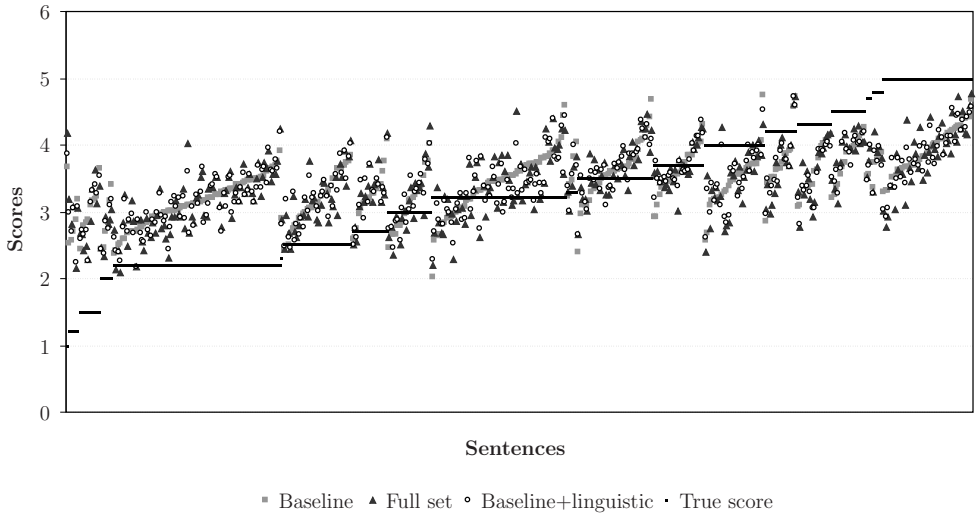


Figure 4.3: Comparison of true versus predicted scores for the three best feature sets. The closer the predicted points are to the horizontal black lines, the better.

We also studied the distribution of predicted values and how they deviate from the true scores. Figure 4.3 compares predictions of the three best performing sets against the gold standard, taking the *baseline* set as a reference.

All models seem to have a similar pattern of behaviour, exhibiting considerable dispersion around the expected values and systematic fluctuation. At times, the models converge in their predictions, especially for true scores ranging from 2.5 to 3, 3.5 to 3.7 and at 5, although they can be quite dissimilar at the beginning of the spectrum.

The fluctuation of predicted values for a single score is also very noticeable, spanning more than one band in many cases. On the other hand, if we consider scores in four bands (1-2, 2-3, 3-4 and 4-5) we find their individual predictions fall consistently into an approximate range from 2 to 4, showing a systematic amplitude. Nevertheless, our error analysis indicates predictions deviate around 0.68 (MAE) and 0.83 (RMSE) absolute points on average.

Figure 4.4 contrasts the distribution of the *linguistic* and *shallow* sets, with the *linguistic* set as reference. Unlike the previous models, these sets exhibit a clearer difference in predictions, with greater dispersion and fewer similar points. One of the reasons for this is that they encode different kinds of information, whose impact on predictions confirms that linguistic and shallow features account for different aspects of translations indeed. On the other hand, the first three models share many features between them, which somehow levels performance, whereas the *linguistic* and *shallow* sets have no features in common.

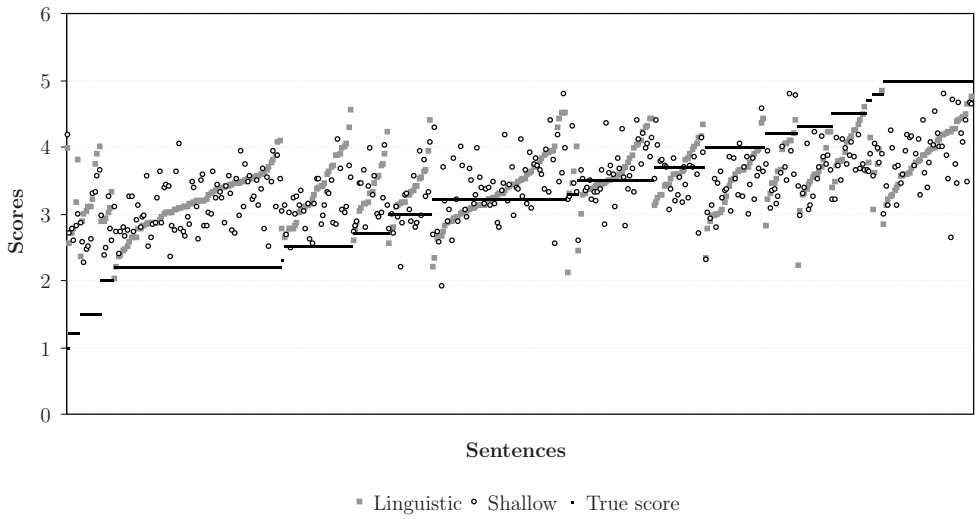


Figure 4.4: Comparison of true versus predicted scores for the linguistic and shallow feature sets. The closer the predicted points are to the horizontal black lines, the better.

As regards the variability of predictions, fluctuation is similar to that in the best three models, although slightly higher. Average MAE and RMSE are 0.70 and 0.80 respectively.

4.2.2 Correlation Analysis

A correlation analysis shows the *baseline* is the best performing set with a Pearson coefficient of 0.566 whereas the *linguistic* set is the worst with 0.456 (Table 4.3). Figure 4.5 includes individual scatter plots for each of the models where the relationship between true scores and predictions can be seen more clearly. These distributions reveal similar moderate dispersion for all models and some difficulty predicting values in the 1-2 range, possibly affected by a lower representation in the training data.

Feature set	**Pearson ↑**
Baseline	**0.566**
Baseline+linguistic	0.555
Full set	0.522
Shallow	0.518
Linguistic	0.456

Table 4.3: Correlation performance.

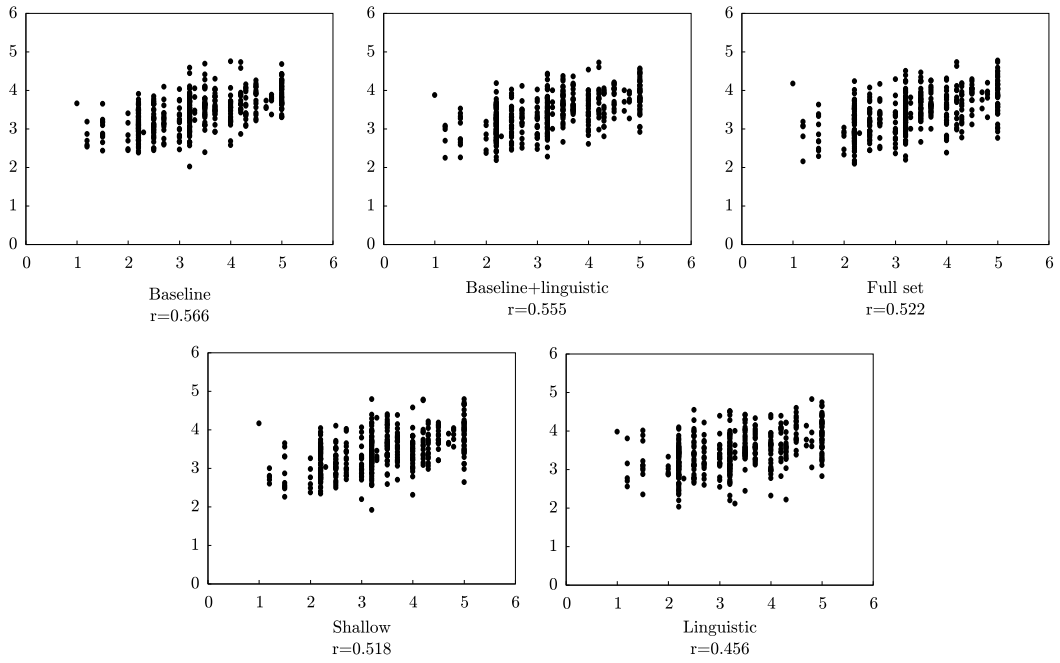


Figure 4.5: Correlation of true scores (x axis) versus predictions (y axis) for the evaluated feature sets.

These plots also show a consistent outlier: a bad translation (with a true score of 1) that is regarded as very good (with predictions greater than 3.5). The original and its translation are as follows:

I won *'t* *give it away* .

* *He ganado* *' t* *darle* .

This is clearly due to faulty tokenisation by the MT system which translated the halves separated by the apostrophe in isolation. As a result, the translations are acceptable for the isolated bits but not for the sentence as a whole. This may reveal that our features act at a very local level and are therefore unable to detect the global mismatch in meaning between the two sentences. More details about this kind of observed phenomena are given in section 4.2.4.

4.2.3 The Role of Linguistic Features

At first glance, the performance of our models might suggest that the integration of linguistic information is not highly beneficial for QE since the gain observed for the sets containing linguistically-motivated features is very modest with regard to the baseline. In addition, the fully *linguistic* set performs poorly when compared to the *shallow* set, which might question the efficiency of this type of feature.

However, we believe this behaviour may be due to two reasons. The first one is that our linguistic features may simply not be expressive enough to capture the differences in quality in our datasets. If this is found to be the case, then the worst-performing features should be replaced with new ones.

Our second hypothesis is that our models may have too many features for too little training data, a fact that is known to affect the performance of learning algorithms.

As a result, we performed a detailed analysis of the contribution of each feature, which is essential to find the most discriminating features and overcome these potential problems in the future.

Our first analysis was aimed at discovering the best and worst performing features. For this purpose, we took the full set of features and trained and tested models using: 1) one feature at a time and 2) all but one feature at a time. In the first case, our models rely on only one feature so the results are direct indicators of individual efficiency. In the second case, models are built using 146 features and show how performance suffers after the deletion of a feature (i.e. how each feature works with others in the set, whether they are found to be redundant or complementary). The ten best and worst features obtained from these experiments are listed in Table 4.4.

These results show that purely statistical shallow features lead the table, which is not surprising given the fact that n-gram language models have been proved to be highly versatile for many NLP applications. Nevertheless, linguistic features make up a significant proportion of these best features, showing that they are as good as many other shallow indicators. This is also backed up by the low proportion of linguistic features among the worst ones, where shallow features take the lead.

This table also allows us to identify the kind of information that seems to be the most and least relevant for our models. It is clear, for instance, that n-gram information on the target text is very helpful, be it linguistic (`target-pos-logprob`) or not (`source-bi-q4`, `target-logprob`), as well as parse tree width (`target-ptree-width`, `t-s-diff-ptree-width`) and noun information (`target-np`, `target-n-pcent`). On the other hand, source trigrams, target named entities (`target-ne-org`, `target-ne-loc`, `target-ne-per`), subject information (`target-s-v-agree-pcent`, `target-zero-subj-pcent`) and type-token ratios (`target-log-ttr`, `target-jarvis-ttr`, `target-corr-ttr`) do not seem to be very informative. Consistently good or bad features are enclosed in boxes.

As a side note, we have also observed that RMSE on the test set for models built on only one feature ranges from 0.912 (best) to 1.000 (worst) while those with all but one feature range from 0.826 (best) to 0.855 (worst). This not only shows that features have variable discriminative power but also reveals that the difference in performance between using one and almost all features is not excessively large (0.086 and 0.145 for the best and worst cases respectively).

Rank	One-feature model		Leave-one-feature-out model	
1	source-bi-q4		source-avg-tok-len	
2	source-logprob		source-bi-q4	
3	target-logprob		target-unknown	Ⓛ
4	source-bi		target-logprob	
5	target-pos-logprob	Ⓛ	t-s-diff-ptree-width	Ⓛ
6	target-pos-logprob-bl	Ⓛ	source-avg-trans-001-freq	
7	source-uni-q4		target-n-pcent	Ⓛ
8	target-ptree-width	Ⓛ	s-t-ratio-pron-pcent	Ⓛ
9	target-np	Ⓛ	target-pos-logprob	Ⓛ
10	source-tri-q4		t-s-token-ratio	
138	target-func-words-pcent	Ⓛ	source-bi-q2	
139	source-avg-trans-02-inv		target-zero-subj-pcent	Ⓛ
140	target-ne-org	Ⓛ	source-tri-q1	
141	target-log-ttr		source-tri-q2	
142	source-tri-q1		target-jarvis-ttr	
143	target-s-v-agree-pcent	Ⓛ	s-t-diff-periods	
144	source-tri-q2		source-bi	
145	target-ne-loc	Ⓛ	target-ne-per	Ⓛ
146	source-tri-q3		target-corr-ttr	
147	source-avg-trans-05-inv		source-tri-q3	

Table 4.4: List of best and worst performing features over the test set. Linguistic features are marked with Ⓛ while repeated features in the two sets are enclosed in boxes.

Our second analysis was aimed at finding the subset of features that would yield optimal performance on the test set, so that we could draw further conclusions. In practice, this was equivalent to finding a realistic lower bound for the error metrics we used, given our features and datasets.

For this analysis, we decomposed the problem into the following three main questions, which are dealt with separately:

1. What set of features would achieve the best performance on the test set?
2. How does this optimal set compare to our existing feature sets?
3. What is the drop in performance when feature selection is performed on the training set?

The first question is essentially a feature selection problem, where the optimal set of features is found by testing our models directly on the test set. Although there are many selection techniques for producing optimal sets of features, the simplest and

Name	Features	Total
Best-test	source-bi-q4, target-pos-logprob, source-avg-tok-len, target-gi, target-unknown, t-s-diff-vp, t-s-diff-ptree-width, target-tok-not-az, t-s-token-ratio, source-tri-q4, source-cont-words, source-ppl, s-t-ratio-pron-pcent, target-np, source-avg-trans-005-freq, target-logprob, target-pos-logprob-bl, t-s-diff-dtree-depth, source-np, source-num-tokens, target-cont-words, source-uni-q3, source-uni-q1, source-uni-q2, source-avg-trans-001-freq, s-t-ratio-tok-not-az, s-t-diff-questm-norm, target-pron-subj-pcent, target-v-pcent, source-ptree-width, s-t-diff-questm, source-avg-trans-02, t-s-diff-ne-per, s-t-diff-periods-norm, s-t-diff-semic-norm, source-ppl-no-smark, s-t-diff-periods	37

Table 4.5: Optimal set of features obtained from feature selection on the test set (in order of selection).

least heuristic approach to find the globally optimal set required building a model for every possible partition of our full feature set, which was not feasible in practice[3]. As a result, we opted for a Sequential Forward Selection method (Alpaydin, 2010, chap. 6) that allowed us to find suboptimal feature sets using substantially fewer iterations[4]. Using this method, we start from an empty set and build models adding one feature at a time, keeping in the set only the features that decrease the error on the evaluation data until no further improvement is possible. Since this method uses a hill climbing strategy, the algorithm is not guaranteed to find a global optimum. However, a local optimum was equally acceptable for our purpose. The resulting feature set found by our method (called *best-test*) is included in Table 4.5.

In order to answer our second question and find out how this set compares to the rest, we built a new regression model using the training data and the features from the *best-test* set. This model was then evaluated on the test set using the same metrics as before, whose values are reported in Table 4.6.

The comparison of these results with our best sets reveal there is a decrease of 10.03% in MAE and 14.63% in MSE with regard to the *baseline* model, 8.32% and 15.63% against the *baseline+linguistic* set, and 9.31% and 19.09% respectively against the *full set*. Improvements in predictions were found to be statistically significant in the first two cases, with 92% confidence (paired t-test, $p < 0.08$) and 95% confidence

Feature set	MAE	MSE	RMSE	Pearson
Best-test	0.618	0.574	0.758	0.647

Table 4.6: Best-test performance.

[3]For 147 features, there are 2^{147} possible feature sets (i.e. models) to test.

[4]For 147 features, the maximum number of sets (i.e. models) that are explored in the worst case is $147 \times (147 + 1)/2 = 10,878$.

Name	Features	Total
Best-cross-validation	target-logprob, target-pos-logprob, source-avg-tok-len, target-tok-not-az, source-logprob, s-t-diff-questm-norm, source-uni, source-ptree-depth, source-num-tokens, source-dtree-width, source-bi-q4, target-pos-ppl, target-n-pcent, s-t-diff-commas-norm, source-tri-q3, source-ppl, source-cont-words-pcent, source-bi-q1	18
Best-train	source-logprob, source-bi-q4, target-pos-logprob, source-num-tokens, source-avg-tok-len, target-tok-not-az, target-logprob, target-num-types, source-ptree-depth, source-ppl, target-s-v-agree-pcent, source-ne, target-jarvis-ttr, target-ttr, source-bi, source-avg-trans-001-freq, source-uni, source-avg-trans-05-freq, source-cont-words-pcent, target-zero-subj-pcent, source-v-pcent, s-t-diff-questm, source-tri, source-ne-other, target-n-pcent, target-ne, target-unknown, source-pp, source-dtree-width, source-avg-trans-001, source-avg-trans-02, target-ne-org, target-dang-det, target-punc, t-s-diff-dtree-depth, source-tri-q1, source-n-pcent, target-cont-words-pcent, target-pron-subj, source-bi-q2, t-s-diff-np, source-avg-trans-02-freq, source-uni-q1, t-s-diff-ptree-depth, target-ne-per, source-avg-trans-05-inv, s-t-ratio-v-pcent, t-s-diff-ne, source-uni-q3, target-dtree-width, target-non-pron-subj, target-punc-pcent, target-s-v-agree, source-tri-q4, target-ttr-bl, source-deixis, source-ne-org, t-s-diff-ptree-width, s-t-token-ratio, t-s-diff-pp, target-num-pcent, source-uni-q4, s-t-ratio-cont-words-pcent, t-s-diff-vp, s-t-ratio-tok-not-az, t-s-diff-deixis, target-corr-ttr, s-t-ratio-n-pcent, t-s-diff-ne-loc, s-t-diff-punc-norm, source-tok-not-az	71

Table 4.7: Optimal sets of features obtained from feature selection on the training set (in order of selection).

(paired t-test, $p < 0.05$) when compared to the *baseline* and *baseline+linguistic* sets respectively. Furthermore, the correlation is also increased by 8.08% over the best correlated set (i.e. the *baseline*). Given that most of these improvements are significant, we can compare the composition of all these sets and draw reliable conclusions regarding the role of our linguistic features.

The first factor we analysed was the proportion of linguistic features in the sets. From the best to the worst models (in terms of MAE), proportions are: 41% for *best-test*, 80% for *baseline+linguistic*, 48% for the *full set* and 0% for the *baseline*. Except for this last set, which shows performance can be comparatively good using only shallow features, the best three performing sets confirm that a significant proportion of linguistic information contributes to improving performance. In addition, this reinforces our hypothesis that a hybrid set of features produces better results than a purely linguistic or shallow set.

The second aspect we studied was the kind of linguistic information that is integrated into the *best-test* set, which helps us gain deeper insight into the most useful indicators. Out of the fifteen linguistic features in this set, six of them appear among the best in Table 4.4 (`s-t-ratio-pron-pcent`, `t-s-`

Feature set	MAE ↓	MSE ↓	RMSE ↓	Pearson ↑
Best-cross-validation	**0.667**	**0.674**	**0.821**	**0.560**
Best-train	0.671	0.687	0.829	0.545

Table 4.8: Performance of best feature sets obtained from cross-validation and the full training set.

`diff-ptree-width`, `target-np`, `target-pos-logprob`, `target-pos-logprob-bl`, `target-unknown`) while another one is not on the list but is closely related to one of the best features (`source-ptree-width` → `target-ptree-width`), confirming they are indeed the most effective linguistic indicators.

The third and last question of our analysis focuses on alternative ways of deriving optimal feature sets from the training data. In particular, we applied the same feature selection method as before but using different strategies on the training set: 1) performing ten-fold cross-validation, and 2) training and testing on the full training set. Features obtained in each case are included in Table 4.7.

Each of these sets was then used to build and test new regression models using the full training and test sets respectively, achieving the results shown in Table 4.8.

Although neither of these models is able to outperform the *best-test* set, they prove to be very good alternatives given the fact that they do not require information from the test set. A quick comparison with the rest of our initial models shows that the *best-cross-validation* set achieves the best MAE values and beats all other hybrid sets as regards overall performance, although these improvements are not statistically significant when compared to the *baseline*, *baseline+linguistic* or *full set*. On the other hand, correlation is slightly lower than that of the baseline set but this fact might be disregarded in favour of better error performance given that the difference is not significant.

Out of the total 18 features in the *best-cross-validation* set, 33% are linguistic and embody some of the most discriminative indicators we discovered in previous analyses, such as `target-n-pcent` and `target-pos-logprob`.

4.2.4 Findings and Challenges

Overall, MT quality estimation seems to be a difficult task and almost as challenging as translation itself. Producing a computational model to assess quality is not only complicated in terms of finding appropriate features and their representations but also in terms of actual correlation with human judgements, since quality is often perceived differently from one person to another. The following sections describe factors that were found to affect our QE models.

4.2.4.1 Human Agreement

In order to analyse human agreement on translation quality, we examined the initial set of translations that was used to compile the training and test data used in our experiments. Each translation was annotated by three human experts, shown in brackets in the examples below. Disagreement was considered to happen whenever the difference between any two annotations for a single translation was greater than one, in which case the translation was excluded from the dataset. Using this data, we could observe that most disagreements were influenced by the following factors:

1. Differences in the structure of sentences and how information is expressed (e.g. active-passive voice shift, variations in predicate-argument structure, etc.). Examples:

   a. *This will let you have two user profiles at once on the same phone.*

   Esto va a permitir que haya dos perfiles de usuario de una vez sobre el mismo teléfono. (4) (3) (5)

   Although the translation is grammatical and conveys the original meaning, there are two noticeable facts that may have an impact on perception: 1) a change in the verbal construction (the indirect object 'you' is dropped and an impersonal construction is used instead) and 2) a few expressions that are unsuitable for this context and make the sentence sound slightly unnatural ('de una vez', 'sobre el mismo teléfono').

   b. *Still, Friday featured the same sort of verbal fireworks that have dominated the talks for the past week.*

   * *Aún así, el viernes aparecía el mismo tipo de fuegos artificiales verbal que han dominado las conversaciones de la pasada semana.*
   (4) (2) (5)

   Here, the translation succeeds at replicating the figurative expression in the original by using a very wise translation of the main verb ('featured' → 'aparecía') but it fails to produce correct inflections for other words. The most interesting aspect of this translation is that the chosen form of the main verb ('aparecía') changes the original predicate-argument structure while still producing a faithful result. However, this seems to be a casual rather than intended effect, which is why it may be perceived with scepticism.

2. Differences in the interpretation of the translation, sometimes even disregarding the original intention.

Examples:

a. *The challenge was to speak like he does.*

El reto era hablar como lo hace. (4) (3) (5)

In this case, the translation is grammatically correct but it misses the original emphasis on 'he', losing the intended reference.

b. *Since Denis Kuljaš was injured, we urgently need another defender.*

* *Desde Denis Kuljaš resultó herido, necesitamos urgentemente otro defensor.* (4) (3) (5)

If 'since' had introduced a point in time, then the translation would only miss 'que' after 'desde' to be completely right, which is why some people would assign a high score to it. However, 'since' is used to introduce a reason in the original sentence, which is why the translation is actually inaccurate.

3. Discrepancies in the way humans regard the different aspects of quality (see section 3.1) and how lenient they are with translation errors. The following are some examples by aspect:

Grammaticality

Pakistani officials unraveling plot to send men to Afghanistan.

* *Funcionarios paquistaníes deshilachado trama enviar a los hombres a Afganistán.* (4) (2) (3)

Substantially different scores assigned despite the evident lack of grammaticality.

Accuracy

a. *The Bulava has generally not lifted off or has been damaged in the air.*

* *Los Bulava en general ha levantado o no se ha visto perjudicado en el aire.* (3) (1) (3)

Apart from a few infelicities, the translation says exactly the opposite of the original sentence. Surprisingly, while one annotator considered it completely wrong for this reason, others gave it a fair score.

b. *New vaccinations*

Nuevas vacunas (5) (3) (5)

Although the translation looks good on the surface, rendering 'vaccinations' as 'vacunas' (instead of 'vacunaciones') introduces a subtle difference in meaning that not all people know or perceive.

c. *It's a real art form, and you need time and steady nerves.*

Es un verdadero arte, y se necesita tiempo y nervios templados.
(3) (4) (5)

Again, this is a case of inaccurate translation that may sound awkward to some but acceptable to others.

Punctuation

Click here to find out more!

* *Clic aquí para averiguar más!* (5) (3) (5)

Proper punctuation is often disregarded. In this case, the opening exclamation mark (¡) is missing at the beginning of the sentence.

4. Unexplained inconsistencies in the annotation process.

Example:

Russian companies Lukoil and Gazprom were the top stakeholders in two of the contracts awarded this weekend.

Las empresas rusas Lukoil y Gazprom fueron las principales partes interesadas en dos de los contratos concedidos este fin de semana.
(5) (5) (3)

While the translation is flawless, one of the annotators gave it only a fair score. The inconsistency of his annotation became evident when he gave the maximum score to exactly the same sentence that was proposed as an alternative.

Out of the 1,000 translation instances in the analysed set, 148 were considered to disagree while 852 were deemed suitable enough for inclusion on the training set, representing 14,8% and 85,2% respectively. These differences in the perception of quality raise the question of whether we can actually evaluate computational models in terms of correlation with human judgements, given that there is no clear consensus on this matter.

4.2.4.2 Datasets

The size of the dataset used to build a model is also a relevant factor and is directly related to the number features and classes. In our case, the total number of training instances (1,832) may not be enough to build a very successful model that differentiates 5 classes using 147 features. Provided a much larger training set was used and the number of classes was reduced, more accurate models could be built, such as the one by Bach, Huang and Al-Onaizan (2011). In cases where larger datasets are not available, feature selection could help reduce dimensionality and minimise the effect of a small training set.

4.2.4.3 Linguistic Resources

Linguistic resources also impose a limitation on the performance of our models. In the first place, resources such as monolingual corpora, dictionaries, spell checkers, grammar checkers, tokenisers, parsers, named entity recognisers and even semantic role labellers may simply not be available for certain languages or be difficult to exploit (because they use different representations, for instance). As a result, the lack of appropriate tools precludes the exploitation of many linguistic indicators and positions shallow features in an advantageous position within the models.

A second concern is the reliability of these resources. On the one hand, they are naturally limited by approaching linguistic analysis in 'hard' numerical terms while, on the other, they tend to produce inaccurate results for ungrammatical input, thus leading to further errors in the computation of features. One such example is the application of constituency parsing to ill-formed translations, which can easily produce an incorrect structure that is further propagated to other features, like the estimation of subject-verb agreement or dangling determiners (see Table 4.9 for an example). It is then expected that the development of robust linguistic processors will help reduce the number of errors in the estimations and therefore lead to a more significant increase in performance.

4.2.4.4 Linguistic Features

Our linguistically-informed models also seem to suggest that many features are limited in their scope. Firstly, most of them act on only one text (either the source or target) and even those aimed at capturing details of the transfer process are based on these isolated estimations. Examples include the differences in noun and verb phrases as well as mismatched named entities, all of which are computed from the individual values on the source and target texts. However, estimating the accuracy of translations in this fashion does not seem very reliable.

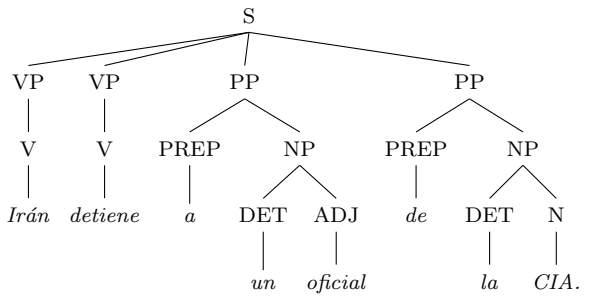
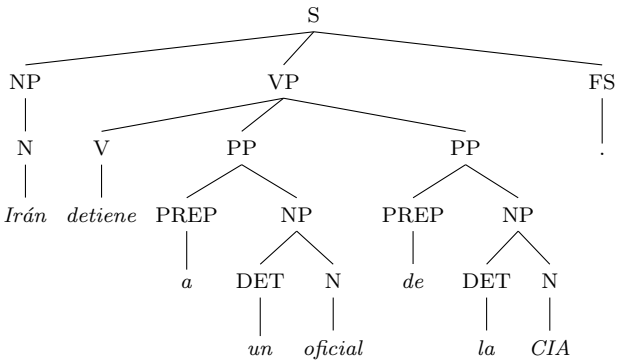
Automatic parsing (FreeLing)	**Features**
	`target-num-tokens` = 8 `target-n-pcent` = 0.125 `target-v-pcent` = 0.25 `target-vp` = 2 `target-ptree-width` = 4 `target-exp-subj` = 0 `target-zero-subj` = 2 `target-s-v-agree` = 0 `target-dang-det` = 1 `target-punc` = 0 `target-ne` = 0 `target-ne-loc` = 0
Human parsing	**Features**
	`target-num-tokens` = 9 `target-n-pcent` = 0.33 `target-v-pcent` = 0.11 `target-vp` = 1 `target-ptree-width` = 3 `target-exp-subj` = 1 `target-zero-subj` = 0 `target-s-v-agree` = 1 `target-dang-det` = 0 `target-punc` = 1 `target-ne` = 1 `target-ne-loc` = 1

Table 4.9: Automatic vs. human parsing of a Spanish sentence and its impact on linguistic of features.

On the one hand, differences between source and target values can be expressed in many ways, for example by computing ordinary subtraction, unsigned subtraction, proportions, etc. In turn, each of these alternatives will yield a specific distribution of values and characterise this aspect differently, causing variations in final predictions. As we mentioned in section 3.2.2, ordinary subtraction was found to be the most effective operation for contrasting features and the one we mostly adopted in our experiments. In a few cases, we decided to keep the original individual values instead of computing contrastive features, in an attempt to let the machine learning algorithm find the most likely relationship between the variables.

The other noticeable effect of estimating accuracy from individual values is that they might not indicate a real match of the evaluated feature. The main reason for this is the inability of contrastive features to check for true linguistic correspondences instead of relying on simple counts. As an example, consider the case of named entity matches and how they can be wrongly estimated for this reason:

a. *The last training session was on Monday night, on the* [$_{\text{LOC}}$ *Manhattan*] *side of the* [$_{\text{ORG}}$ *East River*].

   * *La última sesión de entrenamiento fue el lunes por la noche, en* [$_{\text{LOC}}$ *Manhattan*] *lado del* [$_{\text{OTHER}}$ *Medio*] río.

`t-s-diff-ne`	$= 0$	`t-s-diff-ne-org`	$= -1$
`t-s-diff-ne-loc`	$= 0$	`t-s-diff-ne-other`	$= 1$

The total difference of named entities (`t-s-diff-ne`) is estimated as 0, when in fact the entity *East River* is completely lost.

b. *The* [$_{\text{ORG}}$ *NATO*] *mission officially ended* [$_{\text{ORG}}$ *Oct.*] *31.*

* *La misión de la* [$_{\text{ORG}}$ *OTAN*] *terminó oficialmente* [$_{\text{ORG}}$ *PTUM*]. *31.*

`t-s-diff-ne`	$= 0$	`t-s-diff-ne-org`	$= 0$
`t-s-diff-ne-loc`	$= 0$	`t-s-diff-ne-other`	$= 0$

In this case, not only is *Oct.* wrongly recognised as an entity but it is also lost in the translation despite the value of `t-s-diff-ne`.

Sparsity is also a central issue in the assessment of quality because not all linguistic phenomena occur in all sentences. As a result, many features that would be specially informative may not be applicable to a specific translation pair and be mostly useless. This implies that the learning algorithm should be capable of discerning what features are the most suitable for assessing each specific translation according to the traits it exhibits, but again this may be hindered if no special value is used to signal non-observed phenomena (especially for contrastive features). Some typical examples of sparse features are named entities, numerical expressions, infrequent punctuation marks and non-alphabetic tokens, since they are not expected to occur in most sentences.

Our linguistic features also seem to act at a local level rather than globally. This becomes especially evident in cases where a translation misses the meaning of the original sentence, despite having many aspects in common. This reinforces the idea that more global features are needed to achieve better estimations, such as overall grammar checking, the identification of semantic roles and lexicon accuracy. The following examples show pairs of sentences with many local matches where overall translation quality is low:

a. [$_{\text{PER}}$ *Scalia*] *and* [$_{\text{LOC}}$ *Thomas*] [$_{\text{V}}$ *dine*] *with* [$_{\text{N}}$ *healthcare*] [$_{\text{N}}$ *law*] [$_{\text{N}}$ *challengers*] *as* [$_{\text{N}}$ *court*] [$_{\text{V}}$ *takes*] [$_{\text{N}}$ *case*].

* [$_{\text{ORG}}$ *Scalia*] *y* [$_{\text{ORG}}$ *Thomas*] [$_{\text{V}}$ *cenar*] *con la* [$_{\text{N}}$ *asistencia*] *sanitaria* [$_{\text{N}}$ *ley*] *retadores como* [$_{\text{ORG}}$ *Tribunal*] [$_{\text{V}}$ *tiene*] [$_{\text{N}}$ *caso*].

(too literal)

b. *It* [$_{V}$ *was*] *as big as a* [$_{N}$ *suitcase*], [$_{V}$ *had*] *a small cathode* [$_{N}$ *ray*] [$_{N}$ *tube*] [$_{N}$ *display*], *and* [$_{PRON}$ *I*] [$_{V}$ *fell*] *in* [$_{N}$ *love*] *with* [$_{PRON}$ *it*].

* [$_{V}$ *Fue*] *tan grande como una* [$_{N}$ *maleta*], [$_{V}$ *había*] *un pequeño* [$_{N}$ *rayo*] *catódicos* [$_{N}$ *tubo*] [$_{N}$ *exhibición*], *y* [$_{PRON}$ *me*] [$_{V}$ *cayeron*] *en el* [$_{N}$ *amor*] *con* [$_{PRON}$ *él*].

(bad lexical choice and inability to translate idiomatic expressions)

4.2.4.5 Feature Selection

Our experiments have also shown that feature selection is crucial to finding an optimal set of features for quality estimation (see section 4.2.3), supporting the observations by Specia, Turchi, Cancedda, Dymetman et al. (2009). This not only proved that reducing the feature space was beneficial but it also revealed that linguistic features can be as informative as many shallow features. This also suggests that these features are not interchangeable but actually complementary and should then be carefully combined to achieve optimal results.

In order to further investigate this complementarity, we studied how linguistically-informed feature sets compare to shallow sets in individual sentences. Our aim was to identify whether linguistic features were better at predicting certain types of sentences than shallow features and how this explains complementarity.

Table 4.10 includes the results of comparing our initial feature sets over the 422 test sentences grouped by score classes. Each cell indicates the number of sentences in the class interval that are estimated more accurately on one model than on the other (i.e. how many sentences are predicted with less error than on the other model).

Class	**Full set**	**Base.+ling.**	**Linguistic**	**Linguistic**
[1-2)	**11**	**10**	**11**	**13**
[2-3)	61	62	**74**	**80**
[3-4)	**77**	**79**	**76**	**74**
[4-5]	**67**	59	**69**	61
Total	216	210	230	228
	51.18%	49.76%	54.50%	54.03%
		vs.		
Class	**Shallow**	**Baseline**	**Shallow**	**Baseline**
[1-2)	5	6	5	3
[2-3)	**73**	**72**	60	54
[3-4)	70	68	71	73
[4-5]	58	**66**	56	**64**
Total	206	212	192	194
	48.82%	50.24%	45.50%	45.97%

Table 4.10: Comparison of sentence prediction accuracy between linguistically-enriched and shallow feature sets.

System ID	MAE	RMSE
⋆ SDLLW_M5PbestDeltaAvg	0.61	0.75
UU_best	0.64	0.79
SDLLW_SVM	0.64	0.78
UU_bltk	0.64	0.79
Loria_SVMlinear	0.68	0.82
UEdin	0.68	0.82
TCD_M5P-resources-only	0.68	0.82
Baseline (17FFs SVM)	0.69	0.82
Loria_SVMrbf	0.69	0.83
SJTU	0.69	0.83
WLV-SHEF_FS	**0.69**	**0.85**
PRHLT-UPV	0.70	0.85
WLV-SHEF_BL	**0.72**	**0.86**
DCU-SYMC_unconstrained	0.75	0.97
DFKI_grcfs-mars	0.82	0.98
DFKI_cfs-plsreg	0.82	0.99
UPC_1	0.84	1.01
DCU-SYMC_constrained	0.86	1.12
UPC_2	0.87	1.04
TCD_M5P-all	2.09	2.32

Table 4.11: Official results for the scoring sub-task of the WMT 2012 Quality Evaluation Shared Task (Callison-Burch, Koehn, Monz, Post et al., 2012). The winning submission is indicated by a ⋆ (result statistically-significant at $p = 0.05$). The systems in the middle grey area are not different from the baseline system at a statistically-significant level ($p = 0.05$).

Results show that the addition of linguistic information is clearly complementary, since, on average, it helps predict 52.37% of sentences more accurately than using shallow features alone. These figures also reveal that linguistically-enriched sets consistently outperform shallow sets at predicting scores in classes 1-2 and 3-4, while success in classes 2-3 and 4-5 is evenly divided. Overall, these numbers not only confirm that the error decreases with the use of linguistic features but also that a higher level of refinement is also achieved.

Finally, given that linguistic features require additional resources, time and effort to be computed, it may be questioned whether they are worth using in QE to achieve only a modest improvement. However, our experiments have proved that a carefully combined set of linguistic and shallow features can indeed achieve a statistically significant improvement and reduce the number of total features considerably when compared to the *full* and *baseline+linguistic* sets.

4.2.5 Comparison with State-of-the-Art Systems

As was mentioned in section 1.3, two earlier versions of our models were submitted to the WMT 2012 Quality Estimation Shared Task, which allowed us to see how

they compared to other systems. Submissions were evaluated according to two different sub-tasks: the scoring of translations on a scale from 1 to 5 and the ranking of translations from the best to the worst. However, only the scoring results are reported in this section.

The performance and ranking of participating systems is included in Table 4.11. Our submitted models were WLV-SHEF_FS (built on the *full set* of features) and WLV-SHEF_BL (built on the *baseline+linguistic* set) but their results in the competition were different from the ones reported in this work because of differences in the implementations of some features. In particular, our new results (see Table 4.2) show that these sets now achieve values of 0.67 and 0.68 for MAE and 0.82 and 0.84 for RMSE respectively (all rounded), although they would still be considered in the grey area (i.e. not statistically significant).

On the other hand, our new feature sets obtained from feature selection yield even better results (see Table 4.8) and would then rank higher among these state-of-the-art systems.

Chapter 5

Conclusions and Future Work

This work has presented a study of linguistic features for estimating machine translation quality. Our approach starts from a review of desirable aspects of translation and derives related computational features that are used to build a supervised machine learning model. Experiments were conducted using an English-Spanish dataset enriched with quality judgements of post-editing requirements on a scale from 1 (re-translation required) to 5 (fit for publication), containing 1,832 sentences for training and 422 for testing. Models were built using SVM epsilon regression with a radial basis kernel function and optimised learning parameters.

In order to evaluate the performance of our proposed features, we trained and compared several regression models using different proportions of shallow and linguistic information. In addition, we studied the best and worst performing features and applied a feature selection algorithm to derive optimal sets.

5.1 Main Observations

Evaluation of our models revealed that linguistically-enriched feature sets are able to reduce MAE when compared to a baseline containing only shallow features, although this improvement is often slight and statistically insignificant. On the other hand, values for MSE and RMSE are lower in the baseline model but this difference is not significant either.

We also observed that the best three performing feature sets (*baseline*, *baseline+ linguistic* and the *full set*) behave similarly whereas a comparison of the linguistic and shallow sets reveal marked differences. We believe this is a clear indicator that these features capture different aspects of translations and are therefore complementary.

In terms of correlation, the best and worst models correspond to the *baseline* and *linguistic* sets respectively. For all sets, however, we found the range of predictions for a single true score fluctuates very similarly.

An analysis of individual features revealed that target PoS n-gram probabilities and phrase structure information are the most relevant indicators whereas source trigrams and target features such as named entities, subject agreements and token-type ratios are the least informative.

Additionally, we applied a feature selection algorithm to training and test data to find optimal feature sets. Results indicate that all the extracted optimal sets include a considerable proportion of linguistic features, confirming the hypothesis that they are complementary to shallow features and contribute to achieving better results. In particular, optimal features derived from the test set produced significant improvements in MAE, MSE and correlation when compared to the best sets.

The results of our experiments suggest that QE is a difficult task, especially when compared to reference-based evaluation. Intuitively, MT evaluation seems easier given that references and translation hypotheses are in the same language and can be compared straightforwardly but this is not possible in QE. In fact, it is very difficult to compare segments in different languages because they use different words, structures or even alphabets. In addition, the kind of linguistic information that can be extracted for each segment depends on the availability of appropriate tools and this can be a problem for low-resource languages.

With regard to our linguistic features, we believe there are many factors that may have led to moderate performance:

1. Human ratings are not always homogeneous and are often based on different perceptions of quality.

2. The five-point scale used for scoring translations might be too fine-grained for our learning algorithm.

3. The small size of the training set may be also be a problem, especially when a large number of features is used.

4. The lack of advanced linguistic processors for the source and target language precluded the implementation of more informative linguistic features, such as semantic roles or well-formedness estimations.

5. Automatic parsers might not be robust enough to parse sentences reliably (especially ungrammatical ones), which may introduce further errors in the estimations.

6. The sparsity of some features may also be problem, since not all sentences exhibit the same linguistic phenomena. This may render some feature useless in many cases, such as named entities or numerical expressions.

7. Most of our linguistic features seem to act at a very local level, therefore being unable to capture more global information like grammatical correctness or preservation of original meaning.

8. Our set of linguistic features may not be large or expressive enough to capture the difference between quality scores.

Although some of these problems can be circumvented by refining the experimental setup, many others require the development of new automatic tools and further empirical analysis of linguistic phenomena, some of which are addressed in our proposals for future research.

5.2 Future Work

The directions for future research described in this section are intended to address the problems evidenced in our current approach and investigate performance in other settings.

The first of our proposals is the design and implementation of new linguistic features. In particular, this includes more global indicators such as sentence grammaticality (which could be approximated by using probabilities from a parser), semantic roles (by projection, for example) and latent semantic analysis, which are expected to model fidelity more reliably. Additional features include, but are not limited to, new language models using large corpora (such as Wikipedia dumps), hybrid n-grams mixing tokens and PoS tags, redundancy checks (e.g by counting repetitions of content words or synonyms) and clause identification (e.g. by using a clause splitter).

Secondly, a study of contrastive features should be conducted in order to determine the most appropriate representation for these features. Although we adopted ordinary subtraction and ratios after a few simple experiments, we believe a more detailed study is necessary to find an optimal representation.

Thirdly, given that it has been observed that users have different views on translation quality and can disagree in many cases, it would be interesting to carry out an experiment to evaluate how users perceive automatic quality predictions. In a typical setting, a user would be given a source text, a machine translation and an automatic quality score and they would have to decide whether they agree or disagree with the estimation. If they disagree, they would have to provide a new score which would then be used to estimate an error metric. Although model performance is usually evaluated in terms of prediction error over the expected scores, we believe this user-driven assessment in real scenarios could bring an equally important insight.

Additionally, other machine learning algorithms should also be applied to see if they yield better prediction performance than SVMs, although some have already been proved to be slightly less successful. In fact, we believe experiments should be replicated using a classification algorithm in order to determine whether it is more suitable than regression for this particular task and score scale.

Improvements on the selection of optimal features should also be investigated. This includes the application of standard feature selection algorithms and efficient ways of discovering optimal features without resorting to test data.

Finally, it would be interesting to evaluate how our feature sets perform with other datasets, especially with segments from different translation systems and alternative score scales.

Bibliography

al Qinai, J., 2000. Translation Quality Assessment. Strategies, Parametres and Procedures. *Meta: Translators' Journal*, 45(3), pp. 497–519.

Albrecht, J. and Hwa, R., 2007a. A Re-examination of Machine Learning Approaches for Sentence-Level MT Evaluation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, pp. 880–887.

Albrecht, J. and Hwa, R., 2007b. Regression for Sentence-Level MT Evaluation with Pseudo References. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, pp. 296–303.

Albrecht, J. and Hwa, R., 2008. The role of pseudo references in MT evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*. Columbus, Ohio: Association for Computational Linguistics, pp. 187–190.

Alpaydin, E., 2010. *Introduction to Machine Learning*. Adaptive Computation and Machine Learning. Cambridge, MA: The MIT Press, 2nd ed.

Amigó, E., Giménez, J. and Verdejo, M.F., 2009. Procesamiento lingüístico en métricas de evaluación automática de traducciones. *Procesamiento del Lenguaje Natural*, 43, pp. 215–222.

Bach, N., Huang, F. and Al-Onaizan, Y., 2011. Goodness: A method for measuring machine translation confidence. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, OR, USA: Association for Computational Linguistics, pp. 211–219.

Banerjee, S. and Lavie, A., 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 65–72.

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S. et al., 2004. Confidence estimation for machine translation. Final Report of Johns Hopkins 2003 Summer Workshop on Speech and Language Engineering, Johns Hopkins University, Baltimore, MD, USA.

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

Brumfit, C., 1984. *Communicative methodology in language teaching: the roles of fluency and accuracy*. Cambridge: Cambridge University Press.

Callison-Burch, C., Koehn, P., Monz, C., Peterson, K. et al., 2010. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*. Uppsala, Sweden: Association for Computational Linguistics, pp. 17–53. Revised August 2010.

Callison-Burch, C., Koehn, P., Monz, C., Post, M. et al., 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Montreal, Canada: Association for Computational Linguistics. (To appear).

Callison-Burch, C., Koehn, P., Monz, C. and Zaidan, O., 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, pp. 22–64.

Callison-Burch, C., Osborne, M. and Koehn, P., 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of EACL*. Association for Computational Linguistics, vol. 2006, pp. 249–256.

Carroll, J.B., 1964. *Language and Thought*. Englewood Cliffs, NJ: Prentice-Hall.

Chang, C.C. and Lin, C.J., 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), pp. 1–27.

Chomsky, N., 1957. *Syntactic Structures*. The Hague: Mouton & Co., 1st ed.

Chomsky, N., 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

Crystal, D., 2010. *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press, 3rd ed.

Doddington, G., 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*. San Diego, CA, USA: Morgan Kaufmann Publishers Inc., HLT '02, pp. 138–145.

Doyle, M.S., 2003. Translation Pedagogy and Assessment: Adopting ATA's Framework for Standard Error Marking. *ATA Chronicle*, 32(11), pp. 21–28. Latest version available at `http://www.atanet.org/certification/aboutexams_error.php`.

Dubay, W.H., 2004. The principles of readability. [Electronic format only]. Costa Mesa, CA: Impact Information. Available at `http://www.nald.ca/fulltext/readab/readab.pdf`.

Dugast, D., 1980. *La statistique lexicale*. Genève: Slatkine.

Felice, M. and Specia, L., 2012. Linguistic features for quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Montreal, Canada: Association for Computational Linguistics. (To appear).

Fu, Y., 1973. General remarks on translation. *Renditions*, 1(1), pp. 4–6. Translated by C. Y. Hsiu.

Gamon, M., Aue, A. and Smets, M., 2005. Sentence-level MT Evaluation Without Reference Translations: Beyond Language Modeling. In *Proceedings of the 10th Meeting of the European Association for Machine Translation (EAMT)*. Budapest, Hungary, pp. 103–111.

Giannakopoulos, G., Karkaletsis, V. and Vouros, G., 2012. Detecting Human Features in Summaries - Symbol Sequence Statistical Normality. In *Proceedings of the 7th Hellenic Conference on Artificial Intelligence (SETN 2012)*. Lamia, Greece. (To appear).

Giménez, J. and Màrquez, L., 2010. Linguistic measures for automatic machine translation evaluation. *Machine Translation*, 24, pp. 209–240.

González-Rubio, J., Ortiz-Martínez, D. and Casacuberta, F., 2010. Balancing user effort and translation error in interactive machine translation via confidence measures. In *Proceedings of the ACL 2010 Conference Short Papers*. Uppsala, Sweden: Association for Computational Linguistics, ACLShort '10, pp. 173–177.

Guiraud, P., 1954. *Les Caractères Statistiques du Vocabulaire*. Paris: Presses Universitaires de France.

Habash, N., 2008. Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. Columbus, OH, USA: Association for Computational Linguistics, HLT-Short '08, pp. 57–60.

Halliday, M.A.K. and Hasan, R., 1976. *Cohesion in English*. London: Longman.

Hardmeier, C., 2011. Improving machine translation quality prediction with syntactic tree kernels. In *Proceedings of the 15th Conference of the European Association for Machine Translation (EAMT 2011)*. Leuven, Belgium, pp. 233–240.

He, Y., Ma, Y., Roturier, J., Way, A. et al., 2010. Improving the post-editing experience using translation recommendation: A user study. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*. Denver, CO, USA, AMTA 2010, pp. 247–256.

He, Y., Ma, Y., van Genabith, J. and Way, A., 2010. Bridging SMT and TM with translation recommendation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, ACL '10, pp. 622–630.

Herdan, G., 1960. *Type-token Mathematics: A Textbook of Mathematical Linguistics*. The Hague: Mouton & Co.

House, J.M., 1977. *A model for translation quality assessment*. Tübingen: TBL-Verlag Gunter Narr.

House, J.M., 1997. *Translation quality assessment: a model revisited*. Tübingen: Gunter Narr Verlag.

Institute of Linguists, 2011. *Diploma in Translation: Handbook for Candidates*. London: Institute of Linguists Educational Trust.

Jarvis, S., 2002. Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, 19(1), pp. 57–84.

Johnson, W., 1944. I. A program of research. *Psychological Monographs*, 56(2), pp. 1–15.

Jurafsky, D. and Martin, J.H., 2009. *Speech and Language Processing*. Upper Saddle River, NJ: Prentice-Hall, 2nd ed.

Koehn, P., 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*. AAMT, Phuket, Thailand: AAMT, pp. 79–86.

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

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C. et al., 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Prague, Czech Republic: Association for Computational Linguistics, pp. 177–180.

Koehn, P., Och, F.J. and Marcu, D., 2003. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. Edmonton, Canada: Association for Computational Linguistics, pp. 48–54.

Kubler, S., McDonald, R., Nivre, J. and Hirst, G., 2009. *Dependency Parsing*. Morgan and Claypool Publishers.

Lin, C.Y. and Och, F.J., 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Barcelona, Spain: Association for Computational Linguistics, ACL '04.

Louwerse, M.M. and Graesser, A.C., 2005. Coherence in discourse. In P. Strazny, ed., *Encyclopedia of linguistics*, Chicago: Fitzroy Dearborn, vol. 1, pp. 216–218.

Mederos Martín, H., 1985. *Procedimiento de cohesión en el español actual*. Ph.D. thesis, Universidad de la Laguna, Tenerife, Spain.

Melamed, I.D., Green, R. and Turian, J.P., 2003. Precision and recall of machine translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers - Volume 2*. Edmonton, Canada: Association for Computational Linguistics, NAACL-Short '03, pp. 61–63.

Munday, J., 2008. *Introducing Translation Studies: Theories and Applications*. London: Routledge, 2nd ed.

Nida, E.A., 1964. *Toward a science of translating*. Leiden: E. J. Brill.

Nida, E.A. and Taber, C.R., 1969. *The theory and practice of translation*. Leiden: E. J. Brill.

Niessen, S., Och, F.J., Leusch, G. and Ney, H., 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*. Athens, Greece, vol. 1, pp. 39–45.

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

Padró, L., Collado, M., Reese, S., Lloberes, M. et al., 2010. FreeLing 2.1: Five Years of Open-source Language Processing Tools. In N.C.C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner and D. Tapias, eds., *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA).

Papineni, K., Roukos, S., Ward, T. and Zhu, W.J., 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Philadelphia, PA, USA: Association for Computational Linguistics, ACL '02, pp. 311–318.

Pighin, D. and Màrquez, L., 2011. Automatic projection of semantic structures: an application to pairwise translation ranking. In *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Portland, OR, USA: Association for Computational Linguistics, SSST-5, pp. 1–9.

Quirk, C.B., 2004. Training a sentence-level machine translation confidence metric. In *Proceedings of the International Conference on Language Resources and Evaluation*. Lisbon, Portugal, vol. 4 of *LREC 2004*, pp. 825–828.

Real Academia Española, 2009. *Nueva gramática de la lengua española*. Madrid: Espasa-Calpe, 3rd ed.

Romero, L.G., 2005. Deixis. In P. Strazny, ed., *Encyclopedia of linguistics*, Chicago: Fitzroy Dearborn, vol. 1, pp. 260–261.

Saers, M. and Wu, D., 2009. Improving phrase-based translation via word alignments from stochastic inversion transduction grammars. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation*. Boulder, CO, USA: Association for Computational Linguistics, SSST '09, pp. 28–36.

Schmid, H., 1995. Improvements In Part-of-Speech Tagging With an Application To German. In *In Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland, pp. 47–50.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. et al., 2006a. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*. Cambridge, MA, USA, AMTA 2006, pp. 223–231.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. et al., 2006b. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*. Cambridge, MA, USA, AMTA 2006, pp. 223–231.

Song, X. and Cohn, T., 2011. Regression and ranking based optimisation for sentence level machine translation evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, WMT '11, pp. 123–129.

Soricut, R. and Echihabi, A., 2010. TrustRank: Inducing Trust in Automatic Translations via Ranking. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, pp. 612–621.

Specia, L., 2011a. Exploiting objective annotations for measuring translation post-editing effort. In *Proceedings of the 15th Conference of the European Association for Machine Translation (EAMT 2011)*. Leuven, Belgium, pp. 73–80.

Specia, L., 2011b. Linguistic Information for Measuring Translation Quality. [Electronic format only]. Presented at the International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT-2011). Available at `http://ixa2.si.ehu.es/lihmt2011/lihmt_specia.ppt`. Barcelona, Spain, 18 November 2011.

Specia, L., Cancedda, N. and Dymetman, M., 2010. A dataset for assessing machine translation evaluation metrics. In N.C.C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner and D. Tapias, eds., *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA).

Specia, L. and Farzindar, A., 2010. Estimating Machine Translation Post-Editing Effort with HTER. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*. Denver, CO, USA, AMTA 2010, pp. 33–41.

Specia, L. and Giménez, J., 2010. Combining Confidence Estimation and Reference-based Metrics for Segment-level MT Evaluation. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*. Denver, CO, USA, AMTA 2010.

Specia, L., Hajlaoui, N., Hallett, C. and Aziz, W., 2011. Predicting machine translation adequacy. In *Machine Translation Summit XIII*. Xiamen, China, pp. 19–23.

Specia, L., Raj, D. and Turchi, M., 2010. Machine translation evaluation versus quality estimation. *Machine Translation*, 24, pp. 39–50.

Specia, L., Turchi, M., Cancedda, N., Dymetman, M. et al., 2009. Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT)*. Barcelona, Spain, pp. 28–35.

Specia, L., Turchi, M., Wang, Z., Shawe-Taylor, J. et al., 2009. Improving the confidence of machine translation quality estimates. In *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII)*. Ottawa, Canada, pp. 136–143.

Stolcke, A., 2002. SRILM—An extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*. Denver, USA, vol. 2, pp. 901–904.

Taulé, M., Martí, M.A. and Recasens, M., 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In N.C.C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner and D. Tapias, eds., *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA).

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A. et al., 1997. Accelerated DP based Search for Statistical Translation. In *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH '97)*. Rhodes, Greece, pp. 2667–2670.

Turian, J.P., Shen, L. and Melamed, D.I., 2003. Evaluation of Machine Translation and its Evaluation. In *Machine Translation Summit IX*. International Association for Machine Translation, pp. 386–393.

Tytler, A.F.L.W., 1791. *Essay on the principles of translation*. London: T. Cadell and W.Davies.

Uribe Mallarino, M.d.R., 2002. Camino de la lectura entre ≪topics≫ y marcas de cohesión: la comprensión lectora en la lengua extranjera con atención al contraste español-italiano. In M. Colombo, G. Iamartino and M. Scaramuzza Vidoni, eds., *Mots Palabras Words. Ensayos.*, Milano, Italy: Led on Line.

van Gelderen, E., 2005. Function words. In P. Strazny, ed., *Encyclopedia of linguistics*, Chicago: Fitzroy Dearborn, vol. 1, pp. 362–364.

Way, A., 2010. Machine Translation. In A. Clark, C. Fox and S. Lappin, eds., *The Handbook of Computational Linguistics and Natural Language Processing*, Chichester, UK: Wiley Blackwell, chap. 19, pp. 531–573.

Weissbort, D. and Eysteinsson, A., eds., 2006. *Translation - Theory and Practice: a historical reader*. Oxford: Oxford University Press.

Wetzel, L., 2009. *Types and Tokens: On Abstract Objects*. The MIT Press.

Witten, I.H., Frank, E. and Hall, M.A., 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, CA: Morgan Kaufmann, 3rd ed.

Xiong, D., Zhang, M. and Li, H., 2010. Error detection for statistical machine translation using linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, ACL '10, pp. 604–611.

Yang, M.Y., Sun, S.Q., Zhu, J.G., Li, S. et al., 2011. Improvement of machine translation evaluation by simple linguistically motivated features. *Journal of Computer Science and Technology*, 26, pp. 57–67.