BAR-ILAN UNIVERSITY

Automatic Detection of Machine Translated Text and Translation Quality Estimation

Roee Aharoni

Submitted in partial fulfillment of the requirements for the Master's Degree in the Department of Computer Science, Bar-Ilan University

Ramat Gan, Israel 2015

This work was carried out under the supervision of Professor Moshe Koppel, Department of Computer Science, Bar-Ilan University.

Acknowledgments

First of all, I would like to thank my supervisor, Prof. Moshe Koppel, for giving me the opportunity to enter the fascinating world of academic research in general and of machine learning and natural language processing specifically. I would not enjoy this great journey as I did without his guidance and support. Second, I would like to thank my family and friends for believing in me, supporting and encouraging me throughout this work. I would also like to thank Dr. Yoav Goldberg for sharing his great advice and Mr. Oren Bernstein for his help with reviewing this work. Finally I would like to thank Dr. Noam Ordan and Prof. Shuly Wintner for their help and feedback on the early stages of this work. This research was funded in part by the Intel Collaborative Research Institute for Computational Intelligence and by the Israeli Ministry of Science. Parts of this work are published in the proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), Baltimore, Maryland, July 2014.

Contents

List of Figures

List of Tables

Abstract

The recent success and proliferation of statistical machine translation (MT) systems raise a number of important questions. Prominent among these is how to automatically estimate the translation quality of such a system in various language pairs and domains, as it is crucial in the ongoing process of developing and training new MT systems. Another important question is how to detect machine translated text in an environment containing both human and MT sentences, as commonly found in web-based textual data. This thesis explores the relation between those two tasks and presents a novel approach for machine translation quality estimation based on the correlation between machine translation detection accuracy and translation quality.

To begin, we define the problem of machine translation (MT) quality estimation (QE). This is the problem of automatically estimating translation quality at the corpus, sentence or word level, without reference translations or any preliminary information on the expected output. Contrast this with the problem of MT evaluation, which relies on such reference translations in order to evaluate the translation quality. MT detection is defined as the problem of automatically recognizing the MT text portions from within a corpus containing both MT and human generated text, mainly at the paragraph or sentence level.

In order to perform MT detection the general features of translated texts are employed, which have been studied widely for many years. Attempts to define their characteristics, often called translation universals, include [38, 6, 2, 15] who showed that the differences between native and translated texts go well beyond systematic translation errors and point to a distinct "translationese" dialect. Other works [5, 26, 21, 25] use text classification techniques in order to distinguish human translated text from native language text at the document or paragraph level, using various linguistic features. Regarding the detection of MT text, Carter and Inkpen [10] conducted detection experiments at the document level, and Arase and Zhou [1] did so at the sentence-level. While previous work has considered MT detection at different levels, the correlation between the quality of the MT text and the ability to detect it has not been studied.

It is hypothesized that the quality of a given MT system can be measured by the accuracy with which a classifer can distinguish between human-generated sentences and sentences generated by that MT system. This work shows that while using common linguistic features, such as frequencies of part-of-speech n-grams and function words, it is possible to train classifiers that distinguish MT text from human-translated or native English text. While this is a straightforward and not entirely novel result, the main contribution of this work is to relativize it. It is shown that the success of such classifiers is strongly correlated with the quality of the underlying MT system. Once a classification experiment is performed, the accuracy of classifying the sentences in the corpus is measured. This accuracy will be shown to decrease as the quality of the underlying MT system increases, by measuring the R^2 correlation coefficient. This correlation is strong enough to propose this accuracy measure as a measure of translation quality.

Various experiments are performed to test the above hypothesis in different settings. First, using several commercial MT systems, the correlation is measured between the ability to detect their output as MT and their translation quality as measured by BLEU [31] or METEOR [12]. It is shown that this correlation is very high, whether or not we use reference translations. In another experiment the same measurement is performed, but with in-house MT systems created using the Moses SMT toolkit. Here again a strong correlation is found between the detection accuracy and the expected translation quality as measured by BLEU or METEOR. The last experiment set measures the correlation between detection accuracy and human quality estimation, which is the gold-standard for measuring translation quality. For this purpose data from the machine translation workshop (WMT) is employed, which is annotated with translation quality according to human judgements. While testing several feature sets, it is shown again that a strong correlation is found between the detection accuracy of the classifier and the expected translation quality, this time based on human evaluation.

This approach has several compelling aspects. First, as an MT evaluation method, it obviates the need for a reference corpus, as is necessary for example for BLEU. This is due to the use of general non-MT sentences, rather than expensive reference translations. This also enables quality estimation on very large data sets and on different domains. This method also helps to find specific issues in the system, as one can examine the sentences that are classified as MT and look for the patterns and features that caused those sentences to be classified as nonhuman. Another benefit of this approach is that it does not require the use of source language sentences, as needed in many MT quality estimation techniques, making it suitable for use with various language pairs without further customization at the source side.

Chapter 1

Introduction

This work focuses on the problem of machine translation (MT) quality estimation (QE). This is the problem of estimating the quality of translated text, whether at the corpus, sentence or word level, without reference translations or any information on the expected output. Contrast this with the problem of MT evaluation, which relies on such reference translations in order to evaluate the MT quality. This work presents a novel method for QE based on first solving another problem related to MT, namely MT detection. In MT detection, the goal is to automatically recognize MT sentences from within a corpus containing both MT sentences and human generated sentences. The problem of MT detection is currently highly relevant, as massive amounts of data are constantly scraped from the web for various use cases, and such distinctions are required for its processing.

If we examine the frameworks of MT quality estimation and MT detection, we notice that they have many features in common. For instance, if we attempt to detect MT sentences in a corpus containing both MT sentences and human-generated sentences, the objective naturally transforms into the detection of incorrect or malformed sentences. These sentences may exhibit several types of issues. Some issues are syntactic, while other sentences exhibit semantic issues: their meaning is absurd or illogical in the relevant context. Clearly, these issues will directly harm the translation quality score for a given MT sentence, as translations that exhibit syntactic or semantic issues are likely to be poor translations, regardless of the source sentence.

This work shows that the quality of a given MT system can be measured by the accuracy with which an MT detection method can distinguish between human-generated sentences and sentences generated by that MT system. First, it is shown that using style-related linguistic features, such as frequencies of part-of-speech n-grams and function words, it is possible to learn classifiers that distinguish machine-translated text from human-translated or native English text. While this is a straightforward and not entirely novel result, the main contribution is to relativize it. It will be shown that the success of such classifiers is strongly correlated with the quality of the underlying MT system. Specifically, given a corpus consisting of both machine-translated English text (English being the target language) and native English text (not necessarily the reference translation of the machine-translated text), the accuracy of the system in classifying the sentences in the corpus as machine-translated or not is measured. This accuracy will be shown to decrease as the quality of the underlying MT system increases. In fact, the correlation is strong enough that it is proposed that this accuracy measure itself can be used as a measure of MT system quality.

This approach has several compelling aspects. First, as an MT evaluation method, it obviates the need for a reference corpus, as is necessary for example for BLEU [31]. This is due to the use of general non-MT sentences, rather than expensive reference translations. This enables quality evaluation on very large test sets and on various domains. This method also helps to find the specific issues in a MT system, as one can look at the sentences that the system classified as MT and examine the patterns and features that caused those sentences to be classified as nonhuman. Another feature of this approach is that it does not require the use of source language sentences, as is needed in many MT quality estimation techniques, making it suitable for use with many language pairs without further customization.

To be able to dive into the details of this work, let us better define the problem, as detailed in figure 1-1. We would like to detect MT sentences which are an output of a specific MT system. To do so, a corpus containing both MT sentences and non-MT sentences

is constructed (The non-MT sentences can be either native human sentences or human-translated sentences). Using this corpus, a feature representation for every sentence is extracted. Using the feature vectors representing the sentences, a classifier is trained to distinguish between MT and non-MT instances. We then measure the accuracy of the classifier on a separate test set or run cross validation, and use the accuracy result as a proxy for the quality of the MT system. In this setting, it is shown that the higher the classification accuracy, the lower the expected translation quality.

The thesis is structured as follows: The next chapter presents relevant previous work in the field. The third chapter describes the details of our approach and methodology. The fourth chapter describes the experiments conducted regarding the detection of machine translation and the use of detection techniques as a QE method. The final chapter offers conclusions and suggestions for future work.

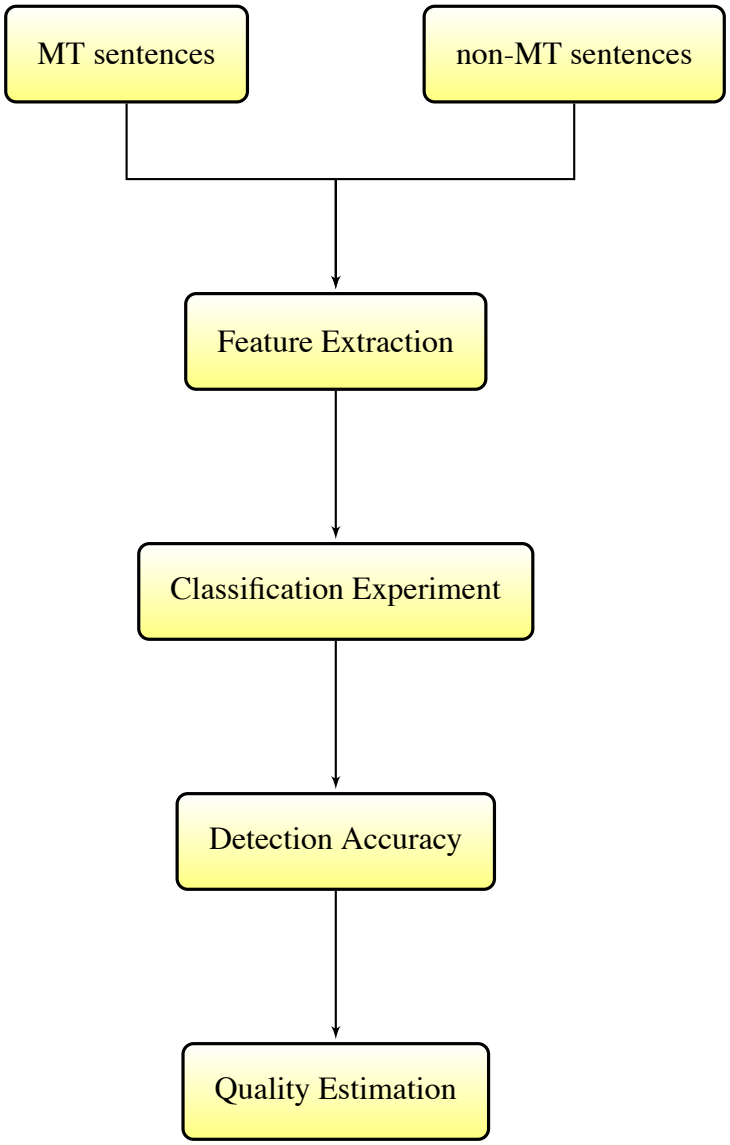


Figure 1-1: Methodological Flow Chart

Chapter 2

Previous Work

2.1 Translationese

The special features of translated texts have been studied widely for many years. Early attempts by Toury [38] to define their characteristics, called "laws of translation" or "translation universals", compared translated texts to native target language texts in order to spot the differentiating characteristics of translations. He proposed the "law of interference", which states that fingerprints of the source language remain in the translation, and the "law of growing standardization", which states that translated sentences tend to follow target language norms unusually closely. Another example of a translation universal is simplification [6]: translated text contains simpler forms of linguistic features than the source text. Gellerstam [16], who introduced the term "translationese", also noted that there are several differences between source and translated texts which aren't necessarily translation errors, but generally characterize translated text. Baker [3] also showed general characteristics of translated corpora, regardless of source language.

Text classification has had many uses in the field of translation studies in recent years, mainly as an empirical method of measuring, proving or contradicting translation universals. Baroni & Brenardini [5] used text categorization techniques in order to distinguish translated text from original text. Specifically, they used a Support Vector Machine (SVM) trained on a collection of articles from an Italian geopolitics journal, and explored a num-

ber of different ways to represent a document (i.e. a journal article) as a feature vector, by varying both the size (unigrams, bigrams, and trigrams) and the type (wordform, lemma, POS tag, mixed) of units encoded as features. In the mixed representation, function words were left in their inflected wordform, whereas content words were replaced by corresponding tags. In the best setting, using a combination of lemma and mixed unigrams, bigrams and POS trigrams, they acquired 86.7% accuracy with 89.3% precision and 83.3% recall on the classification task.

Another case by Van Halteren [40] shows that it is often possible to identify the source language of medium-length speeches in the EUROPARL corpus [23] on the basis of frequency counts of word n-grams. He did so using three classification methods: a marker-based approach, in which n-grams which occurred more often in the training data with a specific source language (SL) than with all other source languages taken together were deemed to be markers for SL; linguistic profiling, in which noticeable over and underuse of specific n-grams were used for classification; and SVM. He examines in detail which positive markers are most powerful for this task, and achieves 87.2%- 96% accuracy depending on the classification method used, thus strengthening the case for the law of interference.

Inspired by Baroni & Bernardini, Kurokawa et al. [26] created a system for automatically detecting whether a piece of text is an original or a translation. Using this system, they detected the directionality of bilingual corpora, and used this information to improve translation performance when training SMT systems. They implemented the detection mechanism as an SVM with a linear kernel, while representing the documents in the same manner as Baroni & Bernardini above (wordform, lemma, POS tag, mixed), gaining 90% accuracy on this task with the Canadian Parliament Hansard corpus. Furthermore, by using the correct direction translated documents as a training corpus for a statistical machine translation system, an improvement of 0.6 BLEU points was achieved.

In order to determine the characteristic features of translationese, Ilisei & Inkpen [21] also trained a system to distinguish between translated and non-translated text, while ex-

amining which features influence the classifiers. They presented a comparison of the classification performance with or without features related to the simplification universal in translation. To do so, they trained an SVM, using general features such as the proportion in texts of grammatical words, nouns, finite verbs, auxiliary verbs, adjectives, adverbs, numerals, pronouns, prepositions, determiners, conjunctions; the proportion of grammatical words to lexical words; and other features, such as average sentence length, parse tree depth, proportion of simple sentences, complex sentences and sentences without any finite verb, ambiguity as measured by the average of senses per word, word length as measured by the proportion of syllables per word, lexical richness, and information load as measured by the proportion of lexical words to tokens. They used several translated corpora in the medical and technical field, resulting a success rate of up to 97.62%.

Investigating other universals using text categorization experiments, Koppel & Ordan [25] show that both interference from a source language spilling over into translation in a source-language-specific way, and general effects of the process of translation that are independent of source language, exist. For the first claim (interference), they used Bayesian logistic regression [30] as the learning method, in order to train a classifier that labels a given text with one of five classes representing the different source languages (Finnish, French, German, Italian and Spanish). To represent the documents, they used a list of 300 function words taken from LIWC [32], wherein each document is represented as a vector of size 300 in which each entry is the frequency of the corresponding word in the document. For the second claim (language independent), they used the same feature space and learning method, but classified the texts as original (O) or translated (T), while training the classifier on a corpus translated from one language, and applying it to a corpus translated from another language. An accuracy of 92.7% was obtained when classifying the source languages of translations from the EUROPARL corpus. Classifying O/T on a cross-language corpus resulted in an accuracy of 96.7%, and similar results were seen on a combination of EUROPARL and articles from the International Herald Tribune and its supplements, translated from Greek, Hebrew and Korean.

A thorough study regarding translation universals by Volansky et al. [41] tested various different hypotheses using supervised machine learning. In practice, they trained SVM's using 32 different linguistically-informed features, and assessed the degree to which different sets of features can distinguish between translated and original texts. They demonstrated that some feature sets are indeed good indicators of translationese, thereby corroborating some hypotheses, whereas others perform much worse (sometimes at chance level), indicating that some assumptions of universality must be reconsidered. The four main translation universals investigated were simplification, explicitation, normalization and interference. For example, the simplification universal was tested using lexical variety, mean word length (in characters), syllable ratio, lexical density, mean sentence length, mean word rank and most frequent words as features for the learning method. Interesting results were obtained, such as contradiction of the simplification assumption that translated text has shorter sentences: in most language pairs used, the sentences in the translated text were actually longer. Another important finding was the fact that interference from source language is by far the strongest, most robust translation universal: use of features relevant to this universal resulted in 85%-100% accuracy in detecting translated sentences.

Popescu [34] presented a set of preliminary experiments which showed that identifying translationese is possible with machine learning methods that work at the character level, specifically methods which use string kernels. For this task, a corpus of literary works was assembled, ranging from the end of the eighteenth century to the beginning of twentieth century: 108 texts which were originally written in English by both American and British authors, while 106 were translated into English - 76 which were originally French and 30 which were German. The classification was done with an SVM using a p-spectrum normalized kernel of length 5. The first experiment performed a cross validation on the entire corpus. The 10-fold cross-validation accuracy was 99.53% and the leave-one-out cross-validation accuracy was 100%, raising suspicions of over-fitting. This was confirmed with a challenging experiment: training on all the French translated text and British original text, while testing the obtained classifier on texts translated from German and American original text, which achieved a poor 45.83% accuracy. To avoid over-fitting, he collected

the French original of all the works of French authors in the corpus, and modified the p-spectrum kernel to exclude all substrings that appear in the reference corpus. Repeating the previous experiment, training on texts from French and British authors and testing on texts from German and American authors with the new kernel, obtained an accuracy of 77.08%.

2.2 Machine Translation Detection

The previous section shows that human translated text can be characterized by special features that makes automatic detection possible for it. This section discusses works that investigate such features within machine translated text. Carter & Inkpen [10] translated the Hansards of the 36th Parliament of Canada using the Bing machine translation service. The resulting corpus was categorized as original and machine translated English and French, labelled hu-e, hu-f, mt-e, and mt-f. They conducted three classification experiments, using SVM as the learning method, and unigrams, average token length, and type-token ratio as features. In the first experiment, training data drawn from the Canadian Hansards was classified using 10-fold cross-validation with LibSVM. An accuracy of 99.89% was achieved overall. To make sure that over-fitting did not occur, a decision tree classifier was trained and examined to verify that there were no extraneous unigrams introduced during data processing that could give the SVM classifier strong hints. The decision tree model appeared to choose common words in both English and French, which suggests that the SVM, despite its strong performance, did not have a trivial task.

In another experiment by Carter & Inkpen [10], the training data drawn from six Government of Canada web sites was classified using 10-fold cross-validation with the previously described setting. The classifier performed well, achieving an average F-measure of 0.98. A final experiment was performed in order to determine whether the somewhat successful Government of Canada prediction models could be applied to find machine translated text on Government of Ontario web sites. The training data drawn from six federal web sites was used to create a model that was tested on Ontario data that was not machine

translated for classification. The results of this last experiment were negative, and demonstrated that the models trained on one set of web sites and its translations are not applicable to web sites in a similar domain with substantially different vocabulary.

Another work in automatic machine translation detection is by Arase & Zhou [1], focusing on the "phrase salad" phenomenon [29] in which each phrase in a sentence is semantically and syntactically correct, but becomes incorrect when combined with other phrases in the sentence. They used three types of features: fluency features, which are language models on both a machine translated corpus and on a human generated corpus, grammaticality features, which are language models of POS tags and function words on both a machine translated corpus and on a human generated corpus, and gappy-phrase features, which are features based on the frequencies of common-word patterns in the text, extracted by the PrefixSpan tool [19]. Using the above features, they trained an SVM classifier to differentiate machine translated text from human generated text in English and Japanese corpora, translated by Google Translate, Bing, and a self trained MT system, resulting in up to 95.8% accuracy at sentence level, using 10-fold cross validation.

While Arase and Zhou [1] considered MT detection at the sentence level, as this work does, they did not study the correlation between the translation quality of the MT text and the ability to detect it. This work shows that such detection is possible with very high accuracy only on low-quality translations.

2.3 Machine Translation Evaluation and Quality Estimation

There are many methods for evaluating the quality of a given machine translation output, which are mainly based on comparing the MT output sentences to human translations of the same source sentences. These methods can be arranged into three families.

The first is the family of precision-based methods. The most prominent method in this family, which became the standard for estimating MT quality in research and industry, is BLEU [31]. This method is based on the assumption that the closer a machine translation is to a professional human translation, the better it is. The main idea in BLEU is to use a weighted average of variable length phrase matches against reference translations, while taking into account the length of the candidate sentences in comparison to the length of the reference sentences.

The NIST metric [13] is based on the BLEU metric, but with some alterations. While BLEU simply calculates n-gram precision, giving equal weight to each one, NIST also calculates how informative a particular n-gram is. That is to say, the rarer an n-gram is, the more weight it is given when it is found to match. For example, if the bigram "on the" correctly matches, it receives lower weight than the correct matching of bigram "interesting calculations", as this is less likely to occur. NIST also differs from BLEU in its calculation of the brevity penalty, insofar as small variations in translation length do not impact the overall score as much.

The second family of MT evaluation methods is the family of F-score based methods. The main method in this family is METEOR [4, 27, 12]. METEOR scores machine translation hypotheses by aligning them to one or more reference translations. Alignments are based on exact, stem, synonym, and paraphrase matches between words and phrases. Segment and system level metric scores are calculated based on the alignments between hypothesis-reference pairs. This metric includes several free parameters that are tuned to emulate various human judgment tasks including WMT ranking and NIST adequacy, and is more complex than BLEU in terms of the required pre-processing per language, which, in turn, may result in a better correlation with human judgements.

The third family of MT evaluation methods is the family of error-rate based methods. The most common metric in this family is the word error rate (WER) metric. This method resembles the Levenshtein distance between the candidate sentence and the reference sen-

tence, except that while the Levenshtein distance works at the character level, WER works at the word level. It was originally used for measuring the performance of speech recognition systems, but has been adopted for use in the evaluation of machine translation. A related metric is the position-independent word error rate (PER), which allows for re-ordering of words and sequences of words between a translated text and a reference translation. Another popular member of this family is the translation edit rate (TER) metric [36], which measures the amount of editing that a human would have to perform to change a system output so it exactly matches a reference translation.

Another relevant topic of increasing interest is the task of translation quality estimation. Different from MT evaluation, quality estimation (QE) systems do not rely on reference translations, but rather estimate the quality of an unseen translated text (document, sentence, phrase) at system run-time. This topic is receiving substantial attention since it was introduced as a shared task in the 7th, 8th and 9th annual Machine Translation Workshops [9, 8, 7].

There are several flavors of tasks in this field. For example, in the 9th annual Machine Translation Workshop [7] there were four different QE tasks: the first was titled "predicting post-editing effort". In this task, the data was labelled with discrete and absolute scores for perceived post-editing effort, according to the following scale: 1, perfect translation, no post-editing needed at all; 2, near miss translation, translation contains maximum of 2-3 errors, and possibly additional errors that can be easily fixed (capitalization, punctuation, etc.); and 3, very low quality translation, cannot be easily fixed. In this task the participants could either label new sentence pairs with one of the three labels, or rank the source-translation pairs according to a ranking of their choice.

The second task was titled "predicting percentage of edits". HTER [36] was used as quality score. This score is the minimum edit distance between the machine translation and its manually post-edited version, and its range is [0, 1] (0 when no edit needs to be made, and 1 when all words need to be edited). Here again, the participants could either label

new sentence pairs with the HTER score, or rank the source-translation pairs according to a ranking of their choice.

The third task was titled "predicting post-editing time". Systems were required to produce for each translation a real valued estimate of the time (in milliseconds) it takes a translator to post-edit the translation. The fourth task was titled "word-level quality estimation". Participants were asked to produce for each token a label that indicates quality at different levels of granularity: binary classification, an OK / bad label, where bad indicates the need for editing the token; level 1 classification, an OK / accuracy / fluency label, specifying coarser level categories of errors for each token, or "OK" for tokens with no error; and multi-class classification, one of the labels specifying the error type for the token (terminology, mistranslation, missing word, etc.).

The most common baseline system in the field is the QUEST system [37], which is used to extract 17 system-independent features from source and translation sentences and parallel corpora. The feature sets it uses are the number of tokens in the source and target sentences, the average source token length, the average number of occurrences of the target word within the target sentence, the number of punctuation marks in the source and target sentences, language model (LM) probability of source and target sentences based on models for the WMT News Commentary corpus, the average number of translations per source word in the sentence as given by IBM Model 1 extracted from the WMT News Commentary parallel corpus, the percentage of unigrams, bigrams and trigrams in frequency quartiles 1 (lower frequency words) and 4 (higher frequency words) in the source language extracted from the WMT News Commentary corpus, and the percentage of unigrams in the source sentence seen in the source side of the WMT News Commentary corpus. These features are used to train a support vector machine (SVM) using a radial basis kernel function from within the SCIKIT-LEARN toolkit. The parameters are optimized via grid search with 5-fold cross validation on the training set. Although the system is referred to as "baseline", it is in fact a strong system and proved as robust across a range of language pairs, MT systems, and text domains for predicting various forms of post-editing effort [9, 8].

As can be seen above, MT evaluation methods require reference translations. Our approach obviates the need for such reference translations which are many times unavailable or hard to achieve. Moreover, it is shown that QE methods use various features which are based on the source (input) and target (output) sentences of a given MT system. Our approach differs from this in that we only use the output sentences of the system, and compare them to random, non-reference sentences, which are easy to obtain. This makes our method independent of the source language involved.

Chapter 3

Our Approach

The following chapter describes the details of our approach. We begin with an overview of the method, followed by a description of the different feature combinations and classification techniques we examined, and conclude with the experimental framework and measurements we used.

3.1 Overview

Our objective is to estimate the quality of a given MT system's output. An MT system's output is a set of sentences produced by translating a set of sentences in one language, the source language (e.g. French), into another language, the target language (e.g. English), using the MT system. The set of translated sentences is called the test set, since it is used to test the quality of the MT system. In this work, a classifier is trained to tell apart the test set sentences from human-generated sentences. This means that in order to use this method, another set of sentences is needed in addition to the test set. This set will contain natural language sentences which are not an output of an MT system. We call this set the human sentence set, and it consists entirely of human-generated sentences in the same language as the test set. It is important to note that the human sentence set does not necessarily consist of reference translations aligned to the test set sentences, but instead contains random natural language sentences which are very easy to acquire. Once those two sets of sentences are obtained, we continue to the next step of our method, feature extraction.

In the feature extraction phase, feature representations are extracted for every sentence in the test set and human sentence set described above. Those features are content-independent linguistic features, which are inspired by the previous work on translationese and MT detection. Once the various features are extracted from every sentence in the data, the most relevant features are selected, as we will describe further. Eventually, a sparse feature vector is created for every sentence in the data using the selected features, which enables us to begin the detection or classification experiment.

In the classification experiment phase, a classifier is trained to detect whether a sentence was generated by an MT system or by a human. This is done in a supervised manner, as labeled data is available: the test set sentences are labeled as MT, and the human sentences as non-MT. With this labeled data, a 10-fold cross-validation classification experiment is performed. In this experiment the data is divided into 10 separate sets, nine sets are used for training a model, and the 10th set is used to test this model; this procedure is repeated 10 times, with each iteration using a different test set, in order to create a diverse experiment and avoid overfitting in the model. Finally, the results of the different iterations are averaged into a single classification accuracy result. With this result in hand, the quality of the translation is estimated. It is hypothesized that a low classification accuracy will signify a high translation quality, as the closer the translations are to human language the harder it will be to the classifier to detect them. When classification accuracy is high, we hypothesize that the quality of the translation will be low, as it is easier for the classifier to tell it apart from the human sentences.

3.2 MT Detection Details

3.2.1 Features

As part of this work, we need to classify sentences as MT or human. To do so, various feature sets are used and tested, inspired by the previous work on translationese and MT

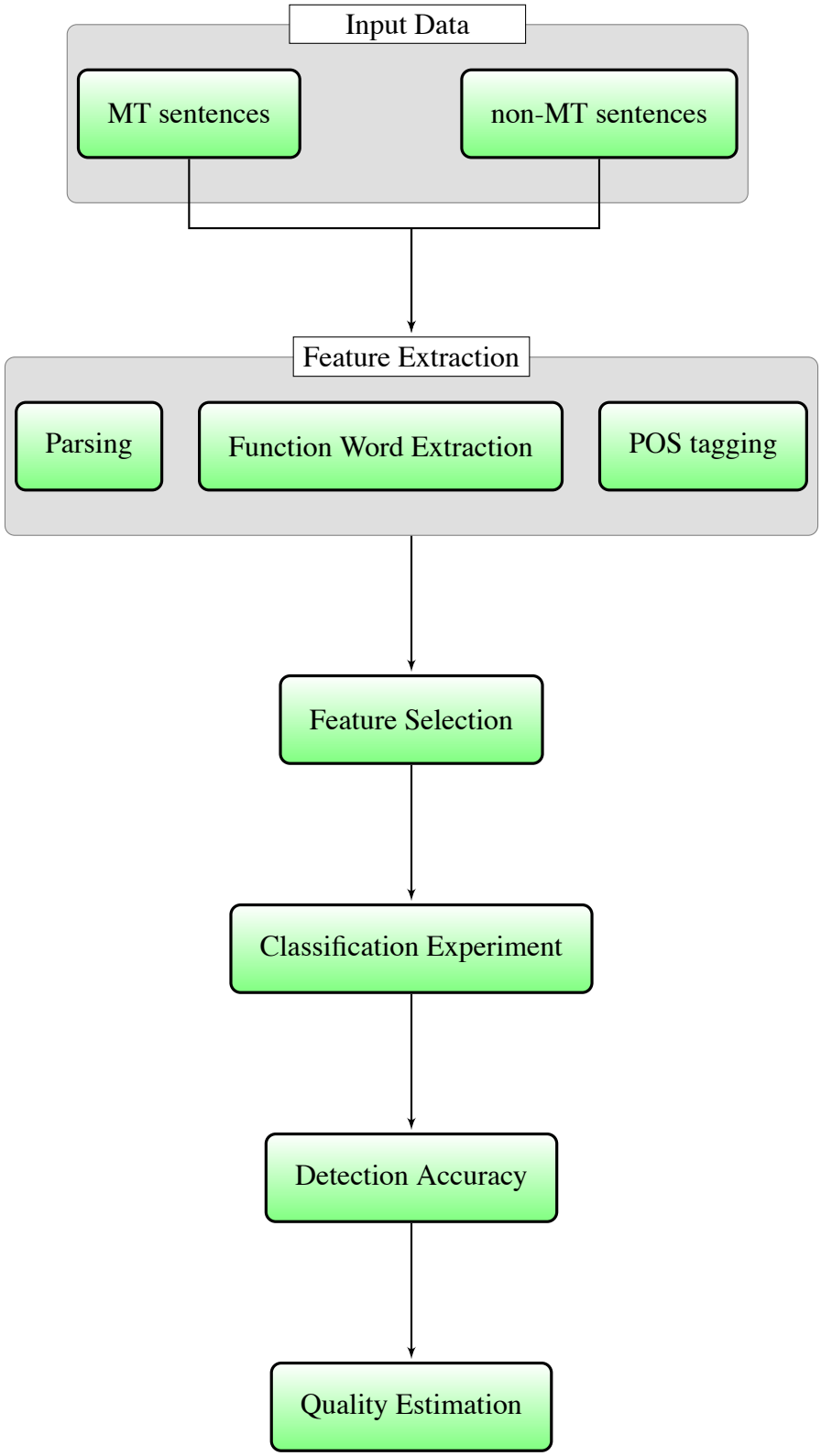


Figure 3-1: Detailed Methodological Flow Chart

detection. The data at the sentence level is very sparse, as there is a relatively small amount of words in a single sentence instance. In order to cope with this sparsity and enable accurate classification, common, content-independent linguistic features are used to model the data. This section will describe these different feature sets and their properties.

Part-of-Speech Tags First, automatic part-of-speech (POS) tags are acquired for each sentence in the data using the Stanford POS tagger [39]. Using these POS tags, a "bag of words"-style vector representation is created, in which each entry represents the presence or absence of a specific POS tag type in the corresponding sentence. Two representations are used: a boolean representation, in which the vector is filled with either 1's or 0's according to the presence or absence of a specific tag, and an implicit count representation, in which each entry in the vector holds the amount of a specific tag-type instances in the sentence.

Function Words Another feature set is based on the presence or absence of each of the 467 function words taken from the LIWC software [32], inspired by the work of Koppel and Ordan [25], which used a similar feature set to classify text chunks as native human text or human-translated text. As in the POS-based feature set, a "bag of words"-style vector representation was used, again with both the boolean representation and the implicit counts representation. The list of function words is detailed in table A.3

Syntactic Features A third feature set is based on the syntactic structure of the sentence. It is created using automatic parse trees tagged by the Berkeley constituency parser [33], from which the one-level context free grammar (CFG) rules are extracted as features. Again, each sentence is represented as a boolean vector in which each entry represents the presence or absence of the CFG rule in the parse-tree of the sentence. Another way to use the syntactic data in a feature representation is to represent only the presence or absence of the different syntactic labels in the sentence parse tree, as described above for POS tags.

Handling Feature Sparsity Since the data is very sparse at the sentence level, a sim-

ple preprocessing procedure is performed on the features in order to improve classification performance. In the chosen approach, only the features that appear at least t times in the entire corpus are considered. A comparison of the classification performance for different sparsity thresholds is detailed in the next chapter, as can be seen in figure 4-9), where it is shown that a low threshold, for example at $t = 5$, gives an inferior classification accuracy when compared to higher thresholds such as $t = 30$ or $t = 50$.

3.2.2 Classifiers

Several classifiers and classifier configurations are used and compared during the experiments in this work. The first is the Naive Bayes classifier, which is considered a baseline given its fast execution and simple approach, followed by Logistic Regression [28], both implemented in the WEKA machine learning toolkit [18]. The support vector machine (SVM) classifier is employed using different implementations. The first is SVM with sequential minimal optimization (SMO) [22], also taken from WEKA. Additional SVM implementations are LibSVM [11] and LibLinear [14], which utilize a linear kernel function. A comparison of the classifiers is described in the next chapter, measuring their performance in the task of classifying whether a sentence is an MT sentence or a human generated sentence. The results are shown in figure B-5 and in table 4.2.

3.3 From Classification to Quality Estimation

As the final goal of this work is to infer a quality estimation score for an MT system, there is a need to convert a classification result in an MT detection experiment into a translation quality estimation score. This section will focus on this step.

The classification accuracy itself can be measured in several ways. The data can be divided into a training set which will be used to train a model and a test set over which a classification experiment will be performed using the learned model. This experiment is where the classification accuracy will be measured. Another option is to perform a k-fold

cross validation experiment and measure the average classification accuracy.

Once the classification experiment is conducted and an accuracy score is measured, we propose it to be used as a proxy for translation quality. This is performed by assuming that lower classification accuracy implies higher translation quality. In this scenario, when a low classification accuracy is obtained, it is harder for the classifier to tell apart the MT sentences from the human sentences, implying that the MT system produces high quality translations that have much in common with human generated sentences. Conversely, high classification accuracy implies the classifier did not confuse MT sentences to be human sentences, meaning that the underlying MT system produces low quality sentences which are very different in their properties from fluent, human generated sentences.

As can be seen, this method does not take the source sentences nor the reference translations into account when evaluating a test set, yet it is still meant to estimate the translation quality. This is explained by the fact that it measures the fluency of the MT sentences, and that there is a strong correlation between the fluency and the overall translation quality, given that the sentences are MT output. The fluency factor is indeed emphasized by this method, since the features used to model a sentence are linguistic, content-independent features, which are designed specifically to capture fluency and non-fluency phenomena. This ensures that non-fluent MT sentences will be classified as MT and fluent high quality MT sentences will be misclassified as human sentences, resulting in a proper quality estimation according to this approach.

In order to measure the accuracy of this approach as a quality estimation method, simple linear regression is obtained over the values of the detection accuracy and the values of the expected translation quality, as measured by BLEU, METEOR or by human judgements. This is done in order to measure the R^2 value, which is a statistical measure of how close the data are to the fitted regression line. This score is also known as the coefficient of determination, as it is a square of the coefficient of multiple correlation or the Pearson correlation coefficient between the predicted (in our case, the detection accuracy score) and

the actual (in our case, the expected translation quality) values in a linear regression model that includes an intercept. The mathematical formula for computing R^2 is:

$$R^2 = \left(\frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}} \right)^2 \tag{3.1}$$

With x denoting the values of the first variable, y denoting the values of the second variable, and n being the amount of (x, y) pairs in the data. The R^2 value is measured in every experiment in this work which compares the detection accuracy to the expected translation quality. This is made in order to confirm our hypothesis regarding the strong correlation between the two.

Chapter 4

Experiments

Various experiments are conducted in order to test our hypotheses in different settings. First, a set of preliminary experiments is made in order to fine-tune the different parameters of our method, such as classifier selection and threshold values for feature selection. The second experiment set tests the ability to use our method in order to estimate the quality of commercial MT systems by comparing its results to those of BLEU [31] and METEOR [27]. The third experiment set does the same with in-house MT systems created using the Moses MT toolkit [24]. The fourth experiment set tests our method as a QE technique in comparison with human quality estimation, using the WMT13' [8] annotated corpus.

4.1 Data Sets

The experiments in this work utilize several data sets of two kinds: the first represents the output of an MT system, and the second represents fluent human generated text. In order to learn the features of poorly or properly translated text, this variety of data sets is used to investigate the general performance of our approach under different parameters and domains. This section will describe the different data sets used in this work and their special properties.

4.1.1 The Commercial MT Systems corpus

The first corpus presented here contains outputs from 6 different commercial MT systems. This data is treated as black box data, since the technical details behind those MT systems (training data, algorithms, etc.) are not disclosed. In order to build this corpus a portion of the Canadian Hansard corpus [17] is used, containing 48,914 parallel sentences translated manually from French to English by professional translators. The French portion of the corpus is then translated using a set of commercial MT systems: Google Translate, Systran, ProMT, Linguatec, Skycode and Trident. The latter five are available at the website http://itranslate4.eu, through which example MT systems built by several european MT companies can be queried in various language pairs. This results in 8 sets of 48,914 sentences each, all parallel to each other, with the first being the native French source sentences, the second being fluent English translation, and the other six being English outputs from each of the black box MT systems for the same source sentences. This data set enables us to explore the difference between MT and human text, and compare these differences with several independent MT systems. It is important to note that for this corpus there are no human quality estimation scores available, so automatic evaluation methods such as BLEU and METEOR are used. Example translations from those systems for the same source sentence are shown in table 4.1.

4.1.2 The In-House MT Systems corpus

As mentioned above, the first corpus contains black box data, meaning the exact details of each MT system are not known to us. In order to explore the effectiveness of our method with respect to the exact algorithms that are used, the exact training data, model parameters etc., seven different statistical machine translation systems are created, while changing several parameters for each system in order to create various translation quality levels. For this task the Moses statistical machine translation toolkit [24] is used. Specifically, these MT systems use a phrase-based translation model and a 5-gram language model using the KenLM toolkit [20]. For parallel training data, a portion of the Europarl corpus [23] is used. In order to create differing quality levels, different amounts of training data are used

MT Engine	Example
Google Translate	"**These** days, **all but one were** subject **to a** vote, **and all had a** direct link **to the** post September 11th."
Moses	"**these** days , **except one were the** subject **of a** vote , **and all had a** direct link **with the after** 11 September ."
Systran	"**From these** days, **all except one were the** object **of a** vote, **and all were** connected **a** direct link **with after** September 11th."
Linguatec	"**Of these** days, **all except one were** making **the** object **of a** vote **and all had a** straightforward tie **with after** September 11."
ProMT	"**These** days, **very** safe **one all** made object **a** vote, **and had a** direct link **with after** September 11th."
Trident	"**From these all** days, **except one** operated object voting, **and all had a** direct rope **with after** 11 septembre."
Skycode	"**In these** days, **all** safe **one** made **the** object **in a** vote **and all had a** direct connection **with him after** 11 **of** September."

Table 4.1: Outputs from several MT systems for the same source sentence (function words marked in bold)

for both the translation model and the language model, as described in table 4.6. Once the seven tailor-made MT systems are ready, each of them is used to translate 20,000 sentences from the Hansard corpus, resulting in 8 sets of 20,000 sentences each, with the first being fluent English, and the other seven being statistical machine translation output with different levels of quality. Here again, there are no human quality estimation scores available for the translations, so automatic evaluation methods such as BLEU and METEOR are used in order to measure the translation quality.

4.1.3 The WMT corpus

While in the case of the in-house MT systems corpus there is absolute control over the details of each MT system, it still lacks of human evaluation for the quality of the translations, which is very useful when evaluating a new quality estimation technique. In order to use such data in our experimental framework, the publicly available data from the annual machine translation workshop (WMT) is used. Since the workshop holds a machine translation competition based on that data, it is evaluated using human judgements. Specifically, the French-English data from the 8th Workshop on Statistical Machine Translation

(WMT13') [8] is used, containing outputs from 13 different MT systems and their human evaluations, all in the news domain. This human evaluation score is based on bootstrap resampling, in which a set of pairwise rankings is randomly sampled from human pairwise rankings (allowing for multiple drawings of the same pairwise ranking). This set is then used to compute the expected wins score and the rank of each system. By repeating this procedure 1000 times, one can determine a range of ranks into which each system falls in at least 95% of the time (i.e., at least 950 times) corresponding to a p-level of $p \leq 0.05$. This corpus contains 3000 output sentences from each of the 13 MT systems presented in the workshop, along with their reference translations and rank, computed as described above. In order to have non-reference human generated sentences suitable for this corpus, 3000 sentences are kept from the newstest 2011-2012 corpora, published in WMT12' [9].

4.2 Preliminary Experiments

4.2.1 Classifier Selection

Several preliminary experiments are conducted in order to tune the various parameters of our method. The first is performed to compare and choose the most appropriate classifier for the task out of several candidates, detailed in section 3.2.2. For this comparison, the WMT corpus is used, containing 3000 sentences from each of the 13 MT systems and another 3000 human non-reference sentences from the newstest corpus. The CFG based feature set is extracted from that dataset as described in section 3.2.1 in order to conduct the classification experiment several times, each time with a different classifier, selected from among the Naive Bayes, LibSVM, SMO, LibLinear and Logistic Regression classifier implementations, as detailed in section 3.2.2. The results are described in table 4.2. As can be seen in the results, SMO, LibLinear and Logistic Regression are better than the rest in terms of classification accuracy, with LibLinear being also the fastest method to train and test. Concluding from the results, the LibLinear SVM implementation is primarily used for the following experiments.

MT System	Naive Bayes	LibSVM	SMO	LibLinear	Logistic Regression
uedin-heafield	66.82	71.99	73.62	73.96	**74.14**
uedin	66.84	71.96	**73.87**	73.66	73.82
online-b	66.46	71.47	73.59	**73.87**	73.69
limsi-soul	66.67	71.44	73.29	**73.47**	73.19
kit	66.62	71.54	**73.89**	73.51	73.52
online-a	66.37	71.86	73.62	**73.99**	73.79
mes-simplified	67.14	72.29	73.62	73.79	**73.89**
dcu	67.19	71.54	**73.96**	73.86	73.91
rwth	67.54	72.01	74.04	74.11	**74.22**
cmu-t2t	67.02	72.96	**74.57**	74.52	74.64
cu-zeman	67.76	73.29	74.87	**75.02**	74.92
JHU	66.96	71.37	73.27	**73.79**	73.67
SHEF	69.47	74.34	76.39	76.67	**76.84**

Table 4.2: Classifier comparison, comparing detection accuracy on systems from WMT13'

4.2.2 Sparsity Threshold

The second preliminary experiment is conducted in order to determine the sparsity threshold to use for feature selection in the classification experiments, meaning the value t, where each feature is used only if it appeared at least t times in the data. In order to find the appropriate value a classification experiment is conducted, again with the WMT data and CFG based features but with different sparsity thresholds, ranging from $t = 5$ up to $t = 50$. The results are shown in table 4.3. As can be seen in the results, a threshold of $t = 30$ gives the best results; for values from $t = 30$ up to $t = 50$ we see a convergence to similar detection accuracy values for most systems, and a minor decrease for few. Therefore, the value $t = 30$ is used for the remaining experiments.

4.3 Commercial MT Systems

4.3.1 Detection Experiments

The second experiment set explores the ability to detect outputs of MT text from different black-box MT systems, in an environment containing both human generated (reference and non-reference) and machine translated text, using the linguistic content-independent

MT System	$t = 5$	$t = 10$	$t = 20$	$t = 30$	$t = 50$
uedin-heafield	59.97	60.96	61.89	62.21	**62.29**
uedin	60.34	61.67	62.74	63.07	**63.54**
online-b	61.84	62.49	63.34	63.76	**64.06**
limsi-soul	61.69	63.21	63.72	63.76	**63.94**
kit	61.17	62.46	63.19	63.41	**63.76**
online-a	62.11	63.26	64.06	**64.67**	64.61
mes-simplified	62.59	63.81	64.89	**65.01**	64.87
dcu	61.74	62.79	63.51	64.06	**64.16**
rwth	62.82	64.09	65.29	**65.36**	65.11
cmu-t2t	63.69	64.72	65.11	65.39	**65.44**
cu-zeman	65.71	66.84	67.71	67.91	**68.11**
JHU	64.34	65.24	66.09	66.19	**66.21**
SHEF	64.42	66.02	66.97	**67.12**	66.92

Table 4.3: Sparsity threshold comparison, comparing detection accuracy on systems from WMT13'

features proposed previously in this work. For this purpose, the commercial MT systems corpus is used, containing 48,914 parallel sentences from French to English which are also translated using the commercial MT systems as described in section 4.1. From this corpus, 20,000 sentences are taken from each system output, and a detection experiment is conducted by labeling these sentences as MT sentences, and another 20,000 sentences, which are the human reference translations, as reference sentences. This creates 7 sets of 40,000 sentences per experiment, one experiment per MT system. A 10-fold cross-validation classification experiment is performed over every one of the 7 sets. These experiments are then performed once more, this time using 20,000 random, non-reference sentences from the Hansard corpus instead of the reference sentences. The detection accuracy results of those experiments are shown in Table 4.4.

As can be seen in the results, MT sentences can be distinguished from human sentences with very high accuracy while using our method - up to 89.36% when running a 10-fold cross-validation experiment. It is also clear that the highest detection accuracy values are obtained when using the combined feature set, with both function word and POS tag based features. Another finding is that it is easier to distinguish machine-translated sentences from non-reference sentences than from the reference sentences set, as might be expected. This is explained by the fact that the data in the reference sentences is far more similar to

Features	Data	Google	Systran	ProMT	Linguatec	Skycode	Trident
mixed	MT + non-ref	**63.34**	**72.36**	**78.2**	**79.57**	**80.9**	**89.36**
mixed	MT + ref	59.51	69.77	75.86	78.11	79.24	88.85
func. w.	MT + non-ref	60.43	69.87	69.78	71.38	75.46	84.97
func. w.	MT + ref	57.27	67.48	67.06	68.58	73.37	84.79
POS	MT + non-ref	60.32	66.61	73	73.9	74.33	79.6
POS	MT + ref	57.21	64.12	70.29	73.06	73.04	78.84

Table 4.4: Classifier detection accuracy over each of the commercial MT system outputs

the MT sentences in its structure and content than the non-reference data, hence making it harder for the classifier to distinguish between the two classes of sentences.

4.3.2 Correlation with Translation Quality

It is hypothesized that using this approach it is possible to detect the MT sentences with high accuracy when the translation quality is low, and with low accuracy when it is high. We measure the translation quality of the commercial MT systems with BLEU and METEOR to examine this hypothesis. This will be the first step towards ranking the translation quality of MT systems using our method, without using reference sentences or source sentences in the process.

In order to measure the correlation between the detection accuracy values and the expected quality as measured by BLEU or METEOR, an R^2 value is obtained using a simple linear regression over the measurements of detection accuracy and BLEU or METEOR score for every MT system, as described in section 3.3. This is performed for each of the three feature set combinations (function words, POS tags and mixed) and the two data combinations (MT vs. reference and MT vs. non reference sentences). The R^2 values measuring the correlation of the detection accuracy with BLEU or METEOR are shown in table 4.5 and the general correlation is detailed in figures 4-1 and 4-2.

Figure 4-1 shows the correlation between the observed detection accuracy for each system, ranging from 57.21% up to 89.36%, and the BLEU score of that system, ranging from 8.39 BLEU points up to 36 BLEU points. Figure 4-2 does similarly with the METEOR

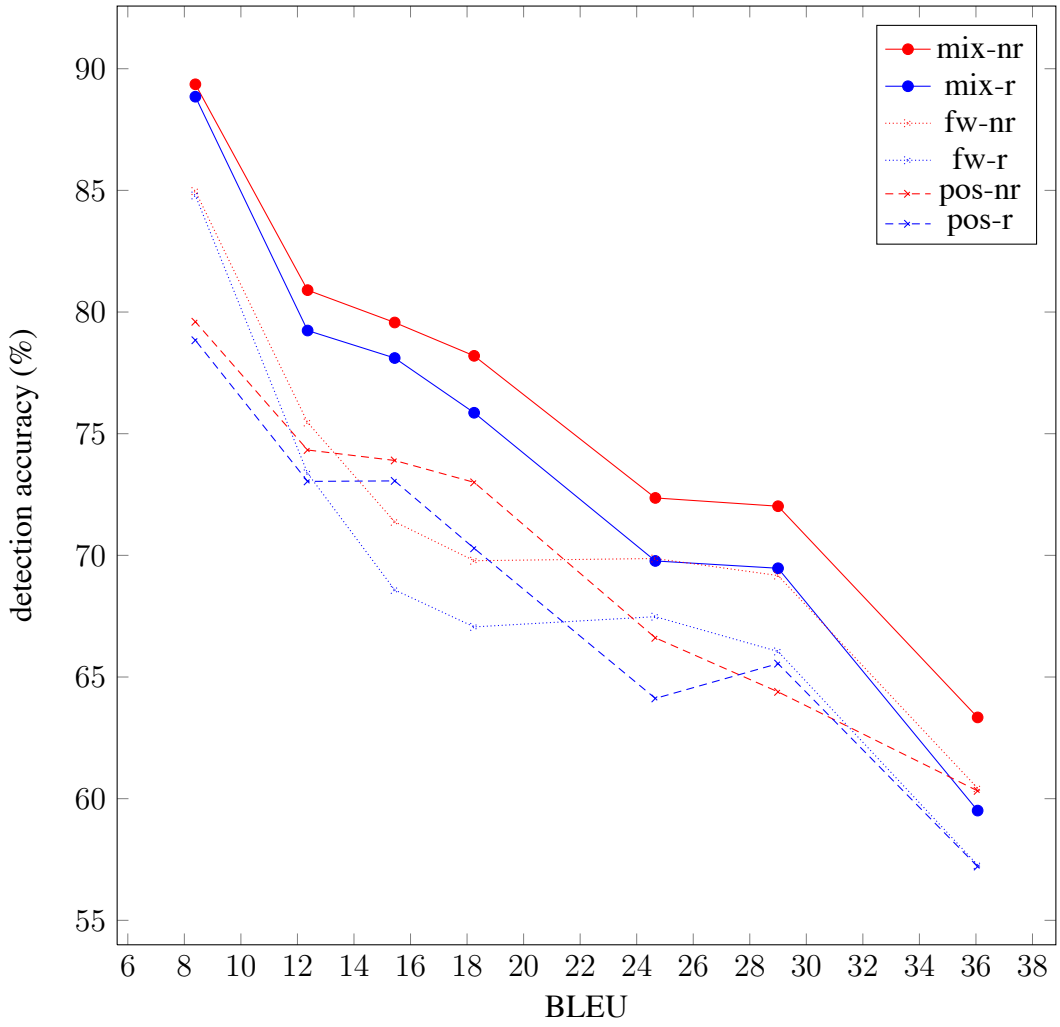


Figure 4-1: Correlation between detection accuracy and BLEU score on commercial MT systems, using POS, function words and mixed features against reference and non-reference sentences.

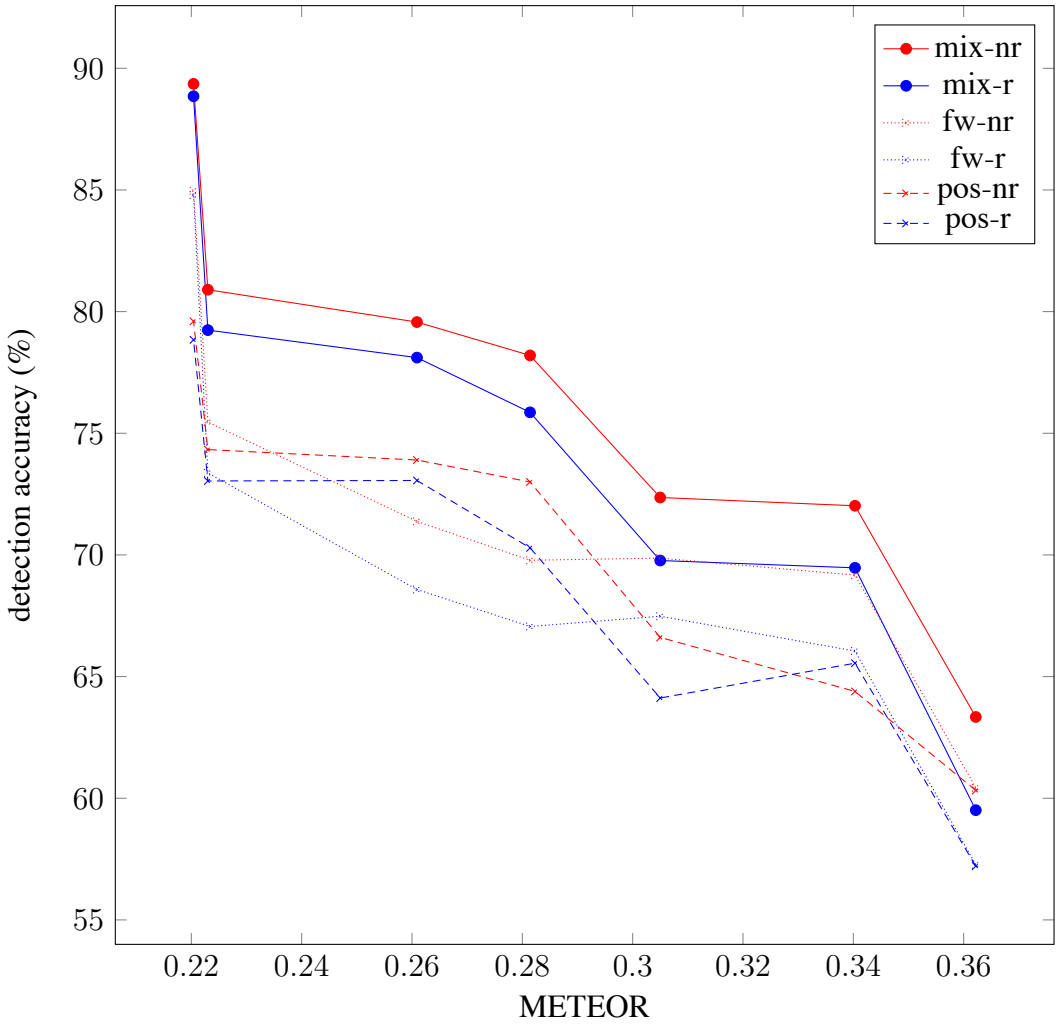


Figure 4-2: Correlation between detection accuracy and METEOR score on commercial MT systems, using POS, function words and mixed features against reference and non-reference sentences.

Features	Data	R^2 w. BLEU	R^2 w. METEOR
mixed	MT + non-ref	0.946	0.871
mixed	MT + ref	0.944	0.869
POS	MT + non-ref	**0.978**	**0.917**
POS	MT + ref	0.948	0.867
func. w.	MT + non-ref	0.798	0.755
func. w.	MT + ref	0.779	0.742

Table 4.5: R^2 values measuring the correlation of the detection accuracy with the BLEU & METEOR scores measured for each of the commercial MT system outputs

scores for those systems, ranging from 0.2204 up to 0.3622. As is evident, regardless of the feature set or non-MT sentences used, the correlation between detection accuracy and BLEU or METEOR scores is very high, as we can also see from the R^2 values in Table 4.5: from 0.779 up to 0.978 for BLEU and from 0.742 up to 0.917 for METEOR. The highest correlation with BLEU at $R^2 = 0.978$ was obtained with POS based features and non-reference sentences, and the same goes for correlation with METEOR at $R^2 = 0.9179$. This confirms the hypothesis regarding the strong correlation between the detection accuracy and the translation quality (as measured here by the automatic evaluation methods) while using our approach over the commercial MT systems output.

4.4 In-House SMT Systems

4.4.1 Detection Experiments

The third experiment set tests our approach with SMT systems created in-house, for which there is control over the internal details and expected overall relative translation quality. This will help to test the hypothesis regarding the correlation between detection accuracy and translation quality, this time in a more general and robust way, as the systems are not black-box as in the previous experiment, but rather controlled in terms of implementation, training data, configuration, and expected translation quality.

In order to do so, the Moses statistical machine translation toolkit [24] is used in to

create the MT systems, as detailed in section 4.1 and in table 4.6. For purposes of classification, similar content-independent features are used as in the previous experiment, based on function words and POS tags, again with SMO-based SVM as the classifier. For data, 20,000 random, non-reference sentences are taken from the Hansard corpus as human sentences, against 20,000 sentences from one MT system per experiment, again resulting in 40,000 sentence instances per experiment. The results of the detection experiments are detailed in table 4.6.

	Parallel	Monolingual	BLEU	METEOR	Detection Accuracy
SMT-1	2000k	2000k	28.54	**0.3485**	**73.10**
SMT-2	1000k	1000k	27.76	0.3434	73.90
SMT-3	500k	500k	**29.18**	0.346	72.33
SMT-4	100k	100k	23.83	0.3219	73.59
SMT-5	50k	50k	24.34	0.3204	74.12
SMT-6	25k	25k	22.46	0.3093	74.78
SMT-7	10k	10k	20.72	0.294	75.98

Table 4.6: Details for in-house Moses based SMT systems

As can be seen in the detection accuracy results, it is also possible to detect machine translated text coming from in-house MT systems, for which we know the general details and implementation. The accuracy ranges between 73.10% and 75.98%, which is a narrower range than the one we saw with the commercial MT systems data. It is also interesting to see the disagreement between BLEU and METEOR regarding the quality of the top-ranked systems.

4.4.2 Correlation with Translation Quality

It interesting to see how the correlation between detection accuracy and translation quality holds in this case, given the narrow detection accuracy range. The relationship between the detection results for each MT system and the BLEU score for that system is shown in figure 4-3, and the same is shown with the METEOR scores for those systems in figure 4-4.

As can be seen in the results in figure 4-3, the correlation between the detection accuracy and the translation quality as measured by BLEU resulted in $R^2 = 0.774$, with

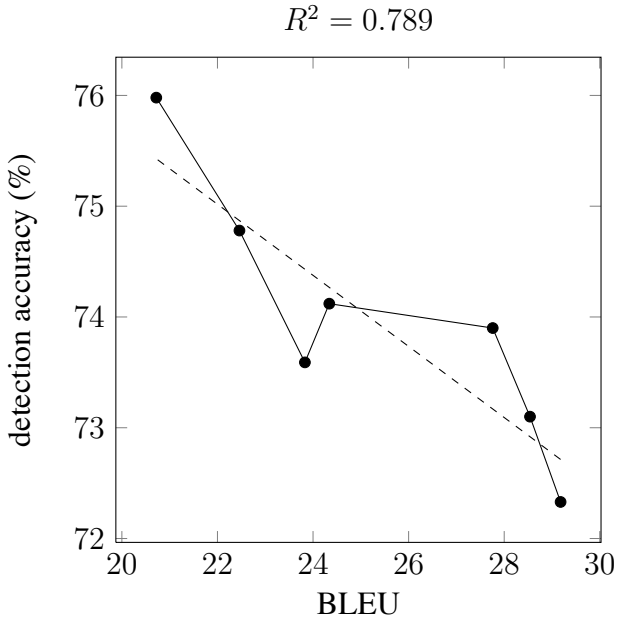


Figure 4-3: Correlation between detection accuracy and BLEU score on in-house Moses-based SMT systems against non-reference sentences using content independent features.

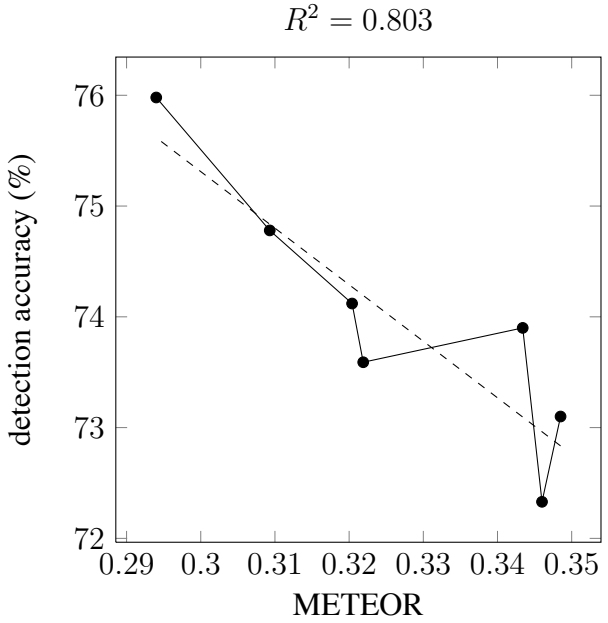


Figure 4-4: Correlation between detection accuracy and METEOR score on in-house Moses-based SMT systems against non-reference sentences using content independent features.

only one outlier breaking the monotony of the correlation. As expected, this correlation is present even though the result range is much narrower then in the previous experiment, with BLEU scores ranging from 20.72 to 29.18, and detection accuracy ranging from 72.33 up to 75.98. When repeating a similar measurement with METEOR, again we find an even higher correlation between the detection accuracy and the expected translation quality, this time resulting in $R^2 = 0.803$. This again strengthens the hypothesis regarding the ability to detect the MT sentences with high accuracy when the translation quality is low, and with low accuracy when the translation quality is high, allowing us to offer this method as a translation quality estimation technique.

4.5 Human Evaluation Experiments

4.5.1 Detection Experiments

As can be seen in the above experiments, there is a strong correlation between the BLEU or METEOR scores and the MT detection accuracy of our method. In fact, results are linearly and negatively correlated with those scores, as can be seen both with commercial systems and with in-house SMT systems. We also wish to consider the relationship between detection accuracy and a human quality estimation score, as it is the gold standard when measuring translation quality. We hypothesize that a similar correlation will be found between detection accuracy and human quality estimation scores, which will validate our approach as a well-functioning quality estimation technique as we propose.

To test this hypothesis, the French-English data from the 8th Workshop on Statistical Machine Translation (WMT13') [8] is used, containing outputs from 13 different MT systems and their human evaluations, as detailed in section 4.1. The same classification experiment is performed as before, with features based on function words and POS tags, and SMO-based SVM as the classifier. The data set consists of 3000 reference sentences taken from WMT13' against the matching 3000 output sentences from one MT system at a time, resulting in 6000 sentence instances per experiment. In the second phase of the experiment,

3000 random, non-reference sentences are taken from the newstest 2011-2012 corpora published in WMT12' [9] against 3000 output sentences from one MT system at a time, again resulting in 6000 sentence instances per experiment. It is important to notice that the data in this experiment set is much smaller than the data in the previous experiments, with only 3000 samples per class in comparison to 20,000 samples per class previously. The detection accuracy results are shown in table 4.7.

MT System	Det. Acc. - Reference	Det. Acc. - Non Reference
uedin-heafield	58.58	74.05
uedin	58.61	73.56
online-b	**57.63**	**73.36**
limsi-soul	58.78	73.63
kit	59.38	73.83
online-a	58.63	74.10
mes-simplified	59.86	74.36
dcu	58.53	73.80
rwth	60.46	74.55
cmu-t2t	61.65	75.10
cu-zeman	62.66	75.65
JHU	60.83	73.71
SHEF	64.10	76.68

Table 4.7: Detection accuracy results on systems from WMT13', using reference and non-reference sentences as human data

As can be seen in the results in table 4.7, it is quite hard for the classifier to distinguish the reference sentences from the MT sentences in this case, as shown by the relatively low detection accuracy ranging in 57.63 up to 64.1. As for the non-reference sentences, the detection accuracy is higher as we saw in the previous experiment sets but narrower in range, with accuracies from 73.36 up to 76.68. Those low detection accuracy scores may also be a result of the significantly smaller dataset in this experiment.

4.5.2 Correlation with Translation Quality

Despite the relatively low detection accuracy results, our goal is to measure the correlation between the detection accuracy and the translation quality of the MT systems. In order to

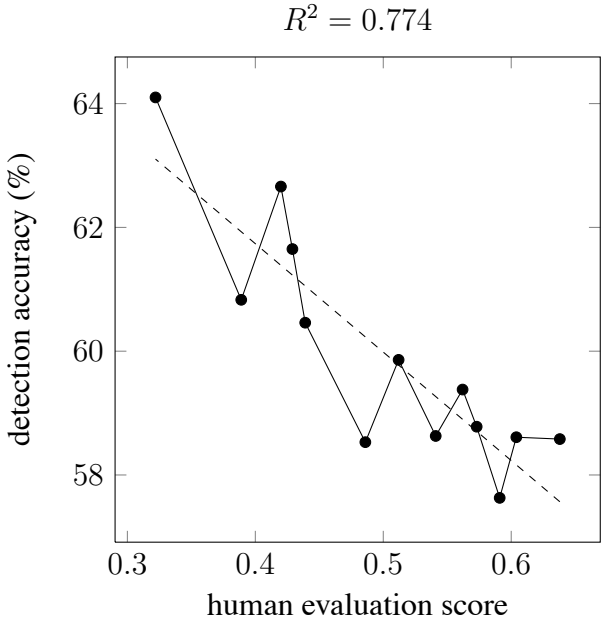


Figure 4-5: Correlation between detection accuracy and human evaluation scores on systems from WMT13' against reference sentences.

do so we calculate the R^2 value as described in section 3.3, this time against the human evaluation scores which are measured as described in section 4.1. The results are shown in figure 4-5 for the reference sentences and in figure 4-6 for the non-reference sentences.

As can be seen in the results in Figure 4-5, the detection accuracy is strongly correlated with the human evaluation scores, yielding $R^2 = 0.774$ when using the reference sentences, despite the low detection accuracy values. To provide another measure of correlation, every pair of data points in the experiment is compared in order to get the proportion of pairs ordered identically by the human judgements and by our method, with a result of 0.846 (66 of 78). Regarding the non-reference data, while applying the same classification method as with the reference sentences, the detection accuracy rises, while the correlation with the translation quality yields $R^2 = 0.556$, as can be seen in Figure 4-6. Here, the proportion of identically ordered pairs is 0.782 (61 of 78). The measurement of the correlation between detection accuracy and quality estimation resulting in $R^2 = 0.556$ does not sit well with our hypothesis, so we seek further for an explanation in the next experiment.

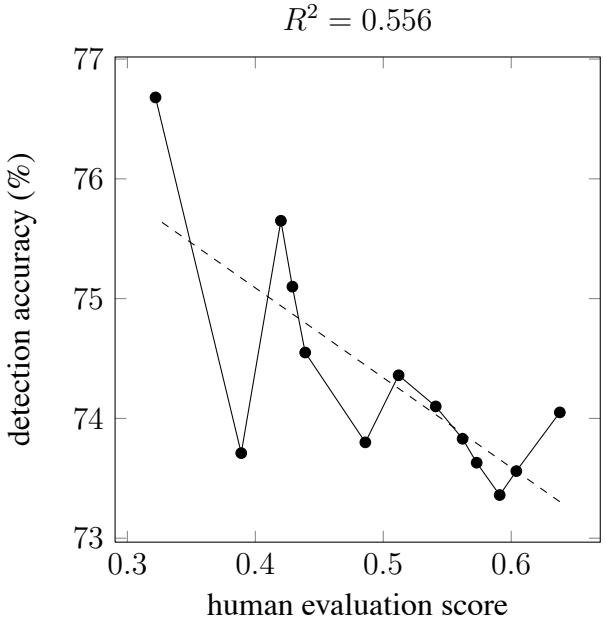


Figure 4-6: Correlation between detection accuracy and human evaluation scores on systems from WMT 13' against non-reference sentences.

4.5.3 Using Syntactic Knowledge

It can be seen that the second leftmost point in Figures 4-5 and 4-6 is an outlier: that is, our method has a hard time detecting sentences produced by this system, although it is not highly ranked by human evaluators. This point represents the Joshua [35] SMT system. This system is syntax-based, which apparently confounds our POS and function word-based classifier, despite its low human evaluation score. It is hypothesized that the use of syntax-based features within our method might improve the result.

To verify this intuition, the same data is taken from WMT13', and parse trees are created using the Berkeley parser [33] for every sentence in the data. The one-level CFG rules are then extracted as features as detailed in section 3.2.1, and a boolean vector is created for every sentence, in which each entry represents the presence or absence of the CFG rule in the parse-tree of the sentence. These features alone are used in this experiment in order

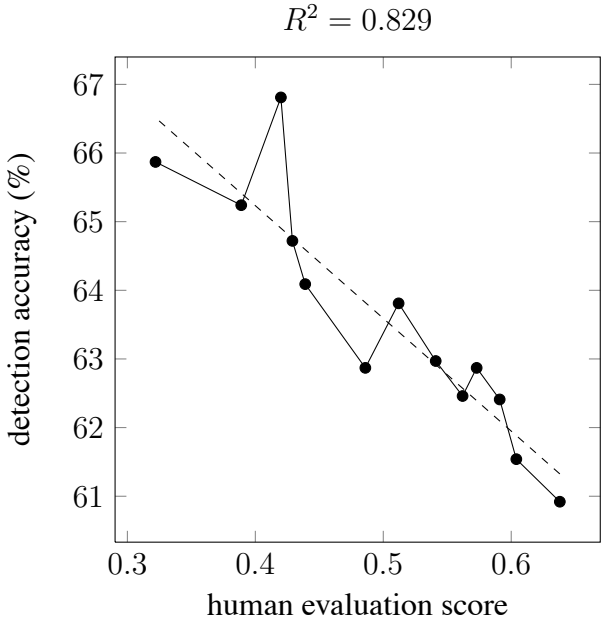


Figure 4-7: Correlation between detection accuracy and human evaluation scores on systems from WMT 13' against non-reference sentences, using the syntactic CFG features described in section 4.2

to prevent the outlier observed in the previous experiment. The results for this experiment are shown in figure 4-7

Using this setup, an $R^2 = 0.829$ is obtained, and the proportion of identically ordered pairs is 0.923 (72 of 78), as shown in Figure 4-7. The increased correlation score is indeed a direct result of the absence of the outlier we saw in the previous experiment. With such high correlation in hand, we verify that our hypothesis regarding the strong correlation between detection accuracy and human evaluation of the translation quality holds as expected.

We further explore the performance of this feature set while tuning the different parameters of our method, such as classifier type as shown in figure 4-8 and sparsity threshold as shown in figure 4-9, in attempt to increase the correlation of the detection accuracy and the translation quality. Those attempts mainly affected the detection accuracy but did not improve the correlation score.

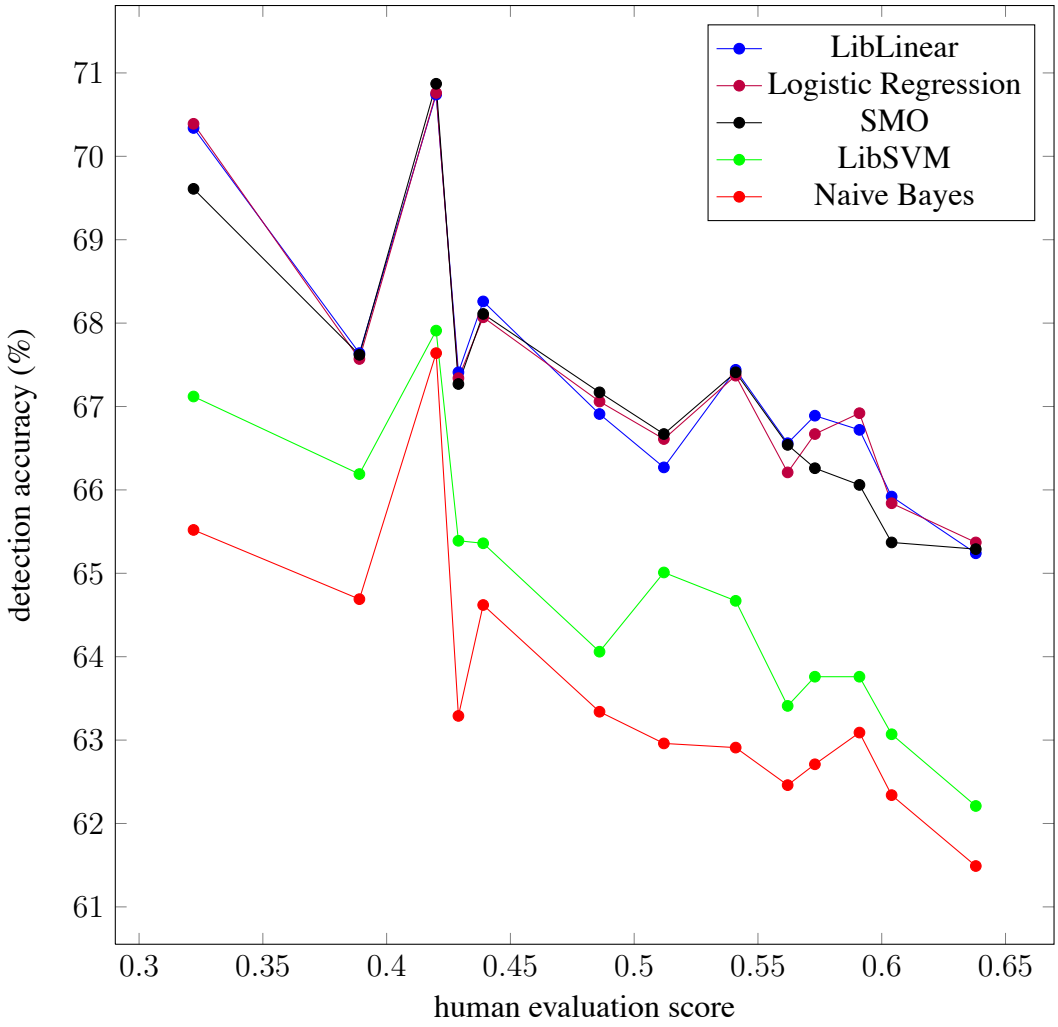


Figure 4-8: Classifier comparison, measuring correlation between detection accuracy and human evaluation scores on systems from WMT13' against non-reference sentences from WMT12', with CFG rules as features and sparsity threshold set at 30.

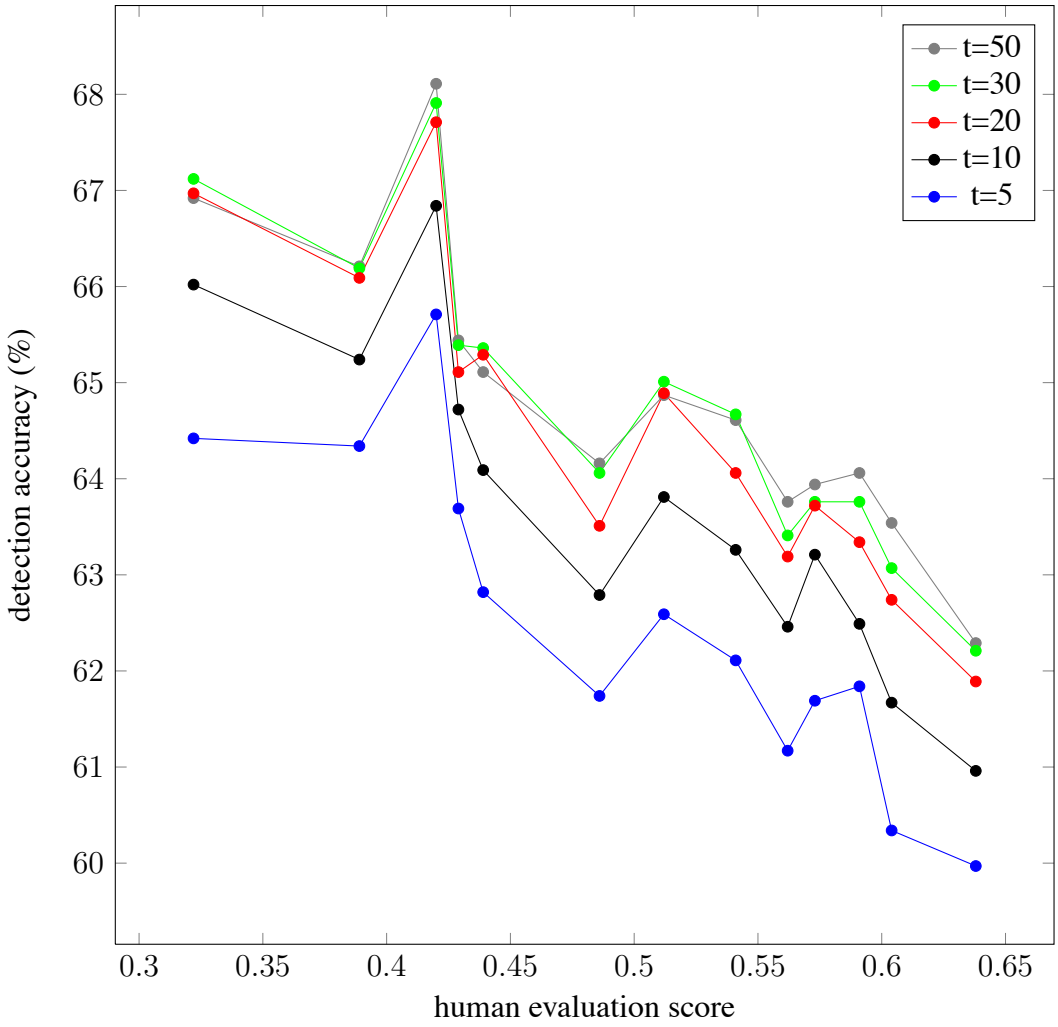


Figure 4-9: Sparsity threshold comparison, measuring correlation between detection accuracy and human evaluation scores on systems from WMT13' against non-reference sentences from WMT12', using the LibSVM classifier and CFG rule features

Chapter 5

Discussion and Future Work

This work has shown that it is possible to detect machine translation from monolingual corpora containing both machine translated text and natural language human generated text. This detection is performed at the sentence level, using a text classification approach based on content-independent linguistic features, many times inspired by the features used in the detection of so called "translationese".

It is also shown that there is a strong correlation between the detection accuracy that can be obtained when detecting a specific MT system output and the BLEU, METEOR, or the human evaluation score of the MT system itself, whether it is a black-box MT system whose details are not familiar to us, an SMT system which we create ourselves, or one of the state-of-the-art systems which participate in a research workshop such as WMT.

An important feature of the result is that this correlation holds whether or not a reference set is used as the human generated data. This suggests that our method can be used as a quality estimation method when no reference sentences are available, such as for resource-poor source languages and for new domains. This also enables us to conduct quality estimation on very large test sets, as no human translation is required in the process.

Further work might include applying our method to other language pairs; while we examined only the French to English language pair, it would be interesting to see the perfor-

mance of this approach on other languages and language types, such as morphologically-rich languages. This is important as we utilize target-language specific feature sets and tools: function word lists, part-of-speech taggers and parsers.

Another important aspect which we did not investigate thoroughly is domain adaptation. As is widely known, domain adaptation has a large influence in the process of building high quality MT systems. As our method requires only random human generated data, it is interesting to see the influence of this data being in-domain or out-of-domain in comparison to the relevant test set.

Another field of growing interest which may benefit from future work related to this method is word-level quality estimation. It is interesting to use the data sets and features we offer in this work while zooming in to classification at the word level, in favor of error detection etc.

As our method does not require human supervision, one can also integrate it into a statistical machine translation system, whether in the decoding phase or in the hypothesis re-ranking phase. Furthermore, additional features and feature selection techniques can be applied, both for improving detection accuracy and for strengthening the correlation with human quality estimation.

Appendix A

Supplementary Tables

Measurement	Naive Bayes	LibSVM	SMO	LibLinear	Logistic Regression
R^2	0.6231	**0.8107**	0.6821	0.6625	0.6647
uedin-heafield	61.49	62.21	65.29	65.24	**65.37**
uedin	62.34	63.07	65.37	**65.92**	65.84
online-b	63.09	63.76	66.06	66.72	**66.92**
limsi-soul	62.71	63.76	66.26	**66.89**	66.67
kit	62.46	63.41	66.54	**66.56**	66.21
online-a	62.91	64.67	67.41	**67.44**	67.37
mes-simplified	62.96	65.01	**66.67**	66.27	66.61
dcu	63.34	64.06	**67.17**	66.91	67.06
rwth	64.62	65.36	68.11	**68.26**	68.07
cmu-t2t	63.29	65.39	67.27	**67.41**	67.34
cu-zeman	67.64	67.91	**70.87**	70.74	70.76
JHU	64.69	66.19	67.62	**67.64**	67.57
SHEF	65.52	67.12	69.61	70.34	**70.39**

Table A.1: Classifier comparison, comparing detection accuracy on systems from WMT13', including the R^2 coefficient describing the correlation of the detection accuracy with human quality estimations, using CFG rules as features.

a	back	does	exeunt	hasn't	its	ninthly
about	be	doesn't	exit	hast	itself	no
above	bear	doing	fact	hath	last	nobody
according	because	don't	fair	have	lastly	none
accordingly	been	done	far	haven't	later	noone
actual	before	dost	farewell	having	less	nor
actually	being	doth	few	he	let	not
after	below	doubtful	fewer	he'd	let's	nothing
afterward	beneath	doubtfully	fifteen	he'll	like	now
afterwards	beside	down	fifteenth	he's	likely	nowhere
again	besides	due	fifth	hence	many	o
against	better	during	fifthly	her	matter	occasionally
ago	between	e.g.	fiftieth	here	may	of
ah	beyond	each	fifty	hers	maybe	off
ain't	bid	earlier	finally	herself	me	oft
all	billion	early	first	him	might	often
almost	billionth	eight	firstly	himself	million	oh
along	both	eighteen	five	his	millionth	on
already	bring	eighteenth	for	hither	mine	once
also	but	eighth	forever	ho	more	one
although	by	eighthly	forgo	how	moreover	only
always	came	eightieth	forth	how's	most	or
am	can	eighty	fortieth	however	much	order
among	can't	either	forty	hundred	must	other
an	cannot	eleven	four	hundredth	mustn't	others
and	canst	eleventh	fourteen	i	my	ought
another	certain	else	fourteenth	i'd	myself	our
any	certainly	enough	fourth	i'm	nay	ours
anybody	come	enter	fourthly	i've	near	ourselves
anyone	comes	ere	from	if	nearby	out
anything	consequently	erst	furthermore	in	nearly	over
anywhere	could	even	generally	indeed	neither	perhaps
are	couldn't	eventually	get	instance	never	possible
aren't	couldst	ever	gets	instead	nevertheless	possibly
around	dear	every	getting	into	next	presumable
art	definite	everybody	give	is	nine	presumably
as	definitely	everyone	go	isn't	nineteen	previous
aside	despite	everything	good	it	nineteenth	previously
at	did	everywhere	got	it'd	ninetieth	prior
away	didn't	example	had	it'll	ninety	probably
ay	do	except	has	it's	ninth	quite

Table A.2: List of function words as used for features in the classification experiments

rare	should	ten	thirteen	twentieth	we’re	wil
rarely	shouldn’t	tenth	thirteenth	twenty	we’ve	will
rather	shouldst	tenthly	thirtieth	twice	welcome	wilst
result	similarly	than	thirty	twill	well	wilt
resulting	since	that	this	two	were	with
round	six	that’s	thither	under	weren’t	within
said	sixteen	the	those	undergo	what	without
same	sixteenth	thee	thou	underneath	what’s	won’t
say	sixth	their	though	undoubtedly	whatever	would
second	sixthly	theirs	thousand	unless	when	wouldn’t
secondly	sixtieth	them	thousandth	unlikely	whence	wouldst
seldom	sixty	themselves	three	until	where	ye
seven	so	then	thrice	unto	where’s	yes
seventeen	soever	thence	through	unusual	whereas	yesterday
seventeenth	some	there	thus	unusually	wherefore	yet
seventh	somebody	there’s	thy	up	whether	you
seventhly	someone	therefore	till	upon	which	you’d
seventieth	something	these	tis	us	while	you’ll
seventy	sometimes	they	to	very	whiles	you’re
shall	somewhere	they’d	today	was	whither	you’ve
shalt	soon	they’ll	tomorrow	wasn’t	who	your
shan’t	still	they’re	too	wast	who’s	yours
she	subsequently	they’ve	towards	way	whoever	yourself
she’d	such	thine	twas	we	whom	yourselves
she’ll	sure	third	twelfth	we’d	whose	
she’s	tell	thirdly	twelve	we’ll	why	

Table A.3: List of function words as used for features in the classification experiments, continued

Appendix B

Supplementary Figures

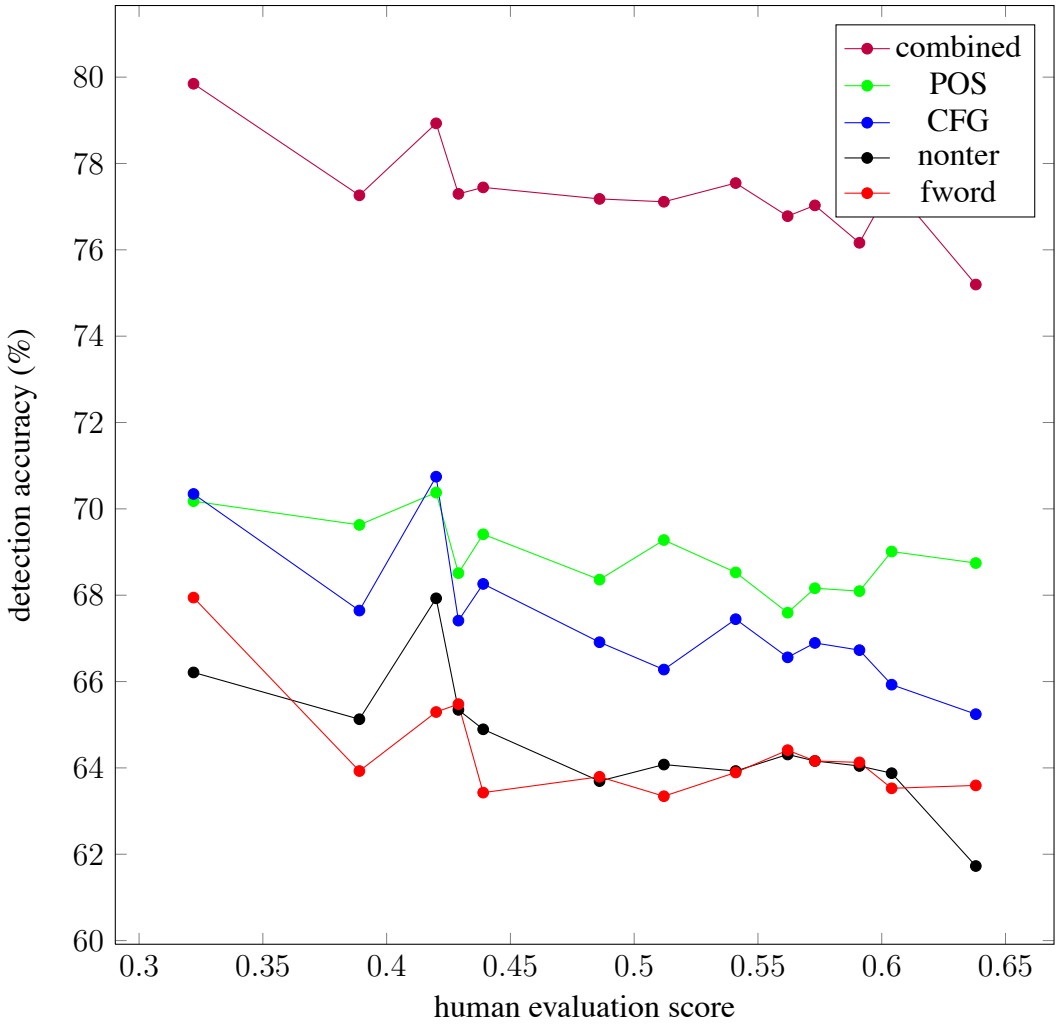


Figure B-1: Feature comparison, measuring correlation between detection accuracy and human evaluation scores on systems from WMT13' against non-reference sentences from WMT12', using the LibLinear classifier with sparsity threshold set at $t = 30$

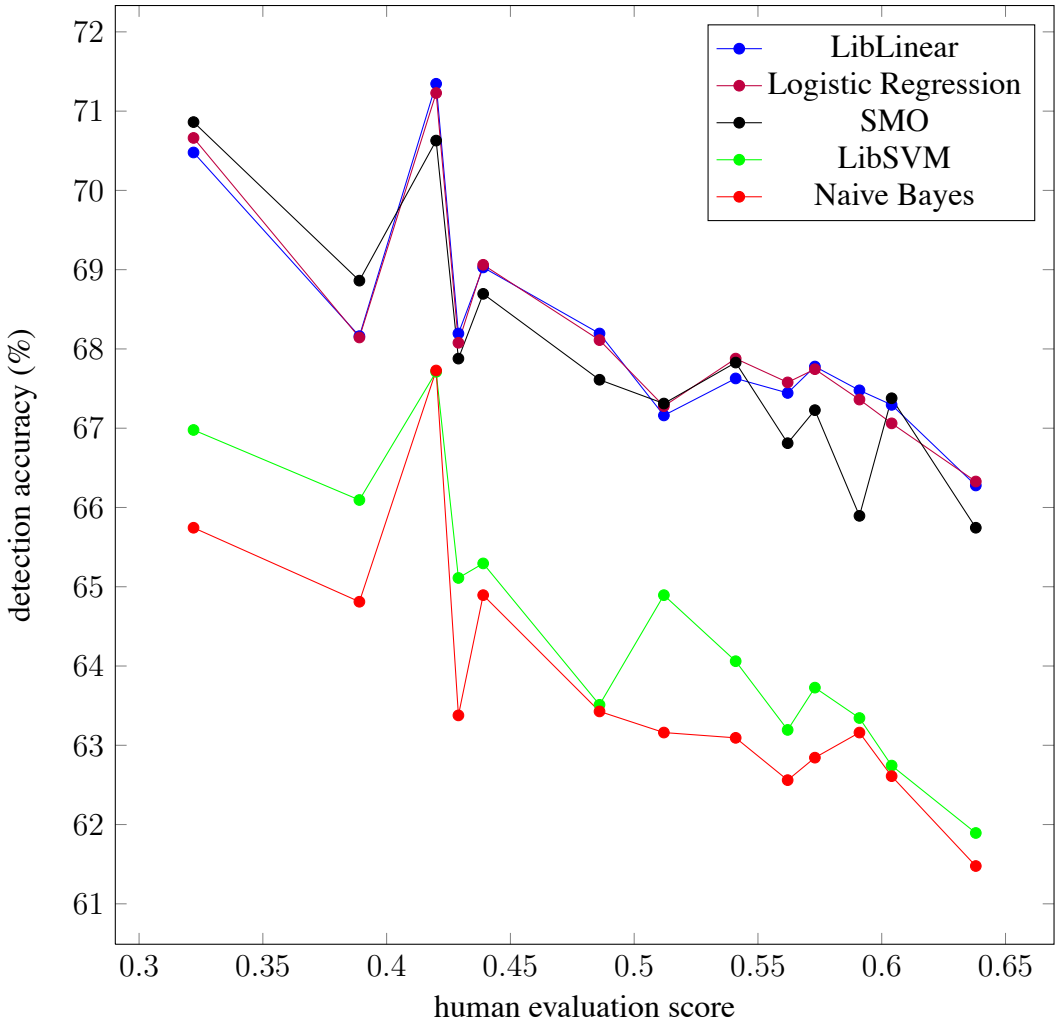


Figure B-2: Classifier comparison, measuring correlation between detection accuracy and human evaluation scores on systems from WMT13' against non-reference sentences from WMT12', with CFG rules as features and sparsity threshold set at 20.

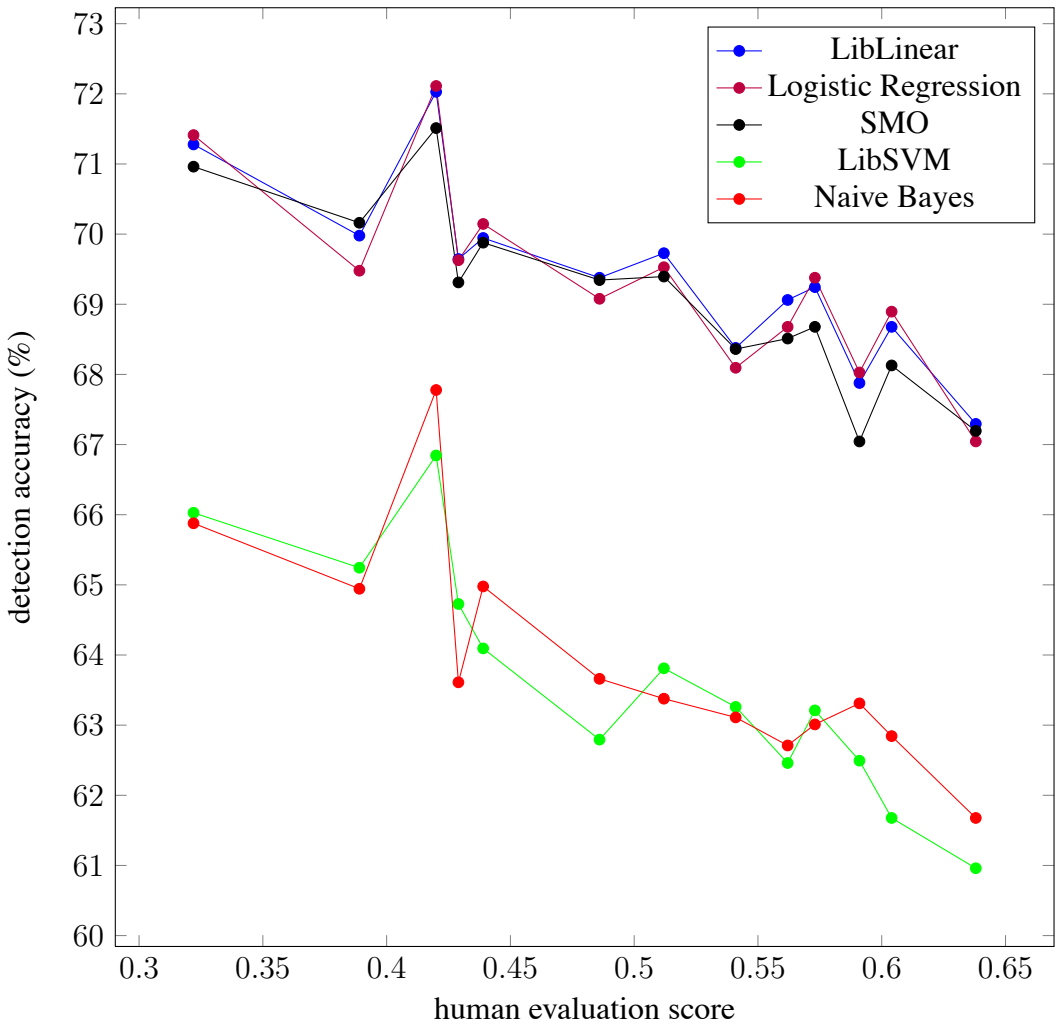


Figure B-3: Classifier comparison, measuring correlation between detection accuracy and human evaluation scores on systems from WMT13' against non-reference sentences from WMT12', with CFG rules as features and sparsity threshold set at 10.

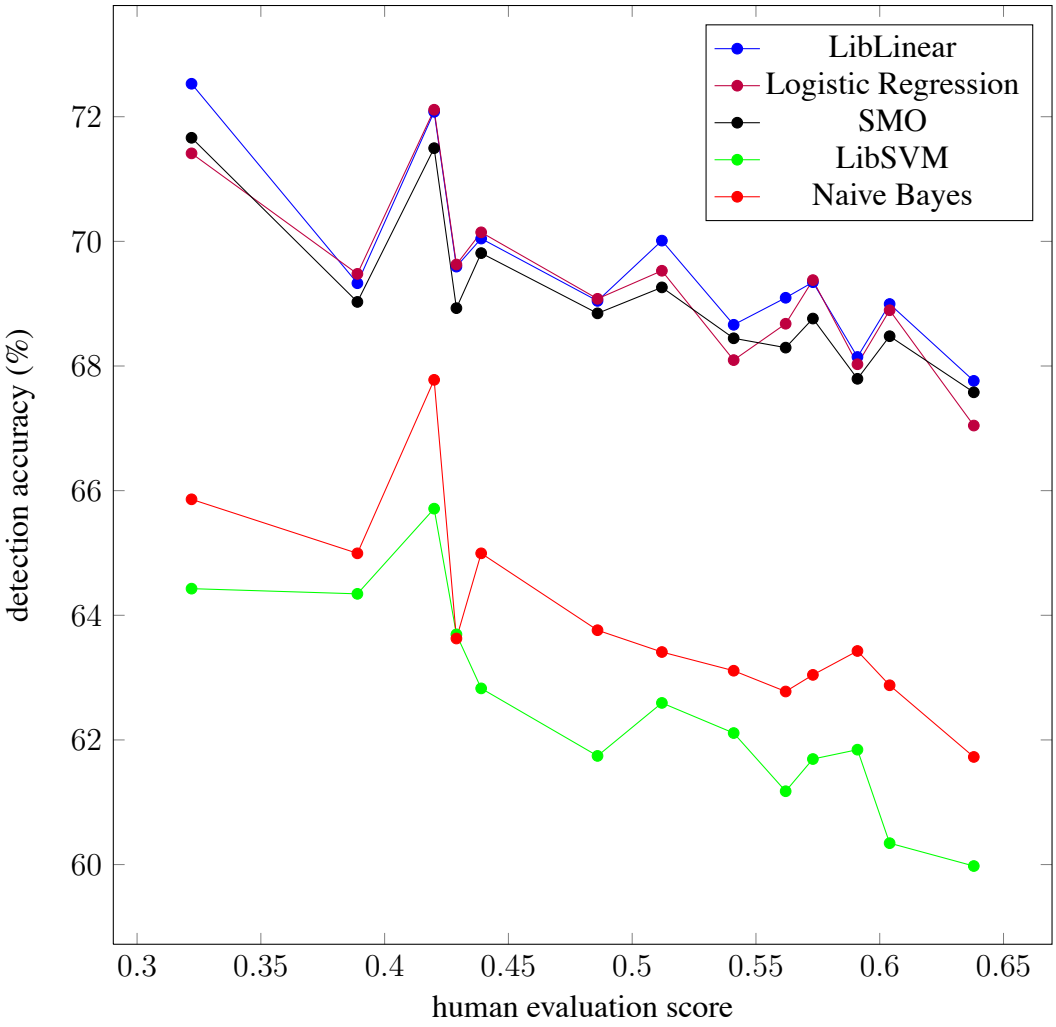


Figure B-4: Classifier comparison, measuring correlation between detection accuracy and human evaluation scores on systems from WMT13' against non-reference sentences from WMT12', with CFG rules as features and sparsity threshold set at 5.

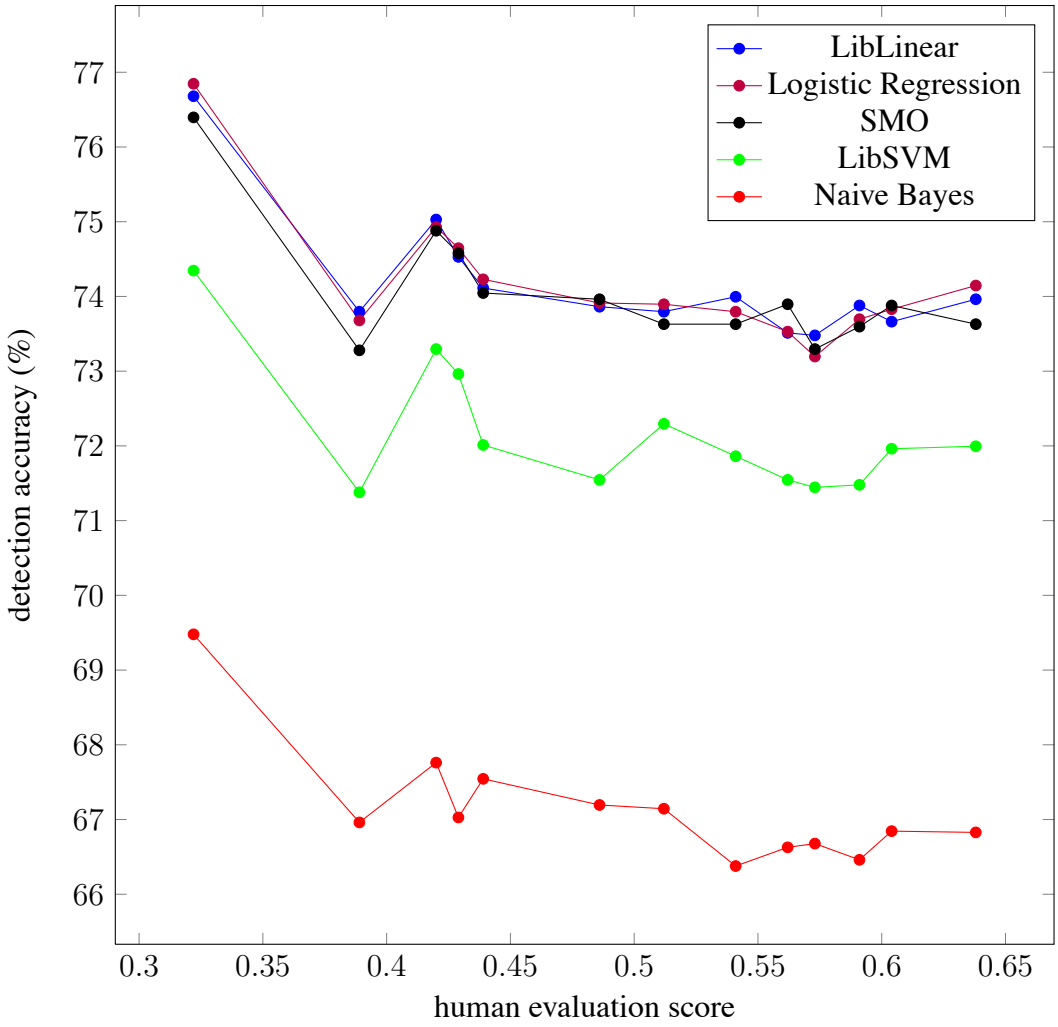


Figure B-5: Classifier comparison, measuring correlation between detection accuracy and human evaluation scores on systems from WMT13' against non-reference sentences from WMT12' with function word and POS features and a sparsity threshold set at $t = 30$.

Figure B-6: BLEU score vs. detection accuracy over the commercial MT systems dataset, using POS unigram features & SVM classifier, MT sentences vs. Reference sentences

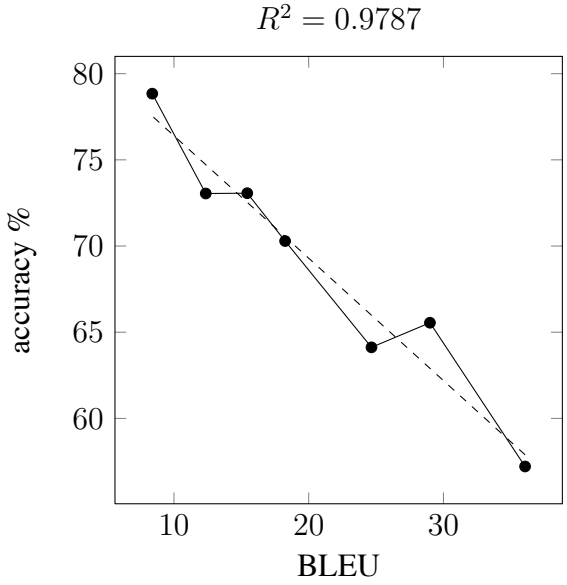


Figure B-7: BLEU score vs. detection accuracy over the commercial MT systems dataset, using word unigram features & Naive Bayes classifier, MT sentences vs. Reference sentences

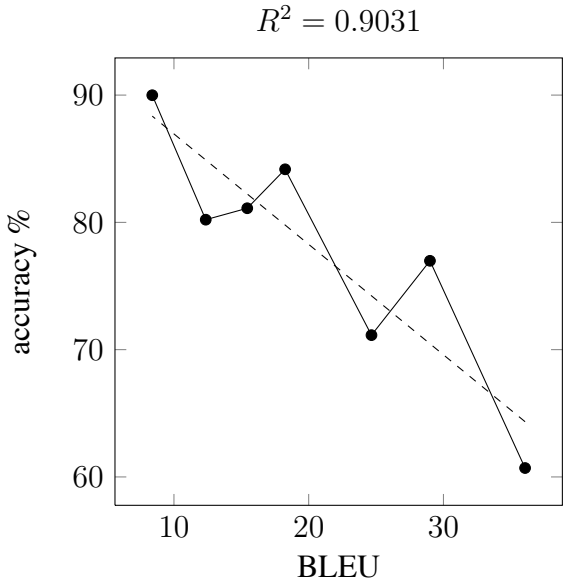


Figure B-8: BLEU score vs. detection accuracy over the commercial MT systems dataset, using POS unigram features & SVM classifier, MT sentences vs. Non-Reference sentences

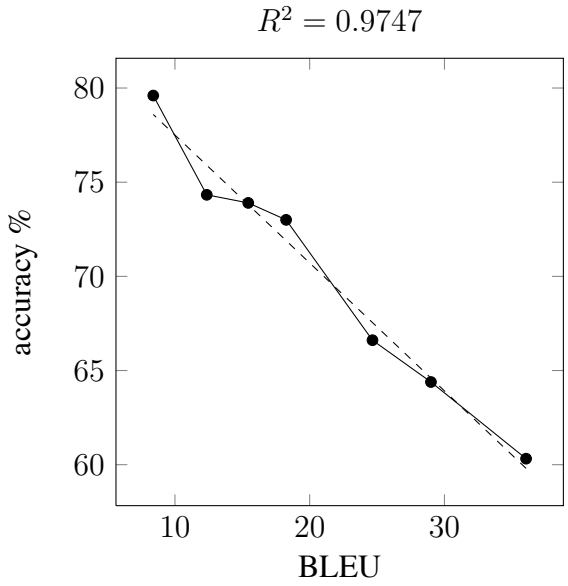


Figure B-9: BLEU score vs. detection accuracy over the commercial MT systems dataset, using word unigram features & Naive Bayes classifier, MT sentences vs. Non-Reference sentences

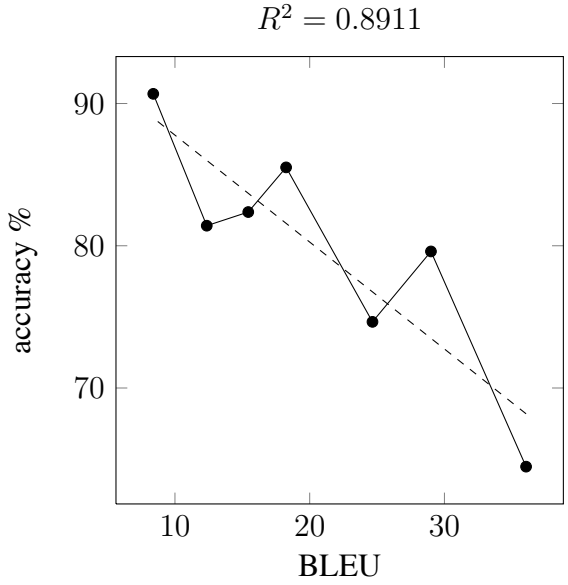


Figure B-10: BLEU score vs. detection accuracy over the commercial MT systems dataset, using function word features & Naive Bayes classifier, MT sentences vs. Non-Reference sentences

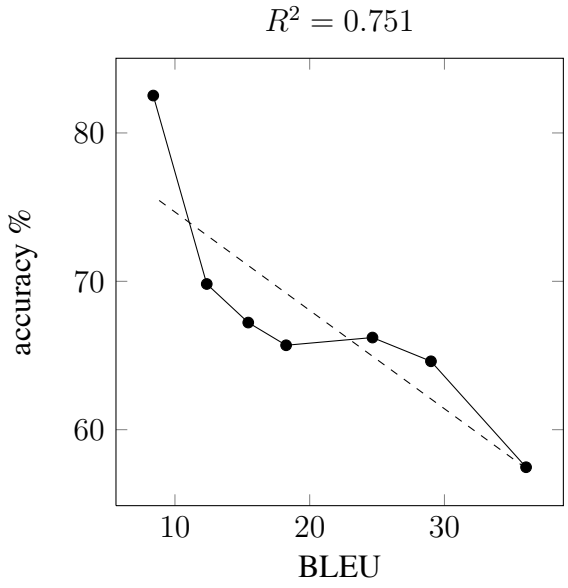


Figure B-11: BLEU score vs. detection accuracy over the commercial MT systems dataset, using function word features & SVM classifier, MT sentences vs. Non-Reference sentences

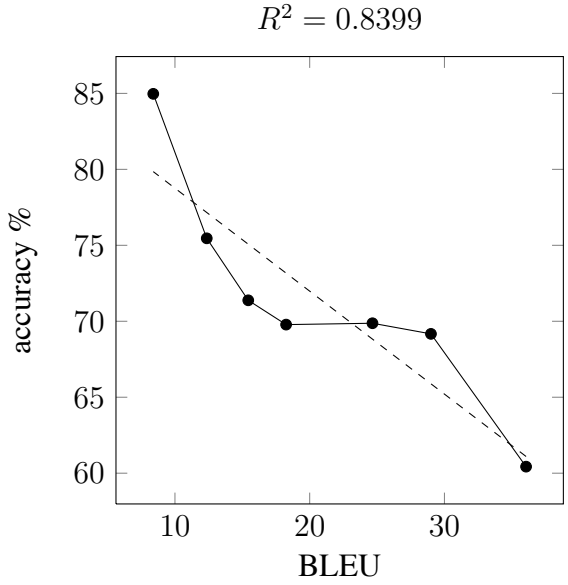


Figure B-12: Correlation between detection accuracy and human evaluation scores on systems from WMT 13' against non-reference sentences, using only tree-based features

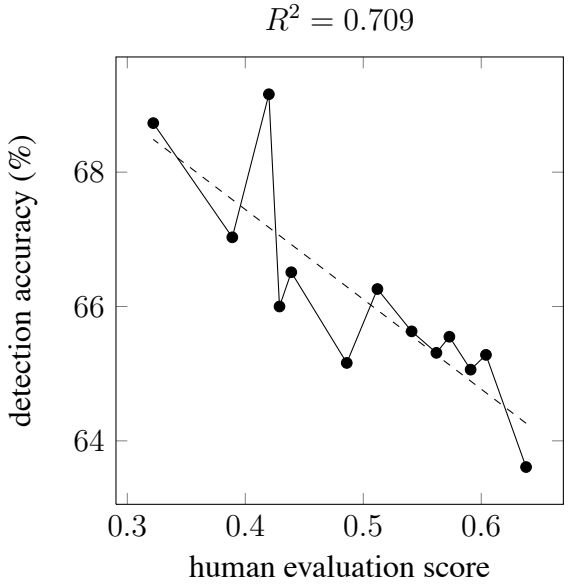


Figure B-13: Correlation between detection accuracy and human evaluation scores on systems from WMT 13' against non-reference sentences, using all tree features besides non-terminal features, and a threshold of 30

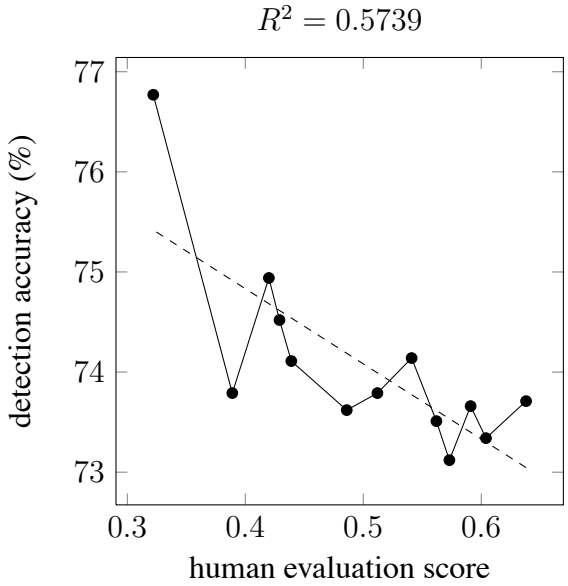


Figure B-14: Correlation between detection accuracy and human evaluation scores on systems from WMT 13' against non-reference sentences, using the parsing-based CFG one-level rules feature set with LibLINEAR classifier

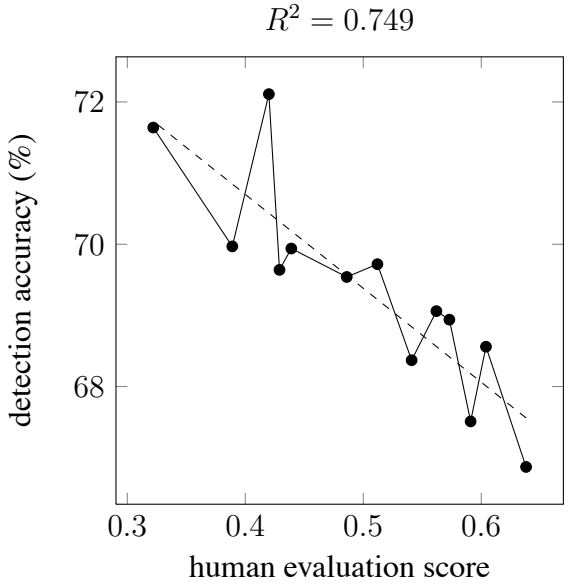


Figure B-15: Correlation between detection accuracy and human evaluation scores on systems from WMT 13' against non-reference sentences, using all the feature sets with a threshold of 30 and LibLINEAR classifier

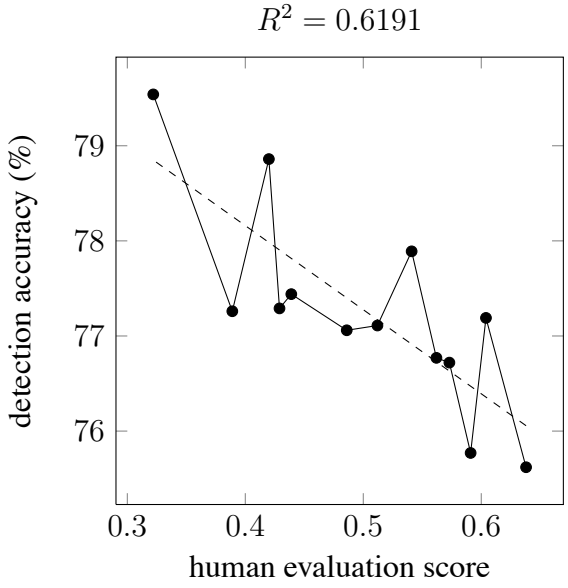


Figure B-16: Correlation between detection accuracy and human evaluation scores on systems from WMT 13' against non-reference sentences, using non-terminal rules feature set

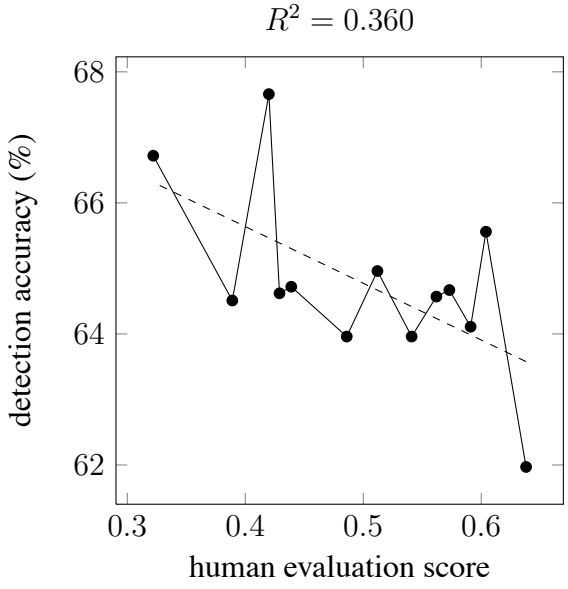
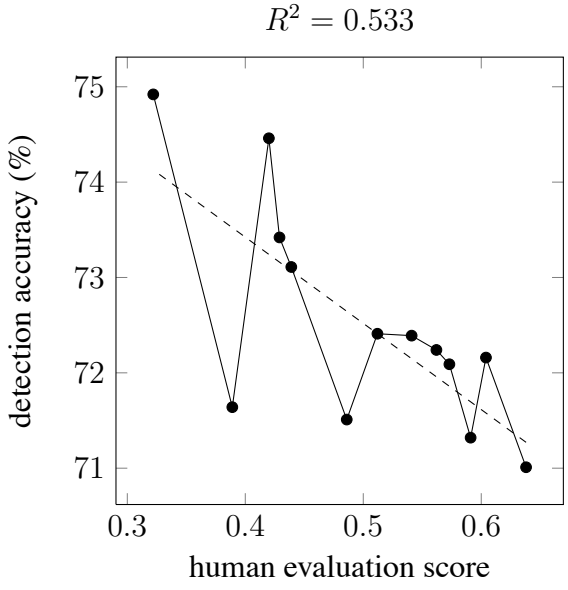


Figure B-17: Correlation between detection accuracy and human evaluation scores on systems from WMT 13' against non-reference sentences, using all the feature sets and a threshold of 20



Bibliography

[1] Yuki Arase and Ming Zhou. Machine translation detection from monolingual web-text. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1597–1607, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[2] Mona Baker. Corpus linguistics and translation studies: Implications and applications. *Text and technology: in honour of John Sinclair*, 233:250, 1993.

[3] Mona Baker. Corpus linguistics and translation studies – implications and applications. In *Text and Technology. In Honour of John Sinclair*, pages 233–250. John Benjamins, 1993.

[4] Satanjeev Banerjee and Alon Lavie. 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*, pages 65–72, 2005.

[5] Marco Baroni and Silvia Bernardini. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *LLC*, 21(3):259–274, 2006.

[6] Shoshana Blum-Kulka and Eddie A. Levenston. Universals of lexical simplification. *Strategies in Interlanguage Communication*, pages 119—139, 1983.

[7] Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[8] Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[9] Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada, June 2012. Association for Computational Linguistics.

[10] Dave Carter and Diana Inkpen. Searching for poor quality machine translated text: Learning the difference between human writing and machine translations. In Leila Kosseim and Diana Inkpen, editors, *Canadian Conference on AI*, volume 7310 of *Lecture Notes in Computer Science*, pages 49–60. Springer, 2012.

[11] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[12] Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*, 2011.

[13] George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc., 2002.

[14] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[15] Martin Gellerstam. Translationese in swedish novels translated from english. In Lars Wollin and Hans Lindquist, editors, *Translation Studies in Scandinavia*, pages 88–95, 1986.

[16] Martin Gellerstam. Translationese in swedish novels translated from english. In *Translation Studies in Scandinavia*. *CWK Gleerup*, pages 88–95, 1986.

[17] Ulrich Germann. Aligned hansards of the 36th parliament of canada release 2001-1a, 2001.

[18] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, 2009.

[19] Jiawei Han, Jian Pei, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, and MC Hsu. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *proceedings of the 17th international conference on data engineering*, pages 215–224, 2001.

[20] Kenneth Heafield. *Efficient Language Modeling Algorithms with Applications to Statistical Machine Translation*. PhD thesis, Carnegie Mellon University, September 2013.

[21] Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. Identification of translationese: A machine learning approach. In Alexander F. Gelbukh, editor, *CICLing*, volume 6008 of *Lecture Notes in Computer Science*, pages 503–511. Springer, 2010.

[22] S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, and K.R.K. Murthy. Improvements to platt's smo algorithm for svm classifier design. *Neural Computation*, 13(3):637–649, 2001.

[23] Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005. AAMT, AAMT.

[24] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *ACL*. The Association for Computer Linguistics, 2007.

[25] Moshe Koppel and Noam Ordan. Translationese and its dialects. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *ACL*, pages 1318–1326. The Association for Computer Linguistics, 2011.

[26] David Kurokawa, Cyril Goutte, and Pierre Isabelle. Automatic Detection of Translated Text and its Impact on Machine Translation. In *Conference Proceedings: the twelvth Machine Translation Summit*, 2009.

[27] Alon Lavie and Abhaya Agarwal. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231. Association for Computational Linguistics, 2007.

[28] Saskia Le Cessie and Johannes C Van Houwelingen. Ridge estimators in logistic regression. *Applied statistics*, pages 191–201, 1992.

[29] Adam Lopez. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):8, 2008.

[30] D. Madigan, A. Genkin, D.D. Lewis, and D. Fradkin. Bayesian multinomial logistic regression for author identification. In *AIP Conference Proceedings*, volume 803, page 509, 2005.

[31] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. Technical report, IBM Research Report, 2001.

[32] J.W. Pennebaker, M.E. Francis, and R.J. Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 2001.

[33] Slav Petrov and Dan Klein. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*, pages 404–411, 2007.

[34] Marius Popescu. Studying translationese at the character level. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, and Nicolas Nicolov, editors, *RANLP*, pages 634–639. RANLP 2011 Organising Committee, 2011.

[35] Matt Post, Juri Ganitkevitch, Luke Orland, Jonathan Weese, Yuan Cao, and Chris Callison-Burch. Joshua 5.0: Sparser, better, faster, server. In *Proceedings of the Eighth Workshop on Statistical Machine Translation, August 8-9, 2013.*, pages 206–212. Association for Computational Linguistics, 2013.

[36] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, pages 223–231, 2006.

[37] Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. Quest - a translation quality estimation framework. In *51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, ACL, pages 79–84, Sofia, Bulgaria, 2013.

[38] Gideon Toury. *In Search of a Theory of Translation*. 1980.

[39] Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *IN PROCEEDINGS OF HLT-NAACL*, pages 252–259, 2003.

[40] Hans van Halteren. Source language markers in europarl translations. In Donia Scott and Hans Uszkoreit, editors, *COLING*, pages 937–944, 2008.

[41] Vered Volansky, Noam Ordan, and Shuly Wintner. On the features of translationese. *Literary and Linguistic Computing*, 2013.

ואילו הם בשפה טבעית, נראה שככל שאיכות התרגום עולה, רמת הדיוק של המסווג יורדת, ולהיפך. למעשה, הקורלציה במקרה זה חזקה מספיק, כך שאנו מציעים להשתמש ברמת הדיוק של המסווג כשיטה להערכת איכות התרגום.

אנו עורכים מגוון ניסויים כדי לבחון את השערותינו במקרים שונים. תחילה, אנו לוקחים מספר מערכות מסחריות לתרגום ממוכן, ומודדים את הקורלציה בין היכולת שלנו לזהות את הפלט שלהן כתרגום ממוכן לבין איכות התרגום הצפויה שלהן, כפי שנקבעה על ידי מדד BLEU או מדד METEOR. אנו מראים כי הקורלציה במקרה זה אכן גבוהה מאוד, בין אם אנו משתמשים בתרגומים לדוגמא או בטקסטים טבעיים רנדומליים. בניסוי נוסף אנו מבצעים את אותה מדידה, אך הפעם על מערכות תרגום ממוכן שאנו יוצרים בעצמנו בצורה מבוקרת, על ידי שימוש בסט הכלים MOSES. במקרה זה אנו שוב מראים קורלציה חזקה בין היכולת שלנו לזהות את המשפטים בצורה נכונה לבין איכות התרגום, כפי שנמדדה על ידי BLEU או METEOR. בניסוי האחרון אנו רוצים למדוד את הקורלציה בין יכולת הזיהוי לבין הערכת איכות תרגום שמבוססת על שיפוט אנושי, שהיא הערכת האיכות המהימנה ביותר. כדי לעשות זאת אנו משתמשים במידע מWMT, הסדנה השנתית לחקר התרגום הממוכן, שתויג בצורה ידנית על ידי קבוצת שיפוט אנושית עם איכות התרגום שהם העריכו. בעוד שבחנו שימוש במספר תכונות לשוניות שונות של הטקסט לצורך הסיווג, שוב מצאנו קורלציה חזקה בין יכולת הזיהוי לאיכות התרגום, הפעם כפי שנמדדה על ידי שיפוט אנושי.

גישה זו מועילה במספר מישורים. תחילה, בתור שיטה למדידת איכות תרגום, היא אינה מצריכה משפטים מתורגמים אנושית, כמו אלה שנזקקים אליהם בBLEU לדוגמא, שכן אנו משתמשים במשפטים רנדומליים בשפה טבעית לצורך השוואה. יכולת זו מאפשרת לערוך בדיקות איכות על קבוצות בדיקה גדולות מאוד, למשל בתחומי עניין (Domains) שונים, בצורה קלה ופשוטה. בנוסף, השיטה שלנו מאפשרת למצוא בעיות ממוקדות במערכת התרגום הממוכן, שכן היא מצביעה על משפטים בעייתיים במיוחד, אלה שסווגו כלא אנושיים. תכונה נוספת של הגישה היא העובדה שהיא לא תלויה בשפת המקור של התרגומים, מה שהופך אותה למתאימה למספר רב של זוגות שפות ללא התאמה מיוחדת.

תקציר

עם ההצלחה והפופולריות של מערכות לתרגום ממוכן סטטיסטי (Statistical Machine Translation) נדרשת התקדמות במספר יכולות חשובות. אחת מהן היא היכולת להעריך בצורה אוטומטית איכות של תרגום ממוכן עבור שפות ותחומי ידע שונים ומגוונים. יכולת זו חשובה במיוחד בתהליך הפיתוח והאימון של מערכות תרגום ממוכן חדשות, במהלכו יש צורך לנטר בצורה תכופה את איכות התרגום הצפויה של תוצרי המערכת. נקודה חשובה נוספת היא היכולת לזהות בצורה אוטומטית תרגום ממוכן מתוך קורפוס המכיל טקסט אנושי לצד תרגום ממוכן, כמו שניתן לראות לרוב במידע שנאסף מהרשת. עבודה זו סוקרת את הקשר בין שתי היכולות הללו ומציגה שיטה חדשה להערכה אוטומטית של איכות תרגום על ידי מדידת הדיוק בו מסווג מצליח להפריד בין תרגום ממוכן וטקסט אנושי.

תחילה נציג את הבעיה של שערוך איכות תרגום ממוכן (Machine Translation Quality Estimation). הבעיה מוגדרת בתור חיזוי איכות התרגום של טקסט שהוא פלט של תרגום ממוכן, בין אם ברמת הקורפוס, המשפט או המילה, וזאת ללא מידע על הפלט הרצוי, וללא שימוש בתרגומים אנושיים לצורך השוואה. לעומת זאת, בהערכת תרגום ממוכן (Machine Translation Evaluation) מסתמכים בעיקר על השוואה לתרגומים אנושיים על מנת לתת ציון לאיכות התרגום. אנו מציגים שיטה חדשה לשערוך איכות תרגום, שמבוססת תחילה על פתרון בעיה אחרת בתרגום ממוכן: זיהוי תרגום ממוכן (Machine Translation Detection). בזיהוי תרגום ממוכן, נרצה לזהות בצורה אוטומטית משפטים שהם פלט של תרגום ממוכן, מתוך קורפוס המכיל משפטים כאלה ומשפטים בשפה טבעית, לסירוגין.

כדי לבצע זיהוי של תרגום ממוכן, נתבונן במאפיינים הכלליים של טקסטים מתורגמים, שנחקרים כבר שנים רבות. נסיונות להגדרת מאפיינים אלה, שנקראים לרוב תכונות התרגום (Translation Universals) כוללים את [38, 6, 2, 15]. ההבדלים שנמצאו בין טקסטים מתורגמים לבין טקסטים מקוריים (Native) הינם מעבר לשגיאות בתרגום, אלא מציגים מעין דיאלקט נוסף, תרגומית. מספר עבודות [5, 26, 21, 25] השתמשו בטכניקות לסיווג טקסט כדי להבדיל בין טקסטים מתורגמים לטקסטים מקוריים, בין אם ברמת המסמך או הפסקה, תוך שימוש במגוון תכונות לשוניות. לגבי זיהוי של תרגום ממוכן, קרטר ואינקפן [10] ערכו ניסויים ברמת המסמך, ואילו אראסה וזו [1] עבדו ברמת המשפט. בעוד העבודות שהוזכרו לעיל עסקו בזיהוי תרגום ממוכן ברמת המשפט, הן לא חקרו את הקשר בין איכות התרגום הממוכן והיכולת לזהותו ככזה.

אנו משערים שאיכות התרגום של מערכת לתרגום ממוכן יכולה להימדד על פי הדיוק שבו שיטה לזיהוי תרגום ממוכן תזהה משפטים של המערכת, מתוך קורפוס המכיל משפטים אלה ומשפטים בשפה טבעית. בגישה זו, נראה תחילה כי באמצעות תכונות לשוניות של הטקסט, כמו תדירויות של חלקי דיבר, מילות קישור וכדומה, ניתן ללמוד מסווגים שיבדילו בין המשפטים שהם פלט של תרגום ממוכן לבין משפטים בשפה טבעית. בעוד שתוצאה זו אינה חדשנית לחלוטין, התרומה שלנו היא להציג אותה ביחס לאיכות התרגום. אנו נראה שהדיוק של מסווגים כמו אלה שהזכרנו לעיל, נמצא בקורלציה חזקה לאיכות התרגום של מערכת התרגום הממוכן המעורבת. אם נערוך ניסויים בהם ננסה לזהות אילו משפטים הם פלט של תרגום ממוכן

עבודה זו נעשתה בהדרכתו של פרופסור משה קופל מהמחלקה למדעי המחשב של אוניברסיטת בר-אילן.

אוניברסיטת בר-אילן

זיהוי אוטומטי של תרגום ממוכן והערכת איכות תרגום

רועי אהרוני

עבודה זו מוגשת כחלק מהדרישות לשם קבלת תואר מוסמך במחלקה למדעי המחשב של אוניברסיטת בר-אילן

רמת גן תשע"ה