Detecting Machine Translated Text using a BERT Based Classifier

Federico Ferlito (s2936860) Remo Sasso (s2965917)
Nino Jansen (s2965690) Boris Marinov (s2957531)

***Abstract*— Machine translation has seen massive improvements over recent years. Especially supervised models are coming ever closer to human parity. Free tools like Google Translate or Deepl allow people to achieve these levels of translations easily, without requiring a human expert. This raises concerns due to the ease of translating and spreading things like 'fake news'. In contrast to corpora in machine translation that often use crawled news like WMT2014, the novel idea of using subtitles is presented. In this study translated subtitle sentences are classified to find out if they are human or machine translated. Subtitles taken from the Opensubtitles corpus are translated from Italian to English and Dutch to English. BERT-models of different sizes are fine-tuned for this classification task on the respective language directions. Results indicate that classification was more successful for the Dutch to English direction compared to the Italian to English, with some improvements seen in larger BERT models. Reasons for the difference could be due to the nature of the data, or more language specific ones too. Nevertheless, the paper presents a promising and novel direction in machine translation classification, with several ideas for future explorations.**

I. INTRODUCTION

The field of machine translation has seen massive advancements in both supervised and unsupervised methods over recent years. The majority of approaches to machine translation have either been statistical (SMT) or neural based (NMT). Supervised methods have shown to be superior to their unsupervised counterparts [3]. A recent line of research [7] has even made the claim that a supervised NMT method has achieved human parity on Chinese to English translations. While their success can not be disputed, the claim was challenged nevertheless in several follow-up studies. Toral et al. [16] address some factors that were not taken into consideration in their claims, such as the language of the original test source side, or the proficiency of the evaluators. Additionally, Laubli et al. [9] look at the wider scope and argue that most comparisons to human levels are done on a sentence-by-sentence basis, while attention should be placed more on a higher document level. These papers approach the arguments from different directions, but in the end, both state human parity claim is overambitious.

The previously mentioned research has focused on supervised methods, where the quality of translations are better than those of unsupervised approaches. Unsupervised methods have come closer to supervised translations in recent research. Artetxe et al. [3] build on their previous work and with their improved unsupervised SMT/NMT hybrid report levels of translation of similar quality to the supervised machine translation winner from WMT2014. Even though these models do not require parallel corpora, they are still limited, as they rely on monolingual corpora containing millions of sentences. To counter this, a study by Gu et al. [6] demonstrates high translation qualities with scarce and limited corpora resources, such as English-Romanian. The combined research in the field and its advancements demonstrates the growing potential of machine translation.

The need for accurate detection of machine-translated text could arise for many reasons. For one, easily available tools for translations, such as Google Translate[1] or DeepL[2], are readily accessible to the public. This raises a concern of harmful behaviour and an especially easy way to spread misinformation in different languages via automatic bots. An ability to detect non-original text could aid the stopping of spam-bots and propagation of "fake news", a topic becoming more relevant with the wider spread of micro-blogs on social media. A similar point can be made about the machine-generated text. Recently, OpenAI released an unsupervised language model called GPT-2 [15], which has raised much attention with its performance and ease of use as it requires no extensive task-specific training. One can easily imagine how machine-generated text could lead to even more harmful cases, where the need for authenticity is even more relevant.

Determining flaws in machine translations can also be used in a more positive manner, such as incorporating these findings as feedback to the system itself. A study by Li et al. [10] demonstrates this by detecting deviations from human translations and using them to improve the original SMT system used for the translations. When determining the performance of the translation system, the Bilingual Evaluation Understudy Score (BLEU) is the standard method used [14]. This metric evaluates how well the generated texts match to reference human translation. Therefore, an iterative approach like this could improve existing methods, since the BLEU score is used during the optimization stages.

The studies mentioned here, as well as those in the following related works section, have mainly focused on popular and widely available corpora for training. For this research, the focus is turned onto a less used text type, which is subtitles for a range of media formats. Subtitles often form short sentences, more associated with conversational contexts, and can also include a lot of colloquial phrases. Another reason for selecting subtitles is the likelihood that they were used in training of the translation systems used. While a popular dataset from the internet, such as the Tatoeba

[1] https://translate.google.com/
[2] https://www.deepl.com/en/translator

multilingual corpus[3], is likely to be included in the training sets for Google Translate, this is highly unlikely for the subtitles. Thus, providing us with a novel and interesting to experiment situation. Further, improving the translation of subtitles via the incorporated feedback could prove to be useful for automatic closed captioning (CC) systems. If the source language is different then the desired CC language, translations are necessary, similar to the subtitle translations used in our study.

To summarise, the research presented here aims to answer the question of whether or not machine translated subtitles can be distinguished effectively from human translated ones. The rest of the paper is as follows: 2) Some related research on the field is presented and used to motivate the methods; 3) An outline of the model used for classification; 4) A descriptions of the methods and experiments; 5) A discussion of the procedure and results; 6) Conclusion and future research ideas.

The code used in this research and a detailed explanation of how to use it are available at `https://github.com/remosasso/Detecting-MT-Text-using-BERT`

II. RELATED WORK

Detection of translated text has a longstanding place in the field of translation studies. Numerous experiments and attempts ([4], [8]) have been conducted in finding a relationship between native and translated texts, with the differences going farther than simple errors. They summarise the idea of a "Translationese" dialect, comprised of subtle differences in the various linguistic features and use of words between the two sets of texts. The features range from word and POS n-grams, proportions of functional words in the sentence, and a general relationship between the different word types in the text. Volanski et. al [17] test several hypotheses, aiming to determine some of the more informative features and challenge some of the previously proposed "Translation Universals". While present in some cases, they oppose the idea that all translations follow the same rules, however, do provide strong discriminative features (such as n-grams lengths) and the importance of text genre and actual language structure.

With the constant improvements in machine translation, the need for accurate and informative detection increases even more. Early successes were recorded by Arase and Zhou [2] who trained a sentence-level classifier to distinguish machine translated and human generated web-page text between English and Japanese. Despite their high accuracy using application-specific features, they did not study the correlation between the translation quality of the machine translated text and the ability to detect it. Aharoni et al. [1] demonstrate that such detection is indeed possible with high accuracy only for low-quality translations. They describe a strong correlation between the detection accuracy and the BLEU score or the human evaluation score of the machine translation itself.

[3]`https://tatoeba.org/eng/`

Li et al. [10] focus their efforts on trying to improve performance of Statistical Machine Translation by detecting poorly translated sentences, staying close to the idea that high-quality translations should be close to those of a professional human translator. Unlike previous attempts that focus on n-gram features, syntax properties and translations coming from a single SMT system, their paper considers deep level linguistic and syntactic structure features derived from parsing trees on texts generated by multiple sourced SMT systems. Their Support Vector Machine classifier, along with a wide selection of parse tree, emotional and lexical features, is able to achieve an acceptable accuracy and is effective in providing feedback for improving the SMT system. Further, the paper proved that human parsing is more balanced in its structure than that of a machine,

While many successes in this line of research are demonstrated, the quality of machine translations are constantly going up, so even more recent attempts at detecting these have been documented. Adding to the idea that most evaluations are done on a sentence level and important relationships are ignored [9]. In [13] a new combined metric for paragraph coherence is proposed. This method integrates a metric for matching similar words at paragraph level based on POS-tag pairs, as well as a matching penalty metric to reduce the effect of unmatched words, proving itself to outperform previous classification accuracy scores. Moreover, their method works equally well for low-resource languages (Japanese) and high-resourced ones (Dutch), further underlying the importance of the coherence measure.

Majority of the papers mentioned here and the general field has made use of carefully selected features of the text and popular machine learning methods used in binary classification tasks (e.g SVMs). Currently to the best of our knowledge, no deep learning methods have been employed to solve the detection of machine translated texts. With the rising popularity of BERT [5] and its proven effectiveness in a wide-domain of tasks, its use here could be worthwhile to investigate. Combining this idea with the reason for selecting subtitles, the project here presents something not investigated in much depth before.

III. MODEL

In this research, we make use of the Bidirectional Encoder Representations from Transformers (BERT) [5]. BERT is an embedding layer that is learned to create deep bidirectional representations from unlabeled texts. It does so by jointly conditioning on both left-to-right and right-to-left context in all layers. The pre-trained models of BERT are trained on an enormous unsupervised plain text corpus with two objectives: *masked word prediction* (MLM) and *next sentence prediction* (NSP). In the MLM task, 15% of the word tokens are masked out of the input sequences and the model learns to predict the masked words. In the NSP task the model is presented with two sentences, say A and B, where the model is required to classify whether sentence B follows from sentence A or that sentence B is just a random sentence. Another interesting aspect of BERT is that it processes all

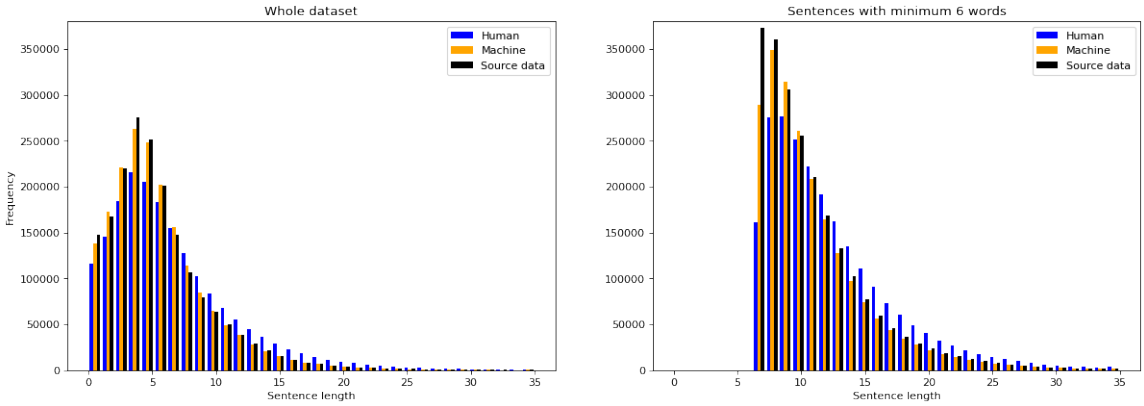

Fig. 1. Distribution of the length of the sentences in the Dutch corpus. The right plot show the distribution after removing the short sentences

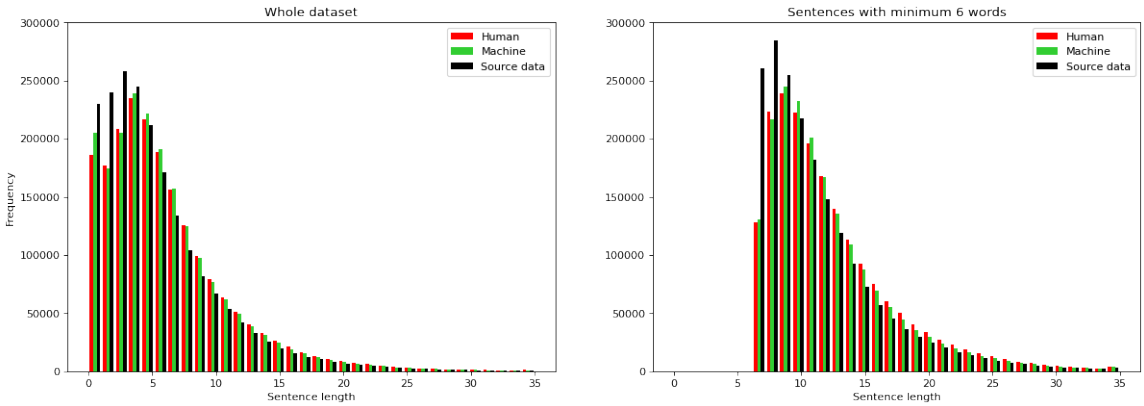

Fig. 2. Distribution of the length of the sentences in the Italian corpus. The right plot show the distribution after removing the short sentences

input tokens in parallel, as its attention architecture processes the whole input sequence at once. An illustration of BERT's architecture can be seen in Figure 3. BERT takes its input token sequence as follows: the first token of every sequence should be a classification token (CLS) and there should be a separation token (SEP) after every sentence. The output embedding corresponding to the classification token then is the sequence embedding that can be used for classifying the whole sequence.

BERT has proven to be an extremely useful and powerful method in Natural Language Processing (NLP) tasks, for which it is considered the state-of-the-art method for many [5]. BERT also provides much ease for researchers, as its pre-trained models[4] can easily be fine-tuned using just one additional layer. In this research we are dealing with a binary text classification task, meaning that only an additional layer of two neurons was necessary to use any of the BERT variations we desired.

BERT comes in several variants, ranging from BERT-tiny (two transformer blocks and 128 hidden neurons per layer) to

[4]https://github.com/google-research/bert

TABLE I

STATISTICS FROM THE DATASET USED

Dataset	**Human sentences**	**Machine sentences**	**Human token count**	**Machine token count**
Dutch	1.86M	1.86M	13.1M	10.9M
Dutch **(short sentences removed)**	2.28M	2.28M	28.9M	24.1M
Italian	2M	2M	13.1M	12.8M
Italian **(short sentences removed)**	1.94M	1.94M	15.5M	24.1M

TABLE II

AN EXAMPLE OF AN ITALIAN SENTENCE, TRANSLATED BOTH BY GOOGLE TRANSLATE AND BY A HUMAN TRANSLATOR

Source sentence	Per nessun motivo parteciperò ad attività riguardanti la Spagna.
Machine translation	For no reason will I participate in activities concerning Spain.
Human translation	To never, under any circumstances, engage in any activity connected, however remotely, with Spain.

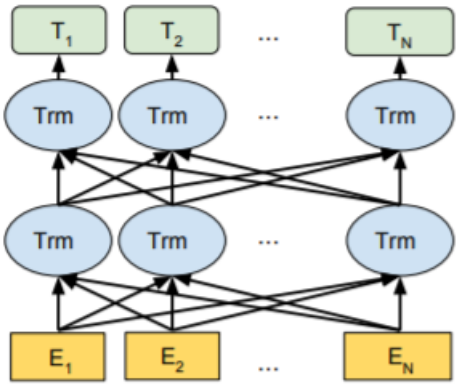

Fig. 3. BERT Architecture. En here represents the n-th token in the input sequence, Trm stands for Transformer block and Tn denotes the corresponding output embedding. (Source: [5])

BERT-large (24 transformer blocks and 1024 hidden neurons per layer). In our research we make use of four variants, which are listed in Table III showing the differences between the models. For many tasks, the larger BERT models have shown to outperform the smaller ones [5]. In this research, we have limited ourselves to use BERT-base as the largest model as the larger variants would have required too large amounts of time and computing power to train them.

TABLE III
OVERVIEW OF THE DIFFERENCES BETWEEN THE BERT MODELS.

BERT model	Nr. transformer blocks	Nr. hidden neurons
Tiny	2	128
Small	4	512
Medium	8	512
Base	12	768

IV. EXPERIMENTS

In our research, we used four different BERT models and four different datasets. This section will describe information concerning the datasets in detail and how they were used together with the models in our experiments.

A. Data

The dataset used in this experiment consists of two parallel corpora made up of movie and TV subtitles, one for English→Italian and one for English→Dutch. It is an extended and improved version of the *OpenSubtitles* collection of parallel corpora, with all subtitles going through several preprocessing steps and being aligned via a time-based approach [11]. To test the quality of the alignments, the same study used the bitexts as training material for a statistical machine translation system, reporting improvements in the BLEU scores compared to previous years.

The Italian→English dataset is made up of 28.8M of subtitles, whereas for the Dutch→English the number is 31M. As these are more than sufficient, and to ensure some randomized order, four smaller sets were created. This was done by creating 200 chunks from each large dataset, containing 200,000 subtitles per chunk. The order of both lists was shuffled in the same way as to keep the alignment accurate. For two of the sets for each language, subtitles of all length are included, whereas for the other two subtitles shorter than 6 words are filtered out. The resulting chunks contained the original human translations, as well as the original subtitles which were to be translated.

Following this, a random selection of approximately 30 chunks determined which subtitles would end up in the final datasets. The 30 chunks containing the original source language were individually translated using Google Translate. Following this, the translated chunks were merged with their human translation counterparts as well as labelled, **0** being a human translation and a **1** referring to a machine translation. The order is alternating, meaning that a human translation is always followed by its machine counterpart. Figure 4 illustrates a short snipped of the subtitle alignment, while Table I provides an overview of the four different datasets.

```
Well, you probably shouldn't say it.    0
Exact. Well, maybe you shouldn't say that.      1
I did everything I could!       0
I did everything I could!       1
Starting with the Olympic monotype class        0
Starting with the single Olympic class ...      1
```

Fig. 4. An example of a small snipped of the data.

B. Preprocessing

Only minor preprocessing was done on the dataset. Some Non-ASCII characters were removed from all sentences, carefully preserving certain encodings as to not lose specific characters, as well as special characters such as hashtags, dashes, etc. This was done to normalize the sentences between the two languages, as it was often the case that a human counterpart sentence would include formatting not present in the source side. No further preprocessing, apart from filtering out short sentences in half the datasets, was done. Consideration of removing stopwords, or lemmatizing words was taken into account, however due to the short nature of the subtitles and not knowing for certain how much importance is placed on each word during translation, it was determined that leaving as much information in each subtitle as possible might be worthwhile.

C. Evaluation

The experiments conducted in this research consist of four different BERT models trained and tested on four different datasets. For the BERT models we used BERT-tiny, BERT-small, BERT-medium and BERT-base. The datasets used, as described earlier, are Italian→English and Dutch→English, of which there is one version which excludes sentences with fewer than 6 words and one which includes all sentences for each. Each model was trained on each of these datasets with a learning rate of 2e-5 and batch sizes of 8. For each trained model, we evaluated them at the point where they had reached a convergence state, meaning the point that there was no notable difference anymore in performance as

training proceeded. This, for instance, was after approximately 450.000 training steps for BERT-tiny and 350.000 training steps for BERT-small. For each of the datasets we took 100.000 sentences as a testing set, consisting of 50.000 machine and human translated sentences.

Note that for training, we use data where first a human translated sentence is presented followed by a machine translated sentence (i.e. not shuffled) in order to feed a balanced amount of data of both classes throughout the whole training procedure. Not doing this often leads to the model overfitting on the class it initially is presented with the most. The evaluation metric used in these experiments is accuracy, simply defined as:

$$\text{accuracy} = \frac{\text{number of correct predictions}}{\text{number of predictions}} \in [0,1] \qquad (1)$$

As there is no comparable research done on this specific task to the best of our knowledge, we consider the baseline in this work to be random chance for a binary classification task i.e. 50% accuracy.

V. RESULTS & DISCUSSION

The results for each of the BERT models for each of the datasets can be seen in Table IV.

There is a clear difference between the classification of the two languages. The accuracies for the Italian to English translations are rather poor and largely around chance level. This is not the case for the Dutch translations, where accuracy improvements can be observed in most levels of the BERT model complexities. This would suggest that the machine translation quality for Italian is higher than that of Dutch, and thus more difficult to classify.

The length of the input sequences appears to make a difference in performance of the models. We observe that for most cases the accuracy is higher for sentences with more than 5 words in contrast to using any sentence, especially for the larger models. This stems from the fact that excluding short sentences provides more context for both the machine translator as well as our models. Having more context allows for more differences in translations between human and machine translated texts. More context and more differences in translations supposedly allows for easier distinction, which would also be the case for a human discriminator. We also observed that for instance for single word inputs, the output probabilities of the models would be identical for any single word.

Another thing which can be observed from the result is slight improvements in some cases of the larger BERT models. This does not come as a surprise, as the larger models contain more parameters meaning that they capture larger amounts of information, allowing for better representations of the input sequences.

The possible shortcomings of the Italian translations are hard to pinpoint. During data preparation a very common problem was observed. Cases where the machine translation would actually be a closer and more accurate fit to the original subtitle, whereas the human translation would contain extra information/words which would be impossible for the machine translation to come up with. This is likely due to the nature of the data itself. An example of this can be seen in Table II, where the human translation has the additional "under any circumstances". More extreme examples were also present, where in some cases the human translations would include a lot of additional text. This is likely due to the nature of the data itself and the setting it is used in. One can imagine many cases where a human input would be needed to fit the flow or sequence of the subtitles so it would fit what is being presented on the screen better.

Another possible hint at a feature that would make distinction more difficult for the Italian data can be seen in Figures 1 and 2. Looking at the distributions of the translated subtitle lengths for both languages, we see larger differences in the Dutch case (top row) compared to the Italian (bottom row). In the Dutch dataset human translations tend to deviate more in length w.r.t. the source data, whereas the machine translations are generally closer to the source data, meaning that generally there is quite some difference between machine and human translated sentences. This isn't the case for the Italian data where a larger overlap in the translation distribution can be seen, thus making them harder to distinguish.

VI. CONCLUSION & FUTURE WORK

In conclusion, this research has presented some novel directions in the field of machine translation classification. This was achieved with the use of a non-standard corpora, as well as relying on deep learning methods for classification. Whereas previous research has focused on identifying specific linguistic features of "Translationese", on which popular machine learning methods (SVMs) were trained, we present an easier and quicker method with the use of BERT.

The research presented aimed solving the question of whether machine translated subtitles can be distinguished from human translated ones. While performing poorly on Italian, simply training BERT with subtitles for Dutch, with no preprocessing or extra features, yielded some interesting results, proving that the classification is indeed possible.

TABLE IV

RESULING ACCURACY SCORES FOR EACH OF THE BERT MODELS ON EACH OF THE DATASETS, WHERE 'LONG' DENOTES THE DATASETS WITH ONLY SENTENCES WITH MORE THAN SIX WORDS AND 'SHORT' DENOTES THE DATASETS WITH ALL SENTENCES

BERT model	IT→EN (long)	IT→EN (small)	NL→EN (long)	NL→EN (small)
Tiny	0.48	0.49	0.66	0.68
Small	0.48	0.46	0.72	0.68
Medium	0.49	0.47	0.73	0.68
Base	0.51	0.49	0.74	0.70

Sticking to the idea of "Translationese" dialects across languages, the results presented here would suggest a stronger dialect in the case of Dutch. In order words, there are more distinct features which would help with classification. Focusing further on the correctly predicted cases and identifying more concrete features of the language-specific dialect could further improve machine translations.

For future works, we propose the following suggestions that were not done due to time restrictions in this work. We suggest that larger BERT models are tried on this task as we expect those to obtain an even higher accuracy. We also suggest using other variants such as ELMO or simply different sets of parameters than used in this work. Another possible improvement would be adding a drop-out layer before the output layer (as done in [12]) such that the models can take longer to converge possibly providing better eventual performance.

Another future idea would be to change the data into small chunks of several subtitles grouped together, instead of simply presenting long and short sentences. This would fit nicely with the already outlined importance of assessing the wider scope [9], as well as the results presented here, as it seems that providing longer sentences (i.e more context) helps the classifier. Again, this outlines the shortcomings and difficulties of machine translation when faced with longer and more convoluted texts. The inclusions of more languages would be interesting too, as it is already evident that language-specific problems can occur.

Finally, finding a way to filter out subtitles which have had additional human "improvisation" would be a good improvement for the direction of this project. Having a dataset where one is sure that the human translations are as close as possible to the source subtitle, without additional information, would provide a basis for more comparable and reliable results.

References

[1] Roee Aharoni, Moshe Koppel, and Yoav Goldberg. Automatic detection of machine translated text and translation quality estimation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 289–295, 2014.

[2] 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, 2013.

[3] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. An effective approach to unsupervised machine translation. *arXiv preprint arXiv:1902.01313*, 2019.

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

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. Universal neural machine translation for extremely low resource languages. *arXiv preprint arXiv:1802.05368*, 2018.

[7] Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*, 2018.

[8] David Kurokawa, Cyril Goutte, Pierre Isabelle, et al. Automatic detection of translated text and its impact on machine translation. *Proceedings of MT-Summit XII*, pages 81–88, 2009.

[9] Samuel Läubli, Rico Sennrich, and Martin Volk. Has machine translation achieved human parity? a case for document-level evaluation. *arXiv preprint arXiv:1808.07048*, 2018.

[10] Yitong Li, Rui Wang, and Hai Zhao. A machine learning method to distinguish machine translation from human translation. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation: Posters*, pages 354–360, 2015.

[11] Pierre Lison and Jörg Tiedemann. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. 2016.

[12] Sushil Shakya Manish Munikar and Aakash Shrestha. Fine-grained sentiment classification using bert. 2019.

[13] Hoang-Quoc Nguyen-Son, Ngoc-Dung T Tieu, Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Identifying computer-translated paragraphs using coherence features. *arXiv preprint arXiv:1812.10896*, 2018.

[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[15] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[16] Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. Attaining the unattainable? reassessing claims of human parity in neural machine translation. *arXiv preprint arXiv:1808.10432*, 2018.

[17] Vered Volansky, Noam Ordan, and Shuly Wintner. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118, 2015.

Appendix

A. Contributions

In the following table you can see how each member of the group contributed to the project.

	Data collection	Code	Report
Federico Ferlito	25%	25%	25%
Nino Jansen	25%	25%	25%
Remo Sasso	25%	25%	25%
Boris Marinov	25%	25%	25%