Automatic Classification of Human Translation and Machine Translation: A Study from the Perspective of Lexical Diversity

Yingxue Fu
School of Computer Science
University of St Andrews
KY16 9SX, UK
yf30@st-andrews.ac.uk

Mark-Jan Nederhof
School of Computer Science
University of St Andrews
KY16 9SX, UK

Abstract

By using a trigram model and fine-tuning a pretrained BERT model for sequence classification, we show that machine translation and human translation can be classified with an accuracy above chance level, which suggests that machine translation and human translation are different in a systematic way. The classification accuracy of machine translation is much higher than of human translation. We show that this may be explained by the difference in lexical diversity between machine translation and human translation. If machine translation has independent patterns from human translation, automatic metrics which measure the deviation of machine translation from human translation may conflate difference with quality. Our experiment with two different types of automatic metrics shows correlation with the result of the classification task. Therefore, we suggest the difference in lexical diversity between machine translation and human translation be given more attention in machine translation evaluation.

1 Introduction

The initial interest in and support for machine translation (MT) stem from visions of high-speed and high-quality translation of arbitrary texts (Slocum, 1985), but machine translation proves to be more difficult than initially imagined. In recent years, progress has been made in MT research and development, and it is claimed that MT achieves human parity in some tasks (Wu et al., 2016; Hassan et al., 2018; Popel et al., 2020). However, these statements are challenged by other researchers and remain open to debate (Läubli et al., 2018; Toral et al., 2018; Toral, 2020).

The typical automatic approach to evaluating MT is to compare a machine translated text with a reference translation. The assumption is that the closer a machine translation is to a professional human translation, the better it is (Papineni et al., 2002). Automatic metrics for MT are developed based on this assumption. Human translation (HT) is treated as gold standard and the deviation from it is transformed into a measure of translation quality of MT.

Many studies have shown that translated texts are different from originally written texts (Baroni and Bernardini, 2006; Ilisei et al., 2010). The typical method used for the identification of translationese is automatic classification of translated texts and originally written texts (Baroni and Bernardini, 2006). There are some studies that compare translation varieties such as professional and student translations and post-edited MT (Kunilovskaya and Lapshinova-Koltunski, 2019; Toral, 2019; Popović, 2020). While surface linguistic features and simple machine learning techniques are capable of classifying translated texts and originally written texts with high accuracy, it is difficult to use the same method to classify translation varieties, with the accuracy being barely over the chance level (Kunilovskaya and Lapshinova-Koltunski, 2019; Rubino et al., 2016).

When comparing translation varieties, MT is used as a translation variety independent of HT or other translation varieties in some studies (Toral, 2019). Different from the conventional practice of MT evaluation that treats HT as the gold standard, some studies adopt a descriptive approach to comparing MT and HT (Bizzoni et al., 2020; Ahrenberg, 2017; Vanmassenhove et al., 2019). Among these studies, Bizzoni et al. (2020) find that MT shows independent patterns of translationese and it resembles HT only partly. This implies that MT may be different from HT in a systematic way, and it remains a question as to whether the deviation of MT from HT is a reliable measure of the quality of MT, and whether the current automatic metrics conflate differences between HT and MT with the quality of MT.

According to research by Toral (2019), translation varieties differ in multiple ways. Based on research by Vanmassenhove et al. (2019), we focus on lexical diversity in our experiments.

We try to answer three questions in this study:

- Can MT and HT be classified automatically with an accuracy above the chance level?
- In what way does lexical diversity influence the classification result?
- Are the results of automatic metrics influenced by the difference in lexical diversity between HT and MT?

2 Related Work

As our study essentially involves comparing translation varieties, we present an overview of previous studies that compare originally written texts and translations, other translation varieties, and HT and MT.

2.1 Comparing Originally Written Texts and Translations

Translated texts show distinctive features which make them different from originally written texts. These features are typically studied under the framework of translationese. Gellerstam (1986) is the first to use this term to refer to the ”fingerprints” that the source text leaves on the translated text. This notion is developed by Baker, who proposes the idea of universals of translation. As suggested by Baker et al. (1993), universals of translation are linguistic features that typically occur in translated texts as opposed to originally written texts, and these features are independent of the specific language pairs. Automatic means to distinguish translated texts and originally written texts have been developed and generally achieve high accuracy (Baroni and Bernardini, 2006; Ilisei et al., 2010; Lembersky et al., 2012; Rabinovich and Wintner, 2015). Meanwhile, computational approaches (Teich, 2003; Volansky et al., 2015) contribute evidence for some translation universals.

2.2 Comparing Translation Varieties

Compared with the considerable amount of research on identifying translationese, the differences between translation varieties are less studied.

Rubino et al. (2016) perform the classification between originally written texts and translations as well as between professional and student translations. They use surface features and distortion features which are inspired by quality estimation tasks, and surprisal and complexity features which are derived from information theory. Their experiment shows that originally written texts and professional translations are different mainly in terms of sequences of words, part-of-speech and syntactic tags, and originally written texts are closer to professional translations than to student translations. While the originally written texts and translations can be classified with high accuracy, automatic classification of different translation varieties is a more challenging task. Professional translations and student translations can only be classified with an accuracy barely above 50%.

This finding is consistent with the result of a study by Kunilovskaya and Lapshinova-Koltunski (2019). While morpho-syntactic features can be used to distinguish translations from non-translations with high accuracy, the performance of the same algorithm on classifying professional and student translations only slightly exceeds the chance level.

The differences of translations authored by human translators with different expertise and native languages are studied by Popović (2020). Similar to other studies on distinguishing originally written texts from translated texts or comparing translation varieties, surface text features at word and part-of-speech levels are used. It concludes by suggesting that detailed information about the reference translation including translator information be provided in the scenario of MT evaluation.

Toral (2019) compares post-edited MT with HT in terms of lexical variety, lexical density, sentence length ratio and part-of-speech sequences. The research shows that post-edited MT has lower lexical diversity and lower lexical density than HT, which is linked to the translation universal of simplification, and post-edited MT is more normalized and has greater interference from the source text (in terms of sentence length and part-of-speech sequences) than HT.

2.3 Comparing MT and HT

While the number of studies on comparing translation varieties is much smaller than on the identification of translationese, there are even fewer studies that explore the differences between MT and HT.

Ahrenberg (2017) compares MT and HT by means of automatically extracted features and statistics obtained through manual examination. By comparing the shifts (i.e. deviation from literal translation) and word order changes, he finds that HT contains twice as many word order changes. Meanwhile, an analysis of the number and types of edits required to give the machine translated text publishable quality is made. He argues that MT is likely to retain interference from the source text even after post-editing, and the machine translated text is more similar to the source text than the human translated text in many ways, including sentence length, information flow and structure.

Research by Vanmassenhove et al. (2019) shows another aspect where MT differs from HT. Three MT systems based on different architectures are trained. The lexical diversity of the translations of the MT systems is measured with three metrics including type/token ratio, Yule's K, and measure of textual lexical diversity (MTLD). It is found that the output of neural machine translation (NMT) systems has a loss of lexical diversity compared with the human translated text. The reason for this phenomenon is that the advantage of NMT systems over statistical machine translation (SMT) systems in terms of learning over the entire sequence is obtained at the expense of discarding less frequently occurring words or morphological forms. This finding is consistent with the research by Toral (2019), who observes that the lexical variety of post-edited MT is lower than of HT and the lexical variety of MT is lower than of post-edited MT, which is attributed to the tendency of MT to choose words used more frequently in the training data (Farrell, 2018).

Bizzoni et al. (2020) study the differences between HT and MT in relation to the original texts. Part-of-speech perplexity and a syntactic distance metric are used to measure the differences between translations in written and spoken forms and produced by different types of MT systems. It is found that MT shows structural translationese, but the translationese of MT follows independent patterns that need further understanding.

3 Experiment

We adopt two approaches for classifying MT and HT: developing a trigram language model with Witten-Bell smoothing and fine-tuning a pretrained BERT model for sequence classification from the Transformers library (Wolf et al., 2020).

3.1 Data

The dataset is from the News commentary parallel corpus v13 (Tiedemann, 2012) provided in the WMT2018 shared task[1]. We use Google Translate[2] to obtain the corresponding machine translation.

The language pairs used in the experiment, the number of sentences for each language pair and the average sentence length for HT and MT are presented in Table 1.

	Number of sentences	**MT avg sentence length**	**HT avg sentence length**
CS-EN	30384	26.33	25.83
DE-EN	30345	26.61	26.15
RU-EN	30387	28.00	27.51

Table 1: Statistics of the dataset: translations from Czech, German and Russian to English.

3.2 Classifying HT and MT

Trigram Model

We train two trigram models on the HT and MT training sets. Let p_{MT} denote the trigram model trained on MT sentences, and p_{HT} the model trained on HT sentences. A sentence s is classified as MT if $p_{MT}(s) > p_{HT}(s)$ and as HT otherwise. If s is from the HT test set and classified as HT, we count it as a success, and the same goes for the case when s is from the MT test set and classified as MT. The classification accuracy is obtained by dividing the number of correct classifications by the total number of sentences in the respective test set. Since the two classes are balanced, accuracy is an appropriate metric. The result is shown in Table 2.

[1] http://www.statmt.org/wmt18/translation-task.html

[2] https://translate.google.co.uk

CS-EN		
Total	MT	HT
0.69	0.79	0.58
DE-EN		
Total	MT	HT
0.66	0.75	0.57
RU-EN		
Total	MT	HT
0.67	0.76	0.58

Table 2: Classification accuracy of the trigram model.

From Table 2 it is clear that HT and MT can be classified automatically with an accuracy above the chance level. However, it is noticeable that MT can be classified with higher accuracy than HT.

Based on research by Vanmassenhove et al. (2019) and Toral (2019), this imbalance in classification accuracy may be partly explained by the higher lexical diversity of HT, so that p_{HT} is a probability distribution over sentences composed of a larger set of words than in the case of p_{MT}, thereby typically assigning a lower probability to any particular sentence, regardless of whether it is from MT or from HT.

From Table 1, it can be seen that the difference in average sentence length between MT and HT is only around 0.5. Therefore, we assume that the influence of sentence length is not significant in this study.

BERT Model

We apply the BERT model on the same dataset, which is divided into training, test and validation sets by the ratio of 70%, 10% and 20%. The sentences are padded to the maximum length of sentences in the dataset. We find that the pretrained BERT model for sequence classification achieves higher accuracy and lower loss in the first epoch. The result is shown in Table 3.

From Table 3, it can be seen that fine-tuning the pretrained BERT model for sequence classification can achieve higher accuracy for this task than the trigram model. Moreover, we can see the same pattern of imbalance in classification accuracy between MT and HT. Similar to the case of the trigram model, we hypothesize that it is because greater lexical diversity makes HT more difficult to classify correctly than MT.

CS-EN		
Total	MT	HT
0.78	0.90	0.66
DE-EN		
Total	MT	HT
0.78	0.87	0.69
RU-EN		
Total	MT	HT
0.78	0.90	0.65

Table 3: Classification accuracy of the BERT model.

3.3 Changing Lexical Diversity

To investigate further whether differences in lexical diversity could be the reason for the observed imbalance in the classification accuracy of MT and HT, we manipulate the lexical diversity of the two. As the lexical diversity of HT is generally higher than of MT (Vanmassenhove et al., 2019; Toral, 2019), we reduce the lexical diversity of HT until it becomes close to or lower than MT, and for comparison, we also reduce the lexical diversity of MT.

Method of Changing Lexical Diversity

Our general strategy of reducing lexical diversity is to replace rare words with words that are close to them in a vector space. First, we find rare words based on the frequency of lemmas in the corpus. Since there are many numerals and proper names and it is difficult to find meaningful candidates to replace them in the vector space, we set token.like_num and token.is_oov in spaCy processing[3] to false. Among the remaining lemmas, those lemmas whose frequency is lower than a threshold will be considered to be rare words. We found that setting the frequency threshold to two is effective in reducing the lexical diversity.

Second, we choose words whose vectors are close to the rare words from the pretrained GloVe embeddings (Pennington et al., 2014), which are computationally less expensive than contextualized word embeddings like BERT. We found that the words which are closest to the rare words are not necessarily the optimal candidates in terms of part-of-speech or meaning, and so we choose the top three most similar words for each rare word. We convert the GloVe vectors into word2vec for-

[3] https://spacy.io

mat with the gensim glove2word2vec API[4] and set restrict_vocab to 30000 in the most_similar function[5] so that the search for the most similar words is limited to the top 30000 words in the pretrained embeddings. The vocabulary size 30000 was determined empirically.

After this step, we apply a check on the fine-grained tags of the rare words and the fine-grained tags of the respective three candidates, the tags being obtained with spaCy [6] and containing more information than the coarse-grained part-of-speech tags from the Universal POS tag set[7]. The candidates with the same tags as the rare words will be chosen. Where there is more than one matched candidate, only the first is chosen, and when there are no matched candidates after the check, the rare words will not be replaced. In this way, we obtain texts with modified lexical diversity. For ease of reference, modified HT texts will be referred to as HT_modf, modified MT texts will be referred to as MT_modf, original HT texts as HT_orig and original MT texts as MT_orig.

To compute the lexical diversity of the texts, based on research by McCarthy and Jarvis (2010) and Vanmassenhove et al. (2019), we choose the measure of textual lexical diversity (MTLD) (McCarthy, 2005), which is reasonably robust to text length difference. We refer those interested in the specific computation and statistical significance of MTLD to McCarthy and Jarvis (2010). The lexical diversity of the texts is presented in Table 4.

MTLD	**Original**	**Modified**
CS_MT	62.02	43.00
CS_HT	63.80	43.04
DE_MT	62.53	42.44
DE_HT	64.59	42.76
RU_MT	61.06	42.66
RU_HT	64.51	43.05

Table 4: MTLD of the original texts and of the modified texts.

From Table 4, it can be seen that the MTLD values of HT texts are generally higher than of MT texts, which is consistent with the result of previous studies (Vanmassenhove et al., 2019, 2021; Toral, 2019). With our method, the difference in MTLD value between MT and HT texts is reduced.

Experimental Result of Trigram Model

We conduct another set of binary classification experiments on the original and modified MT and HT texts paired in different ways. For example, "MT_modf & HT_modf" in the following tables means that the binary classification is performed on the modified MT text and the modified HT text. The result of the trigram model is shown in Table 5. For comparison, the results from Table 2 are repeated in the lines MT_orig & HT_orig.

CS-EN			
Accuracy	Total	MT	HT
MT_orig & HT_orig	0.69	0.79	0.58
MT_modf & HT_modf	0.69	0.77	0.61
MT_orig & HT_modf	0.69	0.56	0.83
DE-EN			
Accuracy	Total	MT	HT
MT_orig & HT_orig	0.66	0.75	0.57
MT_modf & HT_modf	0.67	0.74	0.60
MT_orig & HT_modf	0.67	0.52	0.82
RU-EN			
Accuracy	Total	MT	HT
MT_orig & HT_orig	0.67	0.76	0.58
MT_modf & HT_modf	0.67	0.75	0.59
MT_orig & HT_modf	0.67	0.52	0.82

Table 5: Binary classification of MT and HT by the trigram model under different combinations of MT and HT texts.

From Table 5 in combination with Table 4, we can see that when the difference in lexical diversity between MT and HT becomes smaller, the imbalance in classification accuracy is reduced, and the classification accuracy of MT goes down while the classification accuracy of HT goes up.

Since the lexical diversity of HT is generally higher than MT, we conduct an experiment where the lexical diversity of HT is significantly lower than MT, and the result is shown in the lines MT_orig & HT_modf. Under this condition, the classification accuracy of MT is much lower than HT. In this way, we reverse the previously observed trend that the classification accuracy of MT is higher than HT. Note that the overall classification accuracy does not change much in this experiment.

[4] https://radimrehurek.com/gensim/scripts/glove2word2vec.html

[5] https://radimrehurek.com/gensim/models/word2vec.html

[6] https://spacy.io/api/token#attributes

[7] https://universaldependencies.org/docs/u/pos/

Experimental Result of BERT Model

For fine-tuning the pretrained BERT model for sequence classification, similar experiments were done, with different combinations of MT and HT texts. Accuracies are presented in Table 6.

CS-EN			
Accuracy	Total	MT	HT
MT_orig & HT_orig	0.78	0.90	0.66
MT_modf & HT_modf	0.78	0.89	0.68
MT_orig & HT_modf	0.82	0.91	0.73
DE-EN			
Accuracy	Total	MT	HT
MT_orig & HT_orig	0.78	0.87	0.69
MT_modf & HT_modf	0.78	0.86	0.71
MT_orig & HT_modf	0.81	0.89	0.73
RU-EN			
Accuracy	Total	MT	HT
MT_orig & HT_orig	0.78	0.90	0.65
MT_modf & HT_modf	0.77	0.89	0.65
MT_orig & HT_modf	0.81	0.95	0.68

Table 6: Binary classification of MT and HT by the BERT model under different combinations of MT and HT texts.

Similar to the trigram model, the classification accuracy of HT goes up in the case of CS-EN and DE-EN and the classification accuracy of MT goes down a little, when the lexical diversity of MT and of HT are closer, as shown in the lines MT_modf & HT_modf, and when the lexical diversity of HT is much lower than MT, the classification accuracy of HT goes up, as shown in the lines MT_orig & HT_modf. However, changing the difference in lexical diversity does not tend to decrease the classification accuracy of MT for the BERT model. Recall that with the trigram model, the classification accuracy of HT increases while the classification accuracy of MT decreases. In contrast, with the BERT model, even when the lexical diversity of MT is much higher than HT, the overall classification accuracy and the separate classification accuracies of MT and HT all go up. The difference of the two models in terms of the classification accuracy of MT may be explained by the fact that the pretrained BERT model for sequence classification calculates cross-entropy loss for the classification task[8] while the trigram model results from relative frequency estimation.

3.4 Automatic Metrics

We hypothesize that the performance of the two models in the binary classification task may be reflected in the result of MT metrics that are based on n-gram matching or that use contexualized embeddings.

Since BLEU is a commonly used metric based on n-gram matching, we test the performance of BLEU on the dataset to see if the difference in lexical diversity between MT and HT would influence the result. We calculate the corpus-level BLEU score for MT, as implemented in NLTK[9], using HT as reference. The result is presented in Table 7.

BLEU	MT_orig & HT_orig	MT_modf & HT_modf	MT_orig & HT_modf
CS-EN	0.42	0.46	0.39
DE-EN	0.41	0.45	0.38
RU-EN	0.37	0.40	0.34

Table 7: BLEU score.

As can be seen from Table 7, when the lexical diversity of MT is closest to HT, as shown by the column MT_modf & HT_modf, the MT BLEU score is the highest. When the lexical diversity of the reference is much lower than MT, as is the case in the column MT_orig & HT_modf, the MT BLEU score is the lowest. Much as in the discussion of the results of the trigram model, the difference in lexical diversity between MT and HT is a factor that needs to be taken into account when an n-gram matching based metric like BLEU is used for MT evaluation.

The majority of automatic MT metrics developed in recent years such as BERTScore (Zhang et al., 2019) and Yisi (Lo, 2019) adopt contextualized embeddings. Based on accessibility and performance, we choose MoverScore (Zhao et al., 2019) as an example of a metric that uses BERT representations. Since MoverScore is not a corpus-level metric, we calculate the average

[8] https://github.com/huggingface/transformers/blob/9aeacb58bab321bc21c24bbdf7a24efdccb1d426/src/transformers/modeling_bert.py

[9] https://www.nltk.org/

sentence-level score. The result is presented in Table 8.

Mover-Score	MT_orig & HT_orig	MT_modf & HT_modf	MT_orig & HT_modf
CS-EN	0.57	0.56	0.55
DE-EN	0.57	0.56	0.55
RU-EN	0.52	0.50	0.50

Table 8: MoverScore result for MT.

The MoverScore result in Table 8 shows a different pattern from the BLEU scores. The scores are basically inversely proportional to the overall accuracy of the binary classification task shown in Table 6. As the difference in MoverScore results under different combinations of MT and HT texts is small, more work is needed.

4 Conclusion and Future Work

With the above experiments, we have shown that MT and HT can be classified with an accuracy above the chance level. The trigram model does not involve a machine learning algorithm but is capable of capturing the differences between MT and HT. By fine-tuning the pretrained BERT model for sequence classification, we obtain a higher accuracy for this task.

Similar to the identification of translationese, we may claim that MT and HT belong to different translation varieties. The result serves as supporting evidence for the study by Bizzoni et al. (2020), which maintains that MT only resembles HT in part and often follows independent patterns. This finding calls into question the longstanding assumption in MT evaluation that the more similar an MT output is to a professional human translation, the better it is. If MT and HT are two translation varieties and have different patterns, it leaves room for doubt as to the legitimacy of evaluating MT by its similarity to HT.

Moreover, there is a noticeable imbalance in the classification accuracy of HT and MT. For the trigram model, while more than 70% of the MT test sentences can be classified correctly, fewer than 60% of the HT test sentences are classified correctly. This imbalance also exists in the experiment with the BERT model. Generally speaking, it is easier to correctly classify MT sentences than HT sentences.

Based on previous studies and analysis from the probabilistic perspective, we consider lexical diversity as one of the major reasons for this imbalance in classification accuracy. We change the lexical diversity of the MT and HT texts and conduct another set of experiments with the same models. With the trigram model, if the difference in lexical diversity between MT and HT decreases, the imbalance in classification accuracy between the two is reduced, and we can reverse this imbalance in classification accuracy when the lexical diversity of MT is higher than HT. The result of the experiment with the BERT model shows a different pattern. An increase in classification accuracy of HT is accompanied by an increase in the classification accuracy of MT. This may be explained by the different ways of performing binary classification by the two models.

The performance of automatic MT metrics based on n-gram matching, represented by BLEU in this study, and automatic metrics using BERT representations, such as MoverScore, is related to the result of the binary classification task with the two kinds of models. When the lexical diversity of HT is lower than MT, the MT BLEU score is the lowest and when the lexical diversity of HT is very close to MT, the MT BLEU score is the highest. The evaluation results given by MoverScore are basically inversely proportional to the classification accuracy of the BERT model. Therefore, we suggest the difference in lexical diversity between MT and the reference be given more attention in MT evaluation with automatic metrics.

We are aware that there are other possible factors that may account for the phenomenon that HT is more likely to be classified as MT than the other way around. In our experiment, we only manipulate one factor. In future work, we intend to further study the independent patterns of MT compared with HT and investigate if the differences between MT and HT are related to the quality of MT. As differences in lexical diversity may influence automatic metrics for MT evaluation in different ways, we plan to explore this phenomenon with other metrics, such as COMET (Rei et al., 2020).

References

Lars Ahrenberg. 2017. Comparing machine translation and human translation: A case study. In *RANLP 2017: The First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT)*, pages 21–28. Association for Computational Linguistics.

Mona Baker et al. 1993. Corpus linguistics and translation studies: Implications and applications. *Text and Technology: In honour of John Sinclair*, 233:250.

Marco Baroni and Silvia Bernardini. 2006. 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.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290, Online. Association for Computational Linguistics.

Michael Farrell. 2018. Machine translation markers in post-edited machine translation output. In *Proceedings of the 40th Conference Translating and the Computer*, pages 50–59.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english. *Translation Studies in Scandinavia*, 1:88–95.

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

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2010. Identification of translationese: A machine learning approach. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 503–511. Springer.

Maria Kunilovskaya and Ekaterina Lapshinova-Koltunski. 2019. Translationese features as indicators of quality in english-russian human translation. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 47–56.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38(4):799–825.

Chi-kiu Lo. 2019. Yisi-a unified semantic mt quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513.

Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):1–15.

Maja Popović. 2020. On the differences between human translations. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 365–374.

Ella Rabinovich and Shuly Wintner. 2015. Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Raphael Rubino, Ekaterina Lapshinova-Koltunski, and Josef Van Genabith. 2016. Information density and quality estimation features as translationese indicators for human translation classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 960–970.

Jonathan Slocum. 1985. A survey of machine translation: Its history, current status and future prospects. *Computational linguistics*, 11(1):1–17.

Elke Teich. 2003. *Cross-linguistic variation in system and text: A methodology for the investigation of translations and comparable texts*, volume 5. Walter de Gruyter.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *LREC*, volume 2012, pages 2214–2218.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 273–281.

Antonio Toral. 2020. Reassessing claims of human parity and super-human performance in machine translation at wmt 2019. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 185–194.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.