Improving Semantic Similarity Calculation of Japanese Text for MT Evaluation

Yuki Tanahashi
WANTS Inc.
Nagoya, Japan
tanahashi236@wantsinc.jp

Kyoko Kanzaki
Toyohashi University of Technology
Toyohashi, Japan
kanzaki@cite.tut.ac.jp

Eiko Yamamoto
Gifu Shotoku Gakuen University
Gifu, Japan
eiko@gifu.shotoku.ac.jp

Hitoshi Isahara
Toyohashi University of Technology
Toyohashi, Japan
isahara@tut.jp

Abstract

In recent years, the quality of machine translation has significantly improved, and it is considered that translation of a novel becomes possible. However, when performing translation for novels by a machine translation, it needs an evaluation method that performs comparison using a vector of distributed representations. BERTScore is one of the methods for performing automatic evaluation using a distributed representation. In this research, we examined the optimal setting for applying the BERTScore to the evaluation of translations of Japanese novels, and improved the method by introducing penalties for named entities based on idf values calculated from large corpora. The introduction of the penalty has made it possible to mitigate the false matching of personal names caused by distributed representation. We verified the method by calculating the Pearson correlation between the modified BERTScore and human-rated scores. Furthermore, we set four BERT models and two kinds of corpora to calculate idf value, and investigated which setting is most suitable for evaluation of novel translation. As a result, the setting with the model based on novel corpus, the idf based novel corpus and the penalty had the highest correlation with human-rated scores.

1 Introduction

In recent years, machine translation quality has dramatically improved due to the development of neural translation models that utilize deep learning, such as the sequence transformation model (Sutskever et al., 2014) and the attention model (Dzmitry et al., 2015. Luong et al., 2015), which is an application of the attention mechanism, and improved computer performance. Due to these improvements, not only documents consisting of formal expressions such as patent sentences and academic papers that have been fixed to some extent but also informal expressions such as novels and colloquial expressions could be machine translated. However, previous research has revealed that problems that have not been considered as important in machine translation research so far have a great influence on the learning and results of novel translation. Among them, the variety of text expressions is a serious problem. The problem is that when the author or translator in a novel is different, or even if the same author/translator has a different story speaker, an English sentence is translated into a Japanese sentence with a distinctly different translation but with a similar meaning.

Specific examples are shown below.

English	Japanese
My name is John.	俺はジョンという。 (I am called John.)
My name is Maria	私の名前はマリアよ。 (My name is Maria.)

When a language resource (corpus) containing such parallel translations is used as learning data in neural translation that receives the whole sentence as input and learns so as to maximize the likelihood that a correct word sequence will be output, learning becomes difficult and the output of the translation system becomes unstable. In addition, different expressions cause problems not only in learning but also in translation performance evaluation. In machine translation, BLEU (Papineni et al, 2002) is the de facto standard as an automatic evaluation method for evaluating the performance of translation systems in many previous studies. BLEU is a n-gram matching that scores the translation quality between 0.0 and 1.0 by counting the number of matching words n-grams between the reference sentence that is the human-translated correct data and the sentence to be evaluated output by the translation system. BLEU is used in many machine translation studies because it is a simple and easy-to-interpret method, but it is necessary that the surface text strings of the words in the reference sentence and the sentence to be evaluated are exactly the same.

Therefore, even if two sentences appear to be semantically identical to each other by humans, BLEU gives a low rating if the words used are different (different expressions). In novel translation, where the description of expressions is likely to be different when the translator and the speaker in the story are different for a certain English sentence, even if there is a system that can translate high quality, BLEU will not perform correctly.

As described above, an automatic evaluation method such as BLEU that considers only the surface of a sentence cannot correctly evaluate two sentences that have different expressions but have similar meanings. Therefore, in order to facilitate future novel translation research in Japanese, it is necessary to consider an automatic evaluation method that can accurately evaluate novel sentences from a certain language to Japanese before developing a translation model.

We apply BERTScore (Zhang et al., 2019) to the semantic similarity evaluation of Japanese novels. BERTScore uses BERT (Devlin et al., 2018) which generates a general-purpose linguistic expression for automatic sentence similarity evaluation. In applying the BERTScore, we proposed a modification that reduces the problem of similarity calculation in Japanese novels by imposing the editing distance of word reading (pronunciation) as a penalty. In addition, in order to adapt the BERTScore to novel evaluation, we constructed a BERT pre-learning model using a monolingual novel corpus consisting of sentences collected from the novel posting site. This model, and other existing BERT pre-learning model were applied to BERTScore to investigate which model is most suitable for novel evaluation.

The contributions of this research are the following three points.

1. Investigation of optimal settings for applying BERTScore to Japanese novels
2. Clarification and correction of problems that occur when calculating the similarity of novel Japanese sentences using BERTScore
3. Construction of BERT pre-learning model using large-scale novel corpus

2. Applying BERTScore to Japanese

Zhang et al. conducted experiments on BERTScore using the test set provided in the Metric Shared Task of WMT2017, and confirmed its usefulness in sentence pair evaluation on the English side in several language pairs. However, its usefulness has not yet been verified in Japanese sentences in English-Japanese language pairs. Therefore, in this research, we search for the optimal setting for applying BERTScore to the sentence pairs on the Japanese side in English-Japanese language pairs.

Since this study has the goal of improving the translation evaluation of English-Japanese novel translations, we consider its application especially to the evaluation of Japanese sentences in novel sentences. In other words, we consider a method to correctly evaluate an example in which the reference sentence and the sentence to be evaluated have different expressions but the same meaning. Such differences in expressions can be absorbed to some extent by using a distributed expression vector optimized for the meaning of words created by BERT. However, the use of distributed expressions by BERT causes another problem for the evaluation of sentences that include proper nouns such as person names and place names. In BERT, when considering the meaning of a word, a vector is defined by surrounding words and their arrangement. Consider the following two sentences;

> "Mr. Tanaka bought a bottle of juice."
> "Mr. Sato bought a bottle of juice."

These two sentences represent distinctly different situations for humans. Because the nouns that are the subject are different, i.e. Mr. Tanaka and Mr. Sato, these clearly indicate another person. However, since these sentences have the same sentence structure of "someone", "juice", and "buy", the words around "Tanaka" and "Sato" that correspond to "someone" are the same. Therefore, the vectors of the distributed expressions for the proper nouns "Tanaka" and "Sato" can be relatively close. However, when these sentences appear in a novel, the difference in proper nouns such as a person's name or a place's name greatly affects the reading comprehension of the story. The two sentences above need to be clearly distinguished.

Also, proper names such as person names and place names appear in the corpus less frequently than ordinary nouns and verbs. Therefore, there is a high possibility that it will be out of vocabulary and will be treated as unknown words. The translation system is likely to output incorrect translations for such person names and place names. The possibility of mistranslation is even higher when translating unusual or fictitious names of people or places, or when the translation system itself is learned from a low-resource corpus.

3 Penalty by Edit Distance

In the evaluation of Japanese sentences in novel translation, it is necessary to give a low evaluation value if the reference sentence and the sentence to be evaluated have different corresponding named entity expressions. In particular, we deal with proper nouns in which the flow of the story collapses due to incorrect output of person names and place names among proper expressions.

In this study, the edit distance (Gusfield, 1997) between the tokens that forms the proper nouns in two sentences is calculated, and that value is reflected as a penalty in the final evaluation. The edit distance is a distance indicating how similar two strings are. The editing process that inserts, deletes, and replaces one character in one of the two character strings is repeated until it matches the other character string. Then, the minimum number of processes required to match two character strings is defined as the edit distance.

Since proper nouns appear in the corpus less frequently than general words, in the following steps, low-frequency tokens are assumed and treated as tokens that make up proper nouns. BERTScore calculates idf values to pay attention to the characteristic expressions of sentences. In this study, all tokens whose idf value exceeds the threshold are treated as LFT (Low requent Token). Then, referring to Maximum Similarity, we obtain the paired token of the sentence on the other side that maximizes the cosine similarity for all the low-frequency tokens included in the sentence. For the low-frequency token set thus obtained and the token set of the pair corresponding to the low-frequency token set, the edit distance is calculated after converting them into Japanese pronunciation character, hiragana. As a result, it is possible to give a high editing distance to low frequency words with completely different meanings, and conversely, a low editing distance can be given to the orthographic variants in the same low frequency token, i.e. kanji, katakana, and hiragana in Japanese.

In this study, the edit distance obtained in this way is divided by the longer of the reading lengths of low-frequency tokens, normalized to the range [0,1], and the value is subtracted from 1. Then the edit distance matchs coefficient M expressed between [0,1]. If the concordance coefficient is close to 0, it indicates disagreement, and if it is close to 1, it indicates coincidence. The agreement coefficient M obtained from the edit distance between two sentences is shown in the following Equation.

$$M = 1 - \frac{EditDist(x^{(LFT)}, \hat{x}^{(LFT)})}{max(|x^{(LFT)}|, |\hat{x}^{(LFT)}|)}$$

Here, x(LFT) represents the entire reading string of the low-frequency token in the reference sentence. x^(LFT) represents the entire string of reading of infrequent tokens in the sentence to be evaluated. EditDist(s1,s2) represents the edit distance between sentences s1 and s2.

Then, the matching coefficient is set for all tokens in the evaluation target sentence. A value obtained by dividing the sum of these by the length of the sentence is reflected as a penalty in the precision and recall in BERTScore. However, the concordance coefficient is set to 1 for tokens for which the idf value does not exceed the threshold and is not a low-frequency token. The respective penalties are shown in the following Equations.

$$Pnl^{(P)} = \frac{\sum_{\hat{x}_i \in \hat{x}} p(\hat{x}_i)}{|\hat{x}|}$$

$$Pnl^{(R)} = \frac{\sum_{x_i \in x} p(x_i)}{|x|}$$

$$p(x) = \begin{cases} M & (idf(x) > threshold) \\ 1 & (otherwise) \end{cases}$$

By these penalties, BERTScore is modified as follows.

$$P_{rev} = P_{BERT} \times Pnl^{(P)}$$

$$R_{rev} = R_{BERT} \times Pnl^{(R)}$$

$$F_{rev} = 2\frac{P_{rev} \cdot R_{rev}}{P_{rev} + R_{rev}}$$

This makes it possible to give an appropriate penalty to the similarity evaluation of two sentences that are mostly similar but differ only in proper nouns.

4 Experiments

In this section, we explain the test set and procedure in the experiment to evaluate the performance of the proposed method, and explain various tools related to the experiment. In order to evaluate the performance of the proposed method, we first performed a preliminary experiment under multiple settings and confirmed the setting with the best performance. Then, we conducted a verification experiment comparing it with the existing method to examine whether the proposed method is effective. In order to distinguish the similarity between tokens based on the cosine similarity, the evaluation on how the two sentences are semantically similar, performed by the human or the automatic evaluation system, is called “similarity score”.

4.1 Data

The experimental data and the experimental procedure are based on the evaluation sharing task (Metric Shared Task) (Ma et al, 2018) in WMT. Verification experiments to be described later and comparison experiments with other existing methods were also performed under the same experimental settings.

For the experiment, we extracted 100 sentences, 10 sentences each from 10 novels randomly selected from Project Gutenberg Canada (http://gutenberg.ca/index.html). Then, we asked expert translator to translate the extracted English sentences into Japanese. The 100 translated Japanese sentences obtained were used as the reference sentences in the test set.

When translating, we requested that the translation be performed based on the fact that it was one sentence included in the novel, but on the other hand, we instructed that the information carried over across sentences would not be added. The purpose of this research is to properly evaluate the variety and difference of expressions in novels. It is not intended to evaluate other characteristics included in the novel (such as abbreviations of words and free translations of pronouns considering the preamble). For this reason, the ten sentences selected from one novel were chosen so that the connections of the stories were located far enough away from each other so that they could not be seen as much as possible. We shuffled 100 sentences and asked for translation. We translated these 100 English sentences by Google translation and by our novel translation system using the novel corpus developed by ourselves, i.e. we obtained two types of machine translated Japanese sentences for each English sentence. In the output set obtained, some of the expressions such as personal names, unknown words, and other apparently broken sentences were modified.

The reason for making modifications to translations is that it is difficult for current machine translation systems to translate novel sentences in high quality, so almost all sentences may be classified as low quality by humans. However, in order to properly measure the performance of the proposed method, a high score is given to the evaluation target sentence judged by the human evaluator to be good, and a low score is given to the evaluation target sentence judged to be bad. Even if the proposed method gives a high correlation to a dataset in which all sentences are evaluated low by humans, it cannot be judged that they have been evaluated correctly. Therefore, we intentionally mix high-quality translations to avoid the judgment that all sentences are of low quality. Here, the definition of a good evaluation target sentence is a sentence whose meaning is similar to that of the reference sentence (regardless of difference in expression), and the definition of a bad evaluation target sentence is a sentence whose meaning is different from reference sentence, e.g., a sentence whose subject of action is different.

The reason why two types of sentences are prepared for one reference sentence by two translation systems is to confirm the correlation of evaluations between humans and systems for a large number of variations of expressions, and to check how the evaluations change by correcting the person's name etc. in one sentence.

Since one reference sentence and two evaluation sentences were set for one English sentence, a total of 200 reference sentence and evaluation sentence pairs were obtained. For these pairs, we performed a human evaluation task to evaluate the validity of the content of the reference sentence in the sentence to be evaluated. Using three Japanese native speakers as evaluators, we instructed to score how much the “contents/meanings” of the given pairs match, using a value between 0 and 100.

In scoring, we did not consider the fluency of the sentence to be evaluated, and requested to evaluate in the same way as a fluent sentence if the meaning was unbroken. In addition, even if the reference sentence and the sentence to be evaluated have similar meanings, we asked for a low score if the subject and/or object of the action were different. Since the reference sentence and the sentence to be evaluated are sentences in the novel, different proper nouns indicate different situations.

Normalization was performed to eliminate bias in the scoring of the three evaluators. We calculated the Pearson's correlation coefficient r, which examines the linear correlation between the two variables, and confirmed how well the scores of the evaluators agree. The correlation coefficient r was obtained from following Equation. Here, x represents the average value of the scores evaluated by x.

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

Since the Pearson's correlation coefficient is an index for evaluating the correlation between two variables, we calculated the correlation for every two evaluators. As a result of the calculation, the average value of the correlation was 0.657 to 0.548, and it was confirmed that the evaluations of the three parties were in good agreement. Therefore, we averaged the evaluation scores of each of the three evaluators, and defined them as the correct answer score by human evaluation in the test set.

In the following preliminary and verification experiments, the human correct answer scores obtained by the above procedure are used as references.

For the system scores obtained by the BERTScore and the existing method, we confirmed the agreement between the human evaluation and the evaluation by using Pearson's correlation coefficient.

4.2 Settings

In the BERTScore, which is the basis of the proposed method, the BERT pre-learning model used, the difference in the document set for which the idf value is calculated, and the presence or absence of a penalty affect the score. Therefore, it is necessary to confirm which of these settings has the best score.

In this study, we prepared following four pre-learning models for comparison when generating a token vector using BERT.

✓ Multilingual Model
https://github.com/google-research/bert/

✓ Wikipedia Model
http://nlp.ist.i.kyoto-u.ac.jp/

✓ SNS Model
https://github.com/hottolink/hottoSNS-bert

✓ Novel Model
Novel Model was developed by ourselves.

The BERT pre-learning model with Wikipedia as the learning corpus uses a very large scale of data for distributed representation. Although optimization is possible, almost all sentences in the learning corpus are composed of formal expressions, so they are not compatible with novels. Especially, since the first and second person are rarely included, there is a possibility that learning is not fully optimized for these words that appear frequently in novels.

On the other hand, the BERT pre-learning model, which uses a group of sentences posted on SNS as a learning corpus, has many colloquial sentences, and it is presumed that relatively many first-person and second-person sentences are included compared to Wikipedia sentences. However, due to the characteristics of media such as SNS, it contains unnecessary information that cannot be seen in novels such as emoticons and URLs.

From such a background, texts which includes colloquial expressions rather than Wikipedia and is less than SNS, i.e. positioning model including intermediate expressions in colloquialism, were needed. Therefore, the novel was used as the learning corpus for BERT pre-learning model. We collected 6,876,198 novel sentences from the novel submission site "小説家になろう, (Become a Novelist)". Morphological analysis and subword conversion were applied the sentences. BERT was trained using this learning corpus.

The list that defines the correspondence between each token and idf value is called the idf dictionary. In BERTScore, which is the basis of the proposed method, each sentence included in the reference side that is the correct answer in the sentence pair set to be evaluated is regarded as one document, and the idf value for each token is calculated. However, with this method, if tokens that should be judged to be a unique expression such as a person's name frequently appear on the referrer side, the idf value may be lower than intended and it may not be regarded as a unique expression.

Therefore, in this study, we prepared a special document set separately from the test set, and modified it so that high idf values could be given to unusual words by calculating idf values using them. As a large-scale corpus for idf value calculation, we utilized a Wiki corpus consisting of all 882892 documents acquired from Japanese Wikipedia and 229 documents obtained from novel corpus, and examined how different idf dictionaries affect the score.

The reason why we prepared two kinds of corpora is that the Wiki Corpus has articles that describe various things and can cover a wider range of vocabulary. Novel corpus has many novel-specific expressions such as first person and second person that are rarely seen in Wikipedia and can give a low idf value to the word. Therefore, such words will be not mistakenly processed as a low frequency token.

Throughout the experiment, the threshold value of idf to be processed was set as follows.

1. Sort all token and idf value pairs obtained from large corpus by idf value.
2. Tokens are divided into 70% and 30% of the whole, and 70% tokens are not processed and 30% takens are processed
3. Set the threshold value at the boundary between two groups.

4.3 Preliminary Experiment

In the preliminary experiments, four types of prelearning models, two types of idf dictionaries and the presence/absence of an edit distance penalty for low-frequency words are combined and examined, i.e. a total of 16 experimental environments. The effectiveness of the method was verified by comparing the performance of the experimental environment with the best performance with other existing methods. We set SentBLEU and Quick-Thought (Lajanugen and Honglak, 2018) as the methods to be compared.

We calculated the Pearson's correlation coefficient between the correct (reference) scores by the three evaluators and the predicted scores output by the system, and confirmed how well the human evaluation and the system evaluation agree. Following tables show the result when the idf dictionary was constructed based on the novel corpus, and the result when the idf dictionary was constructed based on the Wiki corpus.

Model	With Penalty	Without Penalty
Multilingual	0.169	0.197
Wikipedia	0.395	0.387
SNS	0.153	0.094
Novel	0.456	0.451

With idf dictionary by novel corpus

Model	With Penalty	Without Penalty
Multilingual	0.303	0.292
Wikipedia	0.405	0.397
SNS	0.094	0.083
Novel	0.438	0.451

With idf dictionary by wiki corpus

From the results of the preliminary experiment, when the Japanese novel sentence is evaluated using the modified BERTScore, using the BERT prelearning model by the novel model, and the penalty with the idf dictionary constructed by the novel corpus has the highest correlation with humans.

5 Consideration

We conducted a comparative evaluation of the proposed method using the setting that achieved the highest performance in a preliminary experiment and other existing methods. Following table shows the results of comparing the Pearson's correlation coefficient with the human evaluation for each of these methods and the proposed method.

Method	Pearson's correlation
Our method	0.456
Sent BLEU	0.093
Quick-Thought	0.067

The proposed method showed higher correlation with human evaluation than other comparison methods.

Next, we check how the presence or absence of a penalty affects the similarity evaluation. Using the novel model and the idf dictionary constructed from the novel, the similarity score with no penalty and the similarity score with penalty were compared with the similarity score by human evaluation. We investigate which sentence is more similar to human evaluation in which sentence.

Table below shows the comparison results. The first column of the table shows which is more similar to the human evaluation between presence and absence of penalty.

Of the two sentences in the second column, the upper sentence is the reference sentence and its English translation, and the lower sentence is the evaluation target sentence and its English translation.

In the third column, base indicates no penalty and penalized indicates penalty. The parentheses following the similarity in the table show the difference between the similarity by the human evaluation and the similarity by the system evaluation. The lower (the smaller the difference), the higher the correlation with the human evaluation.

In the first example of the table, the subject of the reference sentence is “Saito” and the subject of the sentence to be evaluated is “Ozaki”, indicating a completely different person. Therefore, as instructed when creating the data, the average human evaluation for these sentences is 0.477, which is a relatively low score. On the other hand, comparing the system-based scores, the Penalized score, which has a similar score lower due to the penalty, is closer to the manual evaluation than the original Base score. Therefore, the penalty works as intended.

Setting close to human evaluation	Sentence pair	Similarity Value
With penalty	斉藤さんは今最高の環境にいると思った。 (Mr. Saito thought he was in the best environment right now.) 尾崎は自分がこれまでにないほど良い場所にいると思った。 (Ozaki thought he was in a better place than ever before.)	Human: 0.477 (-) Base: 0.514 (0.037) Penalized: 0.492 (0.015)
Without Penalty	あなたは猟師のように槍を握っていなかったから失敗したのよ。 (You failed because you didn't hold your spear like a huntman.) あなたはハンターがすべきように槍を持っていなかった、そしてあなたが逃したのでそれはあなたのせいです。 (You didn't have a spear as a hunter should, and it's your fault because you missed.)	Human: 0.790 (-) Base: 0.617 (0.173) Penalized: 0.605 (0.185)

Contrary to the first example, in the second example of the table, the Base score is closer to the human evaluation than the Penalized score. A closer examination revealed that there was a penalty for the token “ハンター(hunter)” which corresponds to “猟師 (huntsman”). This is because the idf became high as the token “ハンター(hunter)” was not sufficiently included in the novel corpus, and it was considered a penalty target.

Considering this example, we suppose that there are other examples in which penalties are erroneously given because the idf calculated from the corpus is high, even though it is a general token that is not a proper noun. In order to solve this problem, it is conceivable to expand the corpus with a wider range of novel documents, or to not use idf as the penalty granting criterion in the first place. In order to recognize proper nouns more accurately, it is possible to add labels to words in the reference sentence and the sentence to be evaluated by using named entity extraction technique.

Examining the base score that does not give a penalty, the correlation of human evaluation has hardly changed in the novel model and the Wiki model. However, when a penalty is added, the correlation is improved in many models when the idf dictionary based on the novel corpus is used. On the other hand, in the case of using the idf dictionary based on the Wiki corpus, there is not much improvement or rather a decrease compared to the case of using the novel corpus.

This is because the model based on the novel corpus has a low idf value for words often used in novels such as personal pronouns, while the Wiki corpus has a high idf value for these expressions, so that no penalty is imposed. Then, it results improperly penalizing the sentence pair.

6 Conclusion

In this study, we applied BERTScore, which is an automatic evaluation method using distributed expressions of words by BERT, to Japanese novels. When two sentences with different expressions but having similar meanings were compared, it was possible to make a more accurate evaluation than BLEU, which only considers the surface form of words. We investigated which BERT pre-learning model and which idf dictionary should be used when applying BERTScore to Japanese sentences. In addition, we improved the tendency in which high similarity is given to different named entity expressions, which is a problem when evaluating novels, by giving a penalty using the edit distance when calculating the score.

We evaluated the similarity between two sentences using the BERTScore with a penalty, and confirmed the correlation with human evaluation under multiple settings. As a result, it was confirmed that a higher correlation was shown by matching the type of the learning corpus and idf dictionary used for the BERT pre-learning model with the sentence to be evaluated. In addition, a improvement in correlation was seen by giving a penalty.

However, due to the characteristics of the novel corpus used for BERT pre-training and idf dictionary construction, some expressions, such as vector representations of real-world place names, are insufficiently optimized, and idf for general nouns is high.

In the verification experiment, the corrected BERTScore showed higher correlation than SentBLEU, which considers only the surface form of words, and Quick-Thought, which obtained the cosine similarity from the distributed representation of sentences.

In the future, it is conceivable to create a large-scale and reliable Japanese corpus with a score so that the score can be predicted by regression analysis from the BERT distributed expression of two sentences. In BERT, the distributed expression of the obtained sentence was used as an explanatory variable to meet various downstream tasks of natural language processing. It may be possible that the automatic evaluation method can also be handled by setting a task that directly predicts the similarity score from the distributed expression of two sentences.

Acknowledgement

This research is (partially) supported by "Knowledge Hub Aichi", Priority Research Project from Aichi Prefectural Government.

References

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

Bahdanau Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In ICLR 2015.

Dan Gusfield. 1997. Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press, New York, NY, USA.

Logeswaran Lajanugen and Lee Honglak. 2018. An efficient framework for learning sentence representations. In International Conference on Learning Representations.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attentionbased neural machine translation. In Proceedings of the 2015

Conference on Empirical Methods in Natural Language Processing, pp. 1412–1421.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, pp. 671–688, Belgium, Brussels. Association for Computational Linguistics.

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, pp. 311–318. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014.Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems, pp. 3104–3112.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav. Artzi. 2019. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.