Several modifications to BERTScore automatic metric for translation quality evaluating without human-translated references

Vetrov A.A.
NUST "MISIS"
Moscow, Leninskiy prospect, 4
andrej.vetrov@edu.misis.ru

Abstract

In this article we propose several modifications to BERTScore automatic metric for translation quality evaluation (TQE). Experiments were conducted on sentences from fiction texts with pre-trained multilingual BERT as well as with a pair of monolingual BERT. To align monolingual embeddings, an orthogonal transformation based on anchor tokens was used. It was demonstrated that such transformation helps to prevent mismatching issue and shown that this approach gives better results than using embeddings of the multilingual model. In addition to machine translation, several versions of human translation were evaluated, the problems of this approach were listed.

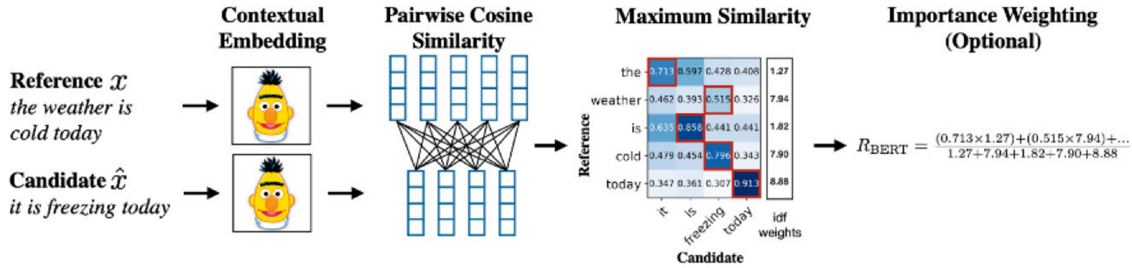
Keywords: BERT, BERTScore, translation quality, machine translation

1 Introduction

To assess machine transaltion (MT), which appeared in the middle of the 20th century, a comparison with the reference translation, performed by people, the so-called "gold standard", was traditionally used. The principles and methods used to transaltion quality evaluating (TQE) of MT are described a lot (Han et al., 2021). Among such methods there are both expert (Freitag et al., 2021) and automatic ones. The purpose of the TQE is to obtain some scores that is as close as possible to the expert's judgments but does not require his involvement. The existing automatic methods can be divided into 2 groups according to the metrics they use. The first group uses metrics such as: BLUE (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and many others. All these metrics cannot be considered as fully automatic since they require a reference translation. Meanwhile their quality depends on the quality of such translation. The second group, which appeared recently, includes YiSi (kiu Lo., 2019), BERTScore (Zhang et al., 2019) and its variations.

Translation of fiction texts causes the greatest difficulties in terms of MT, therefore analyzing the weaknesses of MT on such texts is essential for the future development of MT. Traditionally, translations made by professional translators are considered to be of high quality, but there is always a risk of missing a poor-quality translation made by an amateur translator. Therefore, along with machine TQA, human TQA, using automated methods can be considered an important objective. Therefore, we divided our study into two parts: MTQA and HTQA.

We focus on the BERTScore calculation, which refers to the state-of-the-art in MTQA at the sentence level. Following the approach of Lei Zhou et al. (Zhou et al., 2020) we compute BERTScore without reference translations. First, like the authors, we use a pre-trained multilingual BERT and then use a pair of monolingual BERT models. An orthogonal transformation based on anchor tokens is used to align monolingual embeddings. It is demonstrated that such transformation helps to prevent mismatching issue and shown that this approach gives better results than using embeddings of the multilingual model. To improve the token matching process it is proposed to combine all incomplete tokens into meaningful words and use simple



$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \ , \quad P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \ , \quad F_{\text{BERT}} = 2 \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \ .$$

Figure 1: The BERTScore calculation as given in the original article by Zhang et al.

averaging of corresponding vectors and to calculate BERTScore based on anchor tokens only. Such modifications allow us to achieve a better correlation of the model predictions with human judgments.

2 BERTScore and mismatching issue

The metric was first proposed to assess the quality of text generation with comparing candidate sentences to annotated references (Zhang et al., 2019). For this type of tasks, the metric showed the greatest correlation with the estimates given by humans among other metrics. Metric uses contextual embedding of tokens of the pre-trained BERT model (Devlin et al., 2018). For each token from the candidate sentence, the closest token is determined in the reference sentence, and vice versa, to obtain Recall and Precision, respectively. The obtained values of cosine distances are averaged. Further, Recall and Precision are combined to calculate F1 as shown at Figure 1.

Considering the influence of token frequency, the authors suggested using idf-weighting. Impact of this approach depends on the choice of the original document corpus and slightly affected the results.

Hereafter, Lei Zhou, Liang Ding and Koichi Takeda (Zhou et al., 2020), as participants in the Quality Estimation competition at the 5th conference on MT, tried to adopt BERTScore to unsupervised QE, i.e. without reference translation. They tried the pretrained multilingual BERT and revealed so-called mismatching issue.

The problem appeared strongly for the ru->en direction. The reason was the method of tokenization in the model. BERT uses WordPiece, a tokenizer that splits text either into full words or into piece of words, for which embedding vectors are then calculated. Pre-trained multilingual BERT tokenizer splits ru-words into too small pieces, thus during BERTScore calculation these pieces do not find corresponding en-words in the sentence in English, as in the example from the article Figure 2.

In response to this issue, the authors (Zhou et al., 2020) proposed to use additional token alignment. Practically, that function could align cross-linguistic patterns for the corresponding tokens (ru <-> en) and be applied weight coefficients, depending on the patterns found, to change the weight of distances for certain pairs.

They demonstrated that for 2 out of 6 directions (ne->en, si->en) the proposed unsupervised method performed better compared to the supervised model proposed by the competition organizers, and for the remaining 4 it showed comparable results. In other words, the fully automatic QA method proved to be competitive, and in some cases better, than the non-automatic one. This approach with empirical selection of weights seems rather contrived. Despite this im-

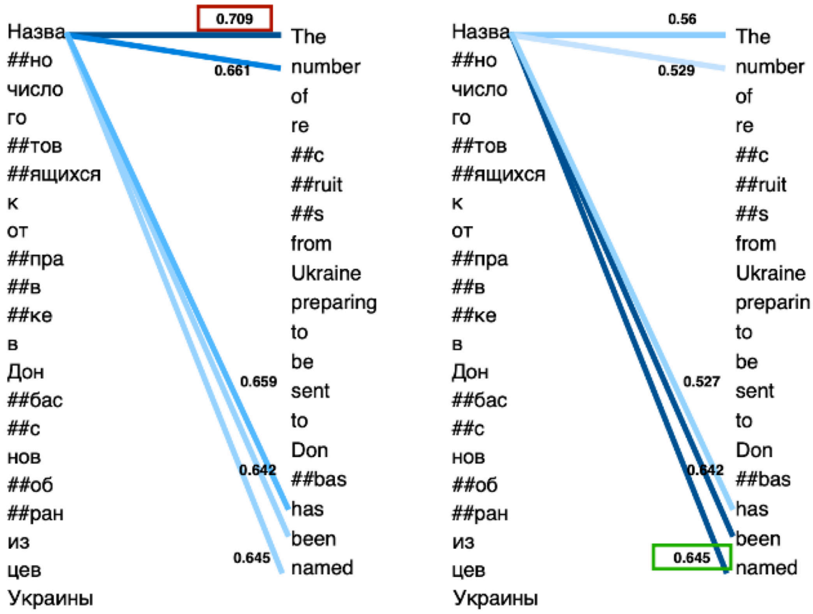


Figure 2: An example of the mismatching issue from the work by Lei Zhou et al., where the Russian token "Назва" is mismatched to the English token "The". After addition weighting authors could improve the matching.

provement, it was for the ru->en direction that the model score turned out to be noticeably worse than for the base supervised model.

3 Data acquisition

As a data source one of the works by an American writer in the mystical thriller genre was taken. About 100 original sentences were randomly selected from the text.

We engaged a linguist to evaluate the translation quality. Since the task of estimating the continuous value of BERTScore is difficult to formalize for human, it was decided to give only relative rank estimates. The rank correlation coefficients (Spearman-ρ, Kendall-τ) were chosen as metrics because these metrics are invariant to any monotonic transformations of the measurement scale. The resulting metrics were calculated as a averages over all test samples.

For MTQA these sentences were translated using the following engines: DeepL, Google Translate, Yandex Translate, Libre Translate. All versions were ranked from 1 to 4 by assessor such that a higher rank corresponds to a better translation quality. Samples with uncertain rank were removed. Finally, we got a dataset of 70 samples.

For HTQA, we took 2 translations made by different humans and 1 MT (Google Translate), all these versions were ranked from 1 to 3 by assessor. Samples with uncertain rank were removed as well, and we got dataset of 58 samples.

To exclude bias in the evaluation in both stages, the assessor was not provided with information on how a particular version of the translation was obtained.

4 Method

4.1 Obtaining a common contextual embedding and solving the mismatching issue

To implement the accurate calculation of BERTScore it is crucial to obtain a common contextual embedding space for tokens in the original English sentences and for tokens of correspondent sentences in Russian.

We tried two approaches:

- Using a multilingual model. (Zhou et al., 2020) followed this path in their work and encountered the “mismatching issue”. Here we used BERT multilingual base model.
- Using 2 monolingual models (Mikolov et al., 2013; Xing et al., 2015). In this case, it is necessary to perform the so-called cross-lingual word alignment for two embeddings. Here we used BERT base model (cased) and DeepPavlov/rubert-base-cased.

There were at least 2 works, there authors aligned the BERT contextual embeddings and exploited the idea to learn a transformation in the anchor space. (Schuster et al., 2019) learned transformation on averaged contextualized embeddings, (Wang et al., 2019) learned transformation directly in the contextual space. We followed the last one to preserve word semantics as much as possible.

In case of using separate monolingual models for 2 languages, the mismatching issue is less apparent, since the tokenizer in each of the models works only with its own language and the division into tokens is more granular as every token is most often a full word, which even more noticeable for the Russian language with its rich morphology.

For example, let’s look at the sentence in Russian and results of WordPiece-tokenization with 2 models:

Вдруг что-то выпало оттуда — большой, неровно сложенный кусок коричневой бумаги.

BERT multilingual tokinizer: ’В’, ’##дру’, ’##г’, ’что’, ’-’, ’то’, ’вы’, ’##пал’, ’##о’, ’от’, ’##ту’, ’##да’, ’[UNK]’, ’большой’, ’,’, ’не’, ’##ров’, ’##но’, ’сл’, ’##ожен’, ’##ный’, ’к’, ’##ус’, ’##ок’, ’кор’, ’##ичне’, ’##вой’, ’бу’, ’##ма’, ’##ги’, ’.’

RuBERT monolingual tokinizer: ’Вдруг’, ’что’, ’-’, ’то’, ’выпало’, ’оттуда’, ’—’, ’большой’, ’,’, ’неров’, ’##но’, ’сложен’, ’##ный’, ’кусок’, ’коричневой’, ’бумаги’, ’.’

This example shows that for multilingual model almost all the words after tokenizer are spitted into 2-3 tokens, as a result, semantics of words is lost and the mismatching issue will be serious. On the contrary, for monolingual model the only small number of words are splitted into separate tokens, and semantics of most words is not violated, so mismatching issue will not be serious.

To improve matching, we did a simple trick: combined all incomplete WorkPiece - tokens into meaningful words:

BERT multilingual tokinizer: ’В’, ’##дру’, ’##г’ -> ’Вдруг’; ’вы’, ’##пал’, ’##о’ -> ’выпало’; ’от’, ’##ту’, ’##да’ -> ’оттуда’; ’не’, ’##ров’, ’##но’ -> ’неровно’; ’сл’, ’##ожен’, ’##ный’ -> ’сложенный’; ’к’, ’##ус’, ’##ок’ -> ’кусок’; ’не’, ’##ров’, ’##но’ -> ’неровно’; ’сложен’, ’##ный’ -> ’сложенный’; ’кор’, ’##ичне’, ’##вой’ -> ’коричневой’; ’бу’, ’##ма’, ’##ги’ -> ’бумаги’

RuBERT monolingual tokinizer: ’неров’, ’##но’ -> ’неровно’; ’сложен’, ’##ный’ -> ’сложенный’

and to obtain the final vectors, we averaged the original vectors.

4.2 Cross-lingual word alignment

In order to align two embeddings derived from the two pre-trained monolingual BERT models it is necessary to perform cross-lingual word alignment, which is based on idea that all common languages share concepts that are grounded in the real world (Mikolov et al., 2013).

The conditions for this alignment can be formulated as follows:

- minimizing the angles between word vectors in two different languages with the same meaning to align word semantics;
- preserving the angles between the word vectors of the original space to save the word relative semantics.

This operation is reduced as orthogonal Procrustes problem (Gower, 1975): for a given matrix A and matrix B one is asked to find an orthogonal transformation matrix Ω that minimizes the distance ΩA to B:

$$\underbrace{argmin}_{\Omega^T\Omega=I} \|\Omega A - B\|_F \tag{1}$$

, where $\|.\|_F$ denotes the Frobenius norm.

This problem has an exact solution found in 1964 (Schönemann, 1966), it uses the operation of singular matrix decomposition.

For our case the ru-embedding is treated as matrix A and the corresponding en-embedding is treated as matrix B. The orthogonality is necessary to ensure that the angles between vectors of the original embedding are not distorted during the transformation. To find Ω, which minimizes the angles between vectors of words with the same meaning in two different languages after the alignment it is necessary to select such word pairs with the same meaning and solve Procrustes problem for them. These word pairs are called anchors, and the found Ω optimizes this transformation for arbitrary ru-word vectors.

Thus, the following operations were performed:

- for all the sentences and for each version of theirs translation, all ru-words were translated using Google Translate (a simple frequency dictionary could also be used);
- the found word translations were compared with the corresponding en-words in the original sentences for full match, in the case of repetition of words in sentences, their sequence was also taken into account, thus identifying the anchor pairs for which transformation is required;
- for the anchor pairs, found in the previous step, A and B matrices were found by concatenating the corresponding ru-word vectors and en-vectors;
- orthogonal matrix Ω was found, which realizes $A \to B$ transformation;
- Ω was optionally used to transform all ru-words vectors of the sentence to the target embedding.

As a result of the above operations, a single agreed ru-en contextual embedding of the source sentence's words and its translations was obtained for all sentences.

4.3 BERTScore calculation methods

BERTScoreF1 was calculated using two methods:

- with all tokens in sentence, as it was in original work (Zhang et al., 2019);
- with anchor tokens only. This modification simplifies and accelerates the calculation, reduces mismatching issue and, as it turned out, finally showed the best results.

Spearman-ρ and Kendall-τ were calculated for each test sample, and finally the results were averaged over all samples.

5 Results

Multilingual BERT embedding showed poor results both for MTQA and HTQA, as the rank correlations for all the approaches are less then zero (Table 1). The most likely reason is mismatching issue. Further we will focus on the results of 2 monolingual BERT embedding.

Table 1: Mean values of correlation coefficients (mean) and standard error of the mean (se) for different assessment, models and BERTScore calculation methods

QA	**Models**	**Method**	**Spearman mean (se)**	**Kendall mean (se)**
MT	multilingval BERT	all words	-0.07 (0.09)	-0.06 (0.08)
MT	multilingval BERT	anchors	-0.04 (0.08)	-0.02 (0.08)
MT	BERT + RUBERT + alignment	all words	0.30 (0.06)	0.25 (0.06)
MT	BERT + RUBERT + alignment	anchors	**0.53 (0.04)**	**0.44 (0.04)**
HT	multilingval BERT	all words	-0.43 (0.08)	-0.36 (0.07)
HT	multilingval BERT	anchors	-0.30 (0.07)	-0.25 (0.07)
HT	BERT + RUBERT + alignment	all words	-0.34 (0.08)	-0.29 (0.07)
HT	BERT + RUBERT + alignment	anchors	0.06 (0.09)	0.05 (0.08)

For the pair of monolingual BERT with word alignment the average values of rank correlation coefficients are strictly above zero for MTQA (Table 1), i.e. BERTScore shows good correlation with human judgments.

For HTQA the values of the rank correlations for all approaches are close to zero (Table 1), which roughly corresponds to a random guess. To clarify the reasons for this result, we analyzed several cases where BERTScore results were opposite to expert estimates of the algorithm's performance (Spearman-$\rho = -1$).

Initial sentence: 'He had learned that lesson yesterday when he had come home to find those terrible pink slips taped up all over the house'
and 3 version of translation:

HT 1: 'Вчера он получил наглядное тому подтверждение, когда вернулся домой и обнаружил эти страшные розовые листки, расклеенные по всему дому.' $\mathrm{BERTScoreF1} = 0.7252$

HT 2: 'Он понял это, когда, вернувшись вчера домой, обнаружил эти жуткие розовые талоны, расклеенные повсюду.' $\mathrm{BERTScoreF1} = 0.7574$

Google Translate: 'Он усвоил этот урок вчера, когда, придя домой, обнаружил эти ужасные розовые бланки, заклеенные по всему дому.' $\mathrm{BERTScoreF1} = 0.7620$

BERTScore seems to be higher for the machine version (Google Translate) and lower for the human versions due to closer in both word count and word semantic. This example indicates the following reasons for poor correlation:

- in human versions of translations, semantic proximity to the original sentence is much harder to establish than in MTs, because these versions include paraphrasing, which is perceived as "artistic language" and is judged higher by humans, despite the greater distance in semantics;
- an assessor appears to be more likely to better score a version that considers the broader context beyond a one sentence, as opposed to the BERT, which considers the context of no more than one sentence.

For the both assessments, the anchors-only method performed noticeably better than the all-words based approach.

6 Conclusion and next steps

Based on the results, we can conclude that BERTScore is well correlates with human judgments for automatic MTQA, but shows poor correlation with human judgments for automatic HTQA.

The most prominent results for MTQA were obtained with the following approaches:

- using pair of pretrained monolingual BERT models without any fine-tuning;
- using word alignment based on anchor words, which is obtained by word translation;
- combining all incomplete WorkPiece tokens into meaningful words and using simple averaging of corresponding vectors;
- BERTScore calculation on anchor words only.

In the future, we plan to perform additional testing for the en->ru direction on publicly available annotated corpora.

For translation directions with linguistically similar language pairs, the mismatching issue will not be so evident due to the effect of big vocabulary overlap(Pires et al., 2019). It can be expected that the same approach for calculating BERTScore can show better results for such languages with both monolingual BERT and multilingual BERT.

Acknowledgements

We thank Irakli Menteshashvili for his help in assessing translations and working with literary sources.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. // *in Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, Michigan, USA.

J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *Computing Research Repository*, arXiv:1810.04805. version 2.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Computing Research Repository*, arXiv.2104.14478. version 1.

J. C. Gower. 1975. *Generalized procrustes analysis*. Psychometrika.

Lifeng Han, Gareth J. F. Jones, and Alan F. Smeaton. 2021. Translation quality assessment: A brief survey on manual and automatic methods. *Computing Research Repository*, arXiv:2105.03311. version 1.

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*, P 507–513, Florence, Italy.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *Computing Research Repository*, arXiv:1309.4168. version 1.

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 (ASL)*, Philadelphia, USA.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *Computing Research Repository*, arXiv:1906.01502. version 1.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. *Computing Research Repository*, arXiv.1902.09492. version 2.

Peter H. Schönemann. 1966. *A generalized solution of the orthogonal procrustes problem*. Psychometrika.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual bert transformation for zero-shot dependency parsing. *Computing Research Repository*, arXiv.1909.06775. version 1.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. *// In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, P 1006–1011, Florence, Italy.

Tianyi Zhang, Varsha Kishore, Kilian Q. Weinberger Felix Wu, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *Computing Research Repository*, arXiv:1904.09675. version 3.

Lei Zhou, Liang Ding, and Koichi Takeda. 2020. Zero-shot translation quality estimation with explicit cross-lingual patterns. *Computing Research Repository*, arXiv:2010.04989. version 1.