REUSE: REference-free UnSupervised quality Estimation Metric

Ananya Mukherjee and **Manish Shrivastava**
Machine Translation - Natural Language Processing Lab
Language Technologies Research Centre
International Institute of Information Technology - Hyderabad
ananya.mukherjee@research.iiit.ac.in
m.shrivastava@iiit.ac.in

Abstract

This paper describes our submission to the WMT2022 shared metrics task. Our unsupervised metric estimates the translation quality at chunk-level and sentence-level. Source and target sentence chunks are retrieved by using a multi-lingual chunker. Chunk-level similarity is computed by leveraging BERT contextual word embeddings and sentence similarity scores are calculated by leveraging sentence embeddings of Language-Agnostic BERT models. The final quality estimation score is obtained by mean pooling the chunk-level and sentence-level similarity scores. This paper outlines our experiments and also reports the correlation with human judgements for en-de, en-ru and zh-en language pairs of WMT17, WMT18 and WMT19 testsets. Our submission will be made available at https://github.com/AnanyaCoder/WMT22Submission_REUSE

1 Introduction

Quality Estimation (QE) is an essential component of the machine translation workflow as it assesses the quality of the translated output without conferring reference translations (Specia et al., 2009; Blatz et al., 2004). High quality reference translations are often hard to find, QE helps to evaluate the translation quality based on the source sentences. Recently QE has emerged as an alternative evaluation approach for NMT systems (Specia et al., 2018). Recently, many researchers have been working on QE, as a part of Quality Estimation Shared Task, several QE systems (Zerva et al., 2021; Lim et al., 2021; Chowdhury et al., 2021; Geigle et al., 2021) were evaluated in WMT conference (Barrault et al., 2021). However, most of the quality estimation systems are supervised i.e., the model regresses on the human judgements. Often, human assessments are not available and it is very difficult to procure high quality human judgements. This motivated our research to emerge with an *Unsupervised Quality Estimation System*. Also, QE is usually performed at different granularity (e.g., word, sentence, document) (Kepler et al., 2019); in this work, we focus on the chunk-level and sentence-level similarity. The final QE score of the target sentence is obtained by mean pooling the chunk similarity scores and sentence similarity scores. Overall, our main contribution is as follows:

- We propose a concept of chunk level similarity i.e., matching the source and target chunks by leveraging multilingual BERT embeddings.
- We release a multilingual chunking model which returns meaningful word group boundaries.
- We present our unsupervised reference free QE metric (REUSE) that estimates the quality of translation by doing a chunk-level and sentence-level comparison with the source.

1.1 Motivation to use chunks

Usually, the words in translated output might not always follow the word sequence of the source text. However, it is observed that few word-groups often occur together irrespective of the order in source.

Figure 1 illustrates two example pairs: English-German (en-de) pair and English-Hindi (en-hi) pair. In the first example pair, the words sequence is not highly altered as English and German belong to the same language family (West Germanic), whereas in en-hi pair we can see a drastic change in the word order as Hindi belongs to a different language family (Indo-Aryan). However, we can observe that few word groups (here we refer as chunk) always occur *together* in both source and target. This phenomenon has motivated our research in the direction of chunk level assessment.

2 REUSE

We propose REUSE, a **RE**ference-free **U**n**S**upervised quality **E**stimation Metric that

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 564–568
Abu Dhabi, December 7–8, 2022.

Figure 1: Illustration of chunk similarity for two example sentences (en-de & en-hi).

evaluates a machine translated output based on the corresponding source sentence regardless of the reference. Figure 2 depicts the high-level architecture of our model. The chunks of source and hypothesis are acquired from the multilingual chunking model. Further chunk-wise subword contextual BERT embeddings are mean-pooled to obtain the chunk-level embeddings. Meanwhile, LaBSE model (Feng et al., 2020) is used for the sentence-level embeddings. Using these embeddings, we compute chunk-level similarity and sentence-level similarity, finally combine them by averaging chunk- and sentence-level similarity scores[1]. We discuss the working details of our system in the following sections.

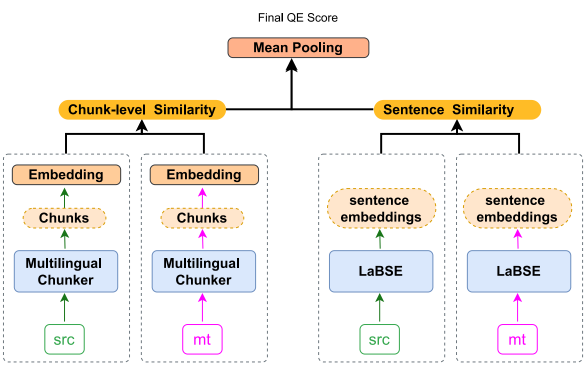

Figure 2: High-level architecture of REUSE model.

[1]REUSE score ranges between 0-1.

2.1 Chunk-level Similarity

We measure the number of matches between source chunks and hypothesis chunks. These matches are obtained by computing a cosine similarity (Foreman, 2014) of the individual chunk embeddings (refer 2.1.2) of source and translation sentence. An all-pair comparison is done to determine the best chunk match. Based on these matches, we compute precision and recall i.e, precision is *count of matches / length of hypothesis* and recall is *count of matches / length of source*. Ultimately, the chunk-level similarity score is calculated as the parameterized harmonic mean (Sasaki, 2007) of precision and recall, assigning more weightage to recall ($\beta = 3$).

2.1.1 Multilingual Chunker

The fundamental innovation in recent neural models lie in learning the contextualized representations by pre-training a language modeling task. Multilingual BERT is one such transformer-based masked language model that is pre-trained on monolingual Wikipedia corpora of 104 languages with a shared word-piece vocabulary. Training the pre-trained mBERT model for a supervised downstream task (finetuning) has dominated performance across a broad spectrum of NLP tasks (Devlin et al., 2018). We leverage this finetuning capability of BERT so as to create a *Multilingual Chunker* model that inputs a sentence and returns a set of divided chunks (word-groups).

We use **BertForTokenClassification** which has BERT (Bidirectional Encoder Representations from Transformers) as its base architecture, with a token classification head on top, allowing it to make predictions at the token level, rather than the sequence level. We use this BertForTokenClassification model and load it with the pretrained weights of "bert-base-multilingual-cased"[2]. We train the token classification head, together with the pretrained weights, using our labelled dataset (chunk annotated data). We employ Cross Entropy as the loss function and Adam optimizer (Kingma and Ba, 2014) with a learning rate of 1e-05.

2.1.2 Chunk Embeddings

Currently, we have word embedding models and sentence embedding models, but there is no specific chunk-level embedding models. Therefore, we embed the chunks leveraging the BERT embeddings by loading the weights of "distiluse-base-multilingual-cased"[3]. For a given sentence, this model return embeddings at a subword-level. To obtain the desired *chunk embeddings*, we perform a chunk to subword mapping and mean-pool the subword embeddings belonging to each chunk.

2.2 Sentence Similarity

To compute similarity at the sentence level, we find the cosine similarity (Foreman, 2014) of source sentence embedding and translation sentence embedding. We use LaBSE (Language Agnostic BERT Sentence Embedding) model to obtain the sentence embeddings. LaBSE model (Feng et al., 2020) is built on BERT architecture and trained on filtered and processed monolingual (for dictionaries) and bilingual training data. The resulting sentence embeddings achieve excellent performance on measures of sentence embedding quality, such as the semantic textual similarity (STS) benchmark and sentence embedding-based transfer learning (Feng et al., 2020).

3 Experiments and Results

3.1 Results on WMT17-19 testset

Each year, the WMT Translation shared task organisers collect human judgements in the form of Direct Assessments. Those assessments are then used in the Metrics task to measure the correlation between metrics and therefore decide which metric works best. Therefore, we estimated the translation quality of about 9K translations from the testset of WMT17 (Bojar et al., 2017), WMT18 (Bojar et al., 2018), WMT19 (Bojar et al., 2019a,b,c) for en-ru, en-de, zh-en language pairs and computed the pearson correlation (Benesty et al., 2009) of human judgements with Chunk-level Similarity scores, Sentence-level Similarity scores and their combination (REUSE). The segment level correlation scores are mentioned in Table 2. It is clearly evident from the correlations that the ensemble of Chunk Similarity model and Sentence Similarity model outperforms the individual models.

[2] https://huggingface.co/bert-base-multilingual-uncased

[3] https://huggingface.co/distiluse-base-multilingual-cased

3.2 WMT22 QE-as-a-metric task submission

Table 1 shows the WMT22 QE-as-a-metric task test-set details for the language pairs we have experimented on.

Language Pair	#Sentences	#Systems
en-ru	36723	88
en-de	82356	91
zh-en	41127	103

Table 1: Data statistics of WMT22 QE-as-a-metric task testset for en-ru, en-de and zh-en pairs.

3.2.1 Segment Level Evaluation

For Segment-level task, we submitted the sentence level scores obtained by our reference free quality estimation metric (REUSE) for en-ru, en-de and zh-en language pairs.

3.2.2 System Level Evaluation

We compute the system-level score for each system by averaging the segment-level scores obtained. A similar method is also used to compute system-level scores based on segment-level human annotations such as DA's and MQM, implying that a metric with a high segment-level correlation should also demonstrate high system-level correlation.

4 Conclusion

In this paper, we describe our submission to the WMT22 Metrics Shared Task (QE-as-a-metric). Our submission includes segment-level and system-level quality estimation scores for sentences of three language pairs Chinese-English (zh-en), English-Russian (en-ru) and English-German (en-de). We evaluate this year's test set using our **unsupervised, reference-free** metric - REUSE, that

WMT test-set	Language Pair	Chunk Similarity using chunker	Sentence Similarity using LaBSE	REUSE (chunk + sentence)
wmt17	zh-en	0.269	0.242	0.316
	en-ru	0.308	0.223	0.337
	en-de	0.280	0.167	0.278
wmt18	zh-en	0.135	0.2	0.210
	en-ru	0.145	0.2	0.213
	en-de	0.306	0.107	0.273
wmt19	zh-en	0.225	0.279	0.3
	en-ru	-0.112	0.144	-0.003
	en-de	0.254	0.131	0.251

Table 2: Correlation with Human Judgements on WMT17, WMT18 and WMT19 testset.

provides a quality estimation score by evaluating a hypothesis against the source sentence. REUSE estimates the translation quality by combining chunk-level similarity score and sentence-level similarity score, leveraging multilingual BERT embeddings. We performed our experiments on testsets of WMT17, WMT18, WMT19 and it has been emperically observed that the combination of *chunk- and sentence-level* similarity scores performed better in terms of agreement with human assessments.

Potential research directions definitely include improving the multilingual chunking model. As part of future work, we aim to further experiment and emerge with such effortless efficient unsupervised approach to estimate the translation quality and exhibit higher agreement with humans.

References

Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors. 2021. *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics, Online.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. *Pearson Correlation Coefficient*, pages 1–4. Springer Berlin Heidelberg, Berlin, Heidelberg.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321, Geneva, Switzerland. COLING.

Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer, editors. 2017. *Proceedings of the Second Conference on Machine Translation, Volume 1: Research Papers*. Association for Computational Linguistics, Copenhagen, Denmark.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors. 2019a. *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*. Association for Computational Linguistics, Florence, Italy.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors. 2019b. *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Association for Computational Linguistics, Florence, Italy.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors. 2019c. *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*. Association for Computational Linguistics, Florence, Italy.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors. 2018. *Proceedings of the Third Conference on Machine*

Translation. Association for Computational Linguistics, Belgium, Brussels.

Shaika Chowdhury, Naouel Baili, and Brian Vannah. 2021. Ensemble fine-tuned mbert for translation quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 897–903, Online. Association for Computational Linguistics.

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

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

John Foreman. 2014. COSINE DISTANCE, COSINE SIMILARITY, ANGULAR COSINE DISTANCE, ANGULAR COSINE SIMILARITY.

Gregor Geigle, Jonas Stadtmüller, Wei Zhao, Jonas Pfeiffer, and Steffen Eger. 2021. Tuda at wmt21: Sentence-level direct assessment with adapters. In *Proceedings of the Sixth Conference on Machine Translation*, pages 911–919, Online. Association for Computational Linguistics.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, António Góis, M. Amin Farajian, António V. Lopes, and André F. T. Martins. 2019. Unbabel's participation in the WMT19 translation quality estimation shared task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 78–84, Florence, Italy. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Seunghyun Lim, Hantae Kim, and Hyunjoong Kim. 2021. Papago's submission for the wmt21 quality estimation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 935–940, Online. Association for Computational Linguistics.

Yutaka Sasaki. 2007. The truth of the f-measure. *Teach Tutor Mater*.

Lucia Specia, Carolina Scarton, and Gustavo Paetzold. 2018. Quality estimation for machine translation. *Synthesis Lectures on Human Language Technologies*, 11:1–162.

Lucia Specia, Marco Turchi, Nicola Cancedda, Nello Cristianini, and Marc Dymetman. 2009. Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Annual conference of the European Association for Machine Translation*, Barcelona, Spain. European Association for Machine Translation.

Chrysoula Zerva, Daan van Stigt, Ricardo Rei, Ana C Farinha, Pedro Ramos, José G. C. de Souza, Taisiya Glushkova, miguel vera, Fabio Kepler, and André F. T. Martins. 2021. Ist-unbabel 2021 submission for the quality estimation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 961–972, Online. Association for Computational Linguistics.