Word-Level Quality Estimation for Machine Translation based on Source-MT Word Alignment

Yizhen Wei[†] **Takehito Utsuro**[†] **Masaaki Nagata**[‡]
[†]Deg. Prog. Sys.&Inf. Eng., Grad. Sch. Sci.&Tech., University of Tsukuba
[‡]NTT Communication Science Laboratories, NTT Corporation, Japan

1 Introduction

Machine translation is developing rapidly nowadays. Among various subtopics of researches of machine translation, quality estimation (referred to as QE) that estimates a translation generated by machine (referred to as MT) is attracting attention.

In this paper, we propose a novel and simple word-level QE method that assists translators by conveying specific instructions for post-editing. We exploit multilingual BERT [1], the pre-trained language model proved to be effective in many natural language understanding tasks as our basic architecture. Besides, we incorporate the source-MT word alignment information into the QE model via two schemes. Therefore, apart from the traditional QE objective, our model is also able to output source-MT word alignment as the by-product. That feature is believed to be a significant improvement because one can realize the specific operations such as replacement, insertion and deletion when source-MT word alignment is known.

For evaluation, we conduct experiments on WMT20 QE shared task 2 word-level post-editing effort [9]. Our method could have ranked at the sixth place on the leaderboard among eight participants and one baseline given by the organizer. With an acceptable performance, we also introduce a demo of user interface illustrating the superiority of our method and how it assists post-editing.

2 Background

2.1 Word-Level QE

Word-level QE requires the model to take a source sentence and an MT generated by a machine translation model as the inputs. The objective is to output the following three types of QE tags like it is shown in Figure 1.

- Source Tags: tags indicating whether a source word is correctly translated or omitted/mistranslated in MT.
- MT Word Tags: tags indicating whether an MT word is a correctly translated one or not.
- MT Gap Tags: tags indicating whether no extra words should be inserted into the gap compared to the correct translation or not.

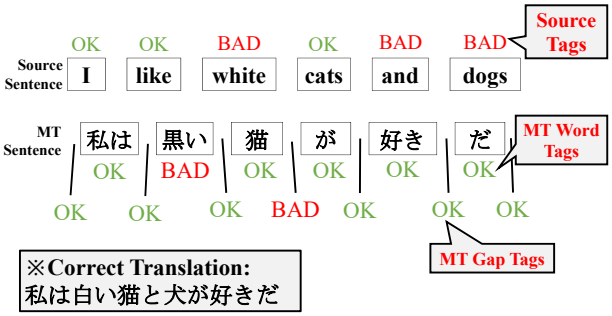


Figure 1: An example of word-level QE.

For convenience, source tags and MT word tags are collectively known as word tags in contrast to gap tags, while MT word tags and MT gap tags are collectively known as MT tags.

2.2 Related Work

Modern word-level QE models are mostly based on neural networks.

Predictor-estimator architecture [3, 4] consists of two stacked modules named predictor and estimator. The predictor is trained on large-scale parallel data and its objective is to predict target words conditioned with unbounded source and target contexts. Kim et al. [4] demonstrates that target word prediction is helpful for QE performance. The estimator, trained for QE objective, is another module that takes the feature vectors generated by the predictor as the inputs. Recent researches like Wang et al. [10] and Wu et al. [11] have proved the effectiveness of the architecture.

QE BERT proposed by Kim et al. [5] is a simple architecture based on BERT [1]. During pre-training, BERT is trained on large-scale parallel data in order to adapt for QE task. During fine-tuning, BERT is then topped by a linear layer followed with softmax function for QE tag classification.

In this paper, we employ a multilingual-BERT as the feature extractor with linear regression top layers for tag prediction. We also incorporate the source-MT word alignment information into our model, which is proved to be helpful to word-level QE. In comparison to the previous researches, our method is believed to be superior in the following aspects.

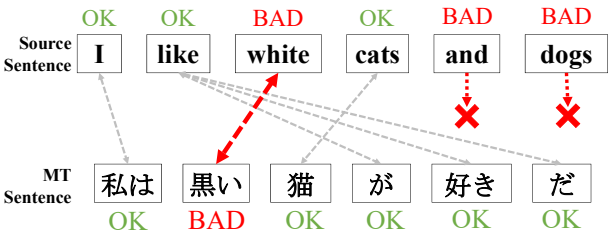


Figure 2: How alignment helps prediction of word tags: "white" and "黒い"(black) are aligned but semantically inequivalent. "and" and "dogs" cannot be aligned to any MT word. As a result, both word should be tagged as **BAD**.

- Our method needs no pre-training on large-scale parallel data.
- Our method can output source-MT word alignment as the by-product.

3 Source-MT Word Alignment Extracted by Multilingual-BERT

Intuitively, knowing source-MT word alignment is helpful to word-level QE, particularly for the prediction of word tags. For example, like it is shown in Figure 2, tags of a pair of words aligned but semantically inequivalent should be tagged as **BAD**. Tags of words which do not have aligned counterparts should be tagged as **BAD** too.

However, obtaining source-MT word alignment is non-trivial because MT is not always a correct translation. We found that statistical word alignment tools like GIZA++ [8] and FastAlign [2] could not handle source-MT alignment extraction well.

As a solution, we follow the neural methods based on multilingual-BERT [1] (referred to as mBERT) proposed by Nagata et al. [6]. The source sentence in which a word is marked by special token (" ¶ ") serves as the query and is concatenated by MT considered as the context. The model is trained to identify one or multiple words in a span in MT that are aligned to the marked source word as shown in Figure 3.

Because of the symmetrical characteristics of word alignment extraction, similar operations will be done again in the opposite direction where the MT with a marked word at the front and the source sentence as context at the rear.

Traditional statistical method models the extraction of word alignment upon a joint distribution. As a result, an incorrect matching might cause a domino effect that generates other incorrect alignments. However, that problem is evaded in the aforementioned mBERT-based method by extracting word alignments individually. Nagata et al. [6] proved that the mBERT-based method significantly outperforms statistical method.

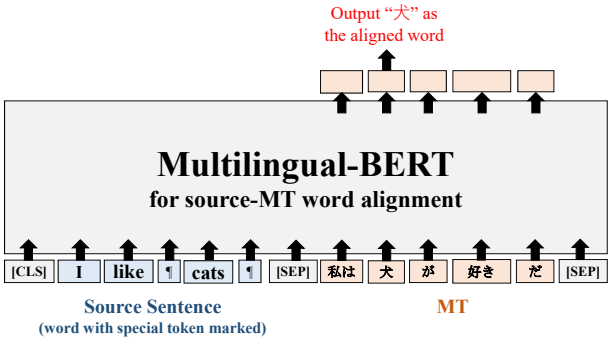


Figure 3: Extracting source-MT word alignment by multilingual-BERT. In this case, although not being semantically equivalent, the word "cats" is still aligned to "犬"(dogs) by mBERT.

Though the situation is slightly different in our case since the MT is not always the correct translation, the method still works giving the credit to the strength in language understanding of mBERT.

4 Word-Level QE based on Source-MT Word Alignment

According to the illustration in Section 3, semantical equivalence still needs to be manually judged to determine word tags.

Instead of utilizing the word alignment explicitly, we designed two schemes implicitly exploiting word alignment information.

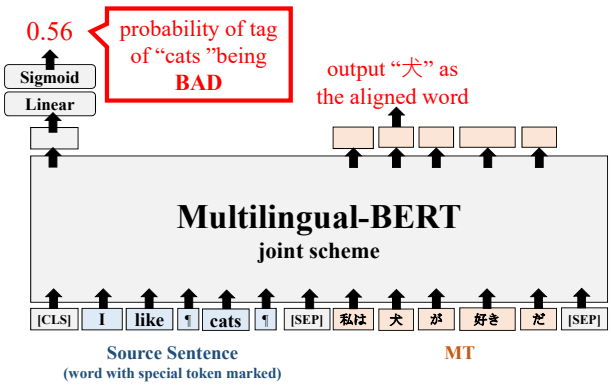
4.1 Joint Scheme

In joint scheme shown in Figure 4(a), a linear layer transfers the output vector of [CLS] token into a scalar value. It is then processed to a probability value through a Sigmoid function. Such a value is considered as the probability of a particular word tag being **BAD**.

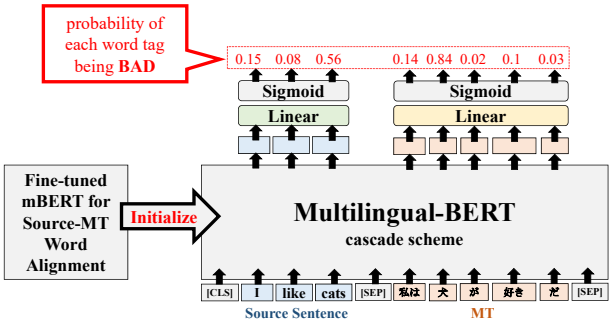
We keep the special tokens in the input sequence to indicate that tag of which word is being calculated. As the input sequence is identical to that for word alignment extraction mentioned above, we jointly trained the model for two objectives (word tags and word alignment) as shown in Figure 4(a). During inference, the model is able to predict the probability of word tags and the aligned words simultaneously.

4.2 Cascade Scheme

In cascade scheme shown in Figure 4(b), the model is trained in two phases. In the first phase, the model is merely trained for source-MT word alignment as introduced in Section 3. Then the model for word tag prediction is initialized with the parameters from the pre-trained model in the first phase. Word tag



(a) Joint scheme: the word tag and the alignment of the marked word is predicted simultaneously.



(b) Cascade scheme: word tags are predicted in a manner of sequence tagging.

Figure 4: Two schemes incorporating source-MT word alignment into QE tag prediction.

prediction in this scheme is modeled as a sequence tagging problem. Two different linear layers respectively transform source and MT output vectors into scalar values as their probabilities of being **BAD**. Because prediction of word tags is a downstream task of source-MT word alignment extraction, we consider this scheme as a cascade.

5 Experiment

5.1 Data and Settings

We adopt the English-German dataset provided by WMT20 quality estimation shared task2[1] for our experiments. There are 7,000, 1,000 and 1,000 groups of data for training, development and testing respectively. In one group of data, there is a source sentence, an MT, a post-edited sentence based on MT (referred to as PE), source tags as well as MT tags. [2] As there is no proper training data for source-MT word alignment, we take PE as a connection. A source word is aligned to an MT word if they are aligned to the same PE word in respective alignments. Both source-PE and PE-MT alignment are extracted by the method introduced in Section 3. For source-PE, we adopt GIZA++ to extract the rough source-PE alignment as training data to fine-tune the mBERT. For PE-MT, we calculate cosine similarities between words from embeddings output by the monolingual BERT of the corresponding language. Then we extract alignments from all word pairs while maximizing the sum of similarities subjecting to the limitation that each PE word aligns to one MT word at most. That is implemented by integer linear programming [7]. The extracted data is utilized to fine-tune a monolingual BERT since PE an MT share a same language.

For the pre-trained model, we adopt *bert-base-multilingual-cased* provided by huggingface[3]. Scripts for QA and sequence tagging from huggingface are modified for our experiments. In both schemes, learning rate is set to 3e-5. The hyper parameters are kept unchanged as the original settings of huggingface. According to our best practice, models of joint scheme and cascade scheme are trained for two epochs and five epochs respectively.

Table 1: Results of QE tags under different schemes.

	Source Tag		
	MCC	**OK-F1**	**BAD-F1**
Base	0.2388	0.7285	0.4718
Joint	0.3229	0.7819	0.5208
Cascade	**0.3467**	**0.8182**	**0.5256**
	MT Tag		
	MCC	**OK-F1**	**BAD-F1**
Base	0.2523	0.7945	0.4579
Joint	0.4453	0.9077	0.5358
Cascade	**0.4514**	**0.9204**	**0.5377**

5.2 Results

As our model output probabilities for prediction of word tags, a threshold is needed to determine the specific tags. We try values from 0.01 to 0.99 with a stride of 0.01 on the development set to search for the optimized threshold. 0.11 and 0.15 are the best values respectively for joint scheme and cascade scheme. The results of the word tags of our experiment are shown in Table 1. According to WMT20 [9], we adopted metrics including F1 scores respectively for **OK** and **BAD** as well as Matthews correlation coefficient (MCC). For baseline system, we adopted OpenKiwi implemented by Unbabel.

Note that MT tags are comprised of MT word tags and MT gap tags. Intuitively, identifying the corresponding gap for a **BAD** source word is effective in predicting gap tags. Unfortunately, word alignment tools including the mBERT-based approach do not

[1] `http://www.statmt.org/wmt20/quality-estimation-task.html`

[2] Source-MT alignment data extracted by FastAlign is given but as mentioned above we gave up using it for its poor quality.

[3] `https://github.com/huggingface/transformers`

Table 2: Results of ablation study. († indicates statistically significant (p<0.01), while ⋆ indicates not statistically significant(p<0.05).)

	Source Tag MCC	MT Tag MCC
Joint w/ weak align	0.3150	0.4338
Joint	0.3229†	0.4453†
Cascade w/o align	0.3397	**0.4569**
Cascade	**0.3467**†	0.4514⋆

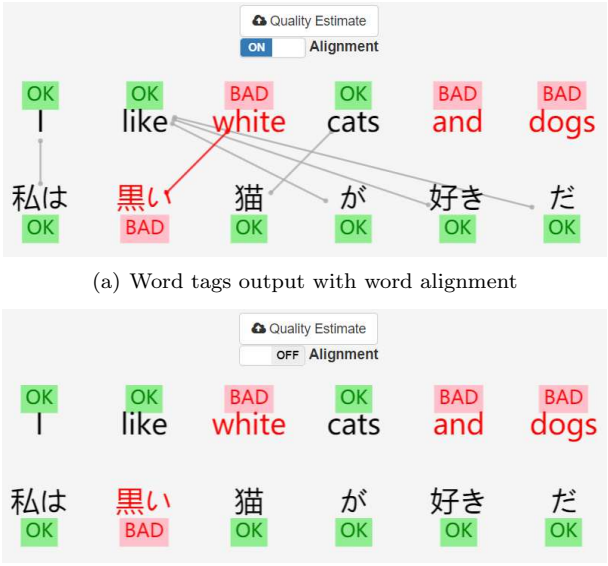
handle the gaps. Therefore, we fail to generate relation between source words and MT gaps according to the method introduced in the previous subsection. As a solution, we tag all gap tags as **OK** because ratio of **BAD** is fairly low in training and development set. Compared to the results of participants of WMT20, our best performance in cascade scheme could have ranked at 6th place either in source tag MCC or MT tag MCC.

We further investigate the effectiveness of source-MT word alignment by multiplying the loss of word alignment with a factor of 0.1 to weaken the influence of the alignment information in joint scheme and initialize with the original mBERT in cascade scheme. Those are referred to as "Joint w/ weak align" and "Cascade w/o align" respectively. The results are shown in Table 2.

For prediction of source tags, a statistically significant difference is confirmed in both schemes when alignment information is involved. For prediction of MT tags in cascade scheme, although the absolute value decreases, there is no statistically significant difference. Consequently, involving source-MT word alignment information in word-level QE could significantly improve the performance. In fact, the training data for source-MT word alignment is automatically generated and is not necessarily gold. If manually labelled source-MT word alignment data is available, a further improvement is expected.

6 A Demo of User Interface

Compared to normal word-level QE models, our model is able to output source-MT word alignment as by-product. In this section, we introduce a demo of user interface which displays the output source-MT word alignment as well as the word tags on the same page.[4] As the task name of WMT20 "post-editing effort" suggests, a significant application for word-level QE is to instruct post-editing. However, current **BAD** tags fail to provide a clear instruction since there are multiple causes. We argue that with additional word alignment information, post-editing will be easier to a certain extent.



(a) Word tags output with word alignment

(b) Word tags output without word alignment

Figure 5: Output of the user interface with/without alignment

We add a switch on the interface shown in Figure 5. In Figure 5(b), when "white" is tagged as **BAD** the user still needs to think whether it is badly translated or its translation is omitted in MT. However, when alignment is displayed on the interface, the user might notice that the word "white" is aligned to "黒い"(black) which is also tagged as **BAD**. Therefore, it is a signal for replacement that "黒い"(black) should be replaced with a correct translation which is "白い". Likewise, "dogs" has not been aligned to any word so that the user only needs to identify a gap where the omitted translation should be inserted. Apparently, our system could merely assist the user to a certain extent since one still needs to identify the gap to insert the omitted translation manually. We would like to research into those problems in our future work.

7 Conclusion

In this paper, we propose a novel method for word-level QE which requires the model to output QE tags for post-editing effort. We incorporate source-MT word alignment into our model by designing two schemes. The effectiveness of incorporating source-MT word alignment is proved. Our model also output the source-MT word alignment as the by-product, which we believe making post-editing easier.

[4]available at `https://wyzypa.cn/tools/qe-demo/`

References

[1] J. Devlin, M. Chang, K. Lee, and Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. 17th NAACL-HLT*, pages 4171–4186, 2019.

[2] C. Dyer, V. Chahuneau, and N. Smith. A simple, fast, and effective reparameterization of IBM model 2. In *Proc. 11th NAACL-HLT*, pages 644–648, 2013.

[3] H. Kim and J. Lee. Recurrent neural network based translation quality estimation. In *Proc. 1st WMT*, pages 787–792, 2016.

[4] H. Kim, J. Lee, and S. Na. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proc. 2nd WMT*, pages 562–568, 2017.

[5] H. Kim, J. Lim, H. Kim, and S. Na. QE BERT: Bilingual BERT using multi-task learning for neural quality estimation. In *Proc. 4th WMT*, pages 85–89, 2019.

[6] M. Nagata, K. Chousa, and M. Nishino. A supervised word alignment method based on cross-language span prediction using multilingual BERT. In *Proc. EMNLP*, pages 555–565, 2020.

[7] M. Nishino, J. Suzuki, S. Umetani, T. Hirao, and M. Nagata. Sequence alignment as a set partitioning problem. *Journal of Natural Language Processing*, 23(2):175–194, 2016.

[8] F. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[9] L. Specia, F. Blain, M. Fomicheva, E. Fonseca, V. Chaudhary, F. Guzmán, and A. Martins. Findings of the WMT 2020 shared task on quality estimation. In *Proc. 5th WMT*, pages 741–762, 2020.

[10] M. Wang, H. Yang, H. Shang, D. Wei, J. Guo, L. Lei, Y. Qin, S. Tao, S. Sun, Y. Chen, and L. Li. HW-TSC's participation at WMT 2020 automatic post editing shared task. In *Proc. 5th WMT*, pages 1054–1059, 2020.

[11] H. Wu, Z. Wang, Q. Ma, X. Wen, R. Wang, X. Wang, Y. Zhang, Z. Yao, and S. Peng. Tencent submission for WMT20 quality estimation shared task. In *Proc. 5th WMT*, pages 1060–1065, 2020.