Confidence Modeling for Neural Machine Translation

Taichi Aida
Nagaoka University of Technology
Nagaoka, Niigata, Japan
aida@jnlp.org

Kazuhide Yamamoto
Nagaoka University of Technology
Nagaoka, Niigata, Japan
yamamoto@jnlp.org

***Abstract*—Current methods of neural machine translation output incorrect sentences together with sentences translated correctly. Consequently, users of neural machine translation algorithms do not have a way to check which outputted sentences have been translated correctly without employing an evaluation method. Therefore, we aim to define the confidence values in neural machine translation models. We suppose that setting a threshold to limit the confidence value would allow correctly translated sentences to exceed the threshold; thus, only clearly translated sentences would be outputted. Hence, users of such a translation tool can obtain a particular level of confidence in the translation correctness. We propose some indices; sentence log-likelihood, minimum variance, and average variance. After that, we calculated the correlation between each index and bilingual evaluation score (BLEU) to investigate the appropriateness of the defined confidence indices. As a result, sentence log-likelihood and average variance calculated by probability have a weak correlation with the BLEU score. Furthermore, when we set each index as the threshold value, we could obtain high quality translated sentences instead of outputting all translated sentences which include a wide range of quality sentences like previous work.**

***Keywords*-machine translation; confidence estimation;**

I. Introduction

Neural machine translation (NMT) [1], which uses neural networks for the purposes of machine translation (MT), is one of the most discussed subjects in the field of natural language processing. NMT has a great impact on society; it is applied to Google Translate and is used by various companies.

In Japan, the Tokyo Olympics will be held in 2020. At this event, many tourists from all over the world are expected to come to Japan, and their language needs to be addressed. Hence, the demand for MT is increasing even more.

MT systems are expected to output high-quality translations. However, recent systems may generate a wide range of quality sentences. Therefore, users cannot ensure which outputted sentences were translated correctly, and which ones were not, until an evaluation method is applied. For a user, it is a serious and time-consuming problem to check all output sentences without information on the quality of outputted translation obtained by a MT system. Although there are some methods to evaluate the translation result, BLEU [2], which is a typical evaluation method, requires a reference translation as an answer, and the evaluation model based on the task called Quality Estimation [3]–[8] uses a large corpus for training.

To solve this issue, we seek to define the confidence in NMT models. Information on the translation confidence would allow understanding whether an output sentence has been translated with a high level of confidence or not. Furthermore, when a user sets a threshold and to prevent a model from outputting sentences with the confidence value lower than the threshold, the user of such a translation tool can have guarantees that the outputted translation is correct and focus on the other translations.

In this paper, we propose the indices to define the confidence value. Next, in order to investigate whether the proposed index can be used for confidence, we correlated it with the BLEU score, the main evaluation method of machine translation. After that, we used a reasonable index of confidence as a threshold, and examined the number of sentences output and the average of BLEU by changing the threshold. Our method requires neither references nor an external large-scale corpus which is necessary to train an evaluation model. We suppose that we can obtain high-quality outputted sentences by setting a threshold to limit the output with the confidence value.

The experimental results show that some indices are correlated with the BLEU score, which means that these indices are appropriate as confidence values. In addition, when we use each index as a threshold value, we can filter high quality sentences from the output, which is unlike the previous works.

II. Related work

A. Confidence Estimation

In statistical machine translation, a confidence estimation task was defined [9]–[11]. In that task, the system labels each word in outputted sentences to correct or incorrect.

To validate the result, a rating scale called Classification Error Rate (CER) had been used:

$$CER = \frac{Count(\text{incorrect-label})}{\sum_{word \in sent} Count(word)} \tag{1}$$

CER, as it is suggested by its name, is defined as a mislabeled percentage of all words labeled by the system.

B. Automatic Evaluation

To evaluate various sentences outputted by machine translation, there are two classical methods. One is a major evaluation method, BLEU. The other is a quality

978-1-7281-5014-7/19/$31.00

estimation method based on using a large parallel corpus.

1) BLEU: BLEU [2] is a representative method for evaluating a translation result using a reference translation. The BLEU score is calculated by examining the number of matches of n-grams between the candidate and the reference translation. Specific formulas are shown below:

$$BLEU = penalty(\prod_{n=1}^{4} p_n)^{1/4} \tag{2}$$

$$penalty = min(1, exp(1 - \frac{len_{ref}}{len_{cand}})) \tag{3}$$

$$p_n = \frac{\sum_{C \in \{Cand\}} \sum_{ngram \in C} Count_{clip}(ngram)}{\sum_{C' \in \{Cand\}} \sum_{ngram' \in C'} Count(ngram')} \tag{4}$$

Finally, the BLEU score is calculated in the range of 0 to 100, and the higher is this value, the closer the output sentence is to the reference translation.

2) Quality Estimation: However, when we use a translation system, we often do not have sources and reference pairs. For this reason, a Quality Estimation (QE) task is defined which implies using only a parallel corpus and not using references.

This is an actively researched approach and is used to train a large amount of external parallel corpus as an evaluation model, and to use the obtained model as a new translation pair for evaluation.

Many works suggest extracting indices and using them as features of support vector regression (SVR) to build the QE model [3]–[6]. There are also other studies that suggest training neural-based QE models separately from translation models, and then these models demonstrated high correlation with manually evaluated test data in QE tasks [7], [8].

However, many QE systems require a large amount of a parallel corpus to train their evaluation model. In other words, it is difficult for these systems to adapt to language pairs with few bilingual corpora such as Asian languages.

Thus, normally, the output sentence is evaluated only from the outside. Therefore, we consider that it may be necessary to perform evaluation from the inside of the model, too.

In the semantic parsing task, the confidence of the outputted sentence was defined [12]. Several indices are proposed in the model and in the corpus used. When the confidence value is limited by a threshold, only the results with high accuracy are outputted.

Moreover, the other research shows that weights that are identified based on the attention mechanism often used in machine translation are appropriate for estimating the confidence [13]. They asserted attention weight, which can be useful for more purposes than just visualizations. They defined two metrics for confidence, coverage deviation penalty and absentmindedness penalty using attention weight α_{ji} between input token j and output token i.

$$CDP = -\frac{1}{J} \sum_{j} \log(1 + (1 - \sum_{i} \alpha_{ji})^2) \tag{5}$$

$$AP = -\frac{1}{I} \sum_{i} \sum_{j} \alpha_{ji} \log \alpha_{ji} \tag{6}$$

In the above equations J and I are the length of input and output sentence. Coverage deviation penalty (CDP) can penalize not only lacking attention but also too much attention per input token. Also, the absentmindedness penalty (AP) can indicate how scattered the attention weights in each token are. Finally, they defined the confidence value:

$$conf_{attn} = CDP + AP_{output} + AP_{input} \tag{7}$$

Inspired by their work, we define the confidence and its indices in NMT models. More information about the proposed approach is described in Section III.

III. Proposal

In Section II, it was noted that many systems evaluate candidates from only outside normally. We think that it is necessary to perform evaluation also from inside of the model.

We propose a method to calculate confidence values inside of the model. Figure 1 gives an overview of our method. As shown in Figure 1, recent NMT models output all translated sentences which include high and low quality sentences. In our process, the model computes the confidence value for each sentence. We believe that we can get high-quality translated sentences by setting a threshold to limit the output with the confidence value. The advantage of this method is that it does not require a reference sentence for the translation, or a large-scale bilingual corpus from outside, as well as training on any other model used for evaluation.

Confidence indices are described in detail below.

A. Sentence log-likelihood

In various sequence-to-sequence (seq2seq) machine translation models, one word generated at each position has the highest probability assigned among several word candidates. Therefore, by taking the sum of log-probabilities of all outputted words and then, taking into account the average number of words, we consider that it may become an index of confidence without affecting the word length:

$$likelihood_{sent} = \frac{\sum_{word \in sent} \log p_{word}}{\sum_{word \in sent} Count(word)} \tag{8}$$

B. Variance

As we mentioned earlier, machine translation models output the results with the probability being the highest one among other candidates. In this index, we calculate the degree of dispersion with probability that a word of the highest probability actually output for the top five candidates in each word of the output sentence:

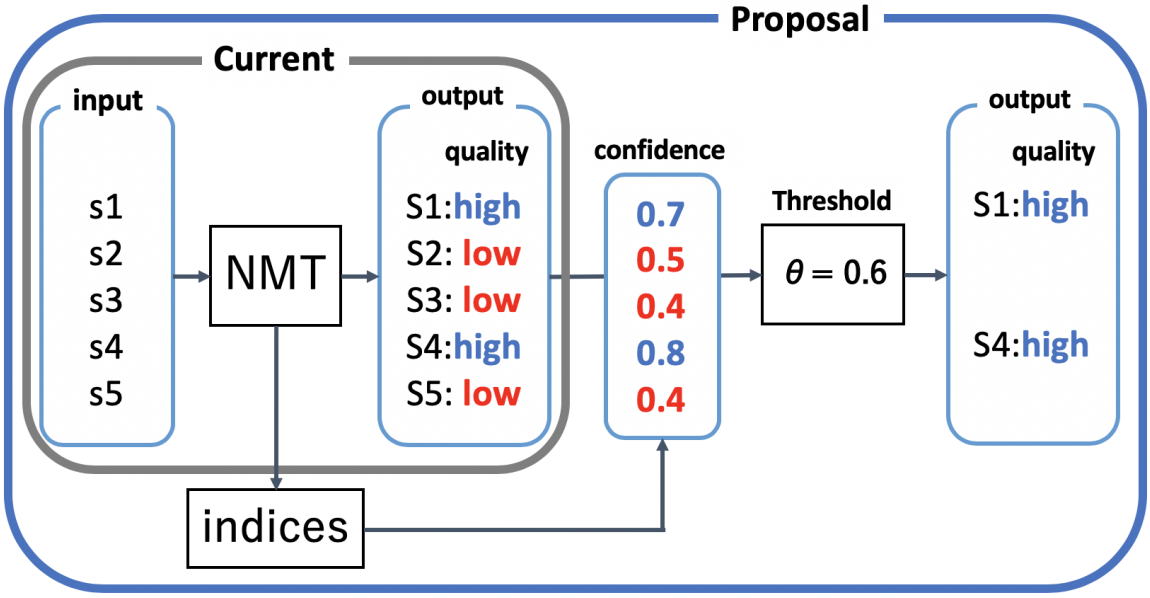


Figure 1: Overview of the proposed method; when sentences $s_{1:5}$ are inputted, current methods output all sentences $S_{1:5}$ which are included both high and low quality sentences. Our method computes confidence value from several indexes in the model and allows the model to output only sentences whose confidence value exceeds a threshold. We suppose that this process can obtain high-quality sentences by using no references and no large amount of parallel corpora to train other models. In this example, the threshold value is 0.6.

$$V_{top} = \frac{1}{4}\sum_{n=2}^{5}(P_n - P_{top})^2 \tag{9}$$

Equation 9 is similar as in case of variance, the higher the V_{top} value, the farther the outputted word's probability outstands from the other candidates, and if the V_{top} value is low, it means the probability of outputted word is closer to them. We think that this index can represent "how much the model got lost when outputting each word."

By using $min(V_{top})$ with the smallest variance V_{top} for each word in the sentence thus obtained, it may be considered as an index of certainty factor.

In addition, we consider that the average of the variance of each word in each sentence can be used as an indices of certainty:

$$Ave(V_{top}) = \sum_{word \in sent} \frac{V_{top}(word)}{Count(word)} \tag{10}$$

We propose these two indices to define the confidence value. This method enables performing the evaluation in the process of generating the output translated sentences even in the environment available for a user (as they have only sentences to be translated) without the need in a large corpus.

IV. Experiment

First, we attempt to conduct this experiment to measure the appropriateness of each index we proposed. As this is a task similar to that of machine translation, the model is first trained on training and development data, thereafter, using the test data, the model translates source language sentences into target language sentences. At the same time, the model calculates a confidence index for each sentence. Candidates are evaluated by the BLEU method. Consequently, the correlation is measured between confidence index and the BLEU score.

Second, we use index as a threshold value, and measure the average BLEU score and the number of outputted sentences when the threshold is changed.

A. Dataset

In order to train and test the translation model, we used ASPEC-JE corpus. ASPEC [14] is constructed by translating scientific papers. There are two language pairs, Japanese-English (ASPEC-JE) and Japanese-Chinese (ASPEC-JC). Details of the number of parallel pairs per a language pair are shown in Table I.

Lang-pair	Train	Dev	DevTest	Test
Ja-En	3,008,500	1,790	1,784	1,812
Ja-Cn	672,315	2,090	2,148	2,107

Table I: ASPEC, Asian Scientific Paper Excerpt Corpus, is constructed by translating scientific papers.

In ASPEC-JE, there are 3 train sets, one of each 1M pairs. They are arranged in descending order of similarity of language pairs, and this time we used only the top 1M pairs of a train-1 set with high similarity between parallel language pairs. Therefore, the actual number of parallel pairs are 1M/1,790/1,812 for Train/Dev/Test.

B. Models and Settings

In this experiment, we used the fairseq [15] transformer. Moreover, for the preprocessing purposes, the proposed

model used the functionality existing in fairseq and made a word of ten times or less appearance frequency into unknown token.

Details of parameters are as follows: lr=0.1, clip_norm=0.1, dropout=0.2, embedding dimension=300, beam_size=5, encoder and decoder layer=4, encoder and decoder attention_heads=5, max_epoch=100, and batch_size=64.

V. Results/Discussion

A. Appropriateness of indices

1) Sentence log-likelihood: Figure 2 shows the distribution of the sentence log-likelihood and the BLEU score in each sentence.

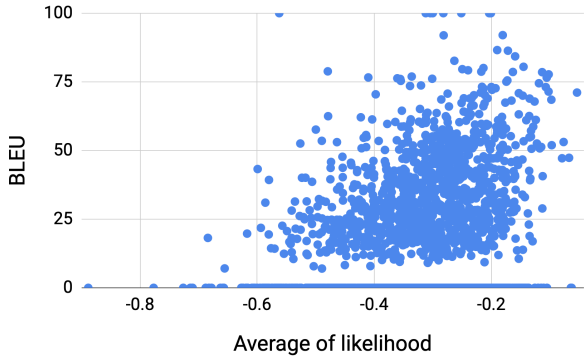


Figure 2: Distribution of log-likelihood and BLEU.

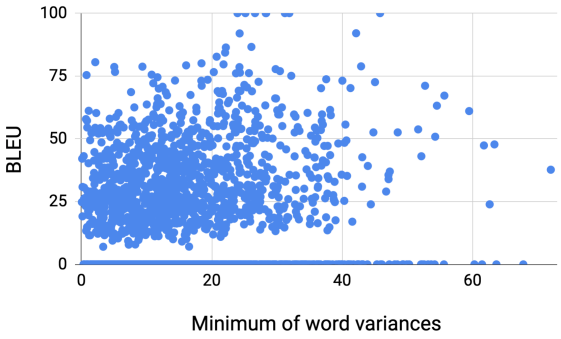
According to Figure 2, it has a triangular distribution, and the Pearson correlation coefficient is 0.308, which means that two indicators have a weak correlation. Therefore, sentence log-likelihood is suitable for confidence indices.

2) Variance: In this section, we analyze the usefulness of variance V_{top}.

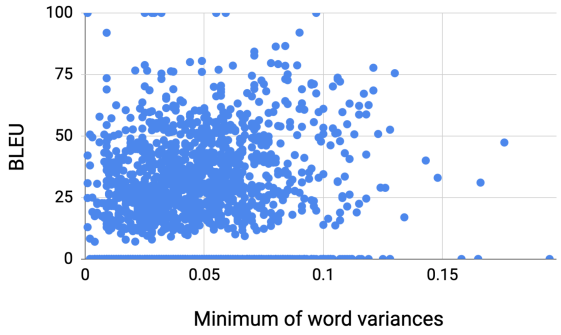
First, we examine the correlation between the minimum of variance each word in the sentence $min(V_{top})$ and the BLEU score. When we calculate the variance, we use both log-probability and probability. The result is shown in Figure 3.

Figure 3 shows that the distribution is concentrated at the bottom left in both cases. The correlation coefficient with BLEU, in this case, is 0.077 for log-probability and 0.112 for probability. This result means that there is no correlation between the minimum variance calculated by either log-probability or probability and the BLEU score. Also, because the log-probability has a value range of 0 to minus infinity, and the variance is squared, we think that the correlation is lower than that of the probability. For that reason, we use only the probability to calculate variance in each word from this experiment. As a result, the minimum variance of each word in the sentence cannot be used for the index of confidence.

Second, we examine the correlation between the average of variance each word in the sentence $Ave(V_{top})$ and the BLEU score using the probability. The result is shown in Figure 4.



(a) Log-probability



(b) Probability

Figure 3: Distribution of the minimum variance of each word in the sentence and BLEU.

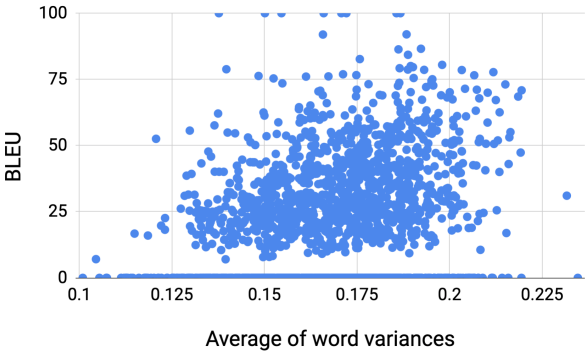
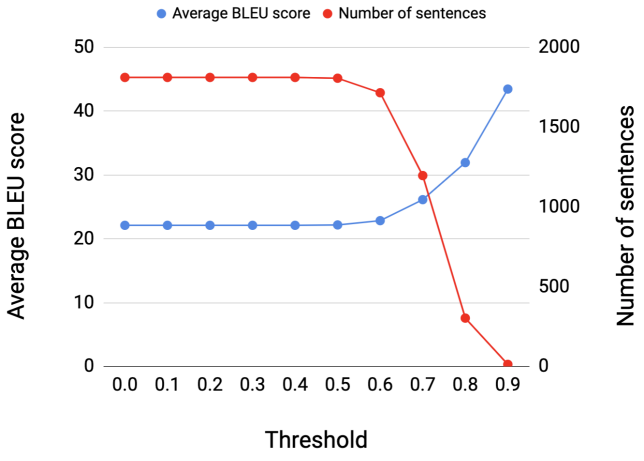


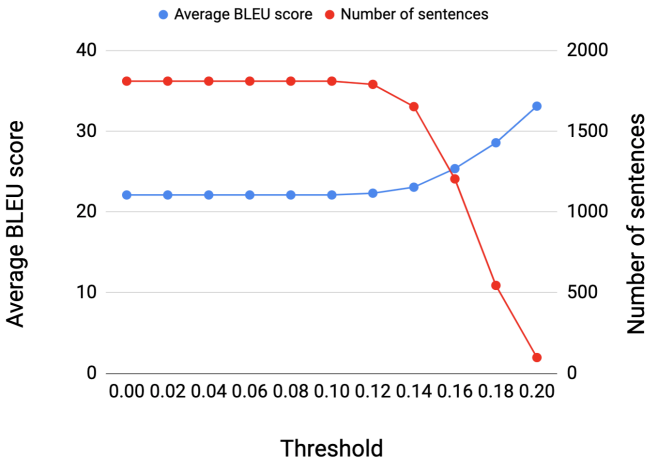
Figure 4: Distribution of average of variance and BLEU.

Similar to the sentence log-likelihood distribution presented in Figure 2, the distribution of the average variance is close to a triangle shape in Figure 4. Also, the Pearson correlation coefficient between the average variance and the BLEU score is 0.2676, which means low correlation between them. This result shows that the average variance $Ave(V_{top})$ calculated by probability can be used as the confidence indices.

According to these experiments, the sentence log-likelihood and the average of word variances of in each sentence can be used for the index of confidence.



(a) Sentence log-likelihood



(b) Average variance

Figure 5: The number of sentences and the BLEU score when the threshold was changed. We used the sentence log-likelihood(left) and the average variance which is calculated by probability(right) as a threshold. If the threshold is high, the number of output sentences is low, but the average of the BLEU score is increased.

B. *Using threshold*

Next, we used sentence log-likelihood and average variance as threshold. We examined the number of outputted sentences and the average value of the BLEU score (Figure 5). When we set no threshold, 1,812 sentences for ASPEC-JE test set were output and the average of the BLEU score is 22.11.

According to Figure 5, it can be seen that increasing the threshold value in both cases reduces the number of output sentences and increases the average of the BLEU score.

1) Sentence log-likelihood: In Figure 5(a), the sentence log-likelihood is distributed from 0.6 to 0.9 when we used exponential to define its value between 0 to 1. When the threshold was set to 0.9, only 13 sentences were output, and the average value of BLEU is 43.45.

2) Average variance: On the other hand, the average variance is distributed from only 0.1 to 0.2(Figure 5(b)). When we set the threshold to 0.2, 98 sentences were output, and the average value of BLEU is 33.12.

From these results, high-quality translated sentences can be obtained by using the sentence log-likelihood or the average variance as a threshold value to limit the output.

VI. Conclusion

In this paper, we analyzed the confidence and corresponding indices used in NMT models. Firstly, we proposed the confidence indices. Secondly, we analyzed the correlation between several indices and the BLEU score to verify if these indices can be used for confidence estimation. Thirdly, we set each index as the threshold value and examined the number of output sentences and the average of the BLEU score. In comparison to the previous work, this method requires neither a reference sentence as the answer to the translation, an external large-scale bilingual corpus, nor training any other model, which is conventionally used for evaluation. As a result, the sentence log-likelihood and the average variance have a weak correlation with the BLEU score, which means these indices can be used as the component of confidence. Furthermore, when we set each index as the threshold value, we could filter high quality sentences from the output, which is unlike the previous works.

VII. Future Work

In the future, we plan to investigate the correlation between other indices and the BLEU score. For example, we consider using the attention weight [13] and unknown words. We think that if the probability of unknown words is higher than for other candidates and the model outputs $< unk >$, the model should have the low confidence about that part. By using this index, we consider that the largest proportion of the $< unk >$ in a statement that corresponds to an output sentence, and the confidence in the output statement in the model is low. We suppose this index can represent "how much the model does not know."

Finally, we plan to consider combining several indices as a confidence factor and collecting only high-quality sentences by using the threshold. We hope that our work will help translators and other users be able to obtain only reliable translations from the machine translation systems.

Acknowledgment

This work was supported in part by JSPS KAKENHI Grants-in-Aid for Challenging Research (Exploratory) Grant ID 17K18481.

References

[1] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3104–3112.

[2] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 311–318.

[3] L. Specia, G. Paetzold, and C. Scarton, "Multi-level translation quality prediction with QuEst++," in *Proceedings of ACL-IJCNLP 2015 System Demonstrations*. Beijing, China: Association for Computational Linguistics and The Asian Federation of Natural Language Processing, Jul. 2015, pp. 115–120.

[4] C. Scarton, D. Beck, K. Shah, K. Sim Smith, and L. Specia, "Word embeddings and discourse information for Quality Estimation," in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 831–837.

[5] K. Shah, F. Bougares, L. Barrault, and L. Specia, "SHEF-LIUM-NN: Sentence level quality estimation with neural network features," in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 838–842.

[6] Z. Chen, Y. Tan, C. Zhang, Q. Xiang, L. Zhang, M. Li, and M. Wang, "Improving Machine Translation Quality Estimation with Neural Network Features," in *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 551–555.

[7] J. Ive, F. Blain, and L. Specia, "deepQuest: A framework for neural-based quality estimation," in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 3146–3157.

[8] J. Wang, K. Fan, B. Li, F. Zhou, B. Chen, Y. Shi, and L. Si, "Alibaba submission for WMT18 quality estimation task," in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, Oct. 2018, pp. 809–815.

[9] J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing, "Confidence estimation for machine translation," in *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. Geneva, Switzerland: COLING, aug 23–aug 27 2004, pp. 315–321.

[10] R. Soricut and A. Echihabi, "TrustRank: Inducing trust in automatic translations via ranking," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, Jul. 2010, pp. 612–621.

[11] N. Ueffing and H. Ney, "Word-level confidence estimation for machine translation using phrase-based translation models," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada: Association for Computational Linguistics, Oct. 2005, pp. 763–770.

[12] L. Dong, C. Quirk, and M. Lapata, "Confidence Modeling for Neural Semantic Parsing," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 743–753.

[13] M. Rikters and M. Fishel, "Confidence through Attention," in *Machine Translation Summit XVI*, Nagoya, Japan, 2017.

[14] T. Nakazawa, M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kurohashi, and H. Isahara, "ASPEC: Asian Scientific Paper Excerpt Corpus," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, N. C. C. Chair), K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Portorož, Slovenia: European Language Resources Association (ELRA), may 2016, pp. 2204–2208.

[15] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A Fast, Extensible Toolkit for Sequence Modeling," in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.