Select the Best Translation from Different Systems without Reference

Jinliang Lu and Jiajun Zhang

University of Chinese Academy of Sciences
National Laboratory of Pattern Recognition, CASIA
lujinliang2019@ia.ac.cn, jjzhang@nlpr.ia.ac.cn

Abstract. In recent years, neural machine translation (NMT) has made great progress. Different models, such as neural networks using recurrence, convolution and self-attention, have been proposed and various online translation systems can be available. It becomes a big challenge on how to choose the best translation among different systems. In this paper, we attempt to tackle this task and it can be intuitively considered as the Quality Estimation (QE) problem that requires enough human-annotated data in which each translation hypothesis is scored by human. However, we do not have rich data with high-quality human annotations in practice. To solve this problem, we resort to bilingual training data and propose a new method of mixed MT metrics to automatically score the translation hypotheses from different systems with their references so as to construct the pseudo human-annotated data. Based on the pseudo training data, we further design a novel QE model based on Multi-BERT and Bi-RNN with a joint-encoding strategy. Extensive experiments demonstrate that our proposed method can achieve promising results for the task to select the best translation from various systems.

Keywords: Machine Translation · Evaluation · Deep Learning

1 Introduction

With the development of neural machine translation (NMT), online machine translation platforms can give users more suitable and fluency translation[26]. Various systems use different translation models, ranging from RNN[2] to Transformer[22]. For a particular sentence, with diversiform decoding methods[12, 16, 27, 28], NMT models will produce translations with different qualities. How to judge which one is more reliable is an ubiquitous but challenging problem as we have no reference in practice.

In this paper, we aim to tackle this task – selecting the best translation from different systems without reference. For this problem, there are some difficulties need to recover. First, although this task can be treated as the well-studied QE problem, it needs enough human-annotated scores as labels while it is hard to get enough high-quality annotated data for training in practice. Second, given the annotated training data, we further need to design a more sophisticated QE

model which can distinguish the difference between similar translations and give accurate scores. To solve these problems, we propose novel methods and make the following contributions:

1. To solve the problem of lack of annotated data, we resort to the large scale bilingual training data and let different translation systems translate the source sentences of the bitext. We propose a new MT metric enriched with the BERT sentence similarity to score the translation hypotheses from different systems and employ the scores to construct the pseudo human annotations.

2. To further improve QE models, we introduce the joint-encoding technique for both source sentence sand its translation hypothesis based on Multi-BERT.

3. We also analyze the reason why joint-encoding with Multi-BERT can bring improvements in cross-lingual tasks.

The extensive experiments show that our method is effective in various real scenarios for the best translation selection. To test the performance of our proposed mixed MT metric, we conduct experiments on WMT 15 metric shared task and the result demonstrates that our mixed metric can get the best correlation with human direct assessment (DA) scores. We also test our QE model on WMT 18 shared task and we observe from the experiments that our model correlates better with sentence-level Terp score than existing QE methods.

2 Data Construction Strategy Based on Mixed Metrics

As we described above, MT evaluation without reference always needs a big amount of annotated data. Even for similar language pairs, current SOTA QE models still need parallel corpus for pre-training to get a better result. Scores judged by bilingual experts can be trusted. However, too much data to label can be time-consuming and impractical. Although WMT provides human DA scores for News Translation task with the quality assurance every year, the annotated data is still insufficient in some language pairs. In order to get enough annotated data, we integrate current outstanding metrics and cosine-similarity of BERT representations(candidate and its reference) smoothly into a new metric using the SVR regression model. The final score can be calculated as

$$score = \sum_{i=0}^{n} \omega_i \varphi(x_i) + b \tag{1}$$

Where n is the number of metrics we fuse, φ is the kernel function, ω_i is the weight for the i-th metric score, b is the bias. The metrics are listed in table 1.

3 Translation Score model with Joint-Encoding

Traditional QE model aims at formulating the sentence level score as a constraint regression problem respectively. One of the representative methods is QuEst++[20], whose feature extractor is rule-based and regression model is a

Table 1. Basic metrics we used for fusion strategy.

ID	Metric
1	BERT-Layer1-12 cosine similarity
2	RUSE[17]
3	BEER[21]
4	CharacTER[23]
5	TERp[18]

SVM. Recently, researchers begin to extract effective features through neural networks, such as POSTECH [9], UNQE[11], Bilingual Expert[7] and deepQuest[8]. In spite that these neural-based feature extractors take the source sentence information into account, their main module is the language model of the target language. Obviously, tokens in the source sentence and its machine translation may not interact with each other, which can be more useful in QE.

With the advent of pre-trained language model like ELMo[14], GPT[15], BERT[6], multilingual version LMs, like Multi-BERT, XLM[10] appeal to our attention. These models are based on Transformer[22], a neural network which can help every token to get attention weights from other tokens. We choose Multi-BERT to do our tasks.

3.1 Cross-lingual Joint Pre-training With Multi-BERT

Even though Multi-BERT is multilingual, in its pre-training process, it is still trained language by language. We aim to adjust the model to be familiar to inputs combined by both source sentence and target sentence. Therefore, we train Multi-BERT with parallel data again through the joint-input way.

Model Architecture and Input Example We keep the architecture of Multi-BERT, whose layer number L=12, hidden state H=768, attention heads A=12. The input representation is also as same as original model. We don't change the position embedding like XLM[10] because we want to emphasize the precedence order of source sentence and its reference or its translation.

Pre-Training Method The training task can also be divided into two parts like BERT[6]. The first one is masked token prediction and the second one is translation prediction. Different from the process of pre-training in BERT, we force that [MASK] can only appear in the target sentence. We hope the model can capture all the source information so that it can predict masked tokens in target sentence easily. The total procedure can be described in figure 1.

3.2 Fine-tune with Multi-BERT for QE

Sentence level QE is a sequence regression task. The basic way to handle sequence regression task is to take the final hidden state for the first token in the input.

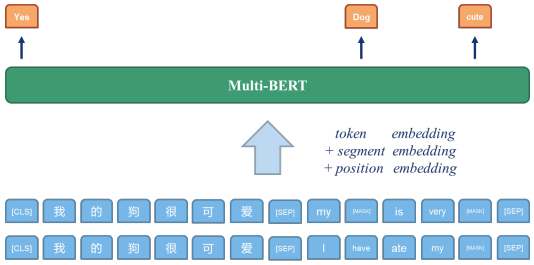


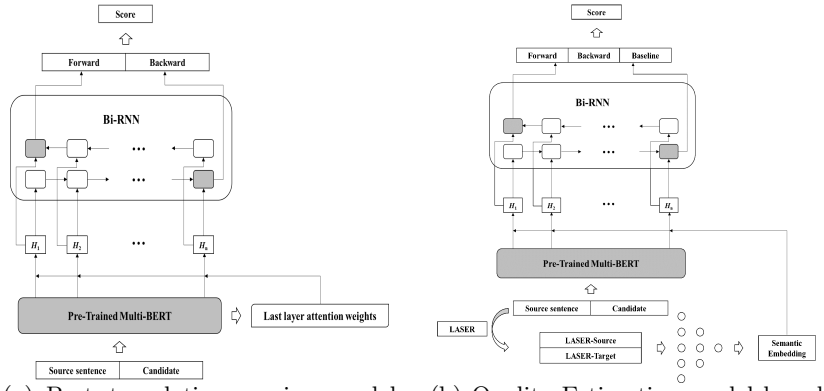
Fig. 1. The training method for Multi-BERT with parallel data

However, for handling long-distance dependency, we apply a single layer Bi-RNN behind BERT. We illustrate the model in figure 2(a).

In the model of Bi-RNN, we set the hidden size $H_2 = 768$ and insure the sequence length is same as Multi-BERT. Finally, we joint the final state from both directions and get a score by a weight matrix.

$$Score_{quality} = W_o \times [\overrightarrow{h_T}; \overleftarrow{h_T}] \tag{2}$$

Where W_o is the weight matrix and $[\overrightarrow{h_T}; \overleftarrow{h_T}]$ is the final states of forward and backward directions.



(a) Best translation scoring model based on Multi-BERT

(b) Quality Estimation model based on Multi-BERT

Fig. 2. Cross-lingual scoring models based on Multi-BERT

4 Experiment

4.1 Select Best Translation Based on Multi-BERT

In this part, we will describe the result of the best translation selection task. The score model we use is as same as what we illustrate in figure 2(a).

Experimental settings We conduct this part experiments in language direction from Chinese to English. First, we collect a group of translations from three different translation systems. One source sentence is aligned to three translations. In order to judge the transfer ability of our model, we also collect samples from WMT 2017 Metric shared task in the language direction from Chinese to English whose distribution is not as same as our data. The basic information of the dataset is listed in table 2.

Table 2. Statistics for the best translation selection task in the language direction zh-en.

Dataset	Samples	Sentence pairs
Training set	361,414	1,084,242
Test-In set	19,017	57,051
Test-Out set (from WMT17)	1,184	3,552

BERT version is BERT-Base, Multilingual Cased: 104 languages, 12-layer, 768-hidden, 12-heads, 110M parameters. We choose GRU as basic unit for Bi-RNN, whose hidden layer is 1, hidden size is 1536. For the pre-training of parallel corpus, we pick up 2M Chinese-English parallel data. The training based on Multi-BERT cost 1 week on a single GPU. In the process of fine-tuning for scoring translations, for all models, the epochs are restricted at 3. Batch size is 32. The learning rate is 2e-5.

Experimental results The experiment result is shown in table 3.

- Para-Trained Multi-BERT: The name of our model described in Section 3.
- No-Trained Multi-BERT: For comparison, we also use the original Multi-BERT to do the experiment.
- LASER-cosine similarity: We use the representation for source sentence and target sentence from LASER. We calculate the similarity of the sentence pair as the quality score of the translation.

We can get conclusions from table 3:

1. Multi-BERT trained with parallel data before being applied into the scoring model can be more accurate in selecting the best translation task.

Table 3. Results of best translation selection task in the language direction zh-en.

Model	Pearson	Spearman	Best Selection Accuracy
Para-Trained Multi-BERT	**0.7246**	**0.6929**	**57.73%**
No-Trained Multi-BERT	0.7109	0.6740	56.21%
LASER-cosine similarity	0.3705	0.3191	38.91%

2. The experimental results show that calculating cosine similarity for the two sentences' embeddings obtained from LASER is not as good as supervised method like fine-tuning by Multi-BERT.

Table 4 and figure 3 show the size of training set can affect the result. With the training set getting bigger, the best translation selection task result gets better. When the training size is enough big, the result becomes stable.

Table 4. Influence of training size on the result of the best translation selection task in the language direction zh-en.

Sentence paris	Pearson	Spearman	Best Selection Accuracy
10k	0.6829	0.6529	55.67%
20k	0.6956	0.6665	55.92%
40k	0.7074	0.6762	56.40%
60k	0.7144	0.6831	57.66%
80k	0.7207	0.6895	57.68%
1M	**0.7246**	**0.6929**	**57.73%**

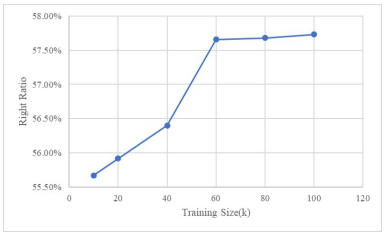


Fig. 3. Influence of training size on the result of best translation selection task in the language direction zh-en

In our common sense, the greater differences among the translations, the easier it is to tell them apart. In order to verify our model has the ability like human, we pick the samples from our test data according to the score gap which can reflect the difference between translations. Then, we calculate the best translation selection accuracy, which was shown in table 5.

Table 5. Influence of score gap on the result of the best translation selection task in the language direction zh-en.

Score Gap	Pearson	Spearman	Best Selection Accuracy
Random	0.7246	0.6929	57.73%
≥ 0.02	0.7212	0.6964	62.79%
≥ 0.04	0.7167	0.6996	66.54%
≥ 0.06	0.7157	0.7030	68.70%
≥ 0.08	0.7246	0.7157	70.86%
≥ 0.10	0.7211	0.7093	**72.90%**

Obviously, our model can get more and more accurate result as the score gap becomes bigger. When the score gap exceeds 0.1, the best translation selection accuracy can be 72.90%. The finding is as same as what we suspect.

In order to observe the transfer ability of our model, we also do the best translation task in Test-Out. As our constructed data's distribution is not as same as the human DA data, we want to see if the result drops greatly when it is tested in the data with different distribution. The result is shown in table 6.

Table 6. Influence of distribution on the result of the best translation selection task in the language direction zh-en.

Test Set	Best Selection Accuracy
Test-In	56.21%
Test-Out	40.70%

From the result shown in table 6, we can see that the result on human DA data is lower. However, it is still higher than 33.33%, the random selection result.

4.2 Mixed Metric for Data Construction

Experimental settings We use the SVR provided in sk-learn. The kernel function we used is RBF and the epsilon we set is 0.01. We obtain the data from WMT 15-17. The training set is the sentence pairs whose target language is English in WMT 16-17 and we use data obtained from WMT 15 for testing.

Experimental results In table 7, we evaluate our mixed metric on two types of correlation index, Pearson and Spearman. Our metric improves the Pearson correlation from 75% to 77%, outperforming RUSE by 4% to 7% accuracy respectively. We get the similar result in Spearman index, which shows that our mixed metric is strongly correspond with human judgment.

Table 7. Segment-level Pearson and Spearman correlation of metric scores and DA human evaluation scores for to-English language pairs in WMT15.

Index	Pearson				Spearman			
Languages	cs-en	de-en	fi-en	ru-en	cs-en	de-en	fi-en	ru-en
Fuse-SVR	**0.760**	**0.772**	**0.772**	**0.755**	**0.752**	**0.746**	**0.757**	**0.727**
RUSE[17]	0.703	0.732	0.707	0.712	0.694	0.708	0.680	0.684
BERT-Layer12	0.550	0.543	0.550	0.531	0.589	0.585	0.612	0.570
characTER[23]	0.552	0.608	0.584	0.629	0.536	0.593	0.542	0.594
BEER[21]	0.555	0.595	0.602	0.621	0.539	0.545	0.552	0.579
TERp[18]	0.485	0.559	0.531	0.569	0.480	0.530	0.482	0.545

4.3 QE model with joint-encoding and LASER cosine similarity

In this part, our QE model is a bit different from what we describe in figure 2(a). We concatenate the LASER cosine similarity into the token level and the baseline feature before the final weight matrix to get a more accurate result. We concatenate the LASER[1] representations of source sentence and its translation. Through a DNN, we can get a fixed dimensional representation of the similarity of cross-lingual sentence pair. The model is shown as figure 2(b).

Experimental settings In the LASER model, DNN output size is 512. We choose GRU as basic unit for Bi-RNN, whose hidden layer is 1, hidden size is 1280. The number of baseline features is 17. We use the parallel data of German and English from WMT, whose total sentence pairs is 2M. BERT version and other settings are same as described in Section 4.1.

Table 8. Results of sentence level QE on WMT 2018 shared task de-en.

Model	Pearson	Spearman	MAE	RMSE
Train+Baseline+LASER	**0.7814**	**0.7427**	**0.0921**	**0.1292**
UNQE[11]	0.7667	0.7261	0.0945	0.1315
Bilingual Expert[7]	0.7631	0.7318	0.0962	0.1328
No-train+Baseline+LASER	0.7533	0.7083	0.0974	0.1359
Split+Concat	0.3853	0.3440	0.1582	0.2049
Baseline-QuEst++[20]	0.3323	0.3247	0.1508	0.1928

Experimental results We conduct the experiment in the language pair: German to English. The result is shown in table 8.

– Train+Baseline+LASER: We add the baseline features and laser features into the model based on Multi-BERT trained by parallel data.

– No-train+Baseline+LASER: Different from the above, we just use the original Multi-BERT.
– Split + Concat: In order to prove the joint-encoding is effective, we put the source sentence and target sentence into Multi-BERT separately and concatenate the outputs from BERT before putting into Bi-RNN.

We can see that our parallel-trained BERT get the best result in WMT 18 QE shared task in DE-EN direction, outperforming Bilingual Expert and UNQE more than 1% in Pearson and Spearman correlation. However, original Multi-BERT cannot surpass Bilingual Expert[7], which shows that trained with parallel corpus by joint-encoding way can help Multi-BERT capture the relationship between source sentence and target sentence accurately. We will explain this finding in Section 5. From the table, We also find that encoding sentence independently by Multi-BERT and then joint the hidden states cannot get a satisfying result. We suspect the reason is that the two sentences cannot interact with each other and a single layer Bi-RNN is not enough to capture their inner relations.

5 Analysis

In this section, we will briefly analyze the influence of joint-encoding pre-training for cross-lingual tasks. We give our explanation in two aspects, cross-lingual word translation accuracy and cross-lingual attention distribution.

5.1 Word Translation Accuracy

Context word embedding can be changed when the same word in different sentences. We suspect that our joint-encoding pre-training strategy can changed the word embedding space to some extent and the words whose semantics are similar in two different languages can be made close to each other. To verify our hypothesis, we acquire the bilingual dictionary MUSE[5] used. We put the words into Multi-BERT and our parallel-trained Multi-BERT to get the word embeddings one by one. As each word is cut into word pieces, we calculate the average of all the word pieces' embeddings as the word embedding.

We calculate cosine-similarity for each word-pair, including internal language words and external language words. We count the number of words of its translations in the top five most similar words, which was list in table 9.

Table 9. The information of word translations at top-5 most similar words list.

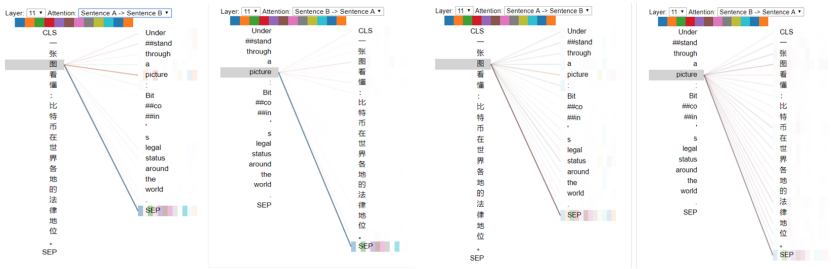
Model	Top@5 num	Total num	Top@5 accuracy
Original Multi-BERT	72	3065	2.349%
Parallel-trained Multi-BERT	1279	3065	41.729%

For words in English or Chinese, using Parallel-Trained Multi-BERT to get the representations, their translations in the other language can appear in the

Top5 most similar word list at a high ratio, 41.729%, which improves greatly than original Multi-BERT. We think that it can be useful in cross-lingual tasks.

5.2 Cross-lingual attention distribution

We also observe the attention weights from source sentence to its reference or translation. Interestingly, we find that words with similar sematic in two languages can mind each other in Parallel-Trained Multi-BERT, as is shown in figure 4(a). However, original Multi-BERT provides attention weights approximately averagely for words as is shown in figure 4(b).



(a) attention weights get from oiginal parallel-trained Multi-BERT

(b) attention weights get from oiginal Multi-BERT

Fig. 4. Cross-lingual attention visualization in different models with joint-encoding

We think that joint-encoding pre-training can also help words in different languages mind each other, especially the words have similar semantic. And this is the second reason we find that joint-encoding is useful in cross-lingual tasks.

6 Related work

To construct enough data, we use the fusion strategy to get a better metric that contains advantages of other metrics. DBMFcomb[24] used the fusion method in WMT 2015. Differently, it is designed to do classification. In 2017, BLEND[13], which was mixed by 57 metric, won the first in WMT 2017 Metric shared task.

In 2014, Zhang et al.[25] proposed bilingually-constrained phrase embeddings to estimate the quality of phrase-level translation. From 2015, Quality Estimation has made great progress. Current baseline model is QuEst++[20]. These years, more and more researchers begin to use neural network to solve the problem. Kim et al. presented POSTECH[9], an estimator-predictor framework based on RNN. UNQE[11] is modified from POSTECH, which combines the estimator and predictor together to help its feature extractor get more useful information for regression. Bilingual Expert[7] is the SOTA model, whose feature extractor is based on Transformer[22].

7 Conclusion

In this paper, we present novel methods to tackle the task of selecting the best translation from different systems without reference. To construct enough annotated data, we design a new MT metric which is mixed with other effective metrics to automatically obtain pseudo human-annotated scores. To improve the QE model, we propose a novel method that uses joint-encoding strategy to handle this kind of cross-lingual task. Experimental results verify the effectiveness of our method in choosing the best translation from various systems. Furthermore, the supplementary experiments and analysis demonstrate the superiority of our proposed mixed MT metric and QE model.

Acknowledgement

The research work descried in this paper has been supported by the National Key Research and Development Program of China under Grant No. 2016QY02D0303, the Natural Science Foundation of China under Grant No. U1836221 and the Beijing Municipal Science and Technology Project under Grant No. Z181100008918017.

References

1. Artetxe, Mikel and Schwenk, Holger: Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. arXiv preprint arXiv:1812.10464 (2018)
2. Bahdanau, Dzmitry and Cho, Kyunghyun and Bengio, Yoshua: Neural machine translation by jointly learning to align and translate. In Proceedings of ICLR (2015)
3. Blatz, John and Fitzgerald, Erin and Foster, George and Gandrabur, Simona and Goutte, Cyril and Kulesza, Alex and Sanchis, Alberto and Ueffing, Nicola: Confidence estimation for machine translation. In Proceedings of COLING (2004)
4. Bojar, Ondřej and Chatterjee, Rajen and Federmann, Christian and Graham, Yvette and Haddow, Barry and Huang, Shujian and Huck, Matthias and Koehn, Philipp and Liu, Qun and Logacheva, Varvara and others: Findings of the 2017 conference on machine translation. In Proceedings of WMT (2017)
5. Conneau, Alexis and Lample, Guillaume and Ranzato, Marc'Aurelio and Denoyer, Ludovic and Jégou, Hervé: Word Translation Without Parallel Data. In Proceedings of ICLR (2018)
6. Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina: Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT (2019)
7. Fan, Kai and Li, Bo and Zhou, Fengming and Wang, Jiayi: "Bilingual Expert" Can Find Translation Errors. In Proceedings of AAAI (2019)
8. Ive, Julia and Blain, Frédéric and Specia, Lucia: DeepQuest: a framework for neural-based quality estimation. In Proceedings of COLING (2018)
9. Kim, Hyun and Jung, Hun-Young and Kwon, Hongseok and Lee, Jong-Hyeok and Na, Seung-Hoon: Predictor-Estimator: Neural quality estimation based on target word prediction for machine translation. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), **17**(1), 3. (2017)

10. Lample, Guillaume and Conneau, Alexis: Cross-lingual Language Model Pretraining. arXiv preprint arXiv:1901.07291 (2019)
11. Li, Maoxi and Xiang, Qingyu and Chen, Zhiming and Wang, Mingwen: A Unified Neural Network for Quality Estimation of Machine Translation. IEICE TRANSACTIONS on Information and Systems, **101**(9), 2417–2421 (2018)
12. Liu, Lemao and Utiyama, Masao and Finch, Andrew and Sumita, Eiichiro :Agreement on targetbidirectional neural machine translation. In Proceedings of NAACL-HLT (2016)
13. Ma, Qingsong and Graham, Yvette and Wang, Shugen and Liu, Qun: Blend: a Novel Combined MT Metric Based on Direct Assessment—CASICT-DCU submission to WMT17 Metrics Task. In Proceedings of WMT (2017)
14. Peters, Matthew and Neumann, Mark and Iyyer, Mohit and Gardner, Matt and Clark, Christopher and Lee, Kenton and Zettlemoyer, Luke: Deep Contextualized Word Representations. In Proceedings of NAACL-HLT (2018)
15. Radford, Alec and Narasimhan, Karthik and Salimans, Time and Sutskever, Ilya: Improving language understanding with unsupervised learning. Technical report, OpenAI (2018)
16. Sennrich, Rico and Haddow, Barry and Birch, Alexandra: Edinburgh neural machine translation systems for wmt 16. In Proceedings of WMT (2016)
17. Shimanaka, Hiroki and Kajiwara, Tomoyuki and Komachi, Mamoru: RUSE: Regressor Using Sentence Embeddings for Automatic Machine Translation Evaluation. In Proceedings of WMT (2018)
18. Snover, Matthew G and Madnani, Nitin and Dorr, Bonnie and Schwartz, Richard: TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate. Machine Translation, **23**(2-3), 117–127 (2015)
19. Specia, Lucia and Turchi, Marco and Cancedda, Nicola and Dymetman, Marc and Cristianini, Nello: Estimating the sentence-level quality of machine translation systems. In Proceedings of EAMT (2009)
20. Specia, Lucia and Paetzold, Gustavo and Scarton, Carolina: Multi-level translation quality prediction with quest++. In Proceedings of ACL-IJCNLP (2015)
21. Stanojević, Miloš and Sima'an, Khalil: BEER 1.1: ILLC UvA submission to metrics and tuning task. In Proceedings of WMT (2015)
22. Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, Łukasz and Polosukhin, Illia: Attention is all you need.In Proceedings of NeurIPS (2017)
23. Wang, Weiyue and Peter, Jan-Thorsten and Rosendahl, Hendrik and Ney, Hermann: Character: Translation edit rate on character level. In Proceedings of WMT (2016)
24. Yu, Hui and Ma, Qingsong and Wu, Xiaofeng and Liu, Qun: Casict-dcu participation in wmt2015 metrics task. In Proceedings of WMT (2015)
25. Zhang, Jiajun and Liu, Shujie and Li, Mu and Zhou, Ming and Zong, Chengqing: Bilingually-constrained Phrase Embeddings for Machine Translation. In Proceedings of ACL (2014)
26. Zhang, Jiajun and Zong, Chengqing: Deep Neural Networks in Machine Translation: An Overview. IEEE Intelligent Systems, **30**(5), 16-25 (2015)
27. Zhou, Long and Zhang, Jiajun and Zong, Chengqing: Synchronous Bidirectional Neural Machine Translation. Transactions of Association for Computational Linguistics (TACL), **7**, 91-105 (2019)
28. Zhou, Long and Zhang, Jiajun and Zong, Chengqing: Sequence Generation: From Both Sides to the Middle. In Proceedings of IJCAI (2019).