A Document-Level Machine Translation Quality Estimation Model Based on Centering Theory

Yidong Chen[1,2(✉)], Enjun Zhong[1,2], Yiqi Tong[1,2,3], Yanru Qiu[1,2], and Xiaodong Shi[1,2]

[1] Department of Artificial Intelligence, School of Informatics, Xiamen University, Xiamen, China
[2] Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan, Ministry of Culture and Tourism, Xiamen, China
[3] Institute of Artificial Intelligence, Beihang University, Beijing, China
{ydchen,mandel}@xmu.edu.cn, {ejzhong,yqtong,yrqiu}@stu.xmu.edu.cn

Abstract. Machine translation Quality Estimation (QE) aims to estimate the quality of machine translations without relying on golden references. Current QE researches mainly focus on sentence-level QE models, which could not capture discourse-related translation errors. To tackle this problem, this paper presents a novel document-level QE model based on Centering Theory (CT), which is a linguistics theory for assessing discourse coherence. Furthermore, we construct and release an open-source Chinese-English corpus at https://github.com/ydc/cpqe for document-level machine translation QE, which could be used to support further studies. Finally, experimental results show that the proposed model significantly outperformed the baseline model.

Keywords: Machine translation · Document-level quality estimation · Centering theory

1 Introduction

Machine translation quality estimation (QE) is a task that aims at automatically estimating the quality of machine translations. Unlike the standard evaluation metrics such as BLEU [15], NIST [4] and METEOR [1], QE models estimate translations without relying on golden references. In the past decade, researches on QE have attracted more and more attentions [7], since QE can be utilized to ensure the diversity and robustness of the NMT systems [25].

Currently, mainstream QE-related researches [2,13,26] mainly focus on sentence-level QE models, which normally ignore the document-level information. While, previous studies [21,23] have shown that document-level information is important for estimating the translation qualities. As shown in Fig. 1, the word

J. Su and R. Sennrich (Eds.): CCMT 2021, CCIS 1464, pp. 1–15, 2021.
https://doi.org/10.1007/978-981-16-7512-6_1

Context source: 五月份，俄罗斯农业监督局宣布，本国农业年度对华粮食出口首次超过100万吨，创新记录。
Context translation: In May, the Russia Agricultural Supervision Bureau announced that for the first time in the agricultural year, Russia's grain exports to China exceeded 1 million tons, setting a new record.

Current source: 该局预测中国可能进入前十大俄罗斯粮食进口国之列。
Current translation: The bureau predicts that China may be among the top ten Russian food importers.

Fig. 1. An example of a translation that is correct in sentence-level but incorrectly in document-level. We use THUMT [20] and 2M Chinese-English parallel data to training the NMT model.

"predicts" in current translation should be "predicted" according to the context, but is wrongly translated into present tense. Obviously, a QE model that does not consider the document-level information could not tell the above-mentioned error.

To alleviate this problem, we propose a document-level QE model called CpQE by introducing Centering Theory (CT) [24] to formulate the sentence relations. Concretely, our CpQE model uses the Preferred Center (Cp), whose meaning could be found in Subsect. 3.1, to represent the context features. Moreover, we adapt a BERT-based [3] sequence labeling model to extract the Cps. In addition, a semi-supervised pseudo-label learning method is adopted to alleviate the low resource problem of Cp extraction.

2 Related Work

Traditional QE works [6,17] used feature engineering to extract features, e.g. QuEst++ [19] design word-, sentence- and document-level features for multi-level QE. Recently, neural QE methods outperformed these hand-craft methods. [16] treated QE as a slot filling problem and proposed a language independent word-level QE system using Recurrent Neural Network (RNN). [14] proposed a stacked model by introducing multi-task learning, which achieved the best result for word-level and sentence-level QE at that time.

More recently, Predictor-Estimator framework [10] was reported superior performance and become a mainstream approach for neural QE. To combine Predictor and Estimator into the architecture, [13] proposed a unified neural network, which were trained jointly to minimize the mean absolute error over the QE training samples. Furthermore, [5] proposed a neural bilingual expert model, which replaced the RNN layers with a novel bidirectional transformer [22] for feature extraction. And [11] apply the pre-trained model, BERT [3], as feature extractor. However, these methods evaluate each translation independently, leading to an inconsistent problem for the evaluation of document-level machine translation.

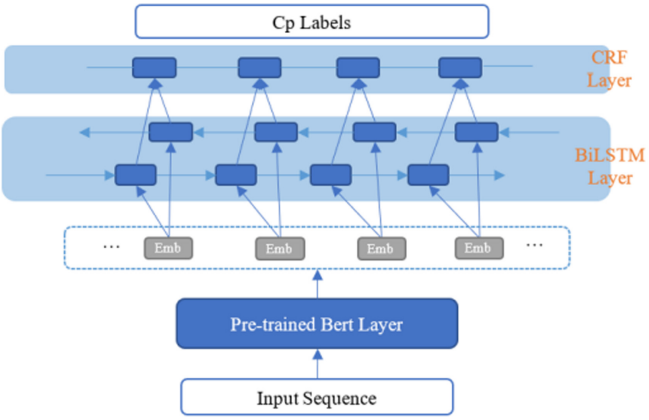

Fig. 2. The overview of Preferred Centering extraction model

3 Centering Theory and Extraction of the Preferred Centers

3.1 Centering Theory and Preferred Centers

Centering Theory (CT) [8,9,24] is a theoretical model about the local coherence of discourses. CT, which can be parameterized and calculated easily compared with other related theories, provides a quantitative standard for evaluating the context consistency of translations. Therefore, in this work, we apply CT to capture the discourse coherence information for document-level QE.

In CT, any entity in a sentence may relate to entities in the following sentences. So an entity is called Forward-looking Center (Cf). And an entity related to entities in the previous sentences is called Backward-looking Center (Cb). Preferred Center (Cp) is the entity that is the most likely one to be associated with a Cb. For example, given a current sentence "Xiao Hong likes to wear a red skirt" and the following sentence "She went shopping today and met Xiao Fang". The entities in the current sentence include "Xiao Hong" and "skirt", so we have Cf = ["Xiao Hong", "skirt"]; and the Cb in following sentence is "she", i.e. Cb = ["she"]. In Cf, the word "Xiao Hong" is the most closely related to the Cb, so "Xiao Hong" is defined as the preferred center. It should be noted that a sentence may contains more than one Cps.

3.2 The Preferred Centers Extraction Model

The conventional methods for extracting Cp are mainly rule-based. While, in this paper, we take this problem as a sequence labeling problem and construct a BERT-BiLSTM-CRF based model to settle it.

Figure 2 presents the overview of our extraction model. The input sentences are encoded by BERT first. Then, the output of BERT are fed to a BiLSTM layer, in which the operations of the LSTM are shown as follows:

Table 1. The format of preferred center annotation.

Chinese example	
current sentence	小_B 明_I 和_O 小_B 红_I 决_O 定_O 去_O 电_B 影_I 院_I ._O (Xiao Ming and Xiao Hong decided to go to the cinema)
following sentence	他们看了一场精彩的电影 (They watched a wonderful movie.)
English example	
current sentence	Brennan_B drives_O an_O Alfa_O Romeo_O ._O
following sentence	She drives too fast.

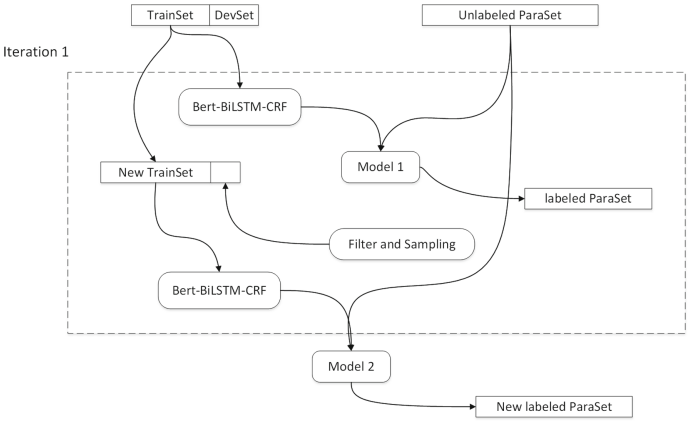

Fig. 3. The pipeline of our semi-supervised training method

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i), \tag{1}$$
$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f), \tag{2}$$
$$c_t = f_t c_{t-1} + i_t \tanh(W_c[h_{t-1}, x_t] + b_c), \tag{3}$$
$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o), \tag{4}$$
$$h_t = o_t \tanh(c_t), \tag{5}$$

where x_t represents the output of BERT. i_t, f_t and c_t are the input gate, forget gate and cell vectors, respectively. o_t is the output gate and h_t is the hidden vector. t represents the t-th cell state of LSTM.

After that, the output of the forward and the backward LSTM are concatenated using (6), as follows:

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{6}$$

Finally, the outputs of BiLSTM are provided to Conditional Random Field (CRF) [12] to decode the Cp labels.

3.3 The Semi-supervised Preferred Center Extraction Method

Since there are no public datasets for Cp extraction, we manually annotated a small-scale Cp extraction dataset. Concretely, the English corpus is annotated in word-level while the Chinese corpus is annotated in character-level. Table 1 shows the format of annotation. Considering that such a small annotated dataset is not enough for training a automatic annotation model, we proposed a semi-supervised method to do so. The training pipeline is shown in Fig. 3.

First, we divided the annotated dataset into training set and development set. Then we trained the BERT-BiLSTM-CRF model with these two sets to get Model 1. After that, we predict the unlabeled parallel corpus with Model 1 to get a labeled dataset. Next, we filtered the labeled data by rules to alleviate the effect of noise. Here are the rules we define:

- Remove the sentences whose ratio of the total length of preferred centers to the total length of sentence is more than 1/4.
- Calculate the maximum similarity between each preferred center and the words in the following sentence. If the similarity is less than 0.5 and such preferred center do not belong to any component of subject, direct object or indirect object, record this preferred center. If the number of such kind of preferred center is greater than or equal to 50% of the number of preferred centers extracted from the sentence, the sentence will be removed.

Roughly, Rule 1 limits the number of preferred centers to avoid selecting excessive entities as the preferred centers for higher recall, and Rule 2 remove the samples which contain ambiguous Cp. For measuring the similarity between words, we use a word2vec model[1] to encode the words into vectors and calculate their cosine similarity:

$$similarity(w_i, w_j) = \frac{emb_i * emb_j}{||emb_i|| * ||emb_j||} \tag{7}$$

where emb_i is the vectorized representation of w_i. If the out-of-vocabulary word can not be found in the following sentence, the similarity is set to be 0, otherwise 1.

After filtering the labeled dataset, the dataset will be randomly sampled to get three sampling datasets. These three datasets will be combined with the initial training set respectively for training three new models. Then we choose the highest recall model on development set as Model 2. Our goal is to obtain comprehensive preferred centers as far as possible so we choose the recall to select the optimal model. So far, we have completed one iteration. The next step is to repeat the previous steps.

4 The Quality Estimation Model

In this section, we present our CT-based document-level QE model. As shown in Fig. 4, we extract the features of preferred centers from two aspects by outer-extractor. First, we get the embeddings of preferred centers in both source and

[1] https://radimrehurek.com/gensim/models/word2vec.html.

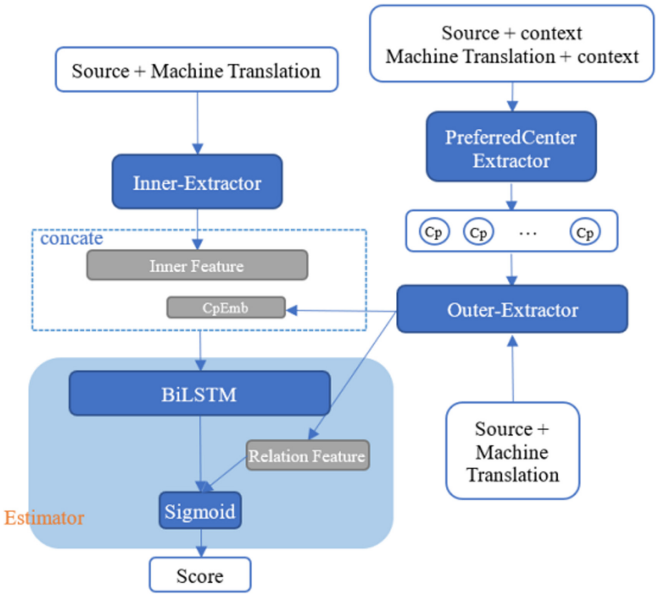

Fig. 4. The overview framework of our CpQE model

target side. Second, compute the consistency between current sentence and context in both source and target side. Finally, the two types of features and the inner sentence features extracted by inner-extractor are passed to the quality evaluator for scoring.

4.1 The Inner-Extractor

As shown in Fig. 5, the encoder of inner-extractor is a standard encoder of transformer [22] and the decoder is bidirectional. The forward self-attention network decodes the target words from left to right, while the backward self-attention network decodes the target words from right to left. The combination of the two self-attention can make the model focus on the whole sentence.

4.2 The Outer-Extractor

Outer-Extractor extract Cp features from two aspects: sentences relation features and embeddings of preferred centers. Sentences relation features can evaluate the coherence between source text and translations. Here we define four rules for designing features:

- The number of preferred centers of current sentence in source and target side and the difference between the numbers.
- The number of preferred centers of previous sentence in source and target side and the difference between the numbers.

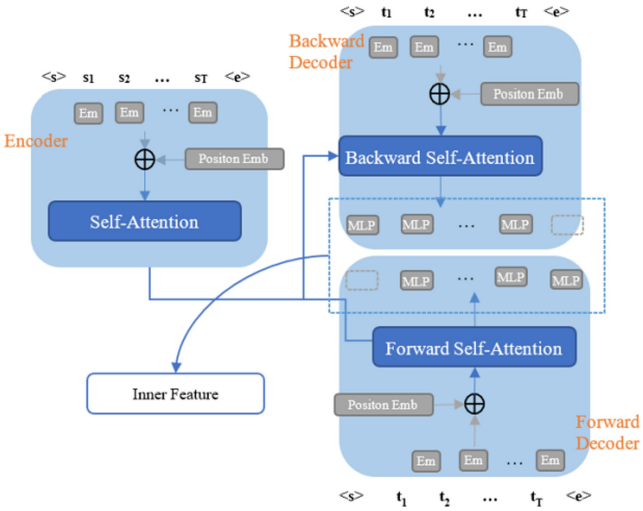

Fig. 5. The architecture of inner-extractor

- The similarity between preferred centers of previous sentence and current sentence in source and target side and the difference between the similarities.
- The similarity between preferred centers of previous sentence and preferred centers of current sentence in source and target side and the difference between the similarities.

Rule 1 and rule 2 focus on the number of preferred centers which can reflect the consistency between source text and translation at some extent. Rule 3 use a quantitative measurement to evaluate the consistency between previous sentence and current sentence. Rule 4 measure the change of entities which reflects the change of topic. If a sentence at the beginning of document, the preferred center of the previous is empty set. The preferred center of the last sentence in document is empty set too. The similarity between the sequence is computed as follow:

$$\begin{aligned} similarity(l_1, l_2) = & \frac{l_{v1} + l_{v2}}{L_1 + L_2} cosine(emb_{w_{v1}}, emb_{w_{v2}}) \\ & + \frac{2}{L_1 + L_2} \sum_{win w_{o1}} f(w, l_2) \end{aligned} \tag{8}$$

$$f(w, l_2) = \begin{cases} 1, w \ in \ l_2, \\ -1, w \ not \ in l_2. \end{cases} \tag{9}$$

where w_{v1} is the word in the sequence 1 which can be found in vocabulary while w_{o1} is the word in the sequence 1 which out of the vocabulary. l_{v1} is the length of w_{v1} and L_1 is the length of the sequence 1. w_{v1} and w_{v2} are calculated by Word2Vec model. According to the four rules, we design 12 features to represent sentence relation information. We provide the running process of outer-extractor on Appendix A.

4.3 The Evaluator

Finally, we provide the features to evaluator. Since the preferred center embedding is a word-level feature, and the local sentence relation feature is for both sentence and context, we integrate the preferred center embedding before BiLSTM. And the sentence relation feature is concatenated with the whole sentence feature output by BiLSTM:

$$\overrightarrow{h_{1:T+n}}, \overleftarrow{h_{1:T+n}} = BiLSTM(f) \tag{10}$$

$$f = [f_{inner}; CpEmb] \tag{11}$$

where T is the length of translation, n is the number of preferred centers. f_{inner} represents the features extracted by inner-extractor. The sentence relation feature can make the evaluator focus on consistency between source text and translation. Finally, sigmoid function is used σ to score the translations:

$$Score = \sigma(w^T[\overrightarrow{h_{1:T+n}}; \overleftarrow{h_{1:T+n}}; f_{outer}]) \tag{12}$$

where w is a trainable parameters, f_{outer} is the features extracted by outer-extractor. The optimization object is calculate as follows:

$$argmin||HTER - Score||_2^2 \tag{13}$$

$$HTER = \frac{N_{edit}}{N_{reference}} \tag{14}$$

where N_{edit} is the number of edits from translation to reference, $N_{reference}$ is the number of words in reference. Human-targeted Translation Edit Rate (HTER) [18] is the widest used metric of QE. Calculation of HTER need to find out the closest reference of the translation, then calculate the edit rate from translation to reference.

5 Experiments

5.1 Metrics

For preferred centers extraction, our goal is to maximize the total number of preferred centers that are correctly tagged by our method, so we use standard Accuracy and Recall score[2] to measure the performance of our BERT-based extraction model.

For quality estimation model, following with previous works such as [5,14], we use Pearson correlation coefficient, which is calculated as follows.

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_X)^2 \sum_{i=1}^{n}(y_i - \mu_Y)^2}} \tag{15}$$

Where n is the number of samples, μ_X and μ_Y denote means of the samples. A larger coefficient represents that X and Y are more correlated.

[2] https://github.com/chakki-works/seqeval.

Table 2. Preferred center extraction performance

Chinese model	Recall	Accuracy	Training set
Rule base model	38.26%	34.53%	–
Model 1	51.74%	47.18%	1000 labeled data
Model 2	57.01%	53.83%	1000 labeled data + 1000 pseudo labeled data
Model 3	60.70%	59.44%	1000 labeled data + 1500 pseudo labeled data
English model	Recall	Accuracy	Training set
Rule based model	40.43%	39.17%	–
Model 1	53.09%	49.32%	1000 labeled data
Model 2	56.84%	56.28%	1000 labeled data + 1000 pseudo labeled data
Model 3	63.61%	61.08%	1000 labeled data + 1500 pseudo labeled data

Table 3. Pearson correlation coefficient of models. CpQE+CpRuled represents the preferred centers are extracted by rule. CpQE+CpSeq represents the preferred centers are extracted by our sequence labeling model.

Model	Sentence testset	Document testset
Baseline	0.6392	0.5536
CpQE+CpRuled	0.6218	0.5911(+0.0375)
CpQE+CpSeq	0.6326	0.6035(+0.0499)

5.2 Dataset Description

Since the lack of document-level QE corpus, we manually annotated an open source Chinese-English document-level dataset[3]. Concretely, our document-level QE corpus is built from the test set of WMT2019 MT automatic evaluation task. We select 996 Chinese source sentences from the corpus, including 112 articles with a text length less than 14 sentences, and the corresponding 1992 sentences of English translations. The 1992 translations are calculated the HTER value to construct our corpus.

For the preferred center extraction experiment, we use our annotated preferred center extraction dataset including 1,432 Chinese sentences and 1,432 English sentences. The Chinese-English parallel corpus comes from FBIS corpus including 10,355 documents and 228,611 sentence pairs are used to generate pseudo labeled data.

For the quality estimation experiment, we use CCMT19 Chinese-English sentence-level translation quality estimation dataset with 11,213 sentences and

[3] Available at https://github.com/ydc/cpqe.

Table 4. Case study results.

Traslation results			
src	中 欧 班列 去程 首次 在 此 进行 班列 宽轨 和 标轨 的 换装 作业		
ref	for the first time, the china-europe train will carry out the reloading operation of the board rail and the standard rail.		
mt1	for the first time, the central european banlei will carry out the replacement of the banliewide rail and the standard rail .		
mt2	for the first time, the central china-europe banlei will carry out the replacement of the banliewide rail and the standard rail .		
Evaluation results			
system	baseline score	QE+CpSqe score	HTER
mt1	0.0897	0.0687	0.2272
mt2	0.0832(-7.25%)	0.0516(-24.89%)	0.1818(-19.98%)

our document-level QE corpus with 1992 sentences. We randomly select 50% sentences to delete or replace 20%–70% of the words and enhance the corpus up to 2,565 sentences. Word2Vec model are trained on 23GB Chinese-English monolingual corpus from Wikipedia and Sohu News. CCMT19 Chinese-English parallel corpus and FBIS Chinese-English corpus are used to train the inner-extractor.

5.3 Preferred Centers Extraction

In this experiment, we use a rule-based method as the baseline. In the rule-based method, Stanfordnlp is used for syntactic analysis. Noun subject, clausal subject, direct object, indirect object are chosen to be preferred centers. The setup of our model is presented in Appendix B.

The experiments results are shown in Table 2. Our semi-supervised training method train model for two iterations on both Chinese and English data. The recall and accuracy of Chinese Model 3 achieve 60.70% and 59.44% respectively. And English Model 3 achieve 63.61% recall and 61.08% accuracy. Both semi-supervised model significantly outperform the rule based model. The performance of each iteration is better than that of last iteration indicating that our proposed semi-supervised method can improve the performance of model. We choose the recall as metrics for the reason that we want to obtain comprehensive preferred centers as far as possible.

5.4 QE Results

In this experiment, we use Transformer-based feature extractor-evaluator as baseline model. Compared with the baseline, our model introduce an inner-extractor. The setup of CpQE model is shown in Appendix C. The result of quality estimation model is shown in Table 3. The Pearson correlation coefficients measure the correlation between model score and HTER. In the sentence-level QE, the difference among the three models is about 0.01. In document-level

QE, our CpQE+CpSeq model achieve the best performance with 0.6035, outperform the baseline by 0.0499. The rule-based Cp extractor with only 40.43% recall but still improve the QE model, indicating that not only preferred centers can improve the documen-level QE, other information also plays a role in the QE model. When the recall of Cp extraction increase, the performance of QE model further improve, which show the effectiveness of preferred centers. In the sentence-level QE, according to the setting of text boundary feature acquisition, the proposed model can not get any hint of the preferred center, which is equivalent to no additional information, so the performance of the model is comparable to that of the baseline model.

5.5 Case Study

As shown in Table 4, we provide the example of CpQE model and baseline model on scoring translation in document-level QE.

In the given example, the word "中欧班列 (china-europe train)" has two meanings. The first one is "the train from China to Europe" and the other one is "the train in central Europe". Since the previous sentences of the same document have mentioned "the train tack from Chengdu, China to Europe", the word in this sentence should be translated into "the train from China to Europe". Unfortunately, the translation output to be evaluated, i.e. mt1, provides an incorrect translation where the word "China" is missed. To test whether our proposed document-level QE system is sensitive to such errors, we simply recover the missing word "China" while ignore other mistakes in mt1 and produce another output, namely mt2. Then we evaluated these two outputs using the baseline model and our model, respectively. Clearly, the evaluation results show that both models indicate the decline of the edition rate. The proportion of the reduction of our model is higher than that of the baseline model, which is consistent with the HTER value, as listed in the fourth column. This results imply that our proposed model is more sensitive to such problems.

6 Conclusion

This research focus on the document-level machine translation quality estimation. Concretely, based on the concept of Preferred Center in the Centering Theory and the evaluation method of local text fluency, we manually annotated a small-scale dataset for Preferred Center extraction. Then, we trained a model to extract Preferred Centers for given texts and combine the extracted Preferred Centers as context information into the Predictor-Estimator model to improve the performance of QE. Furthermore, we construct a document-level Chinese-English QE dataset to measure the performance of our document-level QE models.

Acknowledgements. The authors would like to thank the three anonymous reviewers for their comments on this paper. This research was supported in part by the National Natural Science Foundation of China under Grant Nos. 62076211, U1908216 and 61573294 and the Outstanding Achievement Late Fund of the State Language Commission of China under Grant WT135-38.

A Appendix

Algorithm 1. Running process of outer-extractor

Input: mt mCp src sCp
Output: Emb f_{outer}

do
  for i in range(T) do
    [f_1, f_2, f_3] = $\frac{2}{len(mt)+len(src)}$[len(sCp[i]), len(mCp[i]), len(mCp[i]) - len(sCp[i])]
    if mt[i] is the begining of the document do
      Emb[i] = 0
      f_outer = [$f_1, f_2, f_3, 0, 0, 0, 1, 1, 0, 1, 1, 0$]
      continue
    Emb[i] = [Word2Vec(sCp[i-1], Word2Vec(mCp[i-1])]
    [f_4, f_5, f_6] = $\frac{2}{len(mt)+len(src)}$[len(sCp[i-1]), len(mCp[i-1]), len(mCp[i-1]) - len(sCp[i-1])]
    [f_7, f_8, f_9] = [similarity(sCp[i-1], src[i]), similarity(mCp[i-1], mt[i]), similarity(sCp[i-1], src[i])- similarity(mCp[i-1], mt[i])]
    if mt[i] is the end of the document do
      f_{outer} = [f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9, 1, 1, 0]
      continue
    else do
      [f_{10}, f_{11}, f_{12}] = [similarity(sCp[i-1], sCp[i]), similarity(mCp[i-1], mCp[i]), similarity (sCp[i-1], sCp[i]) - similarity(mCp[i-1], mCp[i])
    f_{outer} = [f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9, f_{10}, f_{11}, f_{12}]
return Emb, f_outer

The input of the outer-extractor is translation sentences mt, the preferred centers of translation sentences mCp, source sentences src and the preferred centers of source sentences sCp. The output of the extractor are embeddings of preferred centers Emb and the sentence relation features f_{outer}. T is the number of sentences in the corpus.

B Appendix

Table 5. Parameter of Bert-BiLSTM-CRF model

Parameter	Value	Describe
batch size	8	Total batch size for training
lr	0.01	The initial learning rate
epoch	10	Total number of training epochs to perform
lstm_size	128	LSTM hidden size
lstm_layers	1	Total number of LSTM layers
optim	Adam	Optimizer type

For preferred center extraction model, we use BERT-Base-Chinese as Chinese pre-trained model and BERT-Base as English pre-trained model. Some hyper-parameters are fixed: decoder layers are 12, hidden size of Bert is 768, the number of heads in multi-head attention is 12. Other parameters are shown in Table 5.

C Appendix

Table 6. Hyper-parameters of baseline predictor

Name	Value	Describe
src vocab size	120000	Size of vocabulary in source language
trg vocab size	120000	Size of vocabulary in target language
hidden size	512	Hidden size of Transformer
layers	2	Numbers of encoders and decoders in Transformer
head nums	8	Number of heads in multi-head attention
dropout	0.1	–
epoch	7	–
batch size	128	–
learning rate	2.0	–
optim	Lazyadam	Optimizer

Table 7. Hyper-parameters of baseline estimator

Name	Value	Describe
src vocab size	120000	Size of vocabulary in source language
trg vocab size	120000	Size of vocabulary in target language
unit nums	128	Unit numbers of BiLSTM
layers	1	Layers of BiLSTM
dropout	0.1	–
epoch	7	–
batch size	128	–
learning rate	2.0	–
optim	Lazyadam	Optimizer

Our CpQE model integrate an outer-extractor compared with baseline model. Other parameters is same as the baseline model. The parameters of baseline is shown in Table 6 and Table 7. The dimension of Word2Vec in outer-extractor is 512.

References

1. Banerjee, S., Lavie, A.: 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, pp. 65–72 (2005)
2. Chen, Z., et al.: Improving machine translation quality estimation with neural network features. In: Proceedings of the Second Conference on Machine Translation, pp. 551–555 (2017)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the Second International Conference on Human Language Technology Research, pp. 138–145 (2002)
5. Fan, K., Wang, J., Li, B., Zhou, F., Chen, B., Si, L.: "bilingual expert" can find translation errors. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 6367–6374 (2019)
6. Felice, M., Specia, L.: Linguistic features for quality estimation. In: Proceedings of the Seventh Workshop on Statistical Machine Translation, pp. 96–103 (2012)
7. Fonseca, E., Yankovskaya, L., Martins, A.F., Fishel, M., Federmann, C.: Findings of the WMT 2019 shared tasks on quality estimation. In: Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), pp. 1–10 (2019)
8. Grosz, B., Joshi, A., Weinstein, S.: Providing a unified account of definite noun phrases in discourse. In: Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics (1983)
9. Grosz, B.J., Joshi, A.K., Weinstein, S.: Centering: a framework for modelling the local coherence of discourse (1995)

10. Kim, H., Jung, H.Y., Kwon, H., Lee, J.H., Na, S.H.: Predictor-estimator: Neural quality estimation based on target word prediction for machine translation. ACM Trans. Asian Low-Resour. Lang. Inf. Process. (TALLIP) **17**(1), 1–22 (2017)
11. Kim, H., Lim, J.H., Kim, H.K., Na, S.H.: QE BERT: bilingual BERT using multi-task learning for neural quality estimation. In: Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), pp. 85–89 (2019)
12. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: probabilistic models for segmenting and labeling sequence data (2001)
13. Li, M., Xiang, Q., Chen, Z., Wang, M.: A unified neural network for quality estimation of machine translation. IEICE Trans. Inf. Syst. **101**(9), 2417–2421 (2018)
14. Martins, A.F., Junczys-Dowmunt, M., Kepler, F.N., Astudillo, R., Hokamp, C., Grundkiewicz, R.: Pushing the limits of translation quality estimation. Trans. Assoc. Computat. Linguist. **5**, 205–218 (2017)
15. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
16. Patel, R.N., et al.: Translation quality estimation using recurrent neural network. arXiv preprint arXiv:1610.04841 (2016)
17. Rubino, R., de Souza, J., Foster, J., Specia, L.: Topic models for translation quality estimation for gisting purposes (2013)
18. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of Association for Machine Translation in the Americas, Cambridge, MA, vol. 200 (2006)
19. Specia, L., Paetzold, G., Scarton, C.: Multi-level translation quality prediction with quest++. In: Proceedings of ACL-IJCNLP 2015 System Demonstrations, pp. 115–120 (2015)
20. Tan, Z., et al.: THUMT: an open-source toolkit for neural machine translation. In: Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (AMTA 2020), pp. 116–122 (2020)
21. Tong, Y., Zheng, J., Zhu, H., Chen, Y., Shi, X.: A document-level neural machine translation model with dynamic caching guided by theme-rheme information. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 4385–4395 (2020)
22. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
23. Voita, E., Sennrich, R., Titov, I.: When a good translation is wrong in context: context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. arXiv preprint arXiv:1905.05979 (2019)
24. Walker, M.A., Joshi, A.K., Prince, E.F.: Centering in naturally-occurring discourse: an overview. In: Centering in Discourse. Citeseer (1998)
25. Yang, S., Wang, Y., Chu, X.: A survey of deep learning techniques for neural machine translation. arXiv preprint arXiv:2002.07526 (2020)
26. Yuan, Y., Sharoff, S.: Sentence level human translation quality estimation with attention-based neural networks. arXiv preprint arXiv:2003.06381 (2020)