Contextual Encoding for Translation Quality Estimation

Abstract

The task of word-level quality estimation (QE) consists of taking a source sentence and machine-generated translation, and predicting which words in the output are correct and which are wrong. In this paper, propose a method to effectively encode the local and global contextual information for each target word using a three-part neural network approach. The first part uses an embedding layer to represent words and their part-of-speech tags in both languages. The second part leverages a one-dimensional convolution layer to integrate local context information for each target word. The third part applies a stack of feed-forward and recurrent neural networks to further encode the global context in the sentence before making the predictions. This model was submitted as the CMU entry to the WMT2018 shared task on QE, and achieves strong results, ranking first in three of the six tracks.

Introduction

Quality estimation (QE) refers to the task of measuring the quality of machine translation (MT) system outputs without reference to the gold translations BIBREF0 , BIBREF1 . QE research has grown increasingly popular due to the improved quality of MT systems, and potential for reductions in post-editing time and the corresponding savings in labor costs BIBREF2 , BIBREF3 . QE can be performed on multiple granularities, including at word level, sentence level, or document level. In this paper, we focus on quality estimation at word level, which is framed as the task of performing binary classification of translated tokens, assigning “OK” or “BAD” labels.

Early work on this problem mainly focused on hand-crafted features with simple regression/classification models BIBREF4 , BIBREF5 . Recent papers have demonstrated that utilizing recurrent neural networks

(RNN) can result in large gains in QE performance BIBREF6 . However, these approaches encode the context of the target word by merely concatenating its left and right context words, giving them limited ability to control the interaction between the local context and the target word.

In this paper, we propose a neural architecture, Context Encoding Quality Estimation (CEQE), for better encoding of context in word-level QE. Specifically, we leverage the power of both (1) convolution modules that automatically learn local patterns of surrounding words, and (2) hand-crafted features that allow the model to make more robust predictions in the face of a paucity of labeled data. Moreover, we further utilize stacked recurrent neural networks to capture the long-term dependencies and global context information from the whole sentence.

We tested our model on the official benchmark of the WMT18 word-level QE task. On this task, it achieved highly competitive results, with the best performance over other competitors on English-Czech, English-Latvian (NMT) and English-Latvian (SMT) word-level QE task, and ranking second place on English-German (NMT) and German-English word-level QE task.

Model

The QE module receives as input a tuple INLINEFORM0 , where INLINEFORM1 is the source sentence, INLINEFORM2 is the translated sentence, and INLINEFORM3 is a set of word alignments. It predicts as output a sequence INLINEFORM4 , with each INLINEFORM5 . The overall architecture is shown in Figure FIGREF2

CEQE consists of three major components: (1) embedding layers for words and part-of-speech (POS) tags in both languages, (2) convolution encoding of the local context for each target word, and (3) encoding the global context by the recurrent neural network.

Embedding Layer

Inspired by BIBREF6 , the first embedding layer is a vector representing each target word INLINEFORM0 obtained by concatenating the embedding of that word with those of the aligned words INLINEFORM1 in the source. If a target word is aligned to multiple source words, we average the embedding of all the source words, and concatenate the target word embedding with its average source embedding. The immediate left and right contexts for source and target words are also concatenated, enriching the local context information of the embedding of target word INLINEFORM2 . Thus, the embedding of target word INLINEFORM3 , denoted as INLINEFORM4 , is a INLINEFORM5 dimensional vector, where INLINEFORM6 is the dimension of the word embeddings. The source and target words use the same embedding parameters, and thus identical words in both languages, such as digits and proper nouns, have the same embedding vectors. This allows the model to easily identify identical words in both languages. Similarly, the POS tags in both languages share the same embedding parameters. Table TABREF4 shows the statistics of the set of POS tags over all language pairs.

One-dimensional Convolution Layer

The main difference between the our work and the neural model of BIBREF6 is the one-dimensional convolution layer. Convolutions provide a powerful way to extract local context features, analogous to implicitly learning INLINEFORM0 -gram features. We now describe this integral part of our model.

After embedding each word in the target sentence INLINEFORM0 , we obtain a matrix of embeddings for the target sequence, INLINEFORM1

where INLINEFORM0 is the column-wise concatenation operator. We then apply one-dimensional convolution BIBREF7 , BIBREF8 on INLINEFORM1 along the target sequence to extract the local context

of each target word. Specifically, a one-dimensional convolution involves a filter INLINEFORM2 , which is applied to a window of INLINEFORM3 words in target sequence to produce new features. INLINEFORM4

where INLINEFORM0 is a bias term and INLINEFORM1 is some functions. This filter is applied to each possible window of words in the embedding of target sentence INLINEFORM2 to produce features INLINEFORM3

By the padding proportionally to the filter size INLINEFORM0 at the beginning and the end of target sentence, we can obtain new features INLINEFORM1 of target sequence with output size equals to input sentence length INLINEFORM2 . To capture various granularities of local context, we consider filters with multiple window sizes INLINEFORM3 , and multiple filters INLINEFORM4 are learned for each window size.

The output of the one-dimensional convolution layer, INLINEFORM0 , is then concatenated with the embedding of POS tags of the target words, as well as its aligned source words, to provide a more direct signal to the following recurrent layers.

RNN-based Encoding

After we obtain the representation of the source-target word pair by the convolution layer, we follow a similar architecture as BIBREF6 to refine the representation of the word pairs using feed-forward and recurrent networks.

Two feed-forward layers of size 400 with rectified linear units (ReLU; BIBREF9);

One bi-directional gated recurrent unit (BiGRU; BIBREF10) layer with hidden size 200, where the forward and backward hidden states are concatenated and further normalized by layer normalization BIBREF11 .

Two feed-forward layers of hidden size 200 with rectified linear units;

One BiGRU layer with hidden size 100 using the same configuration of the previous BiGRU layer;

Two feed-forward layers of size 100 and 50 respectively with ReLU activation.

We concatenate the 31 baseline features extracted by the Marmot toolkit with the last 50 feed-forward hidden features. The baseline features are listed in Table TABREF13 . We then apply a softmax layer on the combined features to predict the binary labels.

Training

We minimize the binary cross-entropy loss between the predicted outputs and the targets. We train our neural model with mini-batch size 8 using Adam BIBREF12 with learning rate INLINEFORM0 and decay the learning rate by multiplying INLINEFORM1 if the F1-Multi score on the validation set decreases during the validation. Gradient norms are clipped within 5 to prevent gradient explosion for feed-forward networks or recurrent neural networks. Since the training corpus is rather small, we use dropout BIBREF13 with probability INLINEFORM2 to prevent overfitting.

Experiment

We evaluate our CEQE model on the WMT2018 Quality Estimation Shared Task for word-level English-German, German-English, English-Czech, and English-Latvian QE. Words in all languages are

lowercased. The evaluation metric is the multiplication of F1-scores for the “OK” and “BAD” classes against the true labels. F1-score is the harmonic mean of precision and recall. In Table TABREF15 , our model achieves the best performance on three out of six test sets in the WMT 2018 word-level QE shared task.

Ablation Analysis

In Table TABREF21 , we show the ablation study of the features used in our model on English-German, German-English, and English-Czech. For each language pair, we show the performance of CEQE without adding the corresponding components specified in the second column respectively. The last row shows the performance of the complete CEQE with all the components. As the baseline features released in the WMT2018 QE Shared Task for English-Latvian are incomplete, we train our CEQE model without using such features. We can glean several observations from this data:

Because the number of “OK” tags is much larger than the number of “BAD” tags, the model is easily biased towards predicting the “OK” tag for each target word. The F1-OK scores are higher than the F1-BAD scores across all the language pairs.

For German-English, English Czech, and English-German (SMT), adding the baseline features can significantly improve the F1-BAD scores.

For English-Czech, English-German (SMT), and English-German (NMT), removing POS tags makes the model more biased towards predicting “OK” tags, which leads to higher F1-OK scores and lower F1-BAD scores.

Adding the convolution layer helps to boost the performance of F1-Multi, especially on English-Czech and

English-Germen (SMT) tasks. Comparing the F1-OK scores of the model with and without the convolution layer, we find that adding the convolution layer help to boost the F1-OK scores when translating from English to other languages, i.e., English-Czech, English-German (SMT and NMT). We conjecture that the convolution layer can capture the local information more effectively from the aligned source words in English.

Case Study

Table TABREF22 shows two examples of quality prediction on the validation data of WMT2018 QE task for English-Czech. In the first example, the model without POS tags and baseline features is biased towards predicting “OK” tags, while the model with full features can detect the reordering error. In the second example, the target word “panelu” is a variant of the reference word “panel”. The target word “znaky” is the plural noun of the reference “znak”. Thus, their POS tags have some subtle differences. Note the target word “zmnit” and its aligned source word “change” are both verbs. We can observe that POS tags can help the model capture such syntactic variants.

Sensitivity Analysis

During training, we find that the model can easily overfit the training data, which yields poor performance on the test and validation sets. To make the model more stable on the unseen data, we apply dropout to the word embeddings, POS embeddings, vectors after the convolutional layers and the stacked recurrent layers. In Figure FIGREF24 , we examine the accuracies dropout rates in INLINEFORM0 . We find that adding dropout alleviates overfitting issues on the training set. If we reduce the dropout rate to INLINEFORM1 , which means randomly setting some values to zero with probability INLINEFORM2 , the training F1-Multi increases rapidly and the validation F1-multi score is the lowest among all the settings. Preliminary results proved best for a dropout rate of INLINEFORM3 , so we use this in all the

experiments.

Conclusion

In this paper, we propose a deep neural architecture for word-level QE. Our framework leverages a one-dimensional convolution on the concatenated word embeddings of target and its aligned source words to extract salient local feature maps. In additions, bidirectional RNNs are applied to capture temporal dependencies for better sequence prediction. We conduct thorough experiments on four language pairs in the WMT2018 shared task. The proposed framework achieves highly competitive results, outperforms all other participants on English-Czech and English-Latvian word-level, and is second place on English-German, and German-English language pairs.

Acknowledgements

The authors thank Andre Martins for his advice regarding the word-level QE task.

This work is sponsored by Defense Advanced Research Projects Agency Information Innovation Office (I2O). Program: Low Resource Languages for Emergent Incidents (LORELEI). Issued by DARPA/I2O under Contract No. HR0011-15-C0114. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.