DirectQE: Direct Pretraining for Machine Translation Quality Estimation

Qu Cui¹, Shujian Huang¹*, Jiahuan Li¹, Xiang Geng¹, Zaixiang Zheng¹, Guopeng Huang², Jiajun Chen¹
¹National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
²Tencent AI Lab, Shenzhen, China

{cuqi, lijh, gx, zhengzx}@smail.nju.edu.cn, {huangsj, chenjj}@nju.edu.cn, donkeyhuang@tencent.com

Introduction

How to evaluate the results of Machine Translation?
- BLEU, TER, Meteor...
- References are hard to get.
- Quality Estimation (QE)

Predict quality labels with only the source sentence and translation.
- Word-level. Each token is tagged as ‘OK’ or ‘BAD’.
- Sentence-level. Each translation is labeled by an HTER score.

Our goal

The results on the pseudo QE development set according to the time steps.

Quality of Pseudo QE Data

• Unlabeled
• Labeled automatically
• Represent rewriting degree (quality related)

Comparison

<table>
<thead>
<tr>
<th>Parallel Data</th>
<th>Pseudo QE Data</th>
<th>Real QE Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; X, Y &gt;</td>
<td>&lt; X, Y, o, q’ &gt;</td>
<td>&lt; X, T, O, q &gt;</td>
</tr>
<tr>
<td>Unlabeled</td>
<td>Labeled automatically</td>
<td>Labeled by human</td>
</tr>
<tr>
<td>Over Million</td>
<td>Representation quality</td>
<td></td>
</tr>
</tbody>
</table>

Controlled Noise Quality Label

Similar format
Larger scale

Experiment

Models
- NMT-based QE (Fan et al. 2018)
- Predictor. Encoder-bi-decoder (512) Estimator. Bi-LSTM
- PML-based QE (Devlin et al. 2018, Wolf et al. 2019)
- Predictor. BERT (768) Estimator. Bi-LSTM
- Direct QE
- Generator. Encoder-decoder (256) Detector. Encoder-decoder (512)

Results

NMT-based PML-based DirectQE-Eqph 53.07 10.99 16.04 50.32 11.76 14.56 15.74 15.02 127.04
PLM-based DirectQE-Eqph 57.55 11.64 17.23 55.88 12.89 14.56 15.48 14.56 137.12
NMT-based DirectQE-Eqph 65.82 10.74 14.74 67.04 10.03 14.48 44.43 43.06

90M

Results of single models (en-de)

Analysis

Data

F1-MULT of our method against the NMT-based method on the data with different error ratios.

DirectQE is better at solving poor translations.

QE performances according to different replacement ratios (Quality).

Quality of Pseudo-translations is important.

The results on the pseudo QE development set according to the time steps.

Diversity of data matters.

Training Objective

The mutual information between the representations and current tokens at the target side.

The detector representation contains more information about the current tokens.

Similarity of representations before and after finetuning.

Pretraining with the QE-like objectives learns more suitable representations for QE tasks.

Approach

Generator
- Generate translations with controlled noise and labels from parallel data.

Training
- Training the word-level rewritter on parallel data. (MLM)
- Generating
- Generating pseudo QE translations.
- Sample tokens according to the probability distribution to replace the original ones.
- Generating pseudo QE labels.
- Word-level. Each token is tagged 0 if it is replaced else 1.
- Sentence-level. Each translation is labeled by the replacement ratio.

Detector
- Trained on pseudo and real QE data with the same QE objectives.