Automatic Machine Translation Evaluation using Source Language Inputs and Cross-lingual Language Model

Presenter: 1 Kosuke Takahashi

1 2 Katsuhito Sudoh

1 Satoshi Nakamura

1: Nara Institute of Science and Technology (NAIST)
2: PRESTO, Japan Science and Technology Agency
Existing metrics based on surface level features

- BLEU [Papineni +, 2002], NIST [Doddington +, 2002], METEOR [Satanjeev +, 2005]
  - Calculate evaluation scores with word matching rate

Problems:
Relying on lexical features
→ cannot appropriately evaluate semantic and syntactic differences
Existing metrics based on embedded representation

- RUSE[Shimanaka +, 2018], BERT regressor[Shimanaka +, 2019]
  - Fully parameterized metrics
    - Use sentence vectors
    - fine-tuned to predict human evaluation scores
  - BERT regressor achieved the SOTA result on WMT17 metrics task in 2019

These metrics provide better evaluation performance than surface level ones.
Proposed multi-reference

Conventional multi-reference

- reference 1
- reference 2
- reference n

system translation (hypothesis)

○ better evaluation
× costly to prepare multiple reference sentences

Proposed Idea

- source sentence (reference 1)
- system translation (hypothesis)
- reference sentence (reference 2)

- system translation (hypothesis)

○ better evaluation
○ a little cost to prepare 2 references for each hypothesis
Architectures of a baseline and proposed models

**Baseline:** BERT regressor

- Hypothesis + Reference

- Sentence-pair encoder

- Hypothesis + Reference

**hypothesis + src/hypothesis + ref**

- Evaluation score

- MLP

- Concatenation $v_{hyp+src}, v_{hyp+ref}$

- Sentence-pair encoder

- Hypothesis + Reference

- Hypothesis + Source

**hypothesis + src + ref**

- Evaluation score

- MLP

- Sentence-pair vector $v_{hyp+src+ref}$

- Sentence-pair encoder

- Hypothesis + Source + Reference
The setting of experiments

- Language model: mBERT, XLM15

- Input: hyp+src/ref, hyp+src+ref, hyp+ref, hyp+src

- Baselines: SentBLEU, BERT regressor (BERT with hyp+ref)

- Data: WMT17 metrics shared task

- Language pairs: {De, Ru, Tr, Zh}-En
## Results: comparison with baselines

<table>
<thead>
<tr>
<th>Metric or language model</th>
<th>Input style</th>
<th>Average score (r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SentBLEU</td>
<td>hyp, ref</td>
<td>48.4</td>
</tr>
<tr>
<td>BERT regressor</td>
<td>hyp+ref</td>
<td>74.0</td>
</tr>
<tr>
<td>(monolingual BERT)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mBERT</td>
<td>hyp+src/hyp+ref</td>
<td>72.6</td>
</tr>
<tr>
<td></td>
<td>hyp+src+ref</td>
<td>68.9</td>
</tr>
<tr>
<td>XLM15</td>
<td>hyp+src/hyp+ref</td>
<td>77.1</td>
</tr>
<tr>
<td></td>
<td>hyp+src+ref</td>
<td>74.7</td>
</tr>
</tbody>
</table>

- Proposed XLM15 with hyp+src/hyp+ref surpassed baseline scores
Results: evaluation performance for each input style

<table>
<thead>
<tr>
<th>language model</th>
<th>input style</th>
<th>average score (r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mBERT</td>
<td>hyp+ref</td>
<td>67.9</td>
</tr>
<tr>
<td></td>
<td>hyp+src</td>
<td>55.9</td>
</tr>
<tr>
<td></td>
<td>hyp+src/hyp+ref</td>
<td><strong>72.6</strong></td>
</tr>
<tr>
<td></td>
<td>hyp+src+ref</td>
<td>68.9</td>
</tr>
<tr>
<td>XLM15</td>
<td>hyp+ref</td>
<td>74.1</td>
</tr>
<tr>
<td></td>
<td>hyp+src</td>
<td>72.8</td>
</tr>
<tr>
<td></td>
<td>hyp+src/hyp+ref</td>
<td><strong>77.1</strong></td>
</tr>
<tr>
<td></td>
<td>hyp+src+ref</td>
<td>74.7</td>
</tr>
</tbody>
</table>

- Using src and ref improve evaluation performance
- hyp+src/hyp+ref was the best input style
Analysis:
scatter plots of evaluation and DA scores

XLM15 hyp+src/ref

Pearson’s correlation score
All: 0.768
DA ≥ 0.0: 0.580
DA < 0.0: 0.529

Low quality translation is hard to evaluate

Note: DA (Direct Assessment) is a human evaluation score
Analysis: The drop rate of Pearson’s correlation score from high DA to low DA range

<table>
<thead>
<tr>
<th>language model</th>
<th>input style</th>
<th>reduction rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT regressor (monolingual BERT)</td>
<td>hyp+ref</td>
<td>16.10</td>
</tr>
<tr>
<td></td>
<td>hyp+src</td>
<td>6.88</td>
</tr>
<tr>
<td></td>
<td>hyp+src/hyp+ref</td>
<td>7.77</td>
</tr>
<tr>
<td></td>
<td>hyp+src+ref</td>
<td>17.51</td>
</tr>
<tr>
<td>mBERT</td>
<td>hyp+ref</td>
<td>22.05</td>
</tr>
<tr>
<td></td>
<td>hyp+src</td>
<td>8.46</td>
</tr>
<tr>
<td></td>
<td>hyp+src/hyp+ref</td>
<td>8.68</td>
</tr>
<tr>
<td></td>
<td>hyp+src+ref</td>
<td>11.12</td>
</tr>
<tr>
<td>XLM15</td>
<td>hyp+ref</td>
<td>14.20</td>
</tr>
<tr>
<td></td>
<td>hyp+src</td>
<td>8.46</td>
</tr>
<tr>
<td></td>
<td>hyp+src/hyp+ref</td>
<td>8.68</td>
</tr>
<tr>
<td></td>
<td>hyp+src+ref</td>
<td>11.12</td>
</tr>
</tbody>
</table>

Note: reduction rate indicates how much evaluation performance is degraded from high to low quality translations.
Summary

- Proposed a MT evaluation metric that utilizes source sentences as pseudo references.
- hyp+src/hyp+ref makes good use of source sentences and is confirmed to improve evaluation performance.
- XLM15 hyp+src/hyp+ref showed the higher correlation with humans than baselines.
- Source information is contributed to stabilize the evaluation of low quality translations.

Future Work

- Experiment with multiple language models and datasets.
- Focus on a better evaluation of low quality translations.