Automatic Detection of Machine Translated Text and Translation Quality Estimation

Roee Aharoni, Moshe Koppel and Yoav Goldberg
Bar Ilan University
ISCOL 2014
Motivation

• Automatic MT evaluation requires human-translated reference sentences

• BLEU (Papineni et al., 2001)

• METEOR (Lavie et al, 2004)

• Reference sentences are “expensive”, especially for new domains and resource-poor languages

• We would like to estimate the quality of a given MT output, without the use of reference sentences
Our Approach

• Classify text, at sentence level, as MT or human

• Use the classification accuracy as a “proxy” for quality estimation

• The more our classifier confuses MT sentences as human sentences, the better the translation quality is
Our Approach

Test Set
- MT sentences
- Human sentences

Sentence Classifier

Quality Estimation
Our Approach

Test Set

| MT sentences |
| Human sentences |

Sentence Classifier

Quality Estimation

“Random”, Non-Reference sentences
Features

• Use common linguistic, domain-independent features to detect MT sentences:
  • Automatic Part of Speech tags
  • Function Words

• Inspired by works on “Translationese” (Koppel and Ordan, 2011) and on Machine Translation Detection (Arase and Zhou, 2013)
Features

Example

"These days, all but one were subject to a vote, and all had a direct link to the post September 11th."

"these days, except one were the subject of a vote, and all had a direct link with the after 11 September."

"From these days, all except one were the object of a vote, and all were connected a direct link with after September 11th."

"Of these days, all except one were making the object of a vote and all had a straightforward tie with after September 11."

"In these days, all safe one made the object in a vote and all had a direct connection with him after 11 of September."

<table>
<thead>
<tr>
<th>Function Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features</td>
</tr>
<tr>
<td>POS tags</td>
</tr>
</tbody>
</table>
Experiments Outline

• Use a linear SVM classifier with the Function-word and POS features to classify human vs. MT

• For a given MT system:
  • Perform a 10-fold cross validation across the different sentences in the test set
  • Measure the correlation of the result with the translation quality (BLEU or human evaluation)
• Examined 7 French-English commercial MT system outputs (Google Translate and 6 others via the itranslate4.eu website)

• Tested 3 different feature settings (POS, function words and both)

• Compared the use of reference and random, non-reference human sentences

• 20,000 sentences per class (human/MT), taken from the Hansard Corpus (Germann, 2001)
Results - Commercial MT Systems

- Very strong reverse correlation with BLEU - $R^2$ from 0.779 up to 0.978
- Up to ~90\% detection accuracy

*Each point represents an MT system*
Results - Commercial MT Systems

- Very strong reverse correlation with BLEU - $R^2$ from 0.779 up to 0.978
- Up to ~90% detection accuracy
- The better the translation quality is, the harder it is to correctly detect it
Experiment II - In-House MT Systems

- Trained 7 French to English phrase-based MT systems, using the Moses SMT toolkit (Koehn et al, 2007)
- Train data (LM + Translation): Europarl corpus (Koehn, 2005)
- Evaluation data: Hansard corpus (Germann, 2001)
- Varied both LM and translation model sizes, resulting in a wide variety of BLEU scores:

<table>
<thead>
<tr>
<th></th>
<th>Parallel</th>
<th>Monolingual</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMT-1</td>
<td>2000k</td>
<td>2000k</td>
<td>28.54</td>
</tr>
<tr>
<td>SMT-2</td>
<td>1000k</td>
<td>1000k</td>
<td>27.76</td>
</tr>
<tr>
<td>SMT-3</td>
<td>500k</td>
<td>500k</td>
<td>29.18</td>
</tr>
<tr>
<td>SMT-4</td>
<td>100k</td>
<td>100k</td>
<td>23.83</td>
</tr>
<tr>
<td>SMT-5</td>
<td>50k</td>
<td>50k</td>
<td>24.34</td>
</tr>
<tr>
<td>SMT-6</td>
<td>25k</td>
<td>25k</td>
<td>22.46</td>
</tr>
<tr>
<td>SMT-7</td>
<td>10k</td>
<td>10k</td>
<td>20.72</td>
</tr>
</tbody>
</table>
Results - In-House MT Systems

- The correlation is consistent among the in-house systems as well
- High correlation with BLEU, using only random, non-reference sentences
Experiment III - Correlation with Human Evaluation

• BLEU scores are nice, but how about correlation with real (human) evaluation?

• Examined 13 French-English MT systems and their human evaluations from WMT13’ (Bojar et al., 2013)

• Used reference sentences and random, non-reference sentences from WMT 12’ (Callison-Burch et al., 2012) as the human data
Results - Correlation with Human Evaluation

Good results with reference sentences

\[ R^2 = 0.774 \]

Using Reference Sentences

“Blunt” outlier with non-reference sentences

\[ R^2 = 0.556 \]

Using non-Reference Sentences
Syntactic Features

- The outlier is an instance of the “Joshua” MT system (Post et al., 2013)
- This system is syntax based, a fact that may have “confused” the classifier
- We hypothesize that using syntax based features in the classifier will help
Syntactic Features

• Parse each sentence using the Berkeley Parser (Petrov and Klein, 2007)

• Extract one level non-terminal CFG rules from each tree

• Use as the only features in the classification task
Results - Correlation with Human Evaluation using syntactic features

- The outlier is gone
- High correlation with human evaluation score - $R^2 = 0.829$ (vs. 0.556 before)
- No use of reference sentences in the process

$R^2 = 0.829$
Why does it work?

• The classifier uses much more data than the standard approaches when evaluating a single sentence

• Our approach measures **fluency**, as we don’t use any reference translations

• There is a strong correlation between fluency and overall translation quality, given the sentences are MT output
Conclusions

• It is possible to detect machine translation in monolingual corpora at sentence level

• Strong correlation resides between detection accuracy and translation quality

• This correlation holds whether or not a reference set is used

• It is possible to estimate translation quality without reference sentences
Future Work

• Apply our methods to other language pairs and domains

• Explore additional features and feature selection techniques

• Integrate our method in a machine translation system (during training or decoding phases)

• Acquire word-level quality estimation
Questions?