Towards Improving Machine Translation and Speech Recognition with Discriminative Fluency Classification

Roee Aharoni, Moshe Koppel
Bar Ilan University
ICRI-CI Retreat, May 24th, 2016
Motivation

- Evaluation of automatic speech recognition (ASR) systems and Machine Translation (MT) systems requires manually aligned speech-to-text or text-to-text examples (Word Error Rate, BLEU)
- Such examples are expensive to obtain, especially for specific domains and resource-poor languages
- We would like to create an evaluation and re-ranking method for MT and ASR which does not depend on such aligned examples
The Generative Approach to Fluency Estimation

- Generative n-gram language models are a core component in today’s state of the art ASR and MT systems (e.g. Baidu’s Deep Speech, Hannun et. al. 2015)

- Usually used as a scoring model in the sequence decoding process - better LM score = more likely hypothesis

- Enables the use of massive amounts of unlabeled data
Our Approach: Discriminative Fluency Estimation

• If we look at the data:
  
  • “Good” ASR outputs:
    
    • GO TO THE LOUNGE
    
    • LOOK FOR MY MOBILE PHONE
    
    • ROBOT CAN YOU TURN THE OVEN ON
  
  • “Bad” ASR outputs:
    
    • ROEDEL CAN YOU OPEN THE WATUSI
    
    • WHO THE GLASS FROM THE SCENE
    
    • ROBOT CAN RETURN MY LAPTOP ONE FTP

• Can we train a classifier to discriminate good ASR/MT outputs from bad ones?
Our Approach: Discriminative Fluency Estimation

• Classify text, at sentence level, into Machine or Human Language

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

• The hypothesis (over a large enough dataset):
  high classification accuracy = bad quality output,
  low classification accuracy = high quality output
Our Approach: Discriminative Fluency Estimation

Test Set

- MT/ASR outputs
- Human sentences

Sentence Classifier

Quality Estimation
Our Approach: Discriminative Fluency Estimation

- Test Set
  - MT/ASR outputs
  - Human sentences

- Sentence Classifier

- Quality Estimation

Can be “Random” Non-Reference sentences
Experiments Outline

• Divide the machine output sentences into variable quality sets (from poor quality to high quality)

• For a given sentence sets:
  • Perform a 10-fold cross-validation experiment using a linear SVM classifier that classifies the sentences into human vs. machine, the machine sentence set vs. a human sentence set
  • Measure the correlation between the classification accuracy and the output quality for the set (WER/BLEU/human evaluation)
MT Experiments
Features - MT

• Use common linguistic, domain-independent features to discriminate MT outputs from human sentences:
  • Function Words
  • Parts of Speech
  • Syntax

• Inspired by works on:
  • “Translationese” (Koppel and Ordan, 2011)
  • Machine Translation Detection (Arase and Zhou, 2013)
Experiment 1 - Commercial MT Systems

- 7 French-English commercial MT system outputs (Google Translate and 6 others via the itranslate4.eu website)
- 3 different feature settings (POS, function words and both)
- Compared 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 negative correlation with BLEU - $R^2$: 0.78 to 0.98
- Up to ~90% detection accuracy

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

- Very strong negative correlation with BLEU - $R^2$: 0.78 to 0.98
- 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

- High correlation with human evaluation score - $R^2 = 0.829$
- No use of reference sentences in the process
ASR Experiments
ASR Experiments Setup

• Formal representation of ASR output and human-generated sentences
  • Lexical features only (most frequent words in corpus)
  • Syntactic and POS features not helpful
• Use SVM to distinguish human sentences from ASR output
• Use 10-fold cross-validation to measure success
• Compare success of discriminative model to WER
Datasets

• ICSI Meeting Transcripts Corpus
  • ~60k sentence transcriptions + corresponding 5-best lists from in-house ASR system
  • ~795k words; ~13k unique words

• TED talks corpus with NAIST ASR outputs
  • 1,770 manually transcribed sentences from TED talks
  • Corresponding 5-best lists produced by NAIST ASR system (Heck 2015)

• ROCKIN Robot Challenge Corpus
  • Robot instructions from 4 competitions
  • <700 transcribed instructions (very small!)
  • 5-best ASR outputs for only ~400 transcribed instructions
Results - ICSI Conversations Dataset

- X axis - WER, Y axis - classification accuracy, each point is a cluster

- Very high correlation ($R^2 = 0.93-0.97$) between classification accuracy and WER on both cluster types (n-best or sorted by WER)
Results - TED talks dataset vs. references

- High correlation (0.92) with WER, even with a much smaller dataset and a very small WER variance (X axis, 0.18-0.22)
Results - TED talks dataset vs. non-reference data

• To explore using non-reference data as the “fluent” part, we took 1770 sentences from another TED talks corpus

• Correlation still holds, but lower - $R^2 = 0.67$ (smaller datasets)
Results - Robot Instructions Dataset

- High correlation ($R^2=0.93$) even with very few examples (382 per class) and a much more specific domain
Results - Using Non-Reference Data

• To explore using non-reference data as the “fluent” part, we took 292 instructions from a different robot competition.

• Correlation still holds, but lower - $R^2 = 0.67$ (smaller datasets, slightly different language).

• Removing all the noun features (which are more domain specific), leaving only verbs and function words, improved correlation to 0.71.
Conclusions

- It is possible to evaluate MT and ASR systems even in the absence of sentence-aligned data.

- This measure correlates with standard evaluation measures that use such data. The correlation holds on large, general domain datasets and on small, domain-specific test sets.

- Future work may include different classification techniques and the development of re-ranking components inspired by this approach.
Thank You