Automatic Grammatical Error Detection of Non-native Spoken Learner English

Kate Knill¹, Mark Gales¹, Potsawee Manakul¹, Andrew Caines²

¹Engineering Department / ALTA Institute
²Computer Science and Technology / ALTA Institute

17 May 2019
Challenge and Advantages of Spoken Language

• Spoken language consists of
  
  Text + Pronunciation + Prosody + Delivery

• Challenge for feedback on “grammatical” errors in spoken language

  Spoken Text ≠ Written Text

  • We don’t speak in sentences, we repeat ourselves, hesitate, mumble etc

  • There is no defined spoken grammar standard

• Advantages of speech

  • There are no spelling or punctuation mistakes

  • We provide additional information within the audio signal
flor company is an engineering company in the Poland we do business the refinery business and the chemical business the job we can offer is an engineering job basically this is the job in the office
Grammatical Error Detection

- Task: given a sentence automatically label each word with
  - $P_{\text{word}}$ (grammar is correct) and $P_{\text{word}}$ (grammar is incorrect)
- Example sentence

  
<table>
<thead>
<tr>
<th>Internet</th>
<th>was</th>
<th>something</th>
<th>amazing</th>
<th>for</th>
<th>me</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(c)</td>
<td>0.02</td>
<td>0.96</td>
<td>0.97</td>
<td>0.97</td>
<td>0.95</td>
</tr>
<tr>
<td>P(i)</td>
<td>0.98</td>
<td>0.04</td>
<td>0.03</td>
<td>0.03</td>
<td>0.05</td>
</tr>
</tbody>
</table>

- Predict prob. distribution $y_i$ for each token $w_1: N = \{w_1, \cdots, w_N\}$
Sequence Labeller

\[ P(y_i|w_{1:N}) = [0.92;0.08] \]

- **Hidden Layer**
- **Backward LSTM**
- **Forward LSTM**
- **Word embedding**
  - \( w_i = \text{dog} \)
- **Character-level Bidirectional LSTM**
  - \( w_{i,1} = d \)
  - \( w_{i,2} = o \)
  - \( w_{i,3} = g \)
## Corpora

- Non-native English learners with grammatical error annotation
  - **BULATS**: free speech with up to 1 minute per response
  - **NICT-JLE**: oral proficiency test interviews
  - **CLC**: range of written exams at different grade levels

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Spoken/Written</th>
<th># Wds</th>
<th># Uniq Wds</th>
<th>Audio</th>
<th>L1s</th>
<th>Grades</th>
</tr>
</thead>
<tbody>
<tr>
<td>BULATS</td>
<td>Spoken</td>
<td>61.9K</td>
<td>3.4K</td>
<td>Yes</td>
<td>6</td>
<td>A1-C2</td>
</tr>
<tr>
<td>NICT--JLE</td>
<td>Spoken</td>
<td>135.3K</td>
<td>5.6K</td>
<td>No</td>
<td>1</td>
<td>A1-B2</td>
</tr>
<tr>
<td>CLC</td>
<td>Written</td>
<td>14.1M</td>
<td>79.1K</td>
<td>No</td>
<td>Many</td>
<td>A1-C2</td>
</tr>
</tbody>
</table>
Data Processing for Spoken GED

- Match data processing in training and testing
  a. Train: Text data - correct spelling errors and remove punctuation and casing
  b. Test: Speech data - convert speech transcriptions to be “like” text

// flor company is an engineering company in the poland
// we do business the refinery business and the chemical business
// the job we can offer is a engineering job
// basically this is the job in the office
GED Using CLC Trained Model

F$_{0.5}$ 57.6%
F$_{0.5}$ 49.7%
F$_{0.5}$ 44.1%
(Small Scale) System Error Analysis

• True precision higher for Spoken BULATS than scores suggest
  • System error (~27%)
    .. and i have to practice more because I have ..
  • Unmarked error (~40%)
    .. so I think you need taxi
  • Next to error tagged word(s) (~27%)
    .. and continue to inform with customer when we have ..

• To provide feedback we need to boost recall of high precision items
  • Issue: lack of labelled learner speech corpora
    • Adapt/“fine-tune” CLC trained system to subset of target speech data
Boosting GED Performance on Spoken BULATS

- Fine-tune CLC system with 80% data, dev 10%, test 10% x10

- Fine-tuning produces significant boost in performance
  - has also learnt some annotator bias e.g. “two thousand eight”
Example of Learner Speech: ASR Transcription

**MANUAL GRAMMATICAL ERRORS**

flor company is an engineering compa- is is is eng- engineering company %hes% %hes% in the in the poland %hes% we do business the ref- refinery business and the chemical business %hes% the job we can offer is a engineering job %hes% basically this is the job in the office

**ASR TRANSCRIPTION ERRORS**

flower companies in joining company is is is engineer engineering company in the in the poll one %hes% we do business there are refinery business in a chemical business %hes% the job we can offer is [del] engineering job %hes% basically this is the job in the office

**ASR “GRAMMATICAL ERRORS” – feedback focused**

// flower companies in engineering company in the poll one
// we do business there are refinery business in a chemical business
// the job we can offer is engineering job
// basically this is the job in the office
BULATS ASR Annotation Error Rates

• Overall: 71751 words   16.5% GER   25.2% WER
## BULATS ASR Word Error Rate

<table>
<thead>
<tr>
<th></th>
<th>#</th>
<th>%WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>71751</td>
<td>25.2</td>
</tr>
<tr>
<td>“Fluent”</td>
<td>52698</td>
<td>19.3</td>
</tr>
<tr>
<td>Grammatical Error</td>
<td>10348</td>
<td>29.1</td>
</tr>
<tr>
<td>Disfluency</td>
<td>2524</td>
<td>36.4</td>
</tr>
</tbody>
</table>
GED on BULATS ASR Transcriptions

- Manual transcriptions used for GE marking and meta-data extraction

- Significantly lower performance than manual transcriptions

![Graph showing precision-recall curves for CLC and Finetune with F0.5 scores of 37.3% and 26.6% respectively.](image)
Conclusions

• Detecting “grammatical” errors in learner speech is hard!
  • As is annotating the errors
• Focus on high precision region for feedback
  • Testing if regions where errors detected are sufficient to provide useful help
• More research required into:
  • Meta-data extraction
  • Boosting training data by mimicking learner speech errors
  • Detecting portions of ASR transcription the system is confident in
Questions?

Thanks to Cambridge English Language Assessment for supporting this research and providing access to the BULATS data.