Towards Learning Orientated Assessment for Non-native Learner Spoken English

Kate Knill
Mark Gales, Kostas Kyriakopolous, Edie Lu, Potsawee Manakul, Andrey Malinin, Anton Ragni. Linlin Wang, Yu Wang
ALTA Institute, Cambridge University Engineering Department

6 March 2019
Learning Oriented Assessment

Saville & Jones
Spoken communication is a very rich communication medium.
Spoken Communication Requirements

• Message Construction should consider:
  • Has the speaker generated a coherent message to convey?
  • Is the message appropriate in the context?
  • Is the word sequence appropriate for the message?

• Message Realisation should consider:
  • Is the pronunciation of the words correct/appropriate?
  • Is the prosody appropriate for the message?
  • Is the prosody appropriate for the environment?
Spoken Communication Requirements

• Message Construction should consider:
  - Has the speaker generated a coherent message to convey?
  - Is the message appropriate in the context?
  - Is the word sequence appropriate for the message?

• Message Realisation should consider:
  - Is the pronunciation of the words correct/appropriate?
  - Is the prosody appropriate for the message?
  - Is the prosody appropriate for the environment?
Spoken Communication Assessment

• Construct
  • Read speech
  • Free speaking – short response to prompt/question; 20s-1min responses
  • Conversation – with examiner or between candidates with examiner
Business Language Testing Service (BULATS) Spoken Tests

• Example of a test of communication skills

  A. Introductory Questions: where you are from
  B. Read Aloud: read specific sentences
  C. Topic Discussion: discuss a company that you admire

D. Interpret and Discuss Chart/Slide: example above

E. Answer Topic Questions: 5 questions about organising a meeting
<table>
<thead>
<tr>
<th>Level</th>
<th>Global Descriptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2</td>
<td>Fully operational command of the spoken language</td>
</tr>
<tr>
<td>C1</td>
<td>Good operational command of the spoken language</td>
</tr>
<tr>
<td>B2</td>
<td>Generally effective command of the spoken language</td>
</tr>
<tr>
<td>B1</td>
<td>Limited but effective command of the spoken language</td>
</tr>
<tr>
<td>A2</td>
<td>Basic command of the spoken language</td>
</tr>
<tr>
<td>A1</td>
<td>Minimal command of the spoken language</td>
</tr>
</tbody>
</table>
Candidate Speech Examples

• Level B2

• Level A2
Talk about the picture
Speaking Auto-marking

Talk about the picture

Construct is unchanged

140
Auto-marker
Auto-marker

Automatic Speech Recognition

"it's very cheaper for customer"
Auto-marker and Feedback

Automatic Speech Recognition → Compute Features → Input Features → Grader → 140

"it's very cheaper …" → Compute Feedback → 140
Automatic Speech Recognition

“it’s very cheaper for customer”
Speech Recognition Challenges

- Non-native ASR highly challenging
  - Heavily accented
  - Pronunciation dependent on L1
- Commercial systems poor!
- 2015 CUED systems

<table>
<thead>
<tr>
<th>Training Data</th>
<th>Word error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native &amp; C-level non-native English</td>
<td>54%</td>
</tr>
<tr>
<td>BULATS speakers</td>
<td>30%</td>
</tr>
</tbody>
</table>
Evaluation set

- 225 non-native speakers from BULATS test
  - 6 L1s: Arabic, Dutch, French, Polish, Thai, Vietnamese
  - Approx. equal distribution over CEFR proficiency levels A1-C
- ASR: long free speaking sections only (C-E)
  - Manual transcriptions (1x annotator)
- Grading: all sections
  - Expert marks
- Pronunciation and grammatical error detection
  - Manual annotations
• Trained on real examination data from over 30 L1s
• Transcriptions – automatic merging 2x crowd-sourced trans; ASR
• Graphemic lexicon
• Kaldi training (sMBR/LF-MMI) and decoding
ASR on Non-native Speech (1)

- Joint DNN+LSTM AM + trigram LM
- AM and LM trained on BULATS data only
- Hours of data varied
- LM trained on transcriptions, interpolated with BNE

<table>
<thead>
<tr>
<th>AM</th>
<th>LM</th>
<th>% WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>300</td>
<td>25.5</td>
</tr>
<tr>
<td>400</td>
<td>400</td>
<td>24.4</td>
</tr>
</tbody>
</table>
## ASR on Non-native Speech (2)

<table>
<thead>
<tr>
<th>AM</th>
<th>LM</th>
<th>% WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDNN-F</td>
<td>trigram</td>
<td>23.4</td>
</tr>
<tr>
<td>+AMI</td>
<td></td>
<td>22.7</td>
</tr>
<tr>
<td>+TS Ensemble</td>
<td></td>
<td>22.3</td>
</tr>
<tr>
<td>+ i-vectors</td>
<td></td>
<td>21.3</td>
</tr>
<tr>
<td>RNNLM+succ</td>
<td></td>
<td>19.5</td>
</tr>
</tbody>
</table>

- 400 hour BULATS training set
Auto-marker: Grader Features

Automatic Speech Recognition

“it’s very cheaper for customer”
Example Auto-marker Feature

Speaking Rate vs Candidate Speaking Band

![Box plot graph showing the distribution of Speaking Rate across different Speaking Bands (pre-A1, A1, A1+, A2, A2+, B1, B1+, B2, B2+, C1, C1+, C2). The y-axis represents cv._mfl_frequency, and the x-axis represents SpeakingBand. Each band shows a box plot with median, interquartile range, and outliers.](image-url)
Auto-marker: Grader

Automatic Speech Recognition → Compute Features → Input Features → Grader → "it's very cheaper for customer"
Gaussian Process Grader

• Powerful non-parametric Bayesian model
• Use variance to provide confidence in prediction
• Limitation – $O(n^2)$ memory usage, $O(n^3)$ computational load
Deep Density Network (DDN) Grader

- Predict score and score variance
- Use predicted variance to provide confidence in prediction
- Model size and computational independent of training data set size
### Grader Performance

<table>
<thead>
<tr>
<th>Grader</th>
<th>PCC</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exam</td>
<td>8.48</td>
<td>0.864</td>
</tr>
<tr>
<td>GP</td>
<td>8.56</td>
<td>0.449</td>
</tr>
<tr>
<td>DDN</td>
<td>0.86</td>
<td>0.420</td>
</tr>
</tbody>
</table>

- Automatic graders match human non-expert examiners
- Combine human and auto grades to boost accuracy
- Regions with limited data a challenge: low A1/C2
Automatic Assessment with Feedback

Automatic Speech Recognition

Compute Features

Features

Grader

140

"Text"

Compute Feedback
Pronunciation Error Detection

• Substitution errors
  • e.g.
    
    Correct: /sh/ /aa/ /p/ /ih/ /ng/
    Actual: /s/ /aa/ /p/ /ih/ /ng/

• Insertion errors
  • Word final e.g.
    
    /k/ /ah/ /ml → /k/ /ah/ /m/ /ah/

• Inter-consonant e.g.
  
  /r/ /iy/ /s/ /ah/ /n/ /t/ → /r/ /iy/ /s/ /ah/ /n/ /eh / /t/

• Deletion errors (final phone, 1\textsuperscript{st}/2\textsuperscript{nd} consonant in consonant pair)
### Common Substitutions by L1

<table>
<thead>
<tr>
<th>L1</th>
<th>Substitution</th>
<th>Correct</th>
<th>Incorrect</th>
<th>% correct pronounced as incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>ch</td>
<td></td>
<td>sh</td>
<td>4.76</td>
</tr>
<tr>
<td></td>
<td>dh</td>
<td></td>
<td>z</td>
<td>3.40</td>
</tr>
<tr>
<td>Dutch</td>
<td>th</td>
<td></td>
<td>t</td>
<td>2.11</td>
</tr>
<tr>
<td></td>
<td>dh</td>
<td></td>
<td>d</td>
<td>1.99</td>
</tr>
<tr>
<td>Thai</td>
<td>dh</td>
<td></td>
<td>d</td>
<td>7.24</td>
</tr>
<tr>
<td></td>
<td>oh</td>
<td></td>
<td>aa</td>
<td>5.21</td>
</tr>
</tbody>
</table>

- Top 2 recurrent substitution errors for speakers in each L1
Pronunciation Error Detection and Feedback

- Use Siamese networks and attention networks to detect bad phones
Grammatical Error Detection: Challenge

- 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
Example of Learner Speech (Manual)

THE VISITOR CAN {F %HES%} GO TO THE RESTAURANT AND HAVE {F %HES%} {REP BUFFET + BUFFET} {F %HES%} AFTER THAT THEY CAN {F %HES%} {F %HES%} HAVE THE {F %HES%} KARAOKE ACTIVITIES TO {F %HES%} {REP GET AC- + GET ACQUAINTED} WITH THE COMPANY

THE VISITOR CAN GO TO THE RESTAURANT AND HAVE BUFFET AFTER THAT THEY CAN HAVE THE KARAOKE ACTIVITIES TO GET ACQUAINTED WITH THE COMPANY
Example of Learner Speech (ASR)

{F %HES%} THE VISITOR CAN {F %HES%} GO TO THE RESTAURANT AND HAVE {F %HES%} WITH A BUDGET {F %HES%} AFTER THAT THEY CAN {F %HES%} HAVE THE {F %HES%} CAN OKAY ACTIVITIES TO {F %HES%} GET A QUITE GET ACQUAINTED WITH THE COMPANY

THE VISITOR CAN GO TO THE RESTAURANT AND HAVE WITH A BUDGET AFTER THAT THEY CAN HAVE THE CAN OKAY ACTIVITIES TO GET A QUITE GET ACQUAINTED WITH THE COMPANY

THE VISITOR CAN GO TO THE RESTAURANT AND WITH A BUDGET AFTER THAT THEY CAN HAVE THE CAN OKAY ACTIVITIES TO GET A QUITE GET ACQUAINTED WITH THE COMPANY
Grammatical Error Detection Sequence Labelling

- 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>Word</th>
<th>$P(c)$</th>
<th>$P(i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet</td>
<td>0.02</td>
<td>0.98</td>
</tr>
<tr>
<td>was</td>
<td>0.96</td>
<td>0.04</td>
</tr>
<tr>
<td>something</td>
<td>0.97</td>
<td>0.03</td>
</tr>
<tr>
<td>amazing</td>
<td>0.97</td>
<td>0.03</td>
</tr>
<tr>
<td>for</td>
<td>0.95</td>
<td>0.05</td>
</tr>
<tr>
<td>me</td>
<td>0.98</td>
<td>0.02</td>
</tr>
<tr>
<td>.</td>
<td>0.99</td>
<td>0.01</td>
</tr>
</tbody>
</table>
P(y_i|w_{1:N}) = [0.92; 0.08]

Hidden Layer

Backward LSTM

Forward LSTM

Word embedding

Character-level Bidirectional LSTM

w_i = dog

w_{i,1} = d

w_{i,2} = o

w_{i,3} = g

Marek Rei
1. Match data processing in training and testing
   a. 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
   b. Train: Text data - correct spelling errors, remove punctuation and casing

2. Train GED sequence labeler
   a. Each word is labelled as correct or incorrect

3. Apply GED model to test data
   a. Predict $P_{word}$ (grammar is correct) and $P_{word}$ (grammar is incorrect)
GED Experimental Corpora

- Cambridge Learner Corpus (CLC)
  - Learner written texts
  - Training: FCE, BULATS, IELTS, CPE, CAE; Test: FCE public evaluation set
- NICT Japanese Learner English (JLE)
  - Manual transcription of interviews from English oral proficiency test
  - 167 GE marked interviews – test on interviewees; grades A1-B2
- CA BULATS Spoken Corpus
  - Multi-level test with single speaker free speaking responses (sections C,D,E)
  - Manual and ASR transcriptions of 226 speakers, 1438 responses
  - Speakers evenly distributed across CEFR grades A1-C
GED Using CLC Trained Model

- Lower $F_{0.5}$ than usual as spelling mistakes and punctuation removed

$F_{0.5} 57.6\%$
GED Using CLC Trained Model

![Graph showing F0.5 scores for Spoken-BULATS, Spoken-NICT-JLE, and Written-CLC-FEC models.]

- **Spoken-BULATS**: F0.5 57.6%
- **Spoken-NICT-JLE**: F0.5 49.7%
- **Written-CLC-FEC**: F0.5 44.1%
• 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
    • Corrupt native speaker transcriptions to make learner errors
    • Adapt/“fine-tune” CLC trained system to subset of target speech data
• 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”
GED on BULATS ASR Transcriptions

- Manual transcriptions used for GE marking and meta-data extraction
- Significantly lower performance than manual transcriptions
  - Focus on boosting high precision region

- $F_{0.5} = 35.7\%$
- $F_{0.5} = 26.5\%$
Conclusions

• Learning Orientated Assessment
  • Assessment, feedback, learning activities synced

• Non-native spoken learner English
  • Large variation due to proficiency, L1 etc
  • Assessment for prompt-response on a par with human examiners
  • Research into feedback on errors progressing
    • Pronunciation
    • Use of English/grammar
Future: Conversational Auto-marking
Questions?

• Thanks to ALTA members especially CUED team, Andrew Caines, Marek Rei, Helen Yannakoudakis

• Email: kate.knill@eng.cam.ac.uk

• Try out Speak&Improve https://speakandimprove.com