Automated Assessment and Feedback: the Role of Spoken Grammatical Error Correction

Kate Knill
ALTA Institute, Cambridge University Engineering Department

14 December 2022
Cambridge Automated Language Teaching and Assessment Institute

- Virtual Institute for cutting-edge research on second language (L2) English assessment
  - Machine Learning and Natural Language Processing
  - Develop technology to enhance assessment and learning
  - Look to benefit learners and teachers worldwide

http://mi.eng.cam.ac.uk/~mjfg/ALTA/index.html
Plus members of the ALTA Spoken Language Processing Technology Project,
Cambridge University Engineering Department
Talk Outline

• Automated Learning and Assessment
• Spoken Grammatical Error Correction (GEC)
• Systems for Spoken GEC
• Using Spoken GEC in Automatic Assessment
• Using Spoken GEC for Learner/Teacher Feedback
• Open Challenges
Growth in Automated Learning and Assessment
Automated Learning and Assessment for English L2

- Close to 2 billion people worldwide use and/or are learning English as a second language
  - Not enough teachers or examiners!
    - Automated assessment and Computer-Aided Language Learning (CALL) systems
    - Data-driven Integrated Learning and Assessment

- Speaking is key skill for communication
  - Many systems ignore or heavily restrict speech input – not testing communication
    - Free speaking
      - Monologic e.g. giving a presentation
      - Dialogic e.g. discussing at a conference
L2 learner speech data is challenging!

- No punctuation/sentences
- ASR Errors
- Hesitations
- Disfluencies

Information encoded in how we speak not just what we say.
Spoken Language Assessment and Feedback Pipeline

Free speaking prompt-response tests – speak for up to 1 minute

Speech Processing

Grade over 1 or several responses

Grader

Score: 3.5
Conf: 90%

Feedback

Analytic – holistic feedback across all speech
Fine-grained – feedback on specific errors in words/phrases
Speaking core skills

- Pronunciation
- Hesitation / Extent
- Coherence / Discourse Management
- Language Resource
- Task Achievement
Language Resource: Grammar and Vocabulary

↑ English Proficiency    ↑ Range    ↑ Accuracy

↑ Communication of Ideas

↑ Task Achievement

Writing\(^1\)

1. https://writeandimprove.com
2. S. Bannò et al, “View-Specific Assessment of L2 Spoken English, INTERSPEECH 2022
Aim of GEC is to produce the grammatically correct sentence from a sentence with mistakes

Original: The cat sit on the mat.
Corrected: The cat sat on the mat.
The cat is sitting on the mat.

1:1 output assumption has relatively small effect on use of GEC

Large pre-trained GEC models available for English L1 text
  • Lots of annotated L2 learner English writing for fine-tuning
  • But can we apply these to speech?
Challenges and advantages of spoken language

• Challenges for assessing and giving feedback on speech “grammatical” errors
  • We don’t speak in sentences
  • We repeat ourselves, hesitate, mumble etc
  • No defined spoken grammar ... however there are phrases that a L1 speaker is highly unlikely to say

• Advantages of speech
  • No spelling or punctuation mistakes: over 25% written L2 English grammatical errors
  • We provide additional information about meaning of the text within the audio signal (and gestures)

  Text + Pronunciation + Prosody (+Delivery)

➢ Spoken Text ≠ Written Text
Original speech: the cat sit- sit on the um mat

Corrected: The cat sat on the mat.
Challenges for automatic Spoken GEC

• ASR errors are likely

the cat *sit* on the *um* mat → the cat *see* on the *um* mat → The cat *saw* on the mat.

  • Increased for learner speech as harder to recognise

• Have to automatically segment data into “sentences”

• Lack of training data

  • Very little annotated speech data – insufficient to train on
    • NICT-JLE¹ - no audio samples
    • Linguaskill Speaking² (internal data set) – approx. 3000 annotated responses
  • Written text data is mismatched to speaking e.g. no disfluencies, stricter grammar, different style

Requirements for Spoken GEC for Assessment and Feedback

• Accurate
  • Emphasis on precision with reasonable recall
  • Feedback only on actual grammatical errors – not errors made by the ASR system
    • Giving feedback on an ASR error is worse than giving no feedback

• Interpretable/explainable
  • Let the learner know why what they did was incorrect

• Reasonable computational load and time
  • Near real-time needed for practice apps
Spoken GEC – End2end?

- E2E not feasible (currently)
  - No paired training data
  - Hard to give feedback to learners

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Audio</th>
<th>Text</th>
<th>DSF</th>
<th>GEC</th>
<th>L2?</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASR-Train(^1)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Switchboard(^2)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>CLC(^3) + BEA(^4)</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

1. Y. Lu et al, “Impact of ASR Performance on Spoken Grammatical Error Detection”, INTERSPEECH 2019
Spoken GEC – Cascade pipeline

- Train/fine-tune each module on different data sets
- Various sources of information for feedback
Spoken GEC – Cascade pipeline challenges

Non-native disfluent speech → ASR → DSF → GEC → Grammatically correct fluent text

- **Error propagation**
  - the cat *sit* on the *um* mat → the cat *sit* on the mat → the cat *sat* on the mat

- **Domain mismatch**
  - lemme in → GEC → *<unk>* in

- **Prosody information loss**
  - i feel *well* pretty bad → DD → i feel *well* pretty bad

➢ Error propagation
➢ Domain mismatch
➢ Prosody information loss
Automatic Speech Recognition (ASR)

• Hybrid ASR¹
  • Acoustic model:
    • Kaldi LF-MMI TDNN-F² on L2 learner speech (ASR-Train) and AMI
  • Language model:
    • Trigram LM trained on ASR L2 manual transcriptions and Broadcast News English
    • su-RNNLM³ trained on semi-supervised ASR L2 transcriptions

• Key features:
  • L2 speech from over 25 L1s
  • Explicitly model partial words and map all hesitations to a single word label
  • Word level confidence scores

1. Y. Lu et al, “Impact of ASR Performance on Spoken Grammatical Error Detection”, INTERSPEECH 2019
3. X. Chen et al, “Future word contexts in neural network language models”, ASRU 2017
Spoken Disfluencies

- Typical disfluency
  
  - reperandum
  - interregnum
  - repair

  I want a taxi [ to Marina Bay uh I mean + to Sentosa ]

  I want a taxi to Sentosa

- Interregnum regions – easy to detect using rule-based methods
- Reparandum regions → focus of automatic disfluency detection
Automatic Disfluency Detection (DD)

- Sequence tagging models

\[
\hat{d}_{1:L} = \arg \max_{d_{1:L}} P(d_{1:L} | w_{1:L}; \theta_{\text{DD}})
\]

- Non auto-regressive tagger

\[
P(d_{1:L} | w_{1:L}; \theta_{\text{DD}}) = \prod_{l=1}^{L} P(d_{l} | d_{1:l-1}, w_{1:L}; \theta_{\text{DD}}) \approx \prod_{l=1}^{L} P(d_{l} | w_{1:L}; \theta_{\text{DD}})
\]

- Performance largely dependent on quality of tagger’s internal feature representations
  - Depends on training criteria and framework

---

Sequence Tagging DD Models

- Recurrent neural network (RNN)$^{1,2}$
  - Reasonable good quality feature reps
  - Training scales with sequence lengths

- Large pre-trained model: BERT$^3$
  - Capable of high quality feature reps
  - Fine-tune BERT using DD objective

Since your going out you should check weather on your phone.

Since you're going out, you should check the weather on your phone.

\[ \hat{y}_{1:N} = \arg\max_{y_{1:N} \in Y} P(y_{1:N} \mid w_{1:L}; \theta_{GEC}) \]
seq2seq GEC

Treat GEC as a machine translation problem i.e. seq2seq

- Encoder-decoder architecture
- RNN-based → Transformer-based encoder-decoder models
  - Large pretrained e.g. T5 ’Text-to-Text Transfer Transformer’

- Gramformer\(^2\): T5-based fine-tuned for GEC on various corpora
  - Can further fine-tune to target domain GEC
  - 223M parameters

- ChatGPT\(^3\): generative autoregressive large language model [CGPT]
  - 175B parameters

---

Sequence tagging GEC

Treat GEC as a sequence tagging problem\textsuperscript{1,2}

- For each token in the source sequence GEC should produce a tag (edit) for required correction operation
  - e.g. GECToR by Grammarly (138M parameters)

GECToR architecture (from O. Skurzhanskyi and K. Omelianchuk\textsuperscript{3})

2. A. Awasthi et al, “Parallel Iterative Edit Models for Local Sequence Transduction”, EMNLP-IJCNLP, 2019
3. O. Skurzhanskyi and K. Omelianchuk, “Building sequence tagging approach to Grammatical Error Correction and Text Simplification”, talk at Riken Japan, 2022

A ten years old boy go school

A \Rightarrow A:
- ten $\Rightarrow$ ten, -:
- years $\Rightarrow$ year, -:
- old $\Rightarrow$ old:
- go $\Rightarrow$ goes, to:
- school $\Rightarrow$ school,

A ten-year-old boy goes to school.

Predicted tags

Error correction linear layer

Error detection linear layer

Encoder (transformer-based)

Input sentence

Corrected sentence

Repeat N times
## SGEC System Corpora

### Table 1: Corpora Overview

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Spoken?</th>
<th>#sents</th>
<th>#words</th>
<th>% disfluencies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLC(^1)</td>
<td>✗</td>
<td>1.9M</td>
<td>25.2M</td>
<td>-</td>
</tr>
<tr>
<td>BEA(^2)</td>
<td>✗</td>
<td>1M</td>
<td>11.5M</td>
<td>-</td>
</tr>
<tr>
<td><strong>Evaluation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FCEtst(^3)</td>
<td>✗</td>
<td>2681</td>
<td>37k</td>
<td>-</td>
</tr>
<tr>
<td>LIN(^4)</td>
<td>✓</td>
<td>3361</td>
<td>38k</td>
<td>5%</td>
</tr>
</tbody>
</table>

2. C. Bryant et al, “The BEA-2019 shared task on grammatical error correction”. BEA Workshop 2019
SGEC training data pre-processing

- Convert written text corpora for training to “speechify” it
  - Correct any spelling errors
  - Remove punctuation
  - Upper (or lower) case all text to match ASR
  - Adding speech-like disfluencies through a masked LM can help (not used here)

It was a cleer, blue, sky.
↓
IT WAS A CLEAR BLUE SKY
SGEC spoken evaluation data

- LIN: Linguaskill Speaking L2 learner speech from multi-level English test
  - 3361 grammatically corrected ‘sentences’ from 972 question responses
  - 340 speakers from over 30 L1s
  - Approximately equal distribution over grades A1-C1
- ASR and Manual transcriptions available
  - With/out disfluencies
  - Manually annotated by English Language iTutoring (ELiT) Ltd

\[\text{Diagram of process with labels:}\]

- Audio
- ASR Transcription
- Manual Transcription (disfluencies removed)
- I Scores
- Holistic Analytic
- II Transcript
- Transcription Disfluences Pronunciation Phrase Boundaries
- Correct Grammar
- III Grammar

30
Phase 2 annotation

Previous Annotations

Answer

Long turn 1

Talk about a training course you attended for your work. You should say: • what the course was about • why you went on the course • what you learnt from it.

GENERALY IN JANUARY I ATTENDED LT PROJECT MANAGEMENT TRAINING. I HESITATION BECAUSE I ATTEND THE CO HESITATION TRAINING BECAUSE MY BOSS HESITATION DIRECTOR KINDA GENERAL RECOMMENDED ME AGAIN TO ATTEND THE CO HESITATION TRAINING COURSE. THE TRAINING COURSE CONSISTS OF TWO PARTS: ONE VEHICLES ONE-DAY COURSE IN THE MORNING WE HAVE SIMILAR SEMINAR TO LEARN ABOUT HESITATION HESITATION BASIC SKILL TO MANAGE LT DEVELOPMENT PROJECT HESITATION FOR EXAMPLE HESITATION MANDATES MANAGE THE VENDORS IF OF CONSULTANTS IMPROVED INVOLVED IN THE LT DEVELOPMENT PROJECT. IN THE AFTERNOON BECAUSE COURSE WE HAVE THE LITTLE SESSION HOW TO MANAGE A DIFFERENT AS OF CONSULTANTS HESITATION TO IMPROVE HESITATION TO MANAGE THE HESITATION IF LT
Annotation phase 2 to phase 3

Long turn 1

Talk about a training course you attended for your work. You should say: • what the course was about • why you went on the course • what you learnt from it.

IN JANUARY I ATTENDED I-T PROJECT MANAGEMENT TRAINING. I ATTEND THE TRAINING BECAUSE MY BOSS DIRECTOR GENERAL RECOMMENDED ME TO ATTEND THE TRAINING COURSE. THE TRAINING COURSE CONSISTS OF TWO PARTS IN ONE-DAY COURSE. IN THE MORNING WE HAVE SEMINAR TO LEARN ABOUT BASIC SKILL TO MANAGE I-T DEVELOPMENT PROJECT FOR EXAMPLE TO MANAGE THE 2 OF CONSULTANTS INVOLVED IN THE I-T DEVELOPMENT PROJECT. IN THE AFTERNOON COURSE WE HAVE THE 2 SESSION HOW TO MANAGE A 2 OF CONSULTANTS TO MANAGE THE I-T.
Annotation phase 3

• Minimal edit corrections made
• Type of error not marked
Measuring SGEC performance

• Align and classify GEC edits with ERRor ANnotation Toolkit (ERRANT)\textsuperscript{1,2}
  • Automatically extracts edits from parallel original and corrected sentences
  • Classifies them according to a dataset-agnostic rule-based framework
  • Facilitates error type evaluation at different levels of granularity

<table>
<thead>
<tr>
<th>Auto:</th>
<th>the</th>
<th>cat</th>
<th>sit</th>
<th>on</th>
<th>mat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference:</td>
<td>the</td>
<td>cat</td>
<td>sat</td>
<td>on</td>
<td>the</td>
</tr>
<tr>
<td>Edit:</td>
<td>R:VERB</td>
<td>M:DET</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• Spoken language differences to written text
  • Alignment will change depending on if Manual or ASR reference
  • Edit classification rules do not take speech effects into account such as disfluencies and partial phrases

1. C. Bryant et al, “Automatic annotation and evaluation of error types for grammatical error correction”, ACL 2017
Spoken GEC Performance

- Manual, Fluent speech transcripts → Gramformer

<table>
<thead>
<tr>
<th>FCEtst</th>
<th>LIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>M² F₀.₅ ↑</td>
<td>56.6</td>
</tr>
</tbody>
</table>

- Spoken GEC can match performance of Written GEC L2 English learner task

- Cascade GEC system: TDNN-F ASR → BERT DSF → Gramformer

<table>
<thead>
<tr>
<th>Input to Gramformer</th>
<th>LIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual</td>
<td>ASR+DD+GEC</td>
</tr>
<tr>
<td>SER ↓</td>
<td>43.3</td>
</tr>
<tr>
<td>TER ↓</td>
<td>8.3</td>
</tr>
</tbody>
</table>

- Need to mitigate for ASR error propagation and domain mismatch
Over prediction of OTHER due to ERRANT missing speech-based rules

* This system did not have DD module
Integrated Spoken GEC Training

- Approaches towards tighter integration
  - Semi-supervised ASR error mitigation
  - Reranking
  - Embedding passing
Using Spoken GEC in Assessment
Holistic “hand-crafted” features-based auto-marker

Automatic Speech Recognition → Compute Features → Input Features → Grader

SGEC

\[ \hat{y} = G(x, \theta) \]

Standard approach to speech auto-marking
View Specific and/or Multi-view auto-marker

1. S. Bannò et al, “View-Specific Assessment of L2 Spoken English”, INTERSPEECH 2022
Grammatical accuracy grader pipeline

(no disfluency detection module used here)
GEC Grader evaluation

- ERRANT label distribution over grades

- Grader RMSE

<table>
<thead>
<tr>
<th>Model</th>
<th>LinBus</th>
</tr>
</thead>
<tbody>
<tr>
<td>holistic</td>
<td>0.406</td>
</tr>
<tr>
<td>GEC</td>
<td>0.495</td>
</tr>
<tr>
<td>text</td>
<td>0.409</td>
</tr>
<tr>
<td>pronunciation</td>
<td>0.451</td>
</tr>
<tr>
<td>5-way multi-view</td>
<td>0.386</td>
</tr>
</tbody>
</table>
Using Spoken GEC for Learner/Teacher Feedback
What and how to present?

A. Filighera et al, "Towards A Vocalization Feedback Pipeline for Language Learners," 2022 ICALT

https://writeandimprove.com

https://liulishuo.com
Improve feedback precision: remove ‘OTHER’

• High number of OTHER tags in manual → automatic labels
  • Ambiguous type, not informative to learners
  • May be a paraphrase
  ➢ Don’t present to learner
Improve feedback precision: confidence filtering

- Sentence level > edit level confidence filtering

\[
\log P = \alpha \log P_{ASR} + \beta \log P_{DSF} + \gamma \log P_{GEC}
\]
**Confidence filtering type analysis**

<table>
<thead>
<tr>
<th>Edit Type</th>
<th>ELIT-ASR-AUTO None</th>
<th>ELIT-ASR-AUTO Sent</th>
</tr>
</thead>
<tbody>
<tr>
<td>M:DET</td>
<td>19.12</td>
<td>27.42</td>
</tr>
<tr>
<td>R:PREP</td>
<td>21.54</td>
<td>33.33</td>
</tr>
<tr>
<td>R:NOUN:NUM</td>
<td>22.58</td>
<td>38.39</td>
</tr>
<tr>
<td>R:VERB:TENSE</td>
<td><strong>24.54</strong></td>
<td>21.40</td>
</tr>
<tr>
<td>U:DET</td>
<td>17.19</td>
<td>21.85</td>
</tr>
<tr>
<td>R:VERB</td>
<td>22.61</td>
<td>24.14</td>
</tr>
<tr>
<td><strong>R:NOUN</strong></td>
<td><strong>27.87</strong></td>
<td>12.46</td>
</tr>
<tr>
<td>M:PREP</td>
<td>12.15</td>
<td>18.32</td>
</tr>
<tr>
<td>R:VERB:FORM</td>
<td>22.37</td>
<td><strong>29.93</strong></td>
</tr>
<tr>
<td>R:VERB:SVN</td>
<td>16.85</td>
<td>34.55</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>20.89</strong></td>
<td><strong>24.28</strong></td>
</tr>
</tbody>
</table>

---

**Fo.5 breakdown – top 10 edit types**

- **IMPROVE** with deterministic correction
- **DEGRADE** with more than 1 feasible correction
Open Challenges for Spoken GEC

- **Increasing Accuracy**
  - Tighter integration of modules
  - Data augmentation
  - Domain adaptation of GEC networks
  - Sequence uncertainty – when was a good/bad GEC sequence generated?

- **Interpretability/explainability**

Saliency: What caused the grammatical error?
ChatGPT: are huge amounts of data and compute “all we need”? 

What about “your manager has asked you for the result of an internal staff survey on travelling to work the first chart shows how staff at your company travelled to work the second chart shows what improvements to company facilities the staff would like to see look at the chart and then talk about the information describing the survey results um by car forty percent by train thirty five thirty five percent by bus ten percent on foot five percent bicycle five percent motor bicycle motorbike five percent free company bus to and from station very many”

Here is a corrected version of the sentence:

"Your manager has asked you for the result of an internal staff survey on travelling to work. The first chart shows how staff at your company travelled to work, and the second chart shows what improvements to company facilities the staff would like to see. Look at the chart and then talk about the information, describing the survey results: by car, 40%; by train, 35%; by bus, 10%; on foot, 5%; by bicycle, 5%; by motorbike, 5%; and a free company bus to and from the station is desired by many."

In this sentence, I made a few changes to improve the grammar and clarity. First, I added the missing article “the” before “first chart” and “second chart”. I also added the article “a” before “free company bus” to make it clear that there is only one bus. I also changed the word “very” to “many” to better match the verb “desired”. Finally, I added a few commas to make the sentence easier to read. Let me know if you have any other questions.
ChatGPT: ASR errors are still an issue

Last one (for now ;-)]: "i think my stall language skill is have fast range rate and experiences for quality control as we working for more than ten years"

Here is a corrected version of the sentence:

"I think my language skills have a fast rate of improvement, and I have a lot of experience in quality control. I have been working in this field for more than ten years."

In this sentence, I made a few changes to improve the grammar and clarity. First, I changed "stall" to "skills" to correctly reflect the subject of the sentence. I also added the article "a" before "fast rate of improvement" to make it clear that we are talking about a specific rate. I also changed "experiences" to "experience" because "experiences" is a plural noun, while "experience" is a singular noun. Finally, I added the phrase "I have been working in this field for more than ten years" to clarify the subject's experience in quality control. Let me know if you have any other questions.

• ASR error: “I think my strongest skill...” → “I think my stall language skill ...”
Conclusions

• Automated assessment and computer aided language learning is growing rapidly
  • Needed to cope with over 2Bn learners/speakers

• Learners increase in grammatical accuracy and range as they progress
  • Automatic Grammatical Error Correction (GEC) is very popular for Writing Assessment and Feedback
  • Automatic Spoken GEC (SGEC) shown to be of use to Speaking Assessment and Feedback

• Lots of open challenges in Automatic SGEC
  • Improving core accuracy whilst handling limited labelled training data
  • Increase robustness to and awareness of ASR errors
  • Add interpretability and explainability to give more in-depth feedback
Thanks

Yiting (Edie) Lu and Stefano Bannò for multiple pictures – in addition to their research with Prof Mark Gales and past and present members of the ALTA SLP Technology Project, Cambridge University Engineering Department.

Diane Nicholls and the humannotar team at ELiT for the Linguaskill annotations.

This presentation reports on research supported by Cambridge University Press & Assessment, a department of The Chancellor, Masters, and Scholars of the University of Cambridge.

Questions?