Automated Marking for Learner Language.

Paula Buttery

The ALTA Institute
Gonville and Caius College
University of Cambridge
Automated scoring and feedback

- Automatically analyse the quality of the written and spoken language components of learner language and provide feedback.

Goals

- Assign scores as reliably as human examiners.
- Provide feedback that enables self learning and promotes individualised teacher interaction.
Advantages of automated assessment and feedback

Advantages for students:
- Immediate scoring response
- Constant availability of online systems promotes language skills development at any time
- Enables self-assessment and self-tutoring which complements teacher contact hours

Advantages for teachers/examiners:
- Reduced teacher/examiner workload—can focus on the more interesting/subtle content
- Cost-effective approach to (large-scale) grading
Advantages of automated assessment and feedback

Consistency and reliability:

- Automation ensures that constant assessment criteria are applied for scoring.
- ALTA automated scorer is realistic: trained on real examination responses.
ALTA institute is experienced in automated assessment

- Successful deployment of automated services for Cambridge examinations:
  - KET (A2: elementary)
  - PET (B1: intermediate)
  - FCE (B2: upper intermediate)
  - CEPT: (A1: beginner – C2: proficient)
  - IELTS (A2: elementary – C2: proficient)

- Write & Improve: online writing assessment tool
  - 3 trials, 10 institutions from 9 countries
How does our technology work?

**VERY brief recap on Machine Learning**

Grades are automatically assigned to writing and spoken language using an automated scorer build using machine learning techniques.

**Successful machine learning requires 3 components:**

- representative training data
- sophisticated methods for extracting features to use as machine learning parameters
- appropriate and sophisticated choice of machine learning algorithm

...let’s look at each component in turn.
CLC provides representative training data

The Cambridge Learner Corpus

- 200,000 exam scripts
- 147 language backgrounds
- 217 countries.

Access to Cambridge examination data and experienced Cambridge examiners:

- Fully representative training data: real written examination scripts, spoken language recordings
- Training scores assigned by actual examiners and score validation
Representing input data as predictive features

- Written exam script or audio
- Feature extractor
- Script or audio represented as features
- Machine Learning Algorithm
- Human IELTS score

- Student text or audio
- Feature extractor
- Student text or audio represented as features
- Automated Scorer
- Machine IELTS score
Representing input data as predictive features

Then some though occurred to me

1. Word sequences
   - (unigram) Then, some, though, occurred, ...
   - (bigram) Then some, some though, though occurred, ...
   - (trigram) Then some though, some though occurred, though occurred to, ...

Then a thought occurred to me

1. Word sequences
   - (unigram) Then, a, thought, occurred, ...
   - (bigram) Then a, a thought, thought occurred, ...
   - (trigram) Then a thought, a thought occurred, thought occurred to, ...
Then some thought occurred to me
Then_RR some_DD though_RR occurred_VVN to_Il me_PPI01

2 Linguistic category sequences
   • RR DD (e.g. Then some)
   • RR VVN (e.g. occurred to)

Then a thought occurred to me
Then_RR a_AT1 thought_NN1 occurred_VVD to_Il me_PPI01

2 Linguistic category sequences
   • RR AT1 (e.g. Then a)
   • RR VVN (e.g. occurred to)
Representing input data as predictive features

Grammatical constructions using Robust Accurate Statistical Parser

- T/frag
- Tph/np
- NPh/np
- NP/a1-cat_np-r
- A1/a
  - Then_RR
  - some_DD
- NP/det_a1-r
- A1/advp_ppart-r
  - A1/a
  - occur+ed_VVN
  - though_RR
- PP/p1
  - P1/p
- P1/p_np-pro
  - to_II
  - I+_PPIO1
Representing input data as predictive features

3. Grammatical constructions using Robust Accurate Statistical Parser

T/txt-sc1
\[ \text{S/adv.s} \]
\[ \text{AP/a1} \]
\[ \text{S/np_vp} \]
\[ \text{V1/v_pp} \]
\[ \text{A1/a} \]
\[ \text{NP/det_n1} \]
\[ \text{V1/v_pp} \]
\[ \text{Then}_{\text{RR}} \]
\[ \text{a_AT1} \]
\[ \text{N1/n} \]
\[ \text{occur+ed}_\text{VVD} \]
\[ \text{P1/p1} \]
\[ \text{thought}_{\text{NN1}} \]
\[ \text{to}_{\text{II}} \]
\[ \text{I+}_\text{PPIO1} \]
Grammatical constructions using Robust Accurate Statistical Parser

- NP/a1-cat_npr (e.g. *Then some though occurred*)
- PP/p1 (e.g. *to me*)
Representing input data as predictive features

4. Errors: error coded data, error rate per document

Then \(<\#ER>\) some\(a<\#/ER>\) \(<\#ER>\) though\(thought<\#/ER>\) \\
\(<\#ER>\) occurred\(occurred<\#/ER>\) to me.

5. Other features: complexity...
Evaluation: almost human-like accuracy

<table>
<thead>
<tr>
<th>Features</th>
<th>human-machine agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>word seq</td>
<td>59.8%</td>
</tr>
<tr>
<td>+linguistics category seq</td>
<td>68.7%</td>
</tr>
<tr>
<td>+syntax structure</td>
<td>72.2%</td>
</tr>
<tr>
<td>+error rate</td>
<td>78.5%</td>
</tr>
<tr>
<td>human-human</td>
<td>79.2%</td>
</tr>
</tbody>
</table>
Successful machine learning requires 3 components:

- representative training data
- sophisticated methods for extracting features to use as machine learning parameters
- appropriate choice of machine learning algorithm
Previous work has used binary classification models.
Previous work has used binary classification models.
Sophisticated Machine Learning algorithm: Ranked SVM

Provides a score on a continuum based on feature predictors:

- Input to algorithm is a ranking of all training documents.
- Algorithm builds a scorer that takes into account the relationships between all training documents.
- The scorer is now placing an unseen document within the training rank.
- Most appropriate algorithm for graded data.
- Patent pending technology.
- Models the fact that some scripts are better than others.

Cambridge ALTA
An appropriate algorithm improves performance

### Best previous algorithm vs. ranking algorithm: example written exam

<table>
<thead>
<tr>
<th>Features</th>
<th>human-machine agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best previous model</td>
<td>75.0%</td>
</tr>
<tr>
<td>Ranking Model</td>
<td>78.5%</td>
</tr>
<tr>
<td>human-human</td>
<td>79.2%</td>
</tr>
</tbody>
</table>
The automated scorer is indistinguishable from a human

<table>
<thead>
<tr>
<th>Rater1</th>
<th>Rater2</th>
<th>Rater3</th>
<th>Rater4</th>
<th>Rater5</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>26</td>
<td>23</td>
<td>25</td>
<td>23</td>
</tr>
<tr>
<td>33</td>
<td>36</td>
<td>31</td>
<td>38</td>
<td>36</td>
</tr>
<tr>
<td>29</td>
<td>25</td>
<td>22</td>
<td>25</td>
<td>27</td>
</tr>
<tr>
<td>24</td>
<td>23</td>
<td>20</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>25</td>
<td>25</td>
<td>22</td>
<td>24</td>
<td>22</td>
</tr>
<tr>
<td>27</td>
<td>26</td>
<td>23</td>
<td>30</td>
<td>24</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>5</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>21</td>
<td>24</td>
<td>21</td>
<td>25</td>
<td>19</td>
</tr>
</tbody>
</table>
Successful machine learning requires 3 components:

- representative training data
- sophisticated methods for extracting features to use as machine learning parameters
- appropriate choice of machine learning algorithm

Human-like accuracy is achievable for holistic scoring of texts

...how do we perform on more fine grained band marking criteria?
## Automated scoring for detailed marking criteria

### Agreement between humans and machine: detailed writing criteria

<table>
<thead>
<tr>
<th>Criteria</th>
<th>human-machine agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>task achievement</td>
<td>73%</td>
</tr>
<tr>
<td>human-human</td>
<td>73%</td>
</tr>
<tr>
<td>coherence &amp; cohesion</td>
<td>76%</td>
</tr>
<tr>
<td>human-human</td>
<td>79%</td>
</tr>
<tr>
<td>lexical resource</td>
<td>82%</td>
</tr>
<tr>
<td>human-human</td>
<td>81%</td>
</tr>
<tr>
<td>grammatical range &amp; accuracy</td>
<td>83%</td>
</tr>
<tr>
<td>human-human</td>
<td>83%</td>
</tr>
<tr>
<td>aggregate score</td>
<td>83%</td>
</tr>
<tr>
<td>human-human</td>
<td>85%</td>
</tr>
</tbody>
</table>
The system is robust to lexical gaming

Validity Tests

Determine the extent to which certain text generation strategies pose a threat to the validity of the system

1. Randomly order:
   a. word unigrams within a sentence
   b. word bigrams within a sentence
   c. word trigrams within a sentence

2. Swap words that have the same linguistic category within a sentence

Cambridge ALTA
The system is robust to lexical gaming

Agreement between humans and machine: cheating strategies

<table>
<thead>
<tr>
<th>Cheating Strategy</th>
<th>human-machine agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomly order unigrams</td>
<td>91.2%</td>
</tr>
<tr>
<td>Randomly order bigrams</td>
<td>91.4%</td>
</tr>
<tr>
<td>Randomly order trigrams</td>
<td>86.7%</td>
</tr>
<tr>
<td>swap same linguistic category words</td>
<td>76.1%</td>
</tr>
</tbody>
</table>
What happens if the learner pastes good sentences?

Good sentences in the wrong order will be incoherent but the holistic scorer will score these texts well.

**human-machine agreement : 16.3%**

**Incremental Semantic Analysis (ISA)**

- Coherence of a text $T$:

$$\text{coherence}(T) = \frac{\sum_{i=1}^{n-1} \max_{k,j} \text{sim}(s_i^k, s_{i+1}^j)}{n-1}$$  \hspace{1cm} (1)

- Underlying idea: the overlap in meaning between adjoining sentences serves as a proxy for local discourse coherence

**final human-machine agreement : 62.6%**
Feedback is provided at several levels of granularity

We've discussed the scoring in some detail... What feedback can we provide to learners to improve their scores?

Feedback is provided at 3 levels of granularity:
- Script-level feedback: assess a text as a whole
- Sentence-level feedback: assess sentences independently
- Word-level feedback: error detection and correction
Introduction
How does automated marking work?
Feedback
Discussion

Text-level feedback
Sentence-level feedback
Word-level feedback
User satisfaction

’Write and Improve’ provides automatic learner feedback

Overall score

A1  A2  B1  B2  C1  C2

Latest score

An overall score indication is assigned on the CEFR level scale. Remember, this score is only a guide. You might not get the same score in an exam. Once you have had a look at your feedback, try to use it to improve your writing, and then submit it again.

Now, improve your answer

I live in Moscow, the capital of the Russian Federation. There are different types of public transport that people use in major cities in my country, including Moscow: bus, train, tram and metro. Furthermore people actively use private transport such as cars, motorbikes, scooters and bicycles.

Tourists can encounter with difficulties such as a traffic jam. Sometimes it is very difficult to come home in rush hour because of traffic jams. Banning cars from the centre of the city solves all these problems.

There are some disadvantages of banning cars. For example, it is difficult for some people, including the aged and invalids, to use public transport. I suppose that banning cars from the centres would be a big problem for them. This idea also would be unpopular because it limits a freedom of choice.

Word count: 178

Save  Save & Submit
’Write and Improve’ provides automatic learner feedback

Detailed feedback (Help)

Sentence feedback gives you an idea of the general quality of each sentence. The colours range from green to red through yellow and orange. Green suggests a well written sentence. Yellow and orange suggests the system believes the sentence is acceptable. Red suggests the sentence may have a few problems.

Language acquisition is the process by which humans acquire the capacity to perceive and comprehend language, as well as to produce and use words and sentences to communicate. Language acquisition is one of the quintessential human traits, because non-humans do not communicate by using language. Language acquisition usually refers to first-language acquisition, which studies infants' acquisition of their native language. This is distinguished from second-language acquisition, which deals with the acquisition (in both children and adults) of additional languages.

The capacity to successfully use language requires one to acquire a range of tools including phonology, morphology, syntax, semantics, and an extensive vocabulary. Language can be vocalized as in speech, or manual as in sign. The human language capacity is represented in the brain. Even though the human language capacity is finite, one can say and understand an infinite number of sentences, which is based on a syntactic principle called recursion. Evidence suggests that every individual has three recursive mechanisms that allow sentences to go indeterminately. These three mechanisms are: relativization, complementation and coordination. Furthermore, there are actually two main guiding principles in first-language acquisition, that is, speech perception...
We can use error rate as a proxy for a category label

**Sentence Evaluation**
Assess and score the quality of individual sentences, independently of their context

**Approach**
- Exploit already available annotated data:
  - Text-level scores and error annotation in CLC
- Combine text-level score and errors per sentence, and create pseudo-gold labels to train a sentence classifier
In the past people didn't have electricity and if they wanted, for example, to read or to cook something they used to light a fire.
You must have a TV because you can learn about what is happening in the world and you can see some places that you haven't been to.
You can enjoy watching a film if you have some free time.
In our daily life, however, we seldom notice how easy a life we've got or, what is more, how difficult our grandparents found it.

In the past the people didn't have electricity and if they wanted for example to read or to cook something they used to do in the fire.
You must have TV because you can listen what it happened in the world and you can watch some places that you didn't go.
You can enjoy your time to watch a film if you have free time.
In our daily life, however, we seldom notice how much convenient life we've got, what is more, how much inconvenient our grandparents had got.
Word level feedback must be precise

Error detection and correction

Ensure high accuracy (precision) and reasonable coverage (recall)

1. Corpus-derived rules
   - Error rules derived from the Cambridge Learner Corpus (CLC)
   - Detect incorrect word sequences (unigrams, bigrams and trigrams)
   - At least 90% incorrect occurrences (recurrent errors in CLC)

2. Electronic dictionary-derived rules (morphology)

Error detection (increase coverage / recall)

Highlight words/phrases that may need attention, but for which the system can’t provide suggestions
Response text

Some people learn a foreign language in order to widen their horizons and etc. Perhaps you prefer to stay on dry land.

Can you sea the see from were you live?

Possible errors

**and** Insertion: This word may not be needed.
**etc.** Substitution: A different word might be better here. Perhaps ‘so on’ is better.
**sea** Confusion: Is this the right word? Did you mean to write ‘see’?
**the** Insertion: This word may not be needed.
**see** Confusion: Is this the right word? Did you mean to write ‘sea’?
Word level feedback: Error detection

As for the last day I would suggest you to take a visit to the local museum and the Castle.

Television has affected the all days life, too.

Therefore I am thinking in selling my car, which has changed my life over the last 10 years.

In our lifes, may be, couldn't be without electricity.

I am replying your letter for the help you're being need.

I hope you found the letter useful and If you have nay
Introduction

How does automated marking work?

Feedback

Discussion

Text-level feedback

Sentence-level feedback

Word-level feedback

User satisfaction

User trials: our feedback is shown to be helpful

- 3000+ submissions in 3 trials
  - Over 600,000 words
  - Average response length: 200 words
- Average number of attempts to improve: 3.2
- Median of number of attempts to improve: 2
- Max number of attempts to improve: 54
- Score given to the last revision is higher than that given to the initial revision in over 80% of the cases
## User satisfaction

<table>
<thead>
<tr>
<th></th>
<th>Trial 1</th>
<th>Trial 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using W&amp;I helps me to write better in English</td>
<td>3.80</td>
<td>3.92</td>
</tr>
<tr>
<td>I find W&amp;I useful for understanding my mistakes</td>
<td>3.74</td>
<td>3.96</td>
</tr>
<tr>
<td>I think the sentence colouring is useful</td>
<td>3.74</td>
<td>4.15</td>
</tr>
<tr>
<td>I think the word-level information [error feedback] is useful</td>
<td>3.86</td>
<td>4.12</td>
</tr>
<tr>
<td>W&amp;I is easy to use</td>
<td>4.45</td>
<td>4.49</td>
</tr>
<tr>
<td>The feedback on my writing is clear</td>
<td>3.80</td>
<td>3.93</td>
</tr>
<tr>
<td>If you have used W&amp;I before, has it improved since the last time?</td>
<td>—</td>
<td>3.86</td>
</tr>
</tbody>
</table>

**Table:** Average feedback scores on a scale from 1 (strongly disagree) to 5 (strongly agree)

- User-driven development between trials
## Evaluation: automated speech scorer

<table>
<thead>
<tr>
<th>Speech Scorer</th>
<th>human-machine agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best model 2014</td>
<td>83.1%</td>
</tr>
<tr>
<td>Best model 2015</td>
<td>83.8%</td>
</tr>
<tr>
<td>human-human</td>
<td>85.0%</td>
</tr>
</tbody>
</table>
Can we assess spoken transcripts using Write and Improve?

Overall score

An overall score indication is assigned on the CEFR level scale. Remember, this score is only a guide. You might not get the same score in an exam. Once you have had a look at your feedback, try to use it to improve your writing, and then submit it again.

Detailed feedback (Help)

Sentence background colours give you an idea of the general quality of each sentence. White suggests a well written sentence. Pale yellow suggests the system believes the sentence is acceptable. Brighter yellows suggest the sentence may have a few problems. A red box indicates that explanations or corrections are available and can be viewed by hovering over the word. An orange box indicates words that might need attention to improve your results, but for which the system doesn't have a suggestion.

Set your sights realistically haven’t you? And there’s a lot of people unemployed and what are you going to do when you eventually leave college if you get there? You’re not gonna step straight into television mm right then, let’s see now what we’re doing, where’s that recipe book for that chocolate and banana cake? Chocolate and banana cake which book was it? Oh right (pause) oh some of these chocolate cakes are absolutely mm, mm. mm right what’s the topping? what’s that icing sugar? cocoa powder and vanilla essence oh luckily I’ve got all those, I think, yes.

Set your sights realistically haven’t you? And there’s a lot of people unemployed and what are you going to do when you eventually leave college if you get there? You’re not gonna step straight into television mm right then, let’s see now what we’re doing, where’s that recipe book for that chocolate and banana cake? Chocolate and banana cake which book was it? Oh right oh some of these
Can we assess spoken transcripts using **Write and Improve**?

- NLP tools are trained on written data—feature extraction will be consistent but difficult to interpret.
- We don’t have a large corpus to help us detect possible errors.
- We can’t give any useful learner feedback.
Detailed spoken language feedback is challenging

Example: providing grammatical feedback on spoken language

Example Spoken Component

da I live in <unknown> err I live in a flat room err it’s about 20 heh err it’s about 200 err uh uh quarter meters and there are 4 room in my room err there are 4 rooms in my flat yeah wo and two one kitchen
Research challenge: segmenting spoken language

da I live in <unknown> err I live in a flat room err it’s about 20 heh err it’s about 200 err uh uh quarter meters and there are 4 room in my room err there are 4 rooms in my flat yeah wo and two one kitchen

da I live in <unknown>
err I live in a flat room
err it’s about 20 heh err it’s about 200 err uh uh quarter meters and there are 4 room in my room err there are 4 rooms in my flat yeah
wo and two one kitchen
da I live in <unknown>
err I live in a flat room
err it's about 20 heh err it's about 200 err uh uh quarter meters and there are 4 room in my room err there are 4 rooms in my flat
yeah
wo and two one kitchen
Research challenge: removing hesitation

da I live in <unknown>
err I live in a flat room
err it’s about 20 heh err it’s about 200 err uh uh quarter meters and there are 4 room in my room err there are 4 rooms in my flat yeah wo and two one kitchen
Research challenge: removing hesitation

da I live in <unknown>
err I live in a flat room
err it’s about 20 heh err it’s about 200 err uh uh quarter meters and there are 4 room in my room err there are 4 rooms in my flat
yeah
wo and two one kitchen

I live in <unknown>
I live in a flat room
It’s about 20 it’s about 200 quarter meters and there are 4 room in my room there are 4 rooms in my flat
yeah
and two one kitchen
I live in <unknown>
I live in a flat room
it’s about 20 it’s about 200 quarter meters and there are 4 room in my room
there are 4 rooms in my flat
yeah
and two one kitchen
Research challenge: dealing with repetitions

I live in <unknown>
I live in a flat room
it’s about 20 it’s about 200 quarter meters and there are 4 room in my room
there are 4 rooms in my flat
yeah
and two one kitchen
Research challenge: dealing with repetitions

I live in <unknown>
I live in a flat room
it’s about 20 it’s about 200 quarter meters and there are 4 room in my room
there are 4 rooms in my flat
yeah
and two one kitchen

I live in <unknown>
I live in a flat room
it’s about 200 quarter meters and there are 4 rooms in my flat
yeah
and one kitchen
Ongoing spoken language research programme

- State-of-the-art automated holistic score for the spoken language
- Models for feedback and error detection require large amounts of good training data—which we are actively accumulating
Ongoing spoken language research programme

Clarity feedback
- **Fluency and Coherence**: we can provide scores for speech rate
- **Pronunciation**: prototypes for feedback for vowel distinction and word stress

Acceptability feedback
- **Lexical Resource**: challenges in high accuracy speech recognition for learner speech
- **Grammatical Range and Accuracy**: challenges in segmentation, disfluency detection, ...
Incremental Dependency Parsing and Disfluency Detection in Spoken Learner English

Russell Moore, Andrew Caines(✉), Calbert Graham, and Paula Buttery

Automated Language Teaching and Assessment Institute,
Department of Theoretical and Applied Linguistics,
University of Cambridge, Cambridge, UK
{rjm49, apc38, crg29, pjb48v}@cam.ac.uk

Abstract. This paper investigates the suitability of state-of-the-art natural language processing (NLP) tools for parsing the spoken language of second language learners of English. The task of parsing spoken learner-language is important to the domains of automated language assessment (ALA) and computer-assisted language learning (CALL). Due to the non-canonical nature of spoken language (containing filled pauses, non-
Speech feedback needs to be multi-faceted

Consider some examples from BNC:

- Where’s that recipe book for that chocolate and banana cake? Chocolate and banana cake which book was it?
- Some of these chocolate cakes are absolutely mm, mm.

All are going to score low on any grammatical acceptability tests but are completely communicative.

- Can you pass me the she-du-lee?
- Are you happy with conTENT (as opposed to the CONtent)

These are grammatical but rather uncommunicative.

We need to start to think of errors as multi-faceted (work with Mike McCarthy — “Infinite shades of grey: What constitutes an error?”).
What contributes to an utterance’s clarity?

- lexical choice
- stress
- intonation
- pronunciation
Fig. 1 Mean values of F1 and F2 for L1 Gujarati (Basic English) speakers for target vowels at vowel midpoint (the 50% temporal position).

Fig. 2 Mean values of F1 and F2 for L1 Gujarati (advanced English) speakers for target vowels at vowel midpoint (the 50% temporal position).
Conclusions

- Automated holistic scoring for written and spoken learner language
- Automated scoring is indistinguishable from human scoring
- Effective scoring of sub-criteria for writing
- Detailed feedback for writing at 3 levels of granularity
- Our research programme develops detailed feedback for spoken language
Thank you!