Automated assessment of English learners' writing: leveraging computational linguistics and machine learning

Helen Yannakoudakis

University of Cambridge
Cambridge English Language Assessment

Joint work:
Ø. E. Andersen, F. Barker, T. Briscoe, T. Parish

Cambridge English

iLexIR
NLP Consultancy

Outline

Cambridge English

iLexIR
NLP Consultancy

The task: automated writing feedback

Automated writing feedback

Automatically evaluate the quality of writing and provide immediate feedback

Challenges

- Provide accurate, effective and detailed feedback
- Provide pedagogically useful feedback like human teachers

Deployment

Advantages

- Prompt detailed feedback
- Promote writing development
- Facilitate self-assessment and self-tutoring
- Application of constant assessment criteria
- Reduced workload
- Cost-effective approach to teaching / grading

Script-level feedback

Text assessment

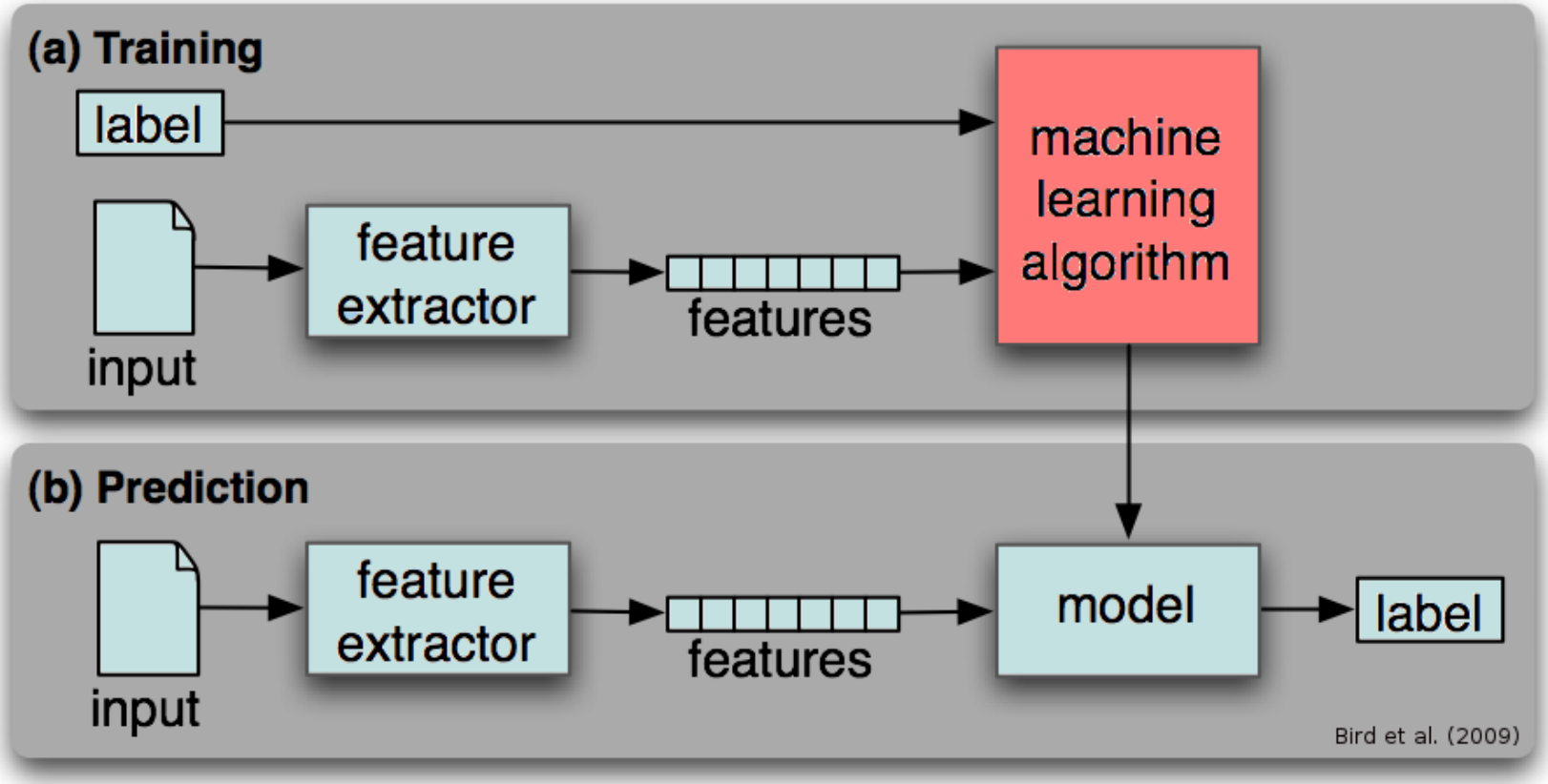
Overall assessment of someone's proficiency by scoring the text as a whole

1. Assess general linguistic competence
   a. Gather annotated data
   b. Identify textual features considered to be proxies for intrinsic qualities of writing competence
   c. Predict score using weighted combination of features (**Machine Learning**)
   d. Evaluate predicted scores
2. Provide scoring feedback

Cambridge English

iLexIR
NLP Consultancy

Script-level feedback: Machine Learning



Script-level feedback: Feature Space

1. Word sequences
   - *belive* (unigram)
   - *suggest idea* (bigram)
   - *the people is* (trigram)

Script-level feedback: Feature Space

1. Word sequences
   - *belive* (unigram)
   - *suggest idea* (bigram)
   - *the people is* (trigram)
2. Part-of-speech (PoS) sequences
   - VV0 VV0 (e.g., *keep develop*)
   - NN2 VVG (e.g., *children smiling*)

Script-level feedback: Feature Space

1. Word sequences
   - *belive* (unigram)
   - *suggest idea* (bigram)
   - *the people is* (trigram)
2. Part-of-speech (PoS) sequences
   - VV0 VV0 (e.g., *keep develop*)
   - NN2 VVG (e.g., *children smiling*)
3. Grammatical constructions
   - V1/modal_bse/+- (e.g., *can only travel in July*)
   - S/pp-ap_s-r (e.g., *for better or worse, he left*)
   - T/txt-frag (e.g., *but know Kim knew*)

Script-level feedback: Feature Space

1. Word sequences
   - *belive* (unigram)
   - *suggest idea* (bigram)
   - *the people is* (trigram)
2. Part-of-speech (PoS) sequences
   - VV0 VV0 (e.g., *keep develop*)
   - NN2 VVG (e.g., *children smiling*)
3. Grammatical constructions
   - V1/modal_bse/+- (e.g., *can only travel in July*)
   - S/pp-ap_s-r (e.g., *for better or worse, he left*)
   - T/txt-frag (e.g., *but know Kim knew*)
4. Error rate & other features (text length, complexity. . .)

Script-level feedback: Evaluation

Correlation between gold scores and the system-predicted scores

Features	**Pearson's correlation r**	**Spearman's correlation ρ**
word seq	0.601	0.598
+PoS seq	0.682	0.687
+text length	0.692	0.689
+syntax	0.714	0.712
+error rate	0.741	0.773
Upper bound	**0.796**	**0.792**

Script-level feedback: Evaluation

Comparison with previous work – Regression vs. Ranking

Model	**Pearson's correlation r**	**Spearman's correlation ρ**
Regression	0.720	0.750
Our model (ranking)	0.750	0.785

Script-level feedback

Overall score

An overall score is assigned on a scale from red to green. Red is for text that looks like it may be at CEFR level B1 or below. Green is for text that shows some evidence of being at CEFR level B2 or above. 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

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?

Word count: 31

Save | Save & Submit

Word-level feedback: error detection and correction

Error detection and correction

Ensure high precision and good coverage

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
2. Electronic dictionary-derived rules

Word-level feedback: error detection and correction

Error detection and correction

1. Strict criteria for rule extraction
2. Reliable rules
3. Few false positives
4. Precision and Recall measured against human annotator: 90% and 30% respectively
5. Precision is more important in terms of learning effect (Nagata and Nakatani, 2010)

Word-level feedback: error detection and correction

Trigrams	**Error**	**Correction**
he] want [to	AGV	wants
to] thanks [all	FV	thank
are] to [old	SX	too
's] interesting [place	MD	an+
is] need [to	MD	a+
Bigrams	**Error**	**Correction**
of] whole	MD	the+
This [why	MV	+is
few] absence	AGN	absences
listening] at	RT	to
Unigrams	**Error**	**Correction**
beloveds	C	beloved
disappointement	S	disappointment
singed	IV	sang

Cambridge English

iLexIR
NLP Consultancy

Word-level feedback: error detection and correction

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'?

Cambridge English

iLexIR
NLP Consultancy

Sentence-level feedback

Sentence evaluation

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

Challenges

- Limited linguistic evidence that can be extracted automatically
- Difficulty in acquiring annotated data

Sentence-level feedback

Previous work

- Content scoring of short answers, ranging from a few words to a few sentences
  (e.g., Attali et al., 2008; Mohler et al., 2011; Ziai et al., 2012)
- Intra-sentential quality (Higgins et al., 2004)
- Writing instruction tools (e.g., Burstein et al., 2003)

Sentence-level feedback

Approach

- Exploit already available annotated data
  - Script-level scores and error annotation in CLC

Sentence-level feedback

Approach

- Exploit already available annotated data
  - Script-level scores and error annotation in CLC
- Evaluate various approaches, two of which are to:

Sentence-level feedback

Approach

- Exploit already available annotated data
  - Script-level scores and error annotation in CLC
- Evaluate various approaches, two of which are to:
  1. Use the script-level model to predict sentence quality scores

Sentence-level feedback

Approach

- Exploit already available annotated data
  - Script-level scores and error annotation in CLC
- Evaluate various approaches, two of which are to:
  1. Use the script-level model to predict sentence quality scores
  2. Combine script-level score and errors per sentence, and create pseudo-gold labels to train a sentence model

Sentence-level feedback

	Model 1	Model 2
Correlation		
r_{errors}	−0.111	−0.750
ρ_{errors}	−0.078	−0.702
Pairwise acc		
Correct	0.608	0.703
Incorrect	0.359	0.204

Model 1: script-level model

Model 2: sentence-level model with pseudo-gold labels: $\frac{\text{score}}{\text{errors}}$

Sentence-level feedback

Model 2: sentence-level model with pseudo-gold labels: $\frac{\text{score}}{\text{errors}}$

Feature set

1. Main verbs, nouns, adjectives, subordinating conjunctions and adverbs
2. Clausal subjects and modifiers
3. Affixes
4. Phrase-structure rules
5. Errors
6. Number of words forming an error

Sentence-level feedback

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 electiity 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 liten what it happend in the world and you can watch some places that you didn't go.
You can enjoy you time to watch a film if you have free time.
In our daily life, however, we seldom notice how much convinient life we've got, what is more, how much inconvinient our grandparents had got.

Write & Improve (www.cambridgeenglish.org/writeandimprovebeta)

Feedback

Overall score

An overall score is assigned on a scale from red to green. Red is for text that looks like it may be at CEFR level B1 or below. Green is for text that shows some evidence of being at CEFR level B2 or above. 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)

Combined | Error feedback | Sentence feedback

Combined feedback allows you to see the information contained in the Sentence feedback and Error feedback together on one page. 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.

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?

Now, improve your answer

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?

Word count: 31

Save | Save & Submit

Trials

- Ten institutions from nine countries
- Eight universities, one secondary school and one private language school
- Between 4 and 8 institutions in each trial
- Each institution participated in two or three trials
- Over 450 students participated, expected to be at or above the upper-intermediate level

Trials

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

User satisfaction

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

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

- User-driven development between trials

Cambridge English

iLexIR
NLP Consultancy

Conclusion

- Feedback at three different levels of granularity
  - Script-level
  - Sentence-level
  - Word-level
- Visualisation displays information in an intuitive and easily interpretable way
- Usefulness and usability of the tool confirmed through questionnaire-based evaluations

Future work

- Improve methodologies used for providing error feedback
- Add further functionality
  - L1-specific feedback
  - Discourse organisation feedback
  - Task achievement feedback

Thank you!