Developing and testing a self-assessment and tutoring system

Øistein E. Andersen
Helen Yannakoudakis
Fiona Barker
Tim Parish

iLexIR
University of Cambridge

Building Educational Applications

NAACL 2013

iLexIR
NLP Consultancy

Outline

iLexIR
NLP Consultancy

The task: automated writing feedback

Automated writing feedback

Automatically evaluate the quality of writing and provide immediate feedback

Challenges

- Accurate and effective feedback
- Provide feedback in similar ways and as usefully as humans typically do

Deployment

Advantages

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

Feedback

Feedback types

- Direct (e.g., error correction)
- Indirect (e.g., underline)

Feedback focus

- Language (e.g., grammar, vocabulary)
- Content (e.g., ideas)

Feedback forms

- Marks/grades
- Diagnostic/corrective (e.g., error feedback)

Script-level feedback

Text assessment

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

1. Assess general linguistic competence
   - Linear rank preference perceptron (Medlock, 2009)
   - Features: lexical and syntactic, as well as errors (Yannakoudakis et al., 2011)
2. Provide scoring feedback

Script-level feedback

Dataset

- First Certificate in English (FCE) exam
- Upper-intermediate level assessment
- Free-text answers annotated with mark in the range 1–40

Evaluation

	r	ρ
Ranking SVMs[a]	0.741	0.773
Ranking perceptron	**0.740**	**0.765**
Upper-bound	0.796	0.792

[a] Yannakoudakis et al., 2011

iLexIR
NLP Consultancy

Script-level feedback

Overall score

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

Change answer

Some people learn a foreign language in order to widen their horizons and etc.

Perhaps you prefer to stay on dry land. Can u sea the see from were you live?

Word count: 31

Save

iLexIR
NLP Consultancy

Word-level feedback

Error detection and correction

Ensure high precision and good coverage

1. Corpus-derived rules (Andersen, 2011)
   - Error rules from the Cambridge Learner Corpus (CLC) (Nicholls, 2003)
   - Detect incorrect unigrams, bigrams and trigrams
   - At least 90% incorrect occurrences
2. Dictionary rules[1] (Andersen, 2011)

[1] Lexical Database developed by the Dutch Centre for Lexical Information (CELEX)

Word-level feedback

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 u sea the see from were you live?

Possible errors

and Insertion: This word may not actually be needed.

etc. Substitution: A different word might work better here. Have you contemplated using 'so on'?

u Register: The word you have chosen might be inappropriate for this writing task. If so, 'you' would be a more conventional choice.

sea Confusion: You may have confused this with a similar-looking word. 'see' seems more likely in this context.

the Insertion: This word may not actually be needed.

see Confusion: You may have confused this with a similar-looking word. 'sea' seems more likely in this context.

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

iLexIR
NLP Consultancy

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., Criterion, Burstein et al., 2003)

Sentence-level feedback

Approach

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

Sentence-level feedback

Approach

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

Sentence-level feedback

Approach

- Exploit already available annotated data
  - Script-level scores and error annotation in FCE
- 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 FCE
- 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
r_g	—	0.550
ρ_g	—	0.646
r_s	0.572	0.385
ρ_s	0.578	0.301
r_e	−0.111	−0.750
ρ_e	−0.078	−0.702
AP	0.393	0.747
Pairwise		
Correct	0.608	0.703
Incorrect	0.359	0.204

Model 1: script-level model

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

iLexIR
NLP Consultancy

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. Error counts[a]
6. Number of words forming an error

[a] Based on corpus and dictionary rules

iLexIR
NLP Consultancy

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.

Self-Assessment and Tutoring (SAT) system

Assessed answer

Overall score

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

Detailed feedback (Help)

Score feedback | Error feedback | Combined

This view combines the information contained in the score feedback and error feedback views. A red box indicates that explanations/corrections are available and can be viewed by hovering over the relevant word.

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

Can u sea the see from were you live?

Change answer

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

Can u sea the see from were you live?
```

Word count: 31

Save

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 total, 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

iLexIR
NLP Consultancy

User satisfaction

	Trial 1	Trial 2
Using the SAT system helps me to write better in English	3.80	3.92
I find the SAT system 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
The SAT system is easy to use	4.45	4.49
The feedback on my writing is clear	3.80	3.93
If you have used the SAT system 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)

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

iLexIR
NLP Consultancy

Future work

- Improve methodologies used for providing feedback
- Add further functionality
  - L1-specific feedback
  - Assessment of clauses and phrases

iLexIR
NLP Consultancy

Thank you!

Acknowledgments

Special thanks to Ted Briscoe and Marek Rei, as well as to the anonymous reviewers, for their valuable contributions at various stages.

iLexIR
NLP Consultancy

Previous work

Examples of existing writing assessment systems

- Criterion (Burstein et al., 2003)
- MY Access! (Elliot, 2003)
- Intelligent Essay Assessor (Landauer et al., 2003)
- ESL Assistant (Gamon et al., 2009)

Word-level feedback

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

iLexIR
NLP Consultancy

Number of revisions per task response

Revisions	Count
1	292
2	272
3	142
4	78
5	50
6	28
7	15
8	25
9	11
10	14
11–15	21
16–20	6
20–	5

iLexIR
NLP Consultancy

Number of words per submission

Words	Count
0– 99	540
100–199	1,294
200–299	928
300–399	201
400–499	67
500–999	26
1,000–	36

iLexIR
NLP Consultancy

Score evolution

Decrease from first to last revision: 12.9%
No change: 4.3%
Increase: 82.8%