A New Dataset and Method for Automatically Grading ESOL Texts

Yannakoudakis et al.

Alex Gamble
29.1.17

Paper Aims

- Present new corpus of ESOL texts.
- Use rank preference learning to automatically assess (AA) quality of scripts.

The Task

- ESOL learners produce essays in response to a given prompt.
- Unlike in assessment of native writers, semantic content of text is less relevant to marking criteria.
    - Accurate use of linguistic constructions awards marks.
    - Assumption made that by modelling documents as features of these linguistic constructions, grading methods can be ‘learned’.

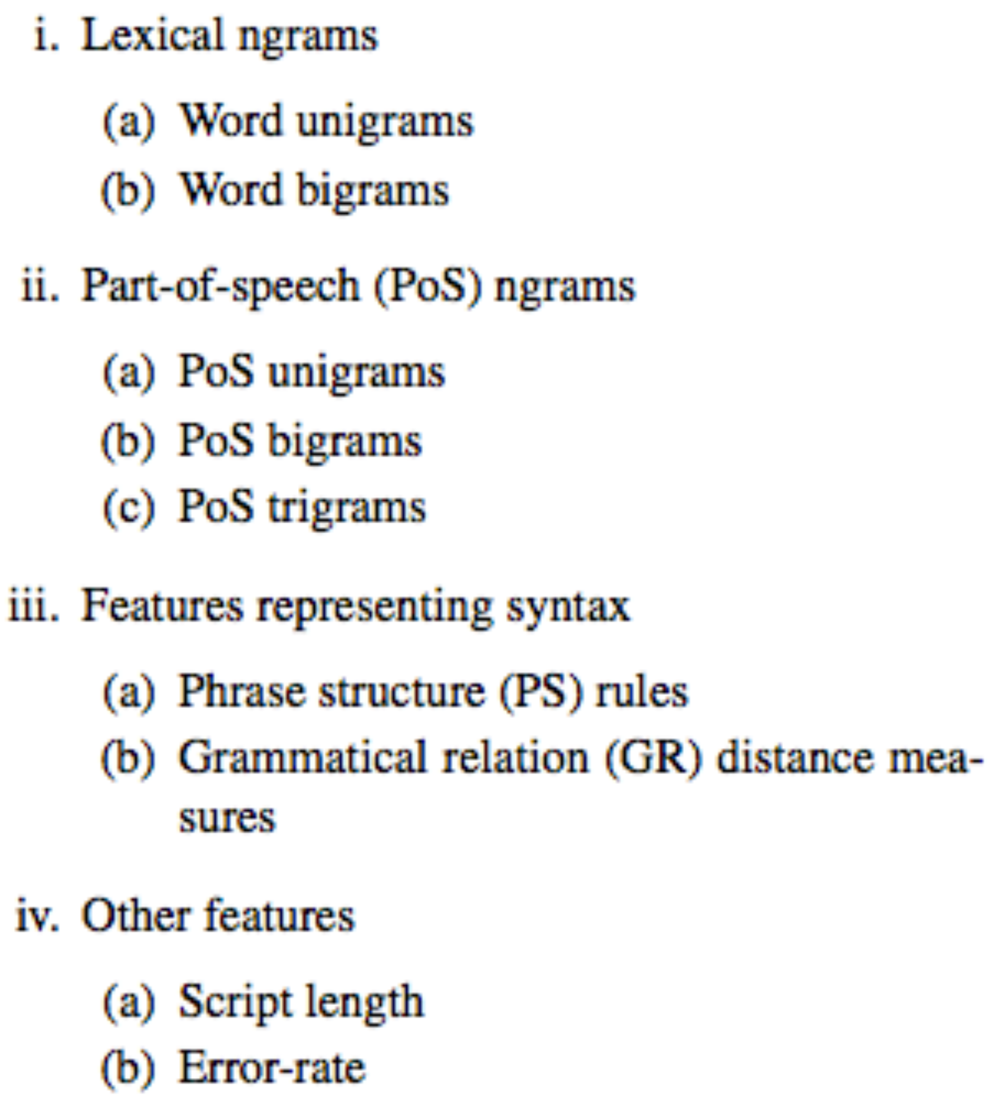
Cambridge Learner Corpus

- Collection of scripts taken from candidates sitting Cambridge ESOL examinations.
- 1,238 total scripts, 200-400 words.
- Annotated with scores 1-40, scaled using RASCH model.

Rank Preference Model

- Trained with rank-preference SVMs on pairwise difference vectors.
- Goal is to maximise the number of correctly ranked pairs.
- Learning to model the grade relationships between scripts.
    - No need for additional mapping of raw classifier output to scoring scale.

Feature Set



i. Lexical ngrams
   (a) Word unigrams
   (b) Word bigrams

ii. Part-of-speech (PoS) ngrams
   (a) PoS unigrams
   (b) PoS bigrams
   (c) PoS trigrams

iii. Features representing syntax
   (a) Phrase structure (PS) rules
   (b) Grammatical relation (GR) distance measures

iv. Other features
   (a) Script length
   (b) Error-rate

Evaluation Measures

- **Pearson’s Product-Moment Coefficient**
  Depicts linear relationships

- **Spearman’s Rank Coefficient**
  Depicts monotonic relationships

Features	Pearson's correlation	Spearman's correlation
word ngrams	0.601	0.598
+PoS ngrams	0.682	0.687
+script length	0.692	0.689
+PS rules	0.707	0.708
+complexity	0.714	0.712
Error-rate features		
+ukWaC LM	0.735	0.758
+CLC LM	0.741	0.773
+true CLC error-rate	0.751	0.789

Table 1: Correlation between the CLC scores and the AA system predicted values.

Ablated feature	Pearson's correlation	Spearman's correlation
none	0.741	0.773
word ngrams	0.713	0.762
PoS ngrams	0.724	0.737
script length	0.734	0.772
PS rules	0.712	0.731
complexity	0.738	0.760
ukWaC+CLC LM	0.714	0.712

Table 2: Ablation tests showing the correlation between the CLC and the AA system.

	CLC	**E1**	**E2**	**E3**	**E4**	**AA**
CLC	-	0.820	0.787	0.767	0.810	0.741
E1	0.820	-	0.851	0.845	0.878	0.721
E2	0.787	0.851	-	0.775	0.788	0.730
E3	0.767	0.845	0.775	-	0.779	0.747
E4	0.810	0.878	0.788	0.779	-	0.679
AA	0.741	0.721	0.730	0.747	0.679	-
Avg	0.785	0.823	0.786	0.782	0.786	0.723

Table 4: Pearson's correlation of the AA system predicted values with the CLC and the examiners' scores, where E1 refers to the first examiner, E2 to the second etc.

Validity Testing

Testing subversion to writers with knowledge of how the automated assessment system works.

i. Randomly order:

   (a) word unigrams within a sentence

   (b) word bigrams within a sentence

   (c) word trigrams within a sentence

   (d) sentences within a script

ii. Swap words that have the same PoS within a sentence

Modification	**Pearson's correlation**	**Spearman's correlation**
i(a)	0.960	0.912
i(b)	0.938	0.914
i(c)	0.801	0.867
i(d)	0.08	0.163
ii	0.634	0.761

Table 6: Correlation between the predicted values and the examiner's scores on 'outlier' texts.

Positives

- Good treatment of task
  - Evaluation against previous discriminative techniques
  - Validity testing against subversion
  - Discussion of outlier texts
  - Ablation testing useful for weighting feature importance.
- Dataset
  - Novel dataset of ESOL texts
  - Clear areas of further research using this dataset.

Criticism

- Grading Scheme
  - Further discussion of how features chosen were motivated from grading scheme would have been useful.
  - More complex grading criteria such as discourse cohesion and relevance to the given prompt were not considered.
  - Validity testing doesn’t consider these areas.