Modeling coherence in ESOL learner texts

Helen Yannakoudakis & Ted Briscoe

University of Cambridge
Computer Lab

Building Educational Applications

NAACL 2012
Outline

1. Introduction
2. Dataset
3. System
4. Conclusions
The Task: Automated Text Scoring (ATS)

Automated Text Scoring (ATS)
Automatically analyse the quality of writing competence and assign a score to a text

Goal
Evaluate writing as reliably as human readers

Challenges
Imitate the value judgements that human readers make when they mark a text
The Task: Automated Text Scoring (ATS)

ATS systems

Marking criteria

- Identify textual features that correlate with intrinsic features of human judgments
- Multiple factors influence the linguistic quality of texts
- Grammar, style, vocabulary usage, topic similarity, discourse coherence and cohesion, etc.
The Task: Automated Text Scoring (ATS)

ATS systems

Discourse coherence & cohesion

Marking criteria

- Identify textual features that correlate with intrinsic features of human judgments
- Multiple factors influence the linguistic quality of texts
- Grammar, style, vocabulary usage, topic similarity, **discourse coherence and cohesion**, etc.
Discourse coherence & cohesion

Mechanisms

- Cohesion: use of cohesive devices that can signal primarily suprasentential discourse relations between textual units (Halliday and Hasan, 1976)
  - Anaphora, discourse markers, etc.
- Local coherence: transitions between textual units
- Global coherence: sequence of topics
Related work

Coherence analysis on:
- News texts
  - e.g., Lin et al. (2011), Elsner and Charniak (2008), Soricut and Marcu (2006), etc.
- Extractive summaries
  - Pitler et al. (2010)
- Learner data
FCE Writing Component

- Upper-intermediate level assessment
- Two tasks eliciting free-text answers, each one between 120 and 180 words
  - e.g. ‘write a short story commencing …’
- Answers annotated with mark (in the range 1–40), fitted to a RASCH model (Fischer and Molenaar, 1995)
- Manually error-coded using a taxonomy of ~80 error types (Nicholls, 2003)
Baseline system

- ATS system described in Yannakoudakis et al. (2011)
- Features focus on lexical and grammatical properties, as well as errors
- Discourse coherence ignored
- Vulnerable to subversion
- Extend with discourse coherence features
Machine Learning

Ranking SVMs

- Address ATS as a ranking learning problem (Joachims, 2002)
- Learn an optimal ranking function that explicitly models the grade relationships between scripts
- Model the fact that some scripts are better than others
‘Superficial’ proxies

- Number of pronouns
‘Superficial’ proxies

- Number of pronouns
- Number of discourse connectives
  - Addition (e.g., additionally)
  - Comparison (e.g., likewise)
  - Contrast (e.g., whereas)
  - Conclusion (e.g., therefore)
‘Superficial’ proxies

- Number of pronouns
- Number of discourse connectives
  - Addition (e.g., additionally)
  - Comparison (e.g., likewise)
  - Contrast (e.g., whereas)
  - Conclusion (e.g., therefore)
- Word length
Models

Lemma/PoS cosine similarity

- Incorporate (syntactic) aspects of text coherence
- Represent sentences using vectors of lemma/PoS-tag counts
- Cosine similarity between adjacent sentences:

\[
\cos(\theta) = \frac{\vec{s}_i \cdot \vec{s}_{i+1}}{\| \vec{s}_i \| \| \vec{s}_{i+1} \|}
\]  

- Coherence of a text \( T \):

\[
\text{coherence}(T) = \frac{\sum_{i=1}^{n-1} \text{sim}(s_i, s_{i+1})}{n - 1}
\]
Incremental Semantic Analysis (ISA)

- Word space model (Baroni et al., 2007)
- Fully-incremental variation of Random Indexing (Sahlgren, 2005)
- Similarity among words measured by comparing their context vectors
- Coherence of a text $T$:

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

- Underlying idea: the degree of semantic relatedness between adjoining sentences serves as a proxy for local discourse coherence
IBM model 1

- **Machine translation:** the use of certain words in a source language is likely to trigger the use of certain words in a target language.

- **In texts:** the use of certain words in a sentence tends to trigger the use of certain words in an adjoining sentence (Soricut and Marcu, 2006).

- **Identification of word co-occurrence patterns across adjacent sentences.**

- **Probability of a text** $T$:

\[
P_{IBM_{dir}}(T) = \prod_{i=1}^{n-1} \prod_{j=1}^{|s_{i+1}|} \frac{\varepsilon}{|s_i| + 1} \sum_{k=0}^{|s_i|} t(s_{i+1}^j|s_k^i)}
\]
IBM model 1

- Machine translation: the use of certain words in a source language is likely to trigger the use of certain words in a target language.
- In texts: the use of certain words in a sentence tends to trigger the use of certain words in an adjoining sentence (Soricut and Marcu, 2006).
- Identification of word co-occurrence patterns across adjacent sentences.
- Probability of a text $T$:

$$P_{IBM_{dir}}(T) = \prod_{i=1}^{n-1} \prod_{j=1}^{\left|s_{i+1}\right|} \frac{\varepsilon}{\left|s_i\right| + 1} \sum_{k=0}^{\left|s_i\right|} t(s_{i+1}^j | s_i^k)$$

Models

Entity-based coherence model

- Measures local coherence on the basis of sequences of entity mentions (Barzilay and Lapata, 2008)
- Learns coherence properties similar to those employed by Centering Theory (Grosz et al., 1995)
- Each text is represented by an entity grid that captures the distribution of discourse entities across sentences

```
LANGUAGE       -   -   X   -   -   -   -   -
COUNTRY        -   -   X   -   -   -   -   -
POINTS         -   -   X   -   -   -   -   -
CHILDREN       -   -   0   X   -   -   -   -
TV             -   -   X   -   -   -   -   -
PROGRAMMES     -   -   S   0   -   -   -   -
1 2 3 4 5 6 7 8
```
Models

Pronoun coreference model

- Unsupervised generative model (Charniak and Elsner, 2009)
- Model each pronoun as generated by an antecedent somewhere in the previous 2 sentences
- Probability of a text: probability of the resulting sequence of pronoun assignments
Models

Discourse-new model

- Discourse-new classifier (Elsner and Charniak, 2008)
- Distinguish NPs whose referents have not been previously mentioned in the discourse from those that have
- Probability of a text: \( \prod_{np:NPs} P(L_{np}|np) \)
Models

Bag-of-Words (BOW)

- Represent a text as a vector \( d \in \mathbb{R}^V \)
- Each word \( v_i \) is associated with a single vector dimension
- Histogram of word occurrences
- Unable to maintain any sequential information
- Unable to capture the semantic transition between different parts of the document
- Partial solution: use \( n \)-grams
Models

Locally-Weighted Bag-of-Words (LoWBOW)

- LoWBOW: sequentially-sensitive alternative to BOW (Lebanon et al., 2007)
- A text is represented by a set of local histograms computed across the whole text, but centered on different locations
- Preserves local contextual information by modeling the text sequential structure
Locally-Weighted Bag-of-Words (LoWBOW) – cont.

Helen Yannakoudakis & Ted Briscoe

Modeling coherence in ESOL learner texts
### Results – examination year 2000

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>(r)</th>
<th>(\rho)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Baseline</td>
<td>0.651</td>
<td>0.670</td>
</tr>
<tr>
<td>1</td>
<td>POS distr.</td>
<td>0.653</td>
<td>0.670</td>
</tr>
<tr>
<td>2</td>
<td>Disc. connectives</td>
<td>0.648</td>
<td>0.668</td>
</tr>
<tr>
<td>3</td>
<td>Word length</td>
<td>\textbf{0.667}</td>
<td>\textbf{0.676}</td>
</tr>
<tr>
<td>4</td>
<td>ISA</td>
<td>\textbf{0.675}</td>
<td>\textbf{0.678}</td>
</tr>
<tr>
<td>5</td>
<td>EGrid</td>
<td>0.650</td>
<td>0.668</td>
</tr>
<tr>
<td>6</td>
<td>Pronoun</td>
<td>0.650</td>
<td>0.668</td>
</tr>
<tr>
<td>7</td>
<td>Disc-new</td>
<td>0.646</td>
<td>0.662</td>
</tr>
<tr>
<td>8</td>
<td>LoWBOW\textsubscript{lex}</td>
<td>\textbf{0.663}</td>
<td>\textbf{0.677}</td>
</tr>
<tr>
<td>9</td>
<td>LoWBOW\textsubscript{POS}</td>
<td>0.659</td>
<td>0.674</td>
</tr>
<tr>
<td>10</td>
<td>IBM model\textsubscript{lex\textsubscript{f}}</td>
<td>0.649</td>
<td>0.668</td>
</tr>
<tr>
<td>11</td>
<td>IBM model\textsubscript{lex\textsubscript{b}}</td>
<td>0.649</td>
<td>0.667</td>
</tr>
<tr>
<td>12</td>
<td>IBM model\textsubscript{POS\textsubscript{f}}</td>
<td>\textbf{0.661}</td>
<td>\textbf{0.672}</td>
</tr>
<tr>
<td>13</td>
<td>IBM model\textsubscript{POS\textsubscript{b}}</td>
<td>0.658</td>
<td>0.669</td>
</tr>
<tr>
<td>14</td>
<td>Lemma cosine</td>
<td>0.651</td>
<td>0.667</td>
</tr>
<tr>
<td>15</td>
<td>POS cosine</td>
<td>0.650</td>
<td>0.665</td>
</tr>
<tr>
<td>16</td>
<td>5+6+7+10+11</td>
<td>0.648</td>
<td>0.665</td>
</tr>
<tr>
<td>17</td>
<td>All</td>
<td>0.677</td>
<td>0.671</td>
</tr>
</tbody>
</table>

**Table:** 5-fold cross-validation performance on texts from year 2000 when adding different coherence features on top of the baseline AA system.
Results – examination year 2001 & outlier scripts

<table>
<thead>
<tr>
<th></th>
<th>$r$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.741</td>
<td>0.773</td>
</tr>
<tr>
<td>ISA</td>
<td>0.749</td>
<td>0.790*</td>
</tr>
<tr>
<td>Upper-bound</td>
<td>0.796</td>
<td>0.792</td>
</tr>
</tbody>
</table>

*Table: Performance on the exam scripts drawn from the examination year 2001. * indicates a significant improvement at $\alpha = 0.05$.

<table>
<thead>
<tr>
<th></th>
<th>$r$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.08</td>
<td>0.163</td>
</tr>
<tr>
<td>ISA</td>
<td>0.400</td>
<td>0.626</td>
</tr>
</tbody>
</table>

*Table: Performance of the ISA AA model on outliers.*
Conclusions & Future Work

- First systematic analysis of different models for assessing discourse coherence on learner data
- Significant improvement over Yannakoudakis et al. (2011)
- ISA, LOWBOW, the POS IBM model and word length are the best individual features
- Local histograms are useful
- Results specific to ESOL FCE texts
- Investigate a wider range of (learner) texts and further coherence models (e.g., Elsner and Charniak (2011a) and Lin et al. (2011)).
Thank you!

**Acknowledgments:** we are grateful to Cambridge ESOL for supporting this research. We would like to thank Marek Rei, Øistein Andersen as well as the anonymous reviewers for their valuable comments and suggestions.