Automatic Paragraph Segmentation with Lexical and Prosodic Features
Catherine Lai, Mireia Farrús and Johanna D. Moore
The Centre for Speech Technology Research, School of Informatics, University of Edinburgh
TAH Research Group, DTIC, Universitat Pompeu Fabra, Barcelona, Spain
cbh@inf.ed.ac.uk; mireia.farrus@upf.edu; j.moore@ed.ac.uk

Introduction
- **Aim**: Recognize spoken document discourse structure that is finer grained than broad topic or story:
  → Paragraph segmentation
- **Applications**:
  → Browsing long audio/video
  → Summarization
  → Improve readability of transcripts
- **Q**: Do features that help high level topic segmentation work at the paragraph level?
  → Lexical coherence/similarity
  → Prosodic features
  → Cue words/discourse markers
- **Q**: How do we combine features?
  → Do we treat lexical and prosodic features as independent?

Data: TED Talks
- 1365 talks, 1156 speakers of English
- Large variance in content, style, accent
- Manual transcription including punctuation and paragraph breaks
- 151829 sentences, 20953 paragraphs, average 7 sentences per paragraph
- Word → sentence → paragraphs timings from Viterbi forced alignment

Features
- **w123**: 1st, 2nd, 3rd word indicators
- **cw**: Knott (1996) cue word at start, middle, end?
- **lm**: average word entropy, sentence probability
- **syntax**: t-phi+ase top level children, branching factor, tree depth, cw
- **pos**: part-of-speech tag counts
- **bow**: bag of words indicators
- **cw**:
  - w123, cw
- **surface**:
  - #words, relative position, final punctuation, cue (previous, target, incomplete), bow, w123
- **durs**:
  - FO, intensity: mean, std, max, min, slope, range x (target, prev, next), (prev, next) drifs, dur
- **lex coh**:
  - LDA, LSA, TKIDF based cosine similarity, TextTiling discourse, lexical chain similarity scores
- **lex base**:
  - pos, surface, lm, syntax → Speller & Lapata (2006)
- **lex all**:
  - lex.coh, lex.base

AdaBoost vs SVM vs BLSTM (K-k)

BLSTM: Pk, WD, K-k
- **Classifier**: Pk, WD, K-k
- **lm**: 0.40 0.40 0.04
- **syntax**: 0.37 0.39 0.11
- **surface-cw-bow**: 0.37 0.39 0.12
- **pos**: 0.36 0.38 0.13
- **bow**: 0.36 0.38 0.17
- **cw**: 0.35 0.37 0.17
- **surface**: 0.33 0.36 0.24
- **durs**: 0.38 0.39 0.13
- **prosody**: 0.34 0.36 0.21
- **lex coh**: 0.38 0.39 0.10
- **lex base**: 0.32 0.35 0.25
- **lex all**: 0.31 0.34 0.28
- **cw+bow**: 0.34 0.37 0.21
- **cw+prosody**: 0.31 0.34 0.28
- **lex all+prosody**: 0.30 0.33 0.31

Task & Metrics
- **Task**: Predict paragraph final sentences
- **Evaluation**: 10-fold cross-validation: 80/10/10
- **Standard Metrics**: for window size k=3 over sequence, calculate penalty average:
  - Pk: penalty 1 for any false positive in window
  - WD: penalty 1 for any mismatch in for window
  - Corrected metric: K-k: chance corrected Pk

Unsupervised Baselines

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Pk</th>
<th>WD</th>
<th>K-k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority</td>
<td>0.41</td>
<td>0.41</td>
<td>0.00</td>
</tr>
<tr>
<td>Random</td>
<td>0.45</td>
<td>0.50</td>
<td>0.03</td>
</tr>
<tr>
<td>TextTiling+lex.chain</td>
<td>0.44</td>
<td>0.45</td>
<td>0.01</td>
</tr>
<tr>
<td>TextTiling+bow</td>
<td>0.44</td>
<td>0.45</td>
<td>0.07</td>
</tr>
<tr>
<td>TextTiling+prosody</td>
<td>0.44</td>
<td>0.46</td>
<td>0.10</td>
</tr>
<tr>
<td>BayesSeg</td>
<td>0.39</td>
<td>0.47</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Feature Fusion
- **Model**: Pk, WD, K-k
- **blstm:decision**: 0.30 0.33 0.31
- **blstm:feature**: 0.30 0.33 0.31
- **blstm:intermediate**: 0.30 0.32 0.32

Conclusions
- Cue word and prosodic features are better indicators of paragraph structure than topical coherence measures
  → Subtler transitions marked by discourse structural linguistic features
- Combination of multiple weak lexical and prosodic signals is necessary
  → Don’t treat lexical features and prosody as independent signals
- Allow low-level interactions and composition across time
- **Future work**: Deeper models, interaction with rhetorical/hierarchical structure, investigate topline human agreement