Coherence in Text

Coherence:
- is a property of well-written texts;
- makes them easier to read and understand;
- ensures that sentences are meaningfully related;
- and that the reader can work out what the linguistic expressions mean.

A coherent text is
- thematically organized;
- temporally organized;
- rather than a random concatenation of sentences.

Automatic models of coherence

Two uses for models of coherence:
- **Discourse segmentation**: Detecting breaks in text where coherence is relatively low
  - Useful in text summarisation, information retrieval, hypertext display...
- **Judgement** of textual coherence:
  - Can help rank the quality of alternative potential output texts/summaries in NLG/summarisation
  - Automatic grading of student essays
Introduction

Text Tiling

Other topic segmentation algorithms

Evaluation of Topic Segmentation

Entity-Based Coherence

Topic Segmentation: The task

- Segment text into non-hierarchical, non-overlapping zones which contain the same subtopic
- Equivalent definition: Detect subtopic shifts (changes of subtopic)
- Can’t we simply use paragraph or section boundaries?
  - Stark (1988) found not all paragraph boundaries reflect topic shifts
  - Paragraph conventions genre-dependent
  - Sections often too large

Factors for Detecting Topic Shifts

Linguistic factors:
- Adverbial clauses, prosodic markers (Brown and Yule)
- Cue phrases (Passonneau and Litman, Beeferman et al., Manning), e.g. oh, well, so, however, . . .
- Pronoun resolution
- Tense and aspect (Webber)

Lexical (co-occurrence) patterns:
- Word overlap or lexical chain overlap between sentences (Skorochod’ko 1979; Hearst 1994, 1997)
- New vocabulary terms (Youmans, 1991)
- Maximise density in dotplots (Reynar, 1994, 1998; Choi, 2000)
- Probabilistic model (Beeferman, Berger, Lafferty, 1999)

Star Gazer Text Structure

<table>
<thead>
<tr>
<th>Para</th>
<th>Subtopics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-3</td>
<td>Intro – the search for life in space</td>
</tr>
<tr>
<td>4-5</td>
<td>The moon’s chemical composition</td>
</tr>
<tr>
<td>6-8</td>
<td>How early earth-moon proximity shaped the moon</td>
</tr>
<tr>
<td>9-12</td>
<td>How the moon helped life evolve on earth</td>
</tr>
<tr>
<td>13</td>
<td>Improbability of the earth-moon system</td>
</tr>
<tr>
<td>14-16</td>
<td>Binary/trinary star systems make life unlikely</td>
</tr>
<tr>
<td>17-18</td>
<td>The low probability of non-binary/trinary systems</td>
</tr>
<tr>
<td>19-20</td>
<td>Properties of earth’s sun that facilitate life</td>
</tr>
<tr>
<td>21</td>
<td>Summary</td>
</tr>
</tbody>
</table>

Term repetition signals topic shift/cohesion
Example Text: “The history of algebra”

1. Algebra provides a generalization of arithmetic by using symbols, usually letters, to represent numbers. For example, it is obviously
2. In about 1100, the Persian mathematician Omar Khayyam wrote a treatise...
3. Boolean algebra is the algebra of sets and of logic. It uses symbols to represent logical statements instead of words. Boolean algebra was formulated by the English mathematician George Boole in 1847. Logic had previously been largely the province of philosophers, but in his book, The Mathematical Analysis of Logic, Boole reduced the whole of classical, Aristotelian logic to a set of algebraic equations. Boole’s original notation is no longer used, and modern Boolean algebra now uses the symbols of either set theory, or propositional calculus.
4. Boolean algebra is an uninterpreted system – it consists of rules for manipulating symbols, but does not specify how the symbols should be interpreted. The symbols can be taken to represent sets and their relationships, in which case we obtain a Boolean algebra of sets. Alternatively, the symbols can be interpreted in terms of logical propositions, or statements, their connectives, and their truth values. This means that Boolean algebra has exactly the same structure as propositional calculus.

TextTiling: The algorithm

Preprocessing: separate texts into pseudo-sentences w tokens long
- Score cohesion b/w pseudo-sentences
- Compare several metrics:
  - Word overlap
  - Vocabulary introduction
  - Vector space distance (not in CL article)
- Find local minima in plot of neighbouring pseudo-sentences scores (“depth scoring”)
- Project boundary onto nearest paragraph boundary

Segment from 51 to 66 about “Boole” and “logic”
- Segment from 67 to 81 about “gates”, “computers” and “Boole”
- Initial segments more general (“century”, “mathematics”)
**TextTiling Algorithm: Shifting window**

- Pseudo-sentences consist of \( w \) tokens (including stop words). Typical \( w = 20 \)
- Blocks consist of \( k \) pseudo-sentences (blocks should approx. paragraphs; often \( k = 6-10 \), but \( k = 2 \) in example)
- Sliding window of 2 blocks
- Compute and plot one or more scores at break between blocks
  - \( 2kw \) tokens are compared at a time
  - Blocks shift one pseudo-sentence at a time
  - You get as many data points as there are pseudo-sentences
  - Each pseudo-sentence occurs in \( 2k \) calculations
- Create two vectors from each block; use non-stoplist-tokens (stemmed)

**Score:** non-normalized inner product of frequencies \( w_{j,b} \) of terms \( t_j \) in left and right term vector \( b_1 = t_{i-k}, \ldots, t_i \) and \( b_2 = t_{i+1}, \ldots, t_{i+k+1} \)

\[
\text{score}(i) = \sum_{j=0}^{\mid T \mid} w_{j,b_1}w_{j,b_2}
\]

(\( T \): set of all tokens)

**TextTiling: Minimal block similarity signals boundary**

**TextTiling: Max. in new vocab. items signals boundary**

**TextTiling: Relative Depth**

- Use relative, not absolute, depth score:
  \[
  \text{Depth}(g_i) = |s_{i-1} - s_i| + |s_{i+1} - s_i| \quad \text{(with } s_{i-1} \text{ and } s_{i+1}\text{ surrounding local maxima; cf. Text 1)}
  \]
Cohesion is relative

- Introductions have many topic shifts → want only strong shifts
- Mid-portion with only minor topic shifts → want also weaker shifts
- Additional low pass filter (Text 2): $\frac{s_{i-1} + s_i + s_{i+1}}{3}$ (because $s_1 - s_2$ should contribute to score at $g_4$)

TextTiling: Boundary determination

- Sort depth scores, determine boundaries:
  - Boundary if $Depth > \mu - \sigma$ (low cutoff; liberal)
  - Boundary only if $Depth > \mu - \frac{\sigma}{2}$ (high cutoff; high P, low R)
- For each gap, assign closest paragraph boundary
- Do not assign close adjacent segment boundaries; 3 pseudosentences must intervene, to avoid sequence of small segments:

TextTiling: Output of depth scorer on “Stargazer” text


Use Church’s (1993) dotplot method (e.g. on the following three concatenated WSJ articles):

- If a word appears both in word positions x and y, then plot $(x,x), (x,y), (y,y), (y,x)$ → the diagonal is always dark
- Dark squares along diagonal indicate regions with many shared words

- Maximise density of regions within squares along the diagonal:
- Density $D = \frac{N}{x^2}$
- $x$: length of a square (in words); $N$: number of points in square
- Use divisive clustering to insert boundaries

Hierarchical clustering: divisive (TopDown) clustering

Given: a set $X = x_1, \ldots, x_n$ of objects;
Given: a function $coh : \mathcal{P} \rightarrow \mathbb{R}$
Given: a function $split : \mathcal{P}(X) \times \mathcal{P}(X)$

```
C := \{X\}(= \{c_1\})
j := 1
while \exists c_i \in C s.t. |c_i| > 1 do
  c_u := \arg\min_{c_v \in C} coh(c_v)
  (c_{j+1}, c_{j+2}) = split(c_u)
  C := C \{c_u\} \cup \{c_{j+1}, c_{j+2}\}
  j := j + 2
end
```

This is a greedy algorithm!

A probabilistic model for topic segmentation

Linear combination of two models:
- A short-range model: trigram language model $P_{trig}(w_2|w_0w_1)$
- A long-range model:
  - Determine trigger pairs $(s,t)$ with high Mutual Information (59,936 pairs found)
  - If $s$ has occurred within the past 500 words, then the probability of $t$ is boosted by a certain factor ($e^{\lambda(s,t)}$, which is estimated by a method called iterative scaling).

Examples of trigger pair boosting (long-range LM)

<table>
<thead>
<tr>
<th>s</th>
<th>t</th>
<th>$e^{\lambda(s,t)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>residues</td>
<td>carcinogens</td>
<td>2.3</td>
</tr>
<tr>
<td>Charleston</td>
<td>shipyards</td>
<td>4.0</td>
</tr>
<tr>
<td>microscopic</td>
<td>cuticle</td>
<td>4.1</td>
</tr>
<tr>
<td>defense</td>
<td>tax</td>
<td>8.4</td>
</tr>
<tr>
<td>tax</td>
<td>Ankara</td>
<td>10.5</td>
</tr>
<tr>
<td>Kurds</td>
<td>Vladimir</td>
<td>14.8</td>
</tr>
<tr>
<td>Steve</td>
<td>Gennady</td>
<td>19.6</td>
</tr>
<tr>
<td>education</td>
<td>tax</td>
<td>20.7</td>
</tr>
<tr>
<td>insurance</td>
<td>education</td>
<td>22.2</td>
</tr>
<tr>
<td>Pulitzer</td>
<td>insurance</td>
<td>23.0</td>
</tr>
<tr>
<td>Yeltsin</td>
<td>Pulitzer</td>
<td>23.6</td>
</tr>
<tr>
<td>sauce</td>
<td>Yeltsin</td>
<td>23.7</td>
</tr>
<tr>
<td>flower</td>
<td>sauce</td>
<td>27.1</td>
</tr>
<tr>
<td>petals</td>
<td>flower</td>
<td>32.2</td>
</tr>
<tr>
<td>scab</td>
<td>petals</td>
<td>103.1</td>
</tr>
</tbody>
</table>
Lexical Chains, Revisited

Consider these two summaries.

**Summary A**
Britain said he did not have diplomatic immunity. The Spanish authorities contend that Pinochet may have committed crimes against Spanish citizens in Chile. Baltasar Garzon filed a request on Wednesday. Chile said, President Fidel Castro said Sunday he disagreed with the arrest in London.

**Summary B**
Former Chilean dictator Augusto Pinochet was arrested in London on 14 October 1998. Pinochet, 82, was recovering from surgery. The arrest was in response to an extradition warrant served by a Spanish judge. Pinochet was charged with murdering thousands, including many Spaniards. Pinochet is awaiting a hearing, his fate in the balance. American scholars applauded the arrest.

Reminder: Lexical Chains

- Sequence of related words in text, spanning short (adjacent sentences) or longer distances (entire text)
- Originally due to Halliday and Hasan (1976)
  - Allowed lexical relations: identity, synonymy, hyponymy, siblings
- **Claim (here):** they capture (some of) the cohesive structure of the text
- (This is on top of the old claim that they provide the right context for WSD – which we know from session 2)
Texttiling with Lexical Chains

- Hearst (1997) also considers lexical chains as a scoring method in the sliding window method (TextTiling).
- High number of lexical chains spanning over a window gap should indicate high coherence.
- However, results are disappointing.

Evaluation of Topic Segmentation

Defining a gold standard
- Hearst (1997): “group opinion” amongst human annotators (3 out of 7)
- 12 magazine articles
- Humans find boundaries at 39% of “allowed” places (paragraph boundaries only)
- Baseline: randomly assign 39% of boundaries

Evaluation: precision and recall

- Measure precision and recall, in comparison to group opinion
- Precision tells us about false positives, recall about false negatives

<table>
<thead>
<tr>
<th></th>
<th>Tiling (VocabIntro)</th>
<th>Tiling (Lexical)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High cutoff</td>
<td>P=.58, R=.64</td>
<td>P=.71, R=.59</td>
</tr>
<tr>
<td>Low cutoff</td>
<td>P=.52, R=.78</td>
<td>P=.66, R=.75</td>
</tr>
<tr>
<td>Judges</td>
<td>P=.83, R=.71</td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>P=.50, R=.51</td>
<td></td>
</tr>
</tbody>
</table>

Evaluation by detecting document boundaries

- Create pseudo document by gluing unrelated documents together; measure how well the original document boundaries are found.
- This evaluation method violates a major assumption of the task:
  - It assumes article boundaries are by definition stronger shifts than within-article subtopic shifts
  - Algorithms is penalized for finding within-article subtopic shifts
- Evaluation of TextTiling on 44 WSJ articles glued together:
Evaluation Metrics for Topic Segmentation

- But: precision and recall are insensitive to near misses
- Improvement: $P_k$ measure (Beeferman et al. 1999)
  - Compute penalties via two probes (moving windows)
  - One probe checks gold standard, the other checks system text
  - Probe width $k$: half the avg. segment size (here $k=4$)
  - If both probes have their ends in the same segment or if both probes have their ends in different ones, score 1 point
  - Divide by number of measurements taken; $P_k$ is in [0..1]

Win$_{diff}$ – Improvement over $P_k$

- Win$_{diff}$ (improvement over $P_k$; Prevner and Hearst 2002): the ends of the two probes must cover the same number of boundaries (0 or more) to score a point.

Entity-based Coherence: Barzilay and Lapata 2005

- Coherence as a model of sequences of entity types in text
- Assume we know whether two linguistic expression co-refer, i.e., talk about the same entity (more about this in session 7)
- Observations from discourse theory:
  - The way entities are introduced and discussed influences coherence (Grosz et al 1995).
  - Salience of entities is related to where in the sentence they occur (Sidner, 1992).
  - Frequency, syntactic position, pronominalisation are relevant coherence properties.

The Entity Grid

1. Former Chilean dictator Augusto Pinochet, was arrested in London on 14 October 1998.
2. Pinochet, 82, was recovering from surgery.
3. The arrest was in response to an extradition warrant served by a Spanish judge.
4. Pinochet was charged with murdering thousands, including many Spaniards.
5. He is awaiting a hearing, his fate in the balance.
6. American scholars applauded the arrest.
Former Chilean dictator Augusto Pinochet, was arrested in London on 14 October 1998. Pinochet, 82, was recovering from surgery. The arrest was in response to an extradition warrant served by a Spanish judge. Pinochet was charged with murdering thousands, including many Spaniards. Pinochet is awaiting a hearing, his fate in the balance. American scholars applauded the arrest.

Columns: entities; lines: sentences
Entity Transitions

Example (transitions of length 2)

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>O</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>d_1</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>d_2</td>
<td>S</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>d_3</td>
<td>O</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>O</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>d_1</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
</tr>
<tr>
<td>d_2</td>
<td>0</td>
<td>0</td>
<td>0.07</td>
</tr>
<tr>
<td>d_3</td>
<td>0.02</td>
<td>0</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Learning a Ranking Function

Training Set

Ordered pairs \((x_{ij}, x_{ik})\), where \(x_{ij}\) and \(x_{ik}\) represent the same document \(d_i\), and \(x_{ij}\) is more coherent than \(x_{ik}\) (assume \(j > k\)).

- Source document and permutations of its sentences.
- Original order assumed coherent.
- Given \(k\) documents, with \(n\) permutations, obtain \(k \cdot n\) pairwise rankings for training and testing.
- Two corpora, Earthquakes and Accidents, 100 texts each.

Linguistic Dimensions

Coreference: Talking about the same entity
- Entities are coreferent if they have (roughly) the same surface form.
- Coreference resolution systems exist (cf. session 7)

Syntax: Does syntactic knowledge matter?
- Use four categories \(\{S, O, X, –\}\).
- Or just two \(\{X, –\}\).

Salience: Are some entities more important than others?
- Discriminate between salient (frequent) entities and the rest.
- Collect statistics separately for each group.

Goal

Find a parameter vector \(\vec{w}\) such that:

\[
\vec{w} \cdot (\Phi(x_{ij}) - \Phi(x_{ik})) > 0 \quad \forall j, i, k \text{ such that } j > k
\]

\(\vec{w}\Phi(x_{ij})\) is a ranking score, such that the violations of pairwise rankings in the training set are minimised.

Use Support Vector Machines (SVMs, Joachims; 2002) to solve this constraint optimization problem.
Introduction

TextTiling

Other topic segmentation algorithms

Evaluation of Topic Segmentation

Entity-Based Coherence

Discourse Representation

Entity Transitions

Evaluation Model

Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Earthquakes</th>
<th>Accidents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coreference + Syntax + Salience +</td>
<td>87.2</td>
<td>90.4</td>
</tr>
<tr>
<td>Coreference + Syntax + Salience −</td>
<td>88.3</td>
<td>90.1</td>
</tr>
<tr>
<td>Coreference − Syntax + Salience +</td>
<td>86.6</td>
<td>88.4**</td>
</tr>
<tr>
<td>Coreference − Syntax + Salience −</td>
<td>83.0**</td>
<td>89.9</td>
</tr>
<tr>
<td>Coreference + Syntax − Salience −</td>
<td>86.1</td>
<td>89.2</td>
</tr>
<tr>
<td>Coreference − Syntax + Salience −</td>
<td>82.3**</td>
<td>88.6*</td>
</tr>
<tr>
<td>Coreference − Syntax − Salience +</td>
<td>83.0**</td>
<td>86.5**</td>
</tr>
<tr>
<td>Coreference − Syntax − Salience −</td>
<td>81.4**</td>
<td>86.0**</td>
</tr>
</tbody>
</table>

Evaluation metric: % correct ranks in test set.

**: sig. different from Coreference + Syntax + Salience +

Entity-based model outperforms LSA.

Linguistically poorer models generally worse.

Omission of coreference causes performance drop.

Syntax and Salience have more effect on Accidents corpus.

In summary, is robust and learns appropriate ranking function.

BUT:

- Entity grid doesn’t contain lexical information.
- Doesn’t contain a notion of global coherence.
- Can’t model multi-paragraph text.

Evaluation Issues:

- Prevner and M. Hearst: “A critique and improvement of an evaluation metric for text segmentation”, Computational Linguistics, 28(1), 2002

Entity-based Coherence:


Topic segmentation algorithms:

- Jurafsky and Martin, chapter 21.1