Lecture 28: Discourse, coherence, cohesion

Henry S. Thompson
With input from Johanna Moore and Bonnie Webber
21 November 2016

1. "If we do not hang together
then surely we must hang separately" (Benjamin Franklin)

Not just any collection of sentences makes a discourse.

- A proper discourse is coherent
- It makes sense as a unit
  - Possibly with sub-structure
- The linguistic cues to coherence are called cohesion

The difference?

**Cohesion**
The (linguistic) clues that sentences belong to the same discourse

**Coherence**
The underlying (semantic) way in which it makes sense that they belong together

2. Linking together

Cohesive discourse often uses lexical chains

- That is, sets of the same or related words (synonyms, antonyms, hyponyms, meronyms, etc.) that appear in consecutive sentences

Longer texts usually contain several discourse segments

- Sub-topics within the overall coherence of the discourse

Intuition: When the topic shifts, different words will be used

- We can try to detect this automatically

*But*, the presence of cohesion does not guarantee coherence
John found some firm ripe apples and dropped them in a wooden bucket filled with water. Newton is said to have discovered gravity when hit on the head by an apple that dropped from a tree.

There are four lexical chains in the above mini-discourse, indicated by the words in red.

- But the two sentences don't actually cohere particularly well.

### 3. Automatically identifying sub-topics/segmenting discourse

Discourse-level NLP can sometimes profit from working with coherent sub-discourses

- So we need an automatic approach to delimiting coherent sub-sequences of sentences

There are several alternative approaches available:

- Segmentation:
  - Look for cohesion discontinuities
- (generative) modelling
  - Find the 'best' explanation

Useful for

- Information retrieval
- Search more generally, in
  - lectures
  - news
  - meeting records
- Summarisation
  - Did we miss anything?
- Information extraction
  - Template filling
  - Question answering

### 4. Finding discontinuities: TextTiling

An unsupervised approach based on lexical chains

- Developed by Marti Hearst

Originally developed and tested using a corpus of scientific papers

- That is, quite lengthy texts, compared to the trivial examples seen in these lectures

Three steps:

1. Preprocess: tokenise, filter and partition
2. Score: pairwise cohesion
3. Locate: threshold discontinuities
5. TextTiling: Preprocessing

In order to focus on what is assumed to matter

- That is, content words

Moderately aggressive preprocessing is done:

- Segment at whitespace
- Down-case
- Throw out stop-words
- Reduce inflected/derived forms to their base
  - Also known as stemming
- Group the results into 20-word 'pseudo-sentences'
  - Hearst calls these token sequences

6. TextTiling: Scoring

Compute a score for the gap between each adjacent pair of token sequences, as follows

1. Merge blocks of $k$ pseudo-sentences on either side of the gap to a bag of words
   - That is, a vector of counts
   - With one position for every 'word' in the whole text
   - Hearst used $k = 6$
2. Compute the normalised dot product of the two vectors
   - The cosine distance
3. Smooth the resulting score sequence by averaging the scores in a window of width $w$
   - Hearst used $w = 3$
   - That is, for a distance $y_i$ Hearst used $y_i' = \frac{y_{i-1} + y_i + y_{i+1}}{3}$ for the smoothed distance

7. TextTiling: Locate

We’re looking for discontinuities

- Where the score drops
- Indicating a lack of cohesion between two blocks

That is, something like this:

The depth score ($s$) at each gap is then given by $s = (y_{i-1} - y_i) + (y_{i+1} - y_i)$

Larger depth scores correspond to deeper 'valleys'
Scores larger than some threshold are taken to mark topic boundaries

- Hearst evaluated several possible threshold values
- Based on the mean and standard deviation of all the depth scores in the document

\[
\text{Liberal} \quad \bar{s} - \sigma
\]

\[
\text{Conservative} \quad \bar{s} - \frac{\sigma}{2}
\]

8. Evaluating segmentation

How well does TextTiling work?

- Here's an illustration from an early Hearst paper

From Hearst, M. A. and C. Plaunt 1993 "Subtopic structuring for full-length document access", in Proceedings of SIGIR 16

- The curve is (smoothed) similarity \((y')\), the vertical bars are consensus topic boundaries from human readers
- How can we quantify this?

Treating this as a two-way forced-choice classification task

- That is, each gap is either a boundary or it isn't

And scoring every gap as correctly or incorrectly classified doesn't work

- Segment boundaries are relatively rare
  - So it's too easy to score well for correctly labelling non-boundary gaps, just by being biased against boundaries
  - The 'block of wood' would do very well by always saying "no"

But counting just correctly labelled boundary gaps seems too strict

- Missing by one or two positions should get some credit

9. Evaluation, cont'd

The WindowDiff metric, which counts only \text{misses} (incorrect classifications) \text{within a window} attempts to address both problems
• It doesn’t give too much credit for correct non-boundary labelling
• It allows certain amount of mis-placing of boundary labels

Specifically, to compare boundaries in a gold standard reference (Ref) with those in a hypothesis (Hyp):
• Each a vector with 1 for a boundary and 0 for non-boundary

We will slide a window of size $k$ over Hyp and Ref comparing the number of boundaries in each
• Define a windowed boundary count $r_i$ in Ref for window size $k$ as $\sum_{j=i}^{i+k-1} \text{Ref}_j$
• And similarly for $h_i$ in Hyp

Then we compare the boundary counts for each possible window position
• That is, $|r_i - h_i|$ for each $i$
  □ This will be 0 if the two agree, positive otherwise
    ■ We count 0 if the result is 0 (correct)
    ■ And count 1 if the result is > 0 (incorrect)

Sum for all possible window positions, and normalise by the number of such positions:
$$\frac{1}{N-k} \sum_{i=1}^{N-k} |r_i - h_i| \neq 0$$
• 0 is the best result
  □ No misses
• 1 is the worst
  □ Misses at every possible window position

10. Evaluation example

An example from J&M with
• $k = 4$ (half the mean width of the gold-standard segments)
• $N = 23$
• a total of 4 misses

![Diagram](image)

Based on Figure 21.2 from Jurafsky and Martin 2009
(The colouring of the rectangles in the bottom row is misleading, I think)
• The resulting score is $\frac{4}{23-4} = 0.21$

The block of wood always guessing "no" would score $\frac{8}{23-4} = 0.42$
• Whereas if we simply counted misses without windowing, both scores would be $\frac{2}{22} = 0.09$

Note that this approach to evaluation is appropriate for any segmentation task where the ratio of candidate segmentation points to actual segments is high
• Sentences in unpunctuated text
• Tone groups in continuous speech
11. Machine learning?

More recently, (semi-)supervised machine learning approaches to uncovering topic structure have been explored.

Over-simplifying, you can think of the problem as similar to POS-tagging.

So you can even use Hidden Markov Models to learn and label:

- There are transitions between topics
- And each topic is characterised by an output probability distribution

But now the distribution governs the whole space of (substantive) lexical choice within a topic:

- Modelling not just one word choice
- But the whole bag of words


12. Topic is not the only dimension of discourse change

Topic/sub-topic is not the only structuring principle we find in discourse:

- Different genres may mean different kinds of structure

Some common patterns, by genre:

**Expository**
- Topic/sub-topic

**Task-oriented**
- Function/precondition

**Narrative**
- Cause/effect, sequence/sub-sequence, state/event

But note that some of this is not necessarily universal:

- Different scholarly communities may have different structural conventions
- Different cultures have different narrative conventions

Cohesion sometimes manifests itself differently for different genres.

13. Functional Segmentation

Texts within a given genre:

- News reports
- Scientific papers
- Legal judgements
- Laws

Generally share a similar structure, independent of topic.
• sports, politics, disasters
• molecular biology, radio astronomy, cognitive psychology

That is, their structure

• reflects the function played by their parts
• in a conventionalised way

14. Example: news stories

The conventional structure is so 'obvious' that you hardly notice it

• Known as the inverted pyramid

In decreasing order of importance

• Headline
• Lead paragraph
  ◦ Who, what, when, where, maybe why and how
• Body paragraphs, more on why and how
• Tail, the least important
  ◦ And available for cutting if space requires it

15. Example: Scientific journal papers

Individual disciplines typically report on experiments in highly conventionalised ways

• Your paper will not be published in a leading e.g. psychology research journal if it doesn't look like this

Front matter
  Title, Abstract

Body
  ◦ Introduction (or Objective), including background
  ◦ Methods
  ◦ Results
  ◦ Discussion
    (or, mnemonically, IMRAD)

Back matter
  Acknowledgements, References

The major divisions (IMRAD) will usually be typographically distinct and explicitly labelled

• Less immediately distinctive, more equivocal, cues give evidence for finer grained internal structure

16. Richer structure

Discourse structure is not (always) just ODTAA

• That is, it's not flat
• "One Damn Thing After Another"

Sometimes detecting this structure really matters
Welcome to word processing:
- That's using a computer to type letters and reports
- Make a typo?
  - No problem
  - Just back up, type over the mistake, and it's gone
- *And, it eliminates retyping
- And, it eliminates retyping