ANLP 2014

Lecture 25: Discourse, coherence, cohesion

Henry S. Thompson
(Based in part on slides by Johanna Moore and Bonnie Webber)
13 November 2014

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 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 an 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.

3. Identifying sub-topics/segmenting discourse

The goal is to delimit coherent sub-sequences of sentences

By division

- Look for cohesion discontinuities

By (generative) modelling

- Find the 'best' explanation

Relevant 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

Three steps:

1. Preprocess: tokenise, filter and partition
2. Score: pairwise cohesion
3. Locate: threshhold 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

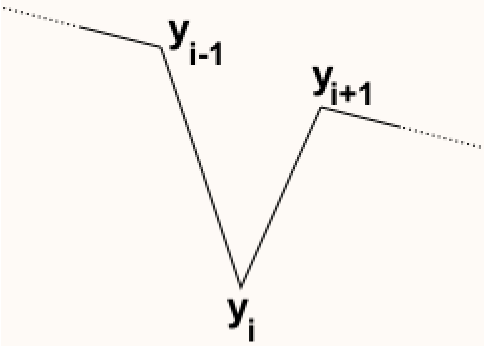
1. Reduce blocks of *k* pseudo-sentences on either side of the gap to a **bag of words**
    - That is, a vector of counts
    - With one position for every 'word' in the whole text
2. Compute the normalised dot product of the two vectors
    - The cosine distance
3. Smooth the resulting score sequence by averaging the scores in a symmetrical window of width *s* around each gap

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** at each gap is then given by $\left(y_{i-1} - y_i\right) + \left(y_{i+1} - y_i\right)$

Larger depth scores correspond to deeper 'valleys'

Scores larger than some threshhold are taken to mark topic boundaries

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

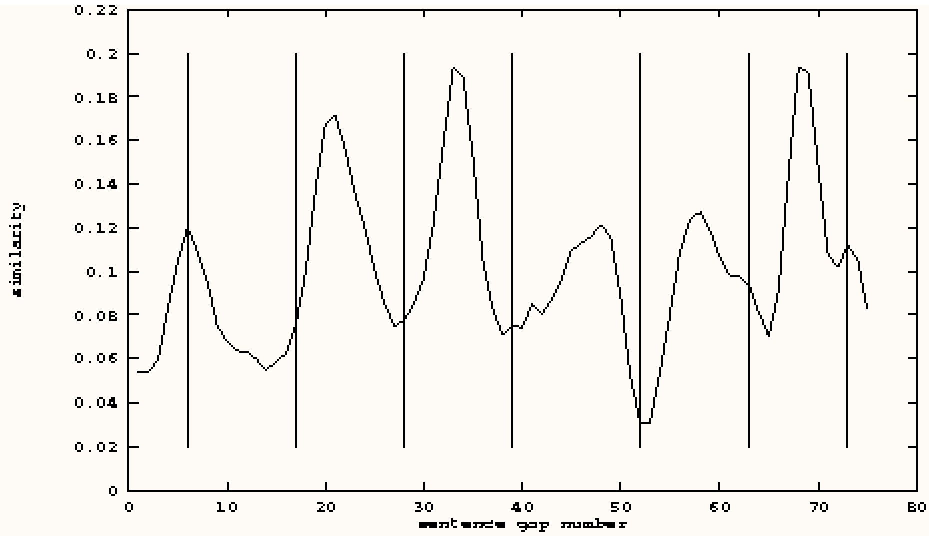
Liberal
: $\bar{s} - \sigma$

Conservative
: $\bar{s} - \frac{\sigma}{2}$

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 depth score, the vertical bars are consensus topic boundaries from human readers
- How can we quantify this?

Just classifying every possibly boundary as correct (Y+Y or N+N) vs. incorrect (Y+N or N+Y) doesn't work

- Segment boundaries are relatively rare
  - So N+N is very common
  - The "block of wood" can do very well by always saying "no"

Counting just Y+Y seems too strict

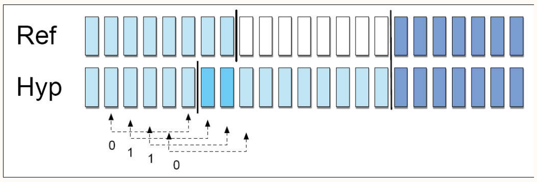
- Missing by one or two positions should get *some* credit

9. Evaluation, cont'd

The **WindowDiff** metric, which counts only **misses** (Y+N or N+Y) *within a window* attempts to address this

Specifically, to compare boundaries in a gold standard reference (**Ref**) with those in a hypothesis (**Hyp)**:

- Slide a window of size k over **Hyp** and **Ref**
- Compare the number of boundaries within the window at each possible position i in **Ref** (r_i) with those in **Hyp** (h_i)
- That is, $|r_i - h_i|$
  - Count 0 if the result is 0 (correct)
  - Count 1 if the result is > 0 (incorrect)



Based on Figure 21.2 from Jurafsky and Martin 2009

- Sum for all possible i, and normalise by the number of possible positions, $N - k$

0 is the best result

- No misses

1 is the worst

- Misses at for every window position

10. 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

See Purver, M. 2011, "Topic Segmentation", in Tur, G. and de Mori, R. *Spoken Language Understanding* for a more detailed introduction

11. Topic is not the only divider

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