Lexical Cohesion in Texts: Extraction Methods and Applications

Beata Beigman Klebanov

Eli Shamir

Cohesion in English, Halliday and Hasan 1976

- Devices that help the text “hang together” as a whole
  - Reference
    - John ate an apple. He liked it.
  - Ellipsis
    - John ate an apple and Jane — a pear.
  - Conjunction
    - John was tired. Moreover, he was hungry.
  - Substitution
    - John went home. Pete did likewise.
  - Lexical cohesion

Lexical Cohesion

- Texture created by use of words with related meanings

- Differently from the other types of cohesion, lexical connections are not explicitly marked in the text but supplied by the reader's background knowledge.

Why is lexical cohesion useful?

- Segmentation — Stokes et al 04
  - into topical sub-units
  - into separate stories in transcribed broadcast news
- Word Sense Disambiguation — Stairmand 97
- Extractive Summarization — Barzilay & Elhadad 97
  - looking for sentences where many strong chains are active, as these connect major issues
- Topical, Stylistic Analysis (e.g. political speeches) — BBK 08-09

What creates lexical cohesion?

A-priori: classical relations

- synonymy: mother / Mom
- hyponymy: barracuda / fish
- antonymy: bad / good

But when readers are asked to group related words in a text

- most have hard-to-characterize relations (Morris & Hirst, 2005)

  born / married homeless / alcoholic date / love

- people don't agree on groupings very well (Morris & Hirst, 2004)

- context changes the perception of relatedness

  (Resnik & Diab, 2000)

Experiment on lexical cohesion between pairs of words

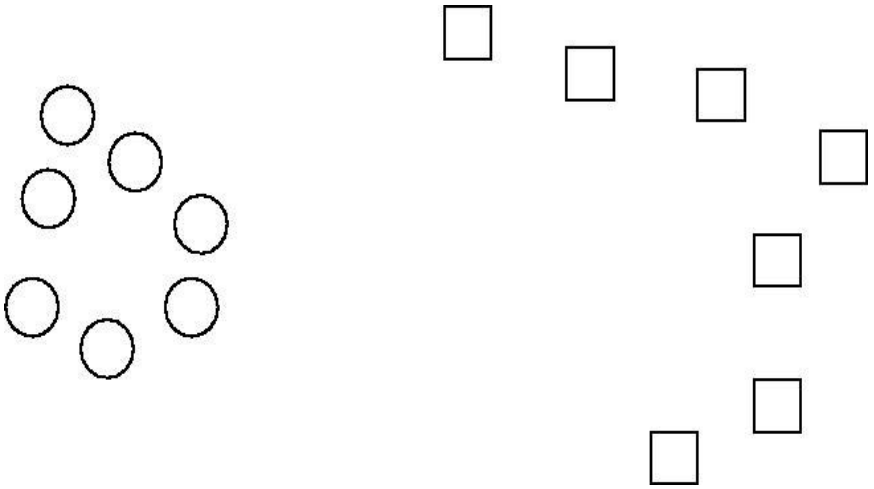
- Question to the subjects:
  - For every word first appearing in the text, is it easy to accommodate into the story so far, according to the commonsense knowledge as you perceive it? If yes, please mark the previous item(s) that help the accommodation, which we call anchors.
  - Commonsense knowledge = knowledge of types, associations, typical situations and even typical utterances

- 22 students marked such structures in 10 texts (400-800 words each)
  - Kappa = 0.45 on whether an item is anchored or not

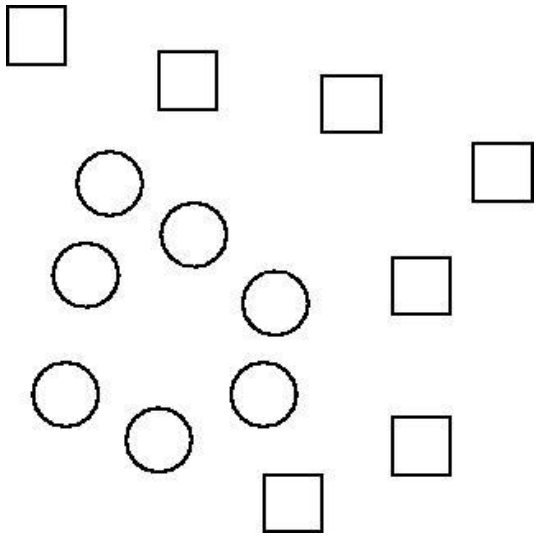
i.e. Yes/No subtask

Beigman Klebanov and Shamir, 2006

Agreement Analysis



Subgroups vary in both Interpretation and consistency



Subgroups vary in consistency around the same interpretation

Beigman Klebanov and Shamir, 2006

Finding Reliable Annotations

- Suppose 22 pseudo-annotators were flipping each his own coin for every item
  - Heads probability induced from actual annotator
- What is the level of agreement for which this scenario is sufficiently improbable?
- Random anchoring hypothesis rejected with 99% confidence for 13 coinciding markups
- Conclusion: For items marked by 13 people or more, at least some annotations are reliably deliberate
  - Core annotations (same-tareget-anchor >= 6)

Beigman Klebanov and Shamir, 2006

Validation Experiment

- Subjects are presented with everything marked by at least one human and some random markups, and asked to cross out things they disagree with.

- Every subject has a yes or no vote per item.

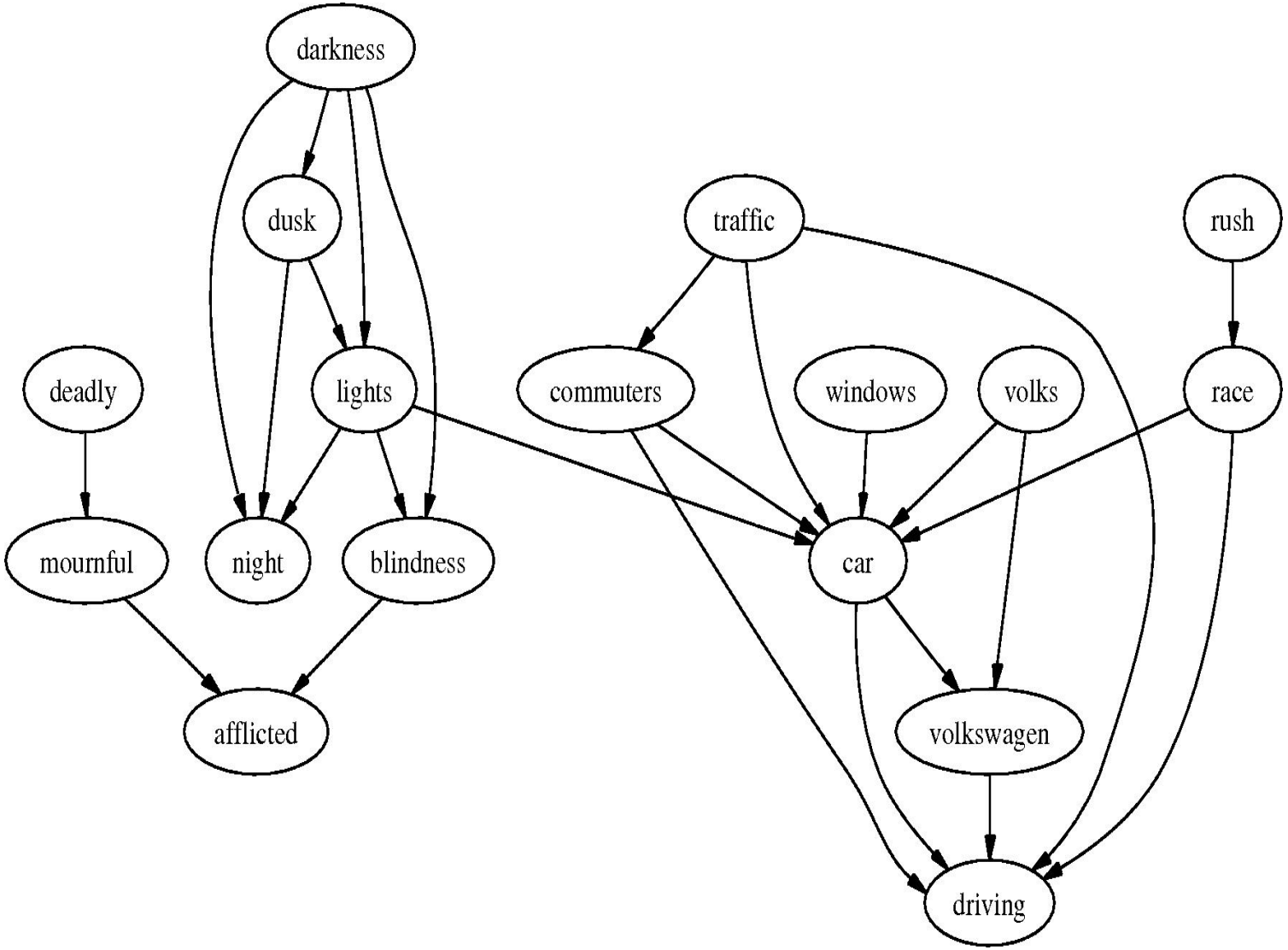
- Random markups: 15% yes

4-5% - All human markups: 62% yes

1% - Core markups: 94% yes

Beigman Klebanov and Shamir, 2006
Beigman Klebanov et al, 2008

Outland Text - Anchor Links



Hard Cases

- The category “lexically cohesive” is radial
  - Clear members at the “center”
  - Controversial members (hard cases) at the “periphery”
  - Radial categories are ubiquitous in human knowledge, from colors to kin relations

- Machine learning from data with hard cases
  - Assuming hard cases have random annotations, many popular machine learning techniques, such as SVM, will in worst case exhibit *hard case bias*, i.e. will make mistakes on the easy, clear cases due to exposure to hard cases during training

Modeling lexical cohesion

To determine that a pair is cohesive, the software uses, among other things:

- Corpus co-occurrence statistics
  - Latent Semantic Analysis trained on Wall Street Journal 1987 articles
- Dictionary definitions
  - WordNet glosses and hierarchy
- Free associations
  - Edinburgh Associative Thesaurus
  - U of South Florida Word Association Norms
    - Dream:fantasy, danger:fear, accident:pain

Supervised learning of lexical cohesion

- Information from various knowledge sources is combined using a decision tree automatically learned from reliable annotations.

- Performance:
  - 53% of core pairs are found by the software (core about 1%)
  - 46% of pairs the software marks as cohesive are also thus marked by at least one human (both about 4-5% of pairs)
  - Remarkable!

Application to Political Science

- From pairs of words to semantic fields
  - Densely interconnected areas in the large graph of cohesive pairs correspond to culturally-based fields of meaning, known as semantic fields
- Semantic fields can capture communicative frames
  - Framing is a way to present an issue such that certain aspects are stressed and not others, so as to promote a particular interpretation or evaluation
    - Example: Using sports vocabulary to talk about a war hides those aspect of war that are not parallel to a game, such as loss of life

Margaret Thatcher 1977

- Major political objectives
  - Improve Tory image with working class voters
  - Project authority and leadership
  - Promote Conservative agenda

- Automatic semantic field analysis of the speech to the Conservative Party conference, 1977
  - Topics of discussion:
    - Tory Thatcher Labour election politics party …
    - money economy wealth economic prosperity …
  - Rhetorical elements: project protective authority
    - threat danger safe risk fear dangerous threaten …
    - strength strong muscle strengthen courage…
  - ???
    - sea boat water sailor fishing ashore fish tide bait

Beigman Klebanov et al, 2008

Fishing and Boating

- Multiple figures draw on this field:
  - Our approach was put very simply by a Chinese philosopher centuries ago. "Govern a great nation," he counselled, "as you would cook a small fish. Don't overdo it."
  - The whole community benefits. When the tide comes in, all the boats rise.
  - People who ask the question are already half way into Labour's trap. They've swallowed the bait and are ripe for the catch.
- Cultural domain intimately familiar to the working class Brits
  - √ Improve communication with working class voters
- "Britain as an island" encourages nationalistic sentiment because it "functions as a metaphor signifying safety, defense against intruders, secludedness, difference" (Mautner 2001)
  - √ Promote Conservative (Euro-sceptic) agenda

Beigman Klebanov et al, 2008

Thank you