Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text (Morris, Hirst, 1991)

M.Sc. Seminar: Discourse Coherence Theories and Modeling

Alexandr Chernov

Department of Computational Linguistics, Saarland University

July 8, 2013

Overview

- Motivation
- Lexical Cohesion
    - Lexical Chains
    - Cohesion and Coherence
- Forming Lexical Chains
- Using Lexical Chains as a Tool
- Conclusion

Motivation

Lexical chains provide a valuable indicator of text structure and also semantic context for interpreting words, concepts, and sentences.

Lexical Cohesion

- Type of cohesion that arises from semantic relationships between words
- Basing on the type of dependency relationship between words 5 basic classes of lexical cohesion are distinguished (Halliday and Hasan)

Classes of lexical cohesion

- Reiteration with identity of reference:

  1. Mary bit into a *peach*.
  2. Unfortunately the *peach* wasn't ripe.

Classes of lexical cohesion

- Reiteration without identity of reference:

1. Mary ate some *peaches*.
2. She likes *peaches* very much.

Classes of lexical cohesion

- Reiteration by means of superordinate:
    1. Mary ate a *peach*.
    2. She likes *fruits*.

Classes of lexical cohesion

- Systematic semantic relation (systematically classifiable):

1. Mary likes *green* apples.
2. She does not like *red* ones.

Classes of lexical cohesion

- Nonsystematic semantic relation (not systematically classifiable):
  1. Mary spent three hours in the *garden* yesterday.
  2. She was *digging* potatoes.

Exercise 1

List of classes:

1. Reiteration with identity of reference.
2. Reiteration without identity of reference.
3. Reiteration by means of superordinate.
4. Systematic semantic relation (systematically classifiable).
5. Nonsystematic semantic relation (not systematically classifiable).

Lexical chain

A sequence of related words in writing, spanning short (adjacent words or sentences) or long distances (entire text).

Example

I like beer. Miller just launched a new *pilsner*. But, because I am a *beer* snob, I am only going to drink pretentious Belgian *ale*.

http://www.lexalytics.com/lexical-chains

Importance of lexical cohesion

1. Lexical chains help in the resolution of ambiguity and in the narrowing to a specific meaning of a word.
2. Lexical chains provide means for the determination of coherence and discourse structure.

Example 1

[gin, alcohol, sober, *drinks*] => noun "drinks" means "alcoholic drinks"

Example 2

[hair, curl, comb, *wave*] => noun "wave" does not mean "a water wave"

Importance of lexical cohesion

- Lexical chains provide means for the determination of coherence and discourse structure:
    1. If a lexical chain ends, it is likely that a linguistic segment ends too (lexical chains tend to indicate the topicality of segments).
    2. If a new lexical chain begins, this is an indication or clue that a new segment has begun.
    3. If an old chain is referred to again, it means that a previous segment is being referred to.

Cohesion and Coherence

- Coherence is a term for making sense; it means there is sense in the text.
- Cohesion is a term for sticking together; it means that the text all hangs together.
- Independent from each other: cohesion can exist among sentences that are not related coherently.

Cohesion != Coherence

Cohesion with NO Coherence:

My favourite color is blue. Blue sports cars go very fast. Driving in this way is dangerous and can cause many car crashes. I had a car accident once and broke my leg. I was very sad because I had to miss a holiday in Europe because of the injury.

http://gordonscruton.blogspot.de/2011/08/what-is-cohesion-coherence-cambridge.html

Cohesion != Coherence

Coherence with NO Cohesion:

My favourite color is blue. I'm calm and relaxed. In the summer I lie on the grass and look up.

http://gordonscruton.blogspot.de/2011/08/what-is-cohesion-coherence-cambridge.html

Cohesion and Coherence

- Both cohesion and coherence are distinct phenomena creating unity in text.
- Cohesion is a useful indicator of coherence.
- Resolution of coreference = identification of coherence (Hobbs).

Finding lexical chains

- Purpose: determination of the text structure.
- The method is useful for texts in any general domain.
- Full understanding of a text is not required.
- The algorithm found well over 90% of the intuitive lexical relations

Forming lexical chains

Looking for candidate words (pronouns, prepositions, auxiliary verbs, and high-frequency words are not considered)

Example

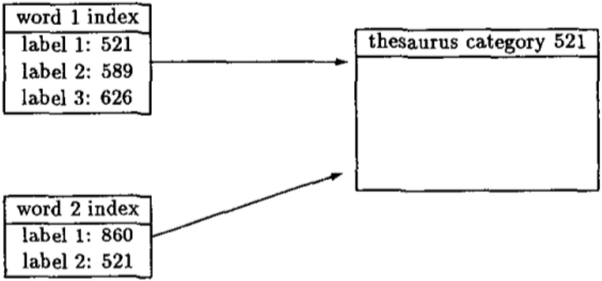
My *maternal grandfather lived* to be *111*. *Zayde* was *lucid* to the *end*, but a few *years before* he *died* the *family assigned* me the *task* of *talking* to him about his *problem* with *alcohol*.

Forming lexical chains

- Building chains using an abridged version of Roget’s Thesaurus.
- 5 types of thesaural relations between words were found to be necessary in forming chains.

Thesaural relation no. 1

- Two words have a category common in their index entries: e.g. *"existence"* and *"being"* both have category *"life"* in their index entries



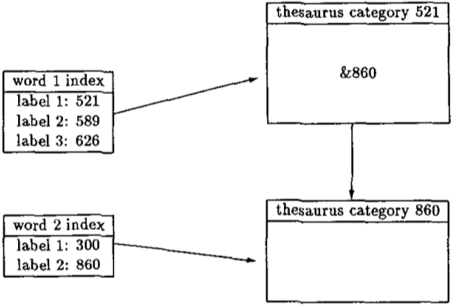
word 1 index
label 1: 521
label 2: 589
label 3: 626

word 2 index
label 1: 860
label 2: 521

thesaurus category 521

Thesaural relation no. 2

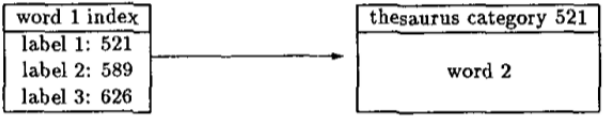
- One word has a category in its index entry that contains a pointer to a category of the other word: e.g. *"airplane"* has in its index entry a category which contains a pointer to another category referring to *"flight"*



Thesaural relation no. 3

- A word is either a label in the other word's index entry (b), or is in a category of the other word: e.g. *"deaf"* has a category containing the word *"hear"* (a)

(a)



word 1 index

label 1: 521

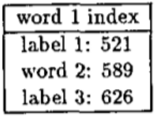
label 2: 589

label 3: 626

thesaurus category 521

word 2

(b)



word 1 index

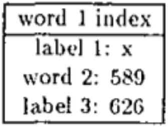
label 1: 521

word 2: 589

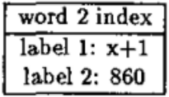
label 3: 626

Thesaural relation no. 4

- Two words are in the same group, and hence are semantically related: e.g. words *"life"* and *"death"* belong to the same group



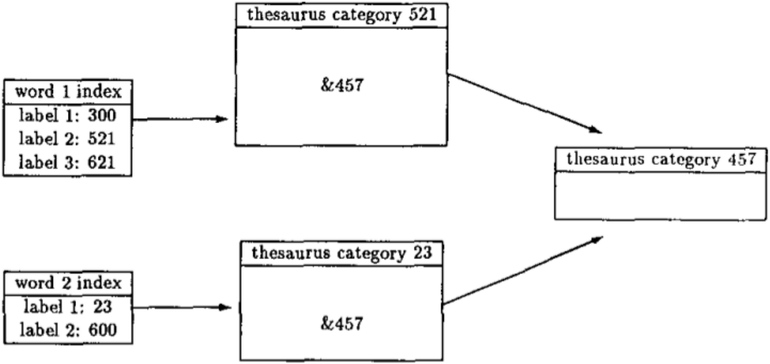
word 1 index
label 1: x
word 2: 589
label 3: 626



word 2 index
label 1: x+1
label 2: 860

Thesaural relation no. 5

- The two words have categories in their index entries that both point to a common category: e.g. *"gentle"* and *"charitable"* point to a common category *"kind"*



word 1 index
label 1: 300
label 2: 521
label 3: 621

thesaurus category 521
&457

word 2 index
label 1: 23
label 2: 600

thesaurus category 23
&457

thesaurus category 457

Chain strength

- Lexical chaining algorithms often produce a much larger number of chains than desired for a particular task (Hollingsworth, 2008).
- Chain strength is used to select the "best" or most relevant chains out of a given set of chains.

Factors contributing to chain strength

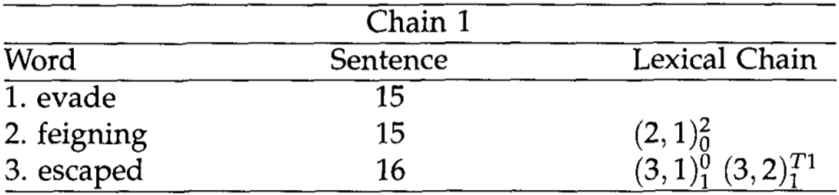
- Reiteration - the more repetitions, the stronger the chain (computed by counting the number of word-tokens of each word-type present in the chain).
- Density - the denser the chain, the stronger it is (the ratio of the number of words in a chain to the number of content words in the text).
- Length - the longer the chain, the stronger it is (the number of word-types it contains) (Hollingsworth, 2008).

Notation and Data Structures

Each lexical relationship in a chain is represented as $(u, v)_x^y$ where:

- u is the current word number,
- v is the word number of the related word,
- x is the transitive distance (0 - no transitive links),
- y is either
  - the number of the thesaural relationship between the 2 words
  - Tq where T stands for transitivity related, q is the word number through which the transitive relation is formed

Lexical chain notation



Chain 1

Word	Sentence	Lexical Chain
1. evade	15	
2. feigning	15	$(2,1)_0^2$
3. escaped	16	$(3,1)_1^0$ $(3,2)_1^{T1}$

Problems during computation of the chains

- General semantic relations between words of similar "feeling": [*hand-in-hand*, matching, whispering, *laughing, warm*]
- Situational knowledge.
- Specific proper names.

Problems during computation of the chains

- General semantic relations between words of similar "feeling": [*hand-in-hand*, matching, whispering, *laughing, warm*]
- Situational knowledge.
- Specific proper names.

Such words are usually not found in the thesaurus

Lexical Chains and Text Structure

A Boeing 777 aircraft that crash-landed at San Francisco airport killing two people did not have mechanical problems, an airline official has said.
The head of the South Korean airline Asiana, Yoon Young-doo, did not rule out human error but said the pilots were experienced veterans.
The witness told: "We heard a ’boom’ and saw the plane disappear into a cloud of dust and smoke".

S1: Boeing 777 aircraft crash-landed San Francisco airport killing two people mechanical problems airline official said
S2: head South Korean airline Asiana Yoon Young-doo rule out human error said pilots experienced veterans
S3: witness told heard ’boom’ saw plane disappear cloud dust smoke

http://www.bbc.co.uk/news/world-us-canada-23216587

Lexical Chains and Text Structure

- Chain 1:
  1. *[Boeing 777, aircraft, crash-landed, airport, airline]*
  2. *[airline, Asiana, pilots, plane, cloud]*
- Chain 2
  1. *[official, said]*
  2. *[head]*
- Chain 3
  1. *[killing, people, problems]*
  2. *[human error]*
- Chain 4
  1. *[witness, 'boom', dust, smoke]*

Exercise 2

Find lexical chains and segments:

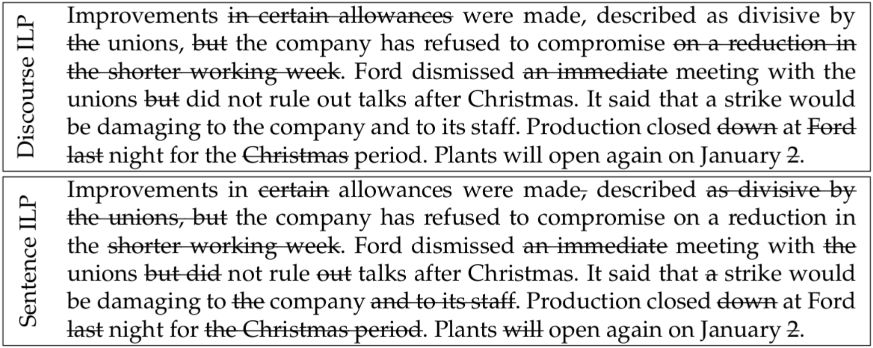
1. Find candidate words (you may use http://thesaurus.com/).
2. Delete "inappropriate" words.
3. Form lexical chains.
4. Find segments.

Lexical chains as a tool

- Provide a good clue for the determination of the intentional structure.
- Can be used to create efficient summarization tools.
- Keywords extraction tool(similar to a brief summary).
- Useful for document clustering

Lexical Chains and Summarization

Discourse Constraints for Document Compression, Clarke and Lapata, 2010



Discourse ILP

Improvements ~~in certain allowances~~ were made, described as divisive by ~~the~~ unions, ~~but~~ the company has refused to compromise ~~on a reduction in the shorter working week~~. Ford dismissed ~~an immediate~~ meeting with the unions ~~but~~ did not rule out talks after Christmas. It said that a strike would be damaging to the company and to its staff. Production closed ~~down~~ at ~~Ford last~~ night for the ~~Christmas~~ period. Plants will open again on January ~~2~~.

Sentence ILP

Improvements in ~~certain~~ allowances were made~~,~~ described ~~as divisive by the unions, but~~ the company has refused to compromise on a reduction in the ~~shorter working week~~. Ford dismissed ~~an immediate~~ meeting with ~~the~~ unions ~~but did~~ not rule ~~out~~ talks after Christmas. It said that ~~a~~ strike would be damaging to ~~the~~ company ~~and to its staff~~. Production closed ~~down~~ at Ford ~~last~~ night for ~~the Christmas period~~. Plants ~~will~~ open again on January ~~2~~.

Conclusions

- Lexical chains correspond closely to the intentional structure.
- Lexical chains appeared to be almost entirely computable with the defined relations.
- Lexical cohesion (and hence this tool) is not domain-specific.
- Lexical chains are useful for finding segments.

Thank you!

Thank you for your kind attention!

Do you have any questions?

References

- Jane Morris and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics Journal. Volume 17, Issue 1, March 1991. P. 21-48
- James Clarke and Mirella Lapata. 2010. Discourse Constraints for Document Compression. Computational Linguistics, 36(3), P. 411-441
- William A. Hollingsworth. Using Lexical Chains to Characterise Scientific Text. PhD thesis, Clare Hall College, University of Cambridge, 2008