Computational Models of Discourse: Preliminaries, Overview

Caroline Sporleder

Universität des Saarlandes

Sommersemester 2008/09

22.04.2008
Preliminaries
Course offered as:

- Proseminar (B.Sc.): 4. Semester
- Hauptseminar (B.Sc.): 6. Semester
- Seminar (M.Sc.)

Scheine/Credits:

- Proseminar: class presentation (20-30 minutes), term paper (10-15 Seiten) ⇒ 5 credit points
- Hauptseminar/Seminar (MSc): class presentation (30-40 minutes), term paper (about 20 pages) ⇒ 7 credit points (4 for presentation, 3 for term paper)
- optional oral exam (10-30 minutes)
- “participation in seminar” (reading, discussions, exercises etc.)

Course content complementary to Helmut Horacek’s Seminar “Diskursphänomene” (i.e. virtually no overlap)
Class Presentations

Preparation
- optionally: preparatory meeting to discuss open questions and presentation draft
- optionally: peer group discussion

Evaluation Criteria

Content
- main aspects of the paper(s) covered
- good explanations of content (examples, graphics)
- discussion of advantages/disadvantages of presented method
- use of additional (printed) references (optionally)

Form
- slides contain references
- slides are well-structured
- the topic was well-presented (interaction with audience etc.)
- the presentation is not too short
- the presentation is not (much) too long
Writing term papers

- answer a concrete research question (or a set of related questions)
- show evidence of independent thinking and reasoning
- independent bibliographic work
- meet academic standards (proper citations/references, well-structured, well-written etc.)

Don’t plagiarise!

- can include some practical work but doesn’t have to
- around 100 hours of work (2-3 weeks full-time)
- 15-20 pages long
- 10-15 references

Please discuss topic with me beforehand!
Course Overview
Why Discourse Processing?

Natural language rarely comes in isolated sentences...

- newspaper articles
- novels
- dialogues
- speeches by politicians
- ...
Why Discourse Processing?

Natural language rarely comes in isolated sentences . . .

- newspaper articles
- novels
- dialogues
- speeches by politicians
- . . .

NLP applications need to be able to deal with discourse . . .
Why Discourse Processing?

Natural language rarely comes in isolated sentences...

- newspaper articles
- novels
- dialogues
- speeches by politicians
- ...

NLP applications need to be able to deal with discourse...

- dialogue systems
- question answering
- text summarisation
- information extraction
- natural language generation
- natural language understanding
- ...

Caroline Sporleder csporled@coli.uni-sb.de

Computational Models of Discourse
Campaigning has closed in Argentina ahead of Sunday’s election to elect a successor to President Nestor Kirchner. The front-runner in the opinion polls is the current first lady, Senator Cristina Fernandez de Kirchner. She praised the economic record of her husband’s government during a rally in Buenos Aires.
Discourse Context Matters

Example: Co-reference Resolution

She praised the economic record of her husband’s government during a rally in Buenos Aires.
Example: Question Answering ("Why was Caesar killed?")

Caesar was proclaimed dictator for life, and he heavily centralised the bureaucracy of the Republic. These events provoked a hitherto friend of Caesar, Marcus Junius Brutus, and a group of other senators, to assassinate the dictator on the Ides of March (March 15) in 44 BC.
Greek officials hope the new site will boost the country’s long campaign for the return of the Elgin Marbles. Crowds of bystanders watched the first of the monuments lifted by cranes at the 2,500-year-old Parthenon. Greece has begun moving the ancient sculptures from the Acropolis in Athens to a new home - a museum at the foot of the hilltop citadel.
Example: Coherence/Text Generation

Greece has begun moving the ancient sculptures from the Acropolis in Athens to a new home - a museum at the foot of the hilltop citadel. Crowds of bystanders watched the first of the monuments lifted by cranes at the 2,500-year-old Parthenon. Greek officials hope the new site will boost the country’s long campaign for the return of the Elgin Marbles.
Draft Schedule

29.04.2009  Introduction Discourse (and Machine Learning)
06.05.2009  Cohesion and Local Coherence
  • Lexical Cohesion, Lexical Chains
  • Focus, Centering
13.05.2009  Text Segmentation
  • TextTiling
  • Other Segmentation Methods
20.05.2009  Applications
  • Automatic Essay Scoring
  • Information Ordering for Text Generation
27.05.2009  Generating Referring Expressions
  • rule-based
  • machine learning
03.06.2009  Co-reference Resolution
  • rule-based
  • supervised machine learning
  • unsupervised machine learning
<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Subtopics</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.06.2009</td>
<td><strong>Discourse Parsing</strong></td>
<td>- Discourse Parsing with RST</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Machine Learning</td>
</tr>
<tr>
<td>17.06.2009</td>
<td><strong>Temporal Ordering</strong></td>
<td></td>
</tr>
<tr>
<td>24.06.2009</td>
<td><strong>Text Summarisation</strong></td>
<td>- lexical chains</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- RST-based</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- multi-document</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- argumentative zoning</td>
</tr>
<tr>
<td>01.07.2009</td>
<td><strong>Sentiment Analysis</strong></td>
<td></td>
</tr>
<tr>
<td>08.07.2009</td>
<td><strong>Dialogue Processing</strong></td>
<td>- classification of dialogue acts</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- dialogue planning</td>
</tr>
<tr>
<td>15.07.2009</td>
<td><strong>Speech, Psycholinguistic Models</strong></td>
<td></td>
</tr>
<tr>
<td>22.07.2009</td>
<td><strong>Recap, Conclusion</strong></td>
<td></td>
</tr>
</tbody>
</table>
Central Questions:

- how can (local) text coherence be modelled?
- when is a text coherent?
John went to his favorite music store to buy a piano. It was a store John had frequented for many years. He was excited that he could finally buy a piano. It was closing just as John arrived.

John went to his favorite music store to buy a piano. He had frequented the store for many years. He was excited that he could finally buy a piano. He arrived just as the store was closing.
Central Questions:

- how can (local) text coherence be modelled?
- when is a text coherent?
Central Questions:

- how can (local) text coherence be modelled?
- when is a text coherent?

Applications:
many . . .

- text generation (concept-to-text, text-to-text, summarisation etc.)
- evaluation of text generating systems (summarisation, machine translation etc.)
- evaluation of human-written text (automatic essay scoring, readability assessment etc.)
chains of semantically related words
⇒ . . . measure lexical cohesion
chains of semantically related words
⇒ . . . measure lexical cohesion

Example
At least eight people have been killed in flooding in the Dominican Republic following torrential rains dumped by Tropical Storm Noel. The deaths were reported in the Dominican capital Santo Domingo, and along the south coast. The centre of the storm had passed by midday on Monday, and was set to head north towards the Bahamas.
chains of semantically related words
⇒ . . . measure lexical cohesion

Example
At least eight people have been killed in flooding in the Dominican Republic following torrential rains dumped by Tropical Storm Noel. The deaths were reported in the Dominican capital Santo Domingo, and along the south coast. The centre of the storm had passed by midday on Monday, and was set to head north towards the Bahamas.
chains of semantically related words
⇒ ... measure lexical cohesion

Example
At least eight people have been killed in flooding in the Dominican Republic following torrential rains dumped by Tropical Storm Noel. The deaths were reported in the Dominican capital Santo Domingo, and along the south coast. The centre of the storm had passed by midday on Monday, and was set to head north towards the Bahamas.
chains of semantically related words
⇒ . . . measure lexical cohesion

Example
At least eight people have been killed in flooding in the Dominican Republic following torrential rains dumped by Tropical Storm Noel. The deaths were reported in the Dominican capital Santo Domingo, and along the south coast. The centre of the storm had passed by midday on Monday, and was set to head north towards the Bahamas.
...models focus shifts

Example

Susan would like to go on a holiday. But she needs to find somebody to do her work while she’s away. She can’t think of anybody to do that. She considered Mike but he’s a bit unreliable. Yesterday he forgot to turn up for an important meeting with a client. The client was very annoyed and said she would never do business with Susan’s company again.
Centering Theory

... models focus shifts

Example

Susan would like to go on a holiday. But she needs to find somebody to do her work while she’s away. She can’t think of anybody to do that. She considered Mike but he’s a bit unreliable. Yesterday he forgot to turn up for an important meeting with a client. The client was very annoyed and said she would never do business with Susan’s company again.
Centering Theory

...models focus shifts

Example
Susan would like to go on a holiday. But she needs to find somebody to do her work while she’s away. She can’t think of anybody to do that. She considered Mike but he’s a bit unreliable. Yesterday he forgot to turn up for an important meeting with a client. The client was very annoyed and said she would never do business with Susan’s company again.
...models focus shifts

**Example**

Susan would like to go on a holiday. But she needs to find somebody to do her work while she’s away. She can’t think of anybody to do that. She considered Mike but he’s a bit unreliable. Yesterday he forgot to turn up for an important meeting with a client. The client was very annoyed and said she would never do business with Susan’s company again.
...identification of (sub-)topic shifts
Penicillin is a group of beta-lactam antibiotics used in the treatment of bacterial infections caused by susceptible, usually Gram-positive, organisms. The discovery of penicillin is usually attributed to Scottish scientist Sir Alexander Fleming in 1928. Fleming noticed a halo of inhibition of bacterial growth around a contaminant blue-green mold Staphylococcus plate culture. Fleming concluded that the mold was releasing a substance that was inhibiting bacterial growth and lysing the bacteria. Common adverse drug reactions associated with use of the penicillins include: diarrhea, nausea, rash, urticaria, and/or superinfection (including candidiasis).
Penicillin is a group of beta-lactam antibiotics used in the treatment of bacterial infections caused by susceptible, usually Gram-positive, organisms. The discovery of penicillin is usually attributed to Scottish scientist Sir Alexander Fleming in 1928. Fleming noticed a halo of inhibition of bacterial growth around a contaminant blue-green mold Staphylococcus plate culture. Fleming concluded that the mold was releasing a substance that was inhibiting bacterial growth and lysing the bacteria. **Common adverse drug reactions associated with use of the penicillins include: diarrhea, nausea, rash, urticaria, and/or superinfection (including candidiasis).**
... indentification of (sub-)topic shifts

Applications

- text retrieval
- preprocessing for other discourse processing tasks (discourse parsing, information extraction, coherence assessment etc.)
Is an essay written well or not?

Example

We as college students, have ahead of us possibly one of the most difficult tasks left to master. It is important to realize that knowledge alone may not lead to the occupation that we desire.

The task of searching for a job may be made somewhat simplified if we take some time to consider the qualities we have demonstrated in the interview, once we leave the office.
What is the best ordering in which to present information (e.g., facts in a database)?

Which ordering makes the text most coherent?
What is the best ordering in which to present information (e.g., facts in a database)?

Which ordering makes the text most coherent?

Example

1. This coin is made of silver.
2. Croton is a Greek colony in South Italy.
3. This coin comes from the archaic period.
4. This coin comes from Croton.
5. Towards the end of the archaic period, coins were used for transactions.
What is the best ordering in which to present information (e.g., facts in a database)?
Which ordering makes the text most coherent?

Example

1. Towards the end of the archaic period, coins were used for transactions.
2. This coin comes from the archaic period.
3. This coin is made of silver.
4. This coin comes from Croton.
5. Croton is a Greek colony in South Italy.
The linguistic form of referring expressions (NPs) depends on the discourse context.

Example: *Noam Chomsky, the linguist, the author of “Syntactic Structures”, the man, he...*

Application:
text generation
He claims record

The 22-year-old computer science undergraduate from Bath is claiming a world record for the longest distance ridden on a unicycle in 24 hours.

A unicycling student covered exactly 282 miles at Aberystwyth University’s athletics track.

Sam Wakeling was aiming to beat the existing record of 235.3 miles.
Example

He claims record

The 22-year-old computer science undergraduate from Bath is claiming a world record for the longest distance ridden on a unicycle in 24 hours.

A unicycling student covered exactly 282 miles at Aberystwyth University’s athletics track.

Sam Wakeling was aiming to beat the existing record of 235.3 miles.
Generating Referring Expressions

Example

**Unicycling student** claims record

A student is claiming a world record for the longest distance ridden on a unicycle in 24 hours.

**Sam Wakeling** covered exactly 282 miles at Aberystwyth University’s athletics track.

The 22-year-old computer science undergraduate from Bath was aiming to beat the existing record of 235.3 miles.
Identify the referent of pronouns, definite NPs etc.
Identify the referent of pronouns, definite NPs etc.

Example

Sophia Loren says she will always be grateful to Bono. The actress revealed that the U2 singer helped her calm down when she became scared by a thunderstorm while travelling on a plane.
Identify the referent of pronouns, definite NPs etc.

Example

Sophia Loren says she will always be grateful to Bono. The actress revealed that the U2 singer helped her calm down when she became scared by a thunderstorm while travelling on a plane.
Identify the referent of pronouns, definite NPs etc.

Example

Sophia Loren says she will always be grateful to Bono. The actress revealed that the U2 singer helped her calm down when she became scared by a thunderstorm while travelling on a plane.
Co-reference Resolution

Identify the referent of pronouns, definite NPs etc.

Example
Sophia Loren says she will always be grateful to Bono. The actress revealed that the U2 singer helped her calm down when she became scared by a thunderstorm while travelling on a plane.

Applications:
- information extraction
- text summarisation
- pre-processing for many other NLP tasks
Discourse Parsing

Identify the discourse structure, i.e. how utterances are linked

Applications:
- anything that involves natural language understanding
- question answering
- information extraction
- text summarisation
- etc.

Caroline Sporleder  
csporled@coli.uni-sb.de  
Computational Models of Discourse
Identify the discourse structure, i.e. how utterances are linked

*John hid Peter’s car keys. He was drunk.*
Identify the discourse structure, i.e. how utterances are linked

*John hid Peter’s car keys. He was drunk.*

⇒ The fact that John was drunk *explains* why he hid Peter’s car keys.
Identify the discourse structure, i.e. how utterances are linked

*John hid Peter’s car keys. He was drunk.*

⇒ The fact that John was drunk *explains* why he hid Peter’s car keys.

*Mary likes chocolate, Maggie likes crisps.*
Identify the discourse structure, i.e. how utterances are linked

*John hid Peter’s car keys. He was drunk.*

⇒ The fact that John was drunk *explains* why he hid Peter’s car keys.

*Marry likes chocolate, Maggie likes crisps.*

⇒ The fact that Maggie likes crisps *contrasts* with Mary’s liking of chocolate.
Identify the discourse structure, i.e. how utterances are linked

*John hid Peter’s car keys. He was drunk.*

⇒ The fact that John was drunk *explains* why he hid Peter’s car keys.

*Mary likes chocolate, Maggie likes crisps.*

⇒ The fact that Maggie likes crisps *contrasts* with Mary’s liking of chocolate.

**Applications:**

anything that involves natural language understanding

- question answering
- information extraction
- text summarisation
- etc.
Temporal Ordering

Identify the temporal ordering of events referred to in discourse
Identify the temporal ordering of events referred to in discourse

Example

1. John arrived at an oasis. He saw the camels around the waterhole.
Identify the temporal ordering of events referred to in discourse

Example

1. John arrived at an oasis. He saw the camels around the waterhole.
2. John arrived at an oasis. He left the camels around the waterhole.
Identify the temporal ordering of events referred to in discourse

Example

1. John arrived at an oasis. He saw the camels around the waterhole.
2. John arrived at an oasis. He left the camels around the waterhole.

Applications:

- information extraction
- question answering
- machine translation
- etc.
Produce a shorter version of an input text (e.g., synthesise various news stories on a topic)

Example

Penicillin is a group of beta-lactam antibiotics used in the treatment of bacterial infections caused by susceptible, usually Gram-positive, organisms. The discovery of penicillin is usually attributed to Scottish scientist Sir Alexander Fleming in 1928. Fleming noticed a halo of inhibition of bacterial growth around a contaminant blue-green mold Staphylococcus plate culture. Fleming concluded that the mold was releasing a substance that was inhibiting bacterial growth and lysing the bacteria. Common adverse drug reactions associated with use of the penicillins include: diarrhea, nausea, rash, urticaria, and/or superinfection (including candidiasis).
Produce a shorter version of an input text (e.g., synthesise various news stories on a topic)

Example

Penicillin is a group of beta-lactam antibiotics. It was discovered by Sir Alexander Fleming in 1928. Common adverse drug reactions associated with its use include: diarrhea, nausea, rash, urticaria, and/or superinfection (including candidiasis).
Argumentative Zoning

Find structure in scientific papers (BACKGROUND, OWN CONTRIBUTION, RELATION TO OTHER WORK, etc.)

Applications:

- summarisation
- question answering
- citation analysis
Researchers in knowledge representation agreed that one of the hard problems of understanding narrative is the representation of temporal information . . .
Recently, Smith has suggested the following solution to this problem . . . But this solution cannot be used to interpret the following example . . .
We propose a solution which circumvents this problem . . .
Researchers in knowledge representation agreed that one of the hard problems of understanding narrative is the representation of temporal information ...

Recently, Smith has suggested the following solution to this problem ... But this solution cannot be used to interpret the following example ...

We propose a solution which circumvents this problem...
Sentiment Analysis

- Is a product review positive or negative?
- What is a given person’s opinion on a particular topic?
- Which parts of a text are objective, which reflect subjective opinions?
Sentiment Analysis

- Is a product review positive or negative?
- What is a given person’s opinion on a particular topic?
- Which parts of a text are objective, which reflect subjective opinions?

Bob Carpenters Amazon Review of Manning & Schütze (1999)

This is the best book I’ve ever read on computational linguistics. It should be ideal for both linguists who want to learn about statistical language processing and those building language applications who want to learn about linguistics. This book isn’t even published and it’s now my most highly used reference book, joining gems such as Cormen, Leiserson and Rivest’s algorithm book, Quirk et al.’s English Grammar, and Andrew Gelman’s Bayesian statistics book (three excellent companions to this book, by the way).
What does Horst Köhler think about free trade?

German President Horst Köhler opened the Hanover Fair on Sunday, criticizing a perceived lurch towards protectionism in some countries. The former head of the International Monetary Fund demanded a "clear-cut timetable" to conclude the current stalled round of talks to encourage free trade around the world.
Sentiment Analysis

- Is a product review positive or negative?
- What is a given person’s opinion on a particular topic?
- Which parts of a text are objective, which reflect subjective opinions?

Tagesschau online 19.04.2009 (modified)

Foreign Minister Frank-Walter Steinmeier underpinned his claim to topple Angela Merkel by presenting his party’s manifesto, which calls for increased taxes on the rich, while cutting them for those less well-off. He also emphasized his party’s aim to defend the rights of employees and their jobs and added that struggling German carmaker Opel, which employs some 26,000 people in Germany alone, had to be saved at all costs.
Dialogues consist of different dialogue acts (OPENING, INFO-REQUEST, COMMIT etc.). For dialogue processing, these need to be identified.

Example

A: I need to travel in May. [ASSERT]
B: And what day in May did you want to travel? [INFO-REQUEST, ACKNOWLEDGEMENT]
A: I need to be there for a meeting that’s from the 12th to the 15th. [ASSERT, ANSWER]
B: And what time would you like to leave Pittsburgh? [INFO-REQUEST, ACKNOWLEDGEMENT]
A: Uh hmm I don’t think there’s many options for non-stop...[CHECK, HOLD]
B: Right. [ACCEPT, ACKNOWLEDGEMENT]
There is three non-stops today. [ASSERT]