Einführung in die Pragmatik und Diskurs:

Computational Discourse Processing

A. Palmer/A. Horbach

Universität des Saarlandes

16 June 2014

Outline

1. Overview
2. Research Themes
3. Modeling Entity-Based Coherence: Entity Grid
4. Modeling Discourse Structure: PDTB

Main readings

- Bonnie Webber, Marcus Egg, and Valia Kordoni, **Discourse structure and language technology**, NLE vol. 18, no. 4, 2012
- Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber, **The Penn Discourse Treebank**, LREC 2004

Optional readings:

- Bonnie Webber and Aravind Joshi, **Discourse Structure and Computation: Past, Present, and Future**, ACL 2010
- Penn Discourse Treebank Annotation Manual
- Regina Barzilay and Mirella Lapata, **Modeling Local Coherence: An Entity-Based Approach**, Computational Linguistics, May 2007

The plan for today

- Overview of *computational* discourse processing
- Research themes in computational discourse processing
- Focus 1: Modeling entity-based coherence
- Focus 2: Modeling discourse structure

Defining discourse

A multi-part definition of discourse.
Following Webber et al., discourse can be thought of as

1. A sequence of sentences
2. which conveys more than its individual sentences through their relationships with one another, and
3. which exploits special features of language that enable discourse to be more easily understood.

A sequence of sentences

Example

If they’re drunk and meant to be on parade and you go to their room and they’re lying in a pool of piss, then you lock them up for a day.

Implementation question: unit of analysis?

Research problem: automatic segmentation

Meaning beyond the individual sentences

Example

Don’t worry about the world coming to an end today. It is already tomorrow in Australia.

Research questions: how to model meaning beyond the sentence? to what extent does it connect to meaning of the sentence? how to model sentence meaning?

Research problem: automatic identification/classification of meaning relations (given particular inventory)

Special features of language

Discourse exploits features of language that let us:

- Talk about topics previously discussed in text
- Indicate relations between states, events, beliefs, etc.
- Change to new topics or resume previous topics

Special features of language 2

Example

The police are not here to create disorder. **They** are here to preserve **it**.

Example

Pope John XXIII was asked ‘How many people work in the Vatican?’ He is said to have replied, ‘About **half**.’

Example

Men have a tragic genetic flaw. **As a result,** they cannot see dirt **until** there is enough of it to support agriculture.

Types of approaches to discourse structure

Linear segmentation

Discourse chunking

Discourse parsing

Some applications

- Summarization
- Information extraction
- Essay analysis and scoring
- Sentiment analysis and opinion mining
- Assessing text quality
- Machine translation
- ...

Outline

1. Overview
2. Research Themes
3. Modeling Entity-Based Coherence: Entity Grid
4. Modeling Discourse Structure: PDTB

What does discourse structure?

Discourse structures are patterns in text.
Different ways of thinking about discourse structure care about different types of elements.

- Entities
- Topics
- Functions
- Eventualities
- Coherence/Discourse/Rhetorical relations

Coreference resolution

- **Entity-level analysis**
- Linking references to common entities
- Cues: anaphoric expressions
  - pronouns
  - demonstratives (e.g. *this movie*)
  - alternate forms of reference (*President Obama, Barack Obama, Obama, President of the US*)
- Supervised learning models work reasonably well ... for English ... in certain types of texts ...

Local coherence: Centering theory

- **Local analysis (words/phrases in pairs of clauses/sentences)**
- Relationships between entities in adjacent utterances
- Coreference is an essential component
- Some small CT-annotated corpora exist
- CT has been used in CL for evaluating coherence

Entities and topical structure

Example

Gliders are aircraft which do not have a motor. They are sometimes called "sailplanes".
Gliders are controlled by their pilots by using control-sticks. Some gliders only carry one person, but some gliders can carry two persons...
Gliders cannot get into the air by themselves. They are pulled into the air by an aircraft with a motor or they are pulled up by motor on the ground.

- entity chains
- lexical cohesion
- lexical chains
- Entity Grid (Barzilay and Lapata)

Topics and structure

- **Text/text-passage level analysis**
- Concerned with *aboutness*
- Topics used to model structure
  - Topic models ~ underlying topics defined in terms of which words are used
  - Topic transitions often co-occur with document-internal boundaries
- Unsupervised models perform well

Functional structure

Different types of functional structure:

- Genre-related structure (e.g. scientific research papers)
- Conventionalized high-level functional structure (e.g. Wikipedia, news)
- Temporal structure
- Narrative structure
- Intentional structure (discourse relations)

Genre

- **Text-level analysis**
- Genre influences various aspects of texts
  - Structure
  - Themes and topics (but != domain)
  - Choice of vocabulary
  - Linguistic register/style
  - ....
- Many different classification schemes

Discourse modes

- **Text-passage level analysis**
- Following Smith 2003 *Modes of Discourse*:
  - Narrative
  - Description
  - Report
  - Information
  - Argument
- Discourse modes “do coherence” in different ways

Narrative structure

- **Eventuality level analysis**
  Structuring by eventualities (events, states, beliefs, etc.) and their spatio-temporal relations

Russian folk tale structure

- an *interdiction is addressed to the protagonist*, where the hero is told not to do something;
- *the interdiction is violated*, where the hero does it anyway;
- *the hero leaves home*, on a search or journey;
- *the hero is tested or attacked*, which prepares teh way for receiving a magic agent or helper.

Temporal structure

- **Eventuality level analysis**

TempEval: three tasks

In Washington today, the Federal Aviation Administration released air traffic control tapes from the night the TWA Flight eight hundred went down.

- Extracting and normalizing time expressions (aka Timex, time stamping)
- Extracting and classifying events
- Identifying temporal relations/links between time expressions and events

Discourse relations and structure

- **Clause/sentence/EDU-level analysis**
- Relations between clauses: causality, temporal structure, etc.
- Higher-level structure: discourse parse for entire texts
- Resources: corpora
  - Penn Discourse Treebank (PDTB)
  - Rhetorical Structure Theory (RST) Bank
  - DISCOR: texts labeled with SDRT structures

Why not address intentional structure?

Structure of discourse relations

Relations holding between the semantic content of two units of discourse.

Example

The kite was created in China, about 2800 years ago. *Later* it spread into other Asian countries, like India, Japan and Korea. *However*, the kite only appeared in Europe by about the year 1600.

- explicit vs. implicit relations
- unit of analysis (arguments)
- sense of the relation

Outline

1. Overview
2. Research Themes
3. Modeling Entity-Based Coherence: Entity Grid
4. Modeling Discourse Structure: PDTB

Entity Grid

Entirely automatic approach for modeling local coherence in a computationally-feasible way.

- Barzilay and Lapata 2008
- Converts text into a set of entity transition sequences
- Uses syntactic, referential, and distributional information/features

Definition

- A **local entity transition** is a sequence [S,O,X,-]*n* that represents entity occurrences and their syntactic roles in *n* adjacent occurrences.
- S=Subject, O=Object, X=Other arguments, -=not present
- Each transition has a probability: frequency of occurrence over total number of transitions of that length.

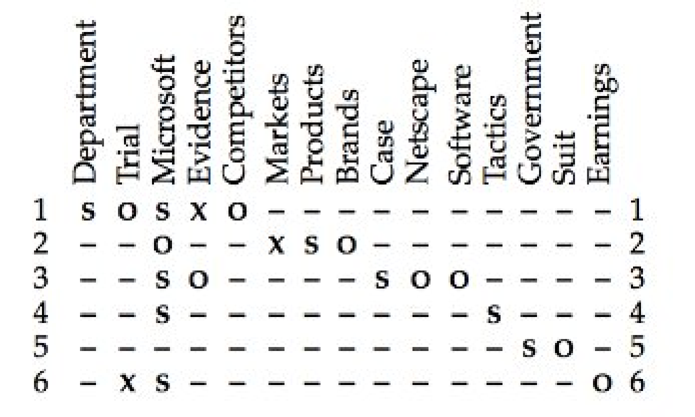
Entity Grid: Example

Entities in text marked with syntactic roles

1. [The justice department]-S is conducting an [anti-trust trial]-O against [Microsoft Corp.]-X with [evidence]-X that [the company]-S is increasingly attempting to crush [competitors]-O.
2. [Microsoft]-O is accused of trying to forcefully buy into [markets]-X where [its own products]-S are not competitive enough to unseat [established brands]-O.
3. [The case]-S revolves around [evidence]-O of [Microsoft]-S aggressively pressuring [Netscape]-O into merging [browser software]-O.
4. [Microsoft]-S claims [its tactics]-S are commonplace and good economically.
5. [The government]-S may file [a civil suit]-O ruling that [conspiracy]-S to curb [competition]-O through [collusion]-X is [a violation of the Sherman Act]-O.
6. [Microsoft]-S continues to show [increased earnings]-O despite [the trial]-X.

Entity Grid: Example

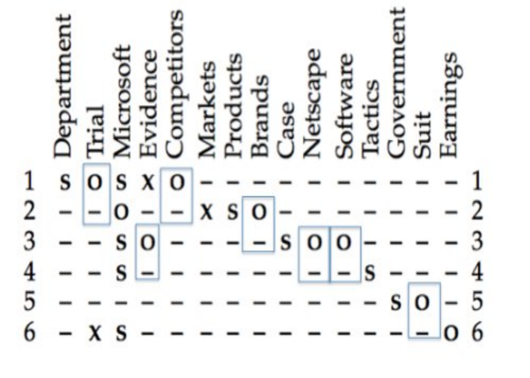
Grid shows which entities occur where, with which role



	Department	Trial	Microsoft	Evidence	Competitors	Markets	Products	Brands	Case	Netscape	Software	Tactics	Government	Suit	Earnings	
1	s	o	s	x	o	–	–	–	–	–	–	–	–	–	–	1
2	–	–	o	–	–	x	s	o	–	–	–	–	–	–	–	2
3	–	–	s	o	–	–	–	–	s	o	o	–	–	–	–	3
4	–	–	s	–	–	–	–	–	–	–	–	s	–	–	–	4
5	–	–	–	–	–	–	–	–	–	–	–	–	s	o	–	5
6	–	x	s	–	–	–	–	–	–	–	–	–	–	–	o	6

Entity Grid: Example

Probability of each transition type is computed, then used as feature



	Department	Trial	Microsoft	Evidence	Competitors	Markets	Products	Brands	Case	Netscape	Software	Tactics	Government	Suit	Earnings	
1	S	O	S	X	O	–	–	–	–	–	–	–	–	–	–	1
2	–	–	O	–	–	X	S	O	–	–	–	–	–	–	–	2
3	–	–	S	O	–	–	–	–	S	O	O	–	–	–	–	3
4	–	–	S	–	–	–	–	–	–	–	–	S	–	–	–	4
5	–	–	–	–	–	–	–	–	–	–	–	–	S	O	–	5
6	–	X	S	–	–	–	–	–	–	–	–	–	–	–	O	6

Probability of [O-] = 7/75 = 0.093

Evaluation

Entity grid approach is evaluated in three applications:

1. Information ordering
2. Evaluation of summary coherence
3. Readability assessment

Outline

1. Overview

2. Research Themes

3. Modeling Entity-Based Coherence: Entity Grid

4. Modeling Discourse Structure: PDTB

Penn Discourse Treebank

Corpus of texts from the Wall Street Journal annotated with discourse relations. Has enabled much research, both empirical analysis and development of systems for automatic analysis.

Example

Slides from Nikos Bampounis