Automatic Textprocessing

Michael Strube
michael.strube @ h-its.org

November 5, 2018

Description

A text is more than a sequence of sentences. To understand a text, one needs to recognize how sentences are connected with each other and why they appear in a particular order. To capture the particular characteristics of text, we developed a range of methods in computational linguistics: local and global coherence models, anaphora and coreference resolution algorithms, methods for recognizing the rhetorical, the temporal, the causal and the argumentative structure of texts. In the module we first talk about linguistically well-founded classical models. Then we continue with more recent machine learning and graph based models before we turn to current neural models for text processing and understanding. The usefulness of such models can only be determined when they are integrated into applications. Hence the module will also deal with evaluating text processing algorithms within applications such as information extraction, machine translation, question answering, readability prediction, essay scoring, etc.

References


selected chapters from:


Journal Papers


Dates, Topics

23.10.2018
Introduction, historical overview . . .
(Stede, 2012; Webber et al., 2012; Webber & Joshi, 2012)
Introduction into discourse structure
(Grosz & Sidner, 1986; Morris & Hirst, 1991; Passonneau & Litman, 1997)

Introduction into local coherence analysis: Information status, coreference

Introduction into centering
(Brennan et al., 1987; Grosz et al., 1995; Tetreault, 2001; Poesio et al., 2004)

30.10.2018
Cancelled, MS @ EMNLP

06.11.2018
Cancelled, MS @ AIPHES Retreat

13.11.2018
Local coherence: Information status
(Nissim et al., 2004; Nissim, 2006; Cahill & Riester, 2009; Riester et al., 2010; Riester & Baumann, 2011; Rahman & Ng, 2011a; Cahill & Riester, 2012; Markert et al., 2012; Eckart et al., 2012; Baumann & Riester, 2013; Hou, 2016)

The structure of documents in different genres
(Teufel & Moens, 2002; Power et al., 2003; Sporleder & Lapata, 2004; Graham et al., 2005; Filippova & Strube, 2006; Chung, 2009; Teufel et al., 2009; Liakata et al., 2010; 2012)

Document structure: Automatic summarizaton
(Liakata et al., 2012; Contractor et al., 2012; Liakata et al., 2013)

to be prepared: info status (Markert et al., 2012) or (Teufel & Moens, 2002, pp.409-424) or (Teufel et al., 2009)

20.11.2018
Cancelled, MS @ HITS GL Retreat

27.11.2018
Topic segmentation: Lexical chains – (Morris & Hirst, 1991; Medelyan, 2007)

optional: (Barzilay & Elhadad, 1997; Hirst & St-Onge, 1998; Barzilay & Elhadad, 1999; Silber & McCoy, 2002; Galley et al., 2003; Stokes et al., 2004; Ye & Chua, 2006;
Ye et al., 2007)

**Topic-Segmentation: Distributional approaches**
(Hearst, 1997; Beeferman et al., 1999; Reynar, 1999; Choi, 2000; Pevzner & Hearst, 2002)

**Probabilistic models for topic segmentation**
(Utiyama & Isahara, 2001; Eisenstein & Barzilay, 2008; Shafiei & Milios, 2008; Chen et al., 2009; Purver, 2011; Du et al., 2013; Simon et al., 2013)

to be prepared:

**4.12.2018**
Cancelled, MS @ HITS SAB Meeting

**11.12.2018**
Attention: MS @ HITS Shareholders Meeting

**Modeling local coherence: Entity Grid with applications (information ordering, evaluating the quality of summaries)**
(Barzilay & Lapata, 2008)

**Modeling local coherence: Entity graph with applications (information ordering, evaluating the quality of summaries)**
(Guinaudeau & Strube, 2013; Petersen et al., 2015)

**Modeling local coherence: Neural models with applications**
(Li & Hovy, 2014; Tien Nguyen & Joty, 2017; Li & Jurafsky, 2017; Joty et al., 2018; Mesgar & Strube, 2018)

**Applications of local coherence models: Information ordering, evaluating the quality of summaries**
(Karamanis et al., 2004; Barzilay & Lapata, 2005; Karamanis, 2007; Elsner et al., 2007; Filippova & Strube, 2007; Karamanis et al., 2009; Pitler et al., 2010; Elsner & Charniak, 2011; Lin et al., 2012; Li & Hovy, 2014)

**Applications of local coherence models: Information retrieval**
(Petersen et al., 2015)

**Applications of local coherence models: Readability** – (Higgins et al., 2004; Pitler & Nenkova, 2008; Feng et al., 2009; Wang et al., 2013; Mesgar & Strube, 2015; 2016; 2018)

**Applications of local coherence models: Essay scoring** – (Miltsakaki & Kukich, 2004; Burstein et al., 2010; Beigman Klebanov & Flor, 2013; Beigman Klebanov et al., 2014; Somasundaran et al., 2014; Mesgar & Strube, 2018)
Applications of local coherence models: Coreference and coherence for automatic translation
(Hardmeier et al., 2013; 2015; Born et al., 2017; Jean et al., 2017)

to be prepared:

18.12.2018
Global coherence: Argumentation structure: Marinco Möbius
(Reed & Long, 1998; Katzav & Reed, 2008; Stab & Gurevych, 2014; Peldszus & Stede, 2015a; 2015b; Habernal & Gurevych, 2017; Stab & Gurevych, 2017; Wachsmuth et al., 2017)

Global coherence: Rhetorical Structure Theory (RST)
(Marcu & Echihabi, 2002; Carlson et al., 2003; Soricut & Marcu, 2003; duVerle & Prendinger, 2009; Subba & Di Eugenio, 2009; Hernault et al., 2012; Feng & Hirst, 2012; Ji & Eisenstein, 2014)

Global coherence: Penn Discourse Treebank
(Prasad et al., 2008; Piter & Nenkova, 2009a; 2009b; Lin et al., 2009; 2010; 2011; Ghosh et al., 2012; Lin et al., 2014; Prasad et al., 2014; Ji et al., 2015; Xue et al., 2016; Mihailov & Frank, 2016)

Global coherence: Applications (automatic summarization, sentiment analysis)
(Marcu, 1997; Maslennikov & Chua, 2007; Clarke & Lapata, 2010; Zirn et al., 2011; Bhatia et al., 2015)

to be prepared:

08.01.2019
Local coherence: Rule-based system for coreference resolution: Ozan Yılmaz
(Raghunathan et al., 2010; Lee et al., 2011; 2013)

Local coherence: Machine learning for coreference resolution: (Soon et al., 2001; Ng & Cardie, 2002; Yang et al., 2008; Ng, 2008; 2010; Fernandes et al., 2012; Durrett & Klein, 2013; Fernandes et al., 2014; Martschat & Strube, 2015)

Local coherence: Deep learning for coreference resolution: (Clark & Manning, 2015; Wiseman et al., 2015; 2016; Clark & Manning, 2016b; 2016a; Lee et al., 2017; Kundu et al., 2018; Zhang et al., 2018)

Local coherence: Evaluating coreference resolution algorithms
(Vilain et al., 1995; Bagga & Baldwin, 1998; Popescu-Belis, 2003; Luo & Zitouni, 2005; Recasens & Hovy, 2011; Pradhan et al., 2014; Tuggener, 2014; Moosavi & Strube, 2016)

Local coherence: Error analysis for coreference resolution – (Uryupina, 2008; Kummerfeld & Klein, 2013; Martschat & Strube, 2014)
Local coherence: Critique on coreference resolution systems – (Moosavi & Strube, 2017; 2018)

to be prepared:

15.01.2019
Local coherence: Coreference resolution in other languages, multilingual coreference resolution – (Kong & Ng, 2013)

optional: (Luo & Zitouni, 2005; Recasens & Martí, 2009; Chen & Ng, 2013; Kong & Ng, 2013; Martins, 2015)

Local coherence: Current work on coreference resolution
(Raghunathan et al., 2010; Sapena et al., 2010; Cai & Strube, 2010; Lee et al., 2011; Rahman & Ng, 2011b; Fernandes et al., 2012; Lee et al., 2013; Fernandes et al., 2014; Durrett & Klein, 2013; 2014; Björkelund & Kuhn, 2014; Martschat & Strube, 2015; Wiseman et al., 2015; Clark & Manning, 2015)

to be prepared: (Lee et al., 2011) or (Fernandes et al., 2012)

22.01.2019
Bias in coreference resolution
(Rudinger et al., 2018; Zhao et al., 2018; Webster et al., 2018)

Local coherence: Bridging
(Clark, 1975; Hahn et al., 1996; Poesio et al., 1997; Vieira & Teufel, 1997; Vieira & Poesio, 2000; Bunescu, 2003; Fan et al., 2005; Lassalle & Denis, 2011; Hou et al., 2013b; 2013a; Rösiger & Teufel, 2014; Hou et al., 2014; 2018; Hou, 2018b; 2018a)

Local coherence: Event coreference resolution
(Bejan & Harabagiu, 2010; Chen & Ji, 2009; Chen et al., 2010b; 2010a; Goyal et al., 2013)

29.01.2019
Cancelled, MS @ FAT*/AAAI

05.02.2019
Cancelled, MS @ AIPHES Review

Optional:
Hierarchical topic segmentation
Applications of topic segmentation: Automatic summarization
(Goldstein et al., 2000; Teufel & Moens, 2002; Narayanan & Harabagiu, 2004; Stokes et al., 2004)

Global coherence: Temporal structure
(Lapata & Lascarides, 2004; 2006; Ng et al., 2013)

Introduction into global coherence analysis
(Hobbs, 1979; 1985; Mann & Thompson, 1988; Knott & Dale, 1994; Webber & Joshi, 1998; Kehler et al., 2008; Webber et al., 2012)

Further Remarks:

Assessment:

1. For each class read the material marked in the schedule as to be prepared. Formulate two questions about the material and send them to me via email until the Monday before the class, 1pm at the latest. Participate actively in the class (important!).

2. After each class write a review (approx. one page) about one of the papers presented. A review consists of a concise summary, a critique of the research being presented, and, if possible, remarks for improvement.

3. Choose a topic in the schedule you want to present in class. Select one or more papers from the reading list. Present this work in the class (30 minutes presentation, 15 minutes discussion).

4. Write a report/an essay towards the end of the term either about the topic you presented in class or about a new topic (6LP: 8-10 pages; 8LP: 12-15 pages). Or: Implement a discourse processing algorithm, evaluate it in comparison to a baseline, and write a very short report (3-4 pages) about it.

Literature: Most papers can be downloaded from the ACL Anthology (http://acl.ldc.upenn.edu/), in particular all papers presented at (E/NA)ACL, Coling and EMNLP conferences, all workshops organized during these conferences and the journals TACL and Computational Linguistics. Papers published through AAAI (AAAI conference, AAAI workshops, AAAI symposia, etc.) are available through the AAAI Digital Library verfügbar (http://www.aaai.org/Library). Other journals are available electronically at the university library (https://www.ub.uni-heidelberg.de/, http://rzblx1.uni-regensburg.de/ezeit/search.phtml?bibid=UBHE&colors=3&lang=de).

Office hours: Right after class or in my office at HITS (https://www.h-its.org/en/).
References


Du, Lan, Wray Buntine & Mark Johnson (2013). Topic segmentation with a structured topic


Kong, Fang & Hwee Tou Ng (2013). Exploiting zero pronouns to improve Chinese corefer-


Pitler, Emily & Ani Nenkova (2008). Revisiting readability: A unified framework for predict-


