Computational Discourse Modeling

LING-765 (Spring 2019)
Mon+Wed, 11:00-12:15, ICC 117

NOTE: We will also have class on **Tuesday, February 19**

Instructor:

Amir Zeldes (Asst. Prof. of Computational Linguistics)
E-Mail: amir.zeldes@georgetown.edu
Website: http://corpling.uis.georgetown.edu/amir
Office: Poulton Hall 243

Summary:

Recent years have seen an explosion of computational work on higher level discourse representations, such as entity recognition, mention and coreference resolution and (shallow) discourse parsing. At the same time, the theoretical status of the underlying categories is not well understood, and despite progress, these tasks remain very much unsolved in practice. This graduate level seminar will concentrate on theoretical and practical models representing how discourse unfolds across sentences as it grows. We will explore cohesion in text by means of discourse relations (e.g. expressing causality, contrastivity), the use of recurring referring expressions, such as mentions of people, things and events, and how these are coded during language processing. We will also study multiple levels of discourse processing in terms of information structure, discourse relations and theories about anaphora, such as Centering Theory and Alternative Semantics. We will then look at computational linguistics implementations of systems for entity recognition, coreference resolution and discourse parsing and explore their relationship with linguistic theory. Over the course of the semester, participants will implement their own coding project exploring some phenomenon within the domain of entity recognition, coreference, discourse relations or a related area. Intermediate programming skills (e.g. in Python) are required, and a previous computational course such as Intro to NLP (LING-362) or Computational Corpus Linguistics (LING-367) is recommended.

Course requirements:

Attendance	
Final project	40%
Assignments	40%
Presentations	10%
Participation	10%

Assignments and final project:

Assignments will include programming assignments, possibly including a brief writing assignment describing the approach. There will be two types of presentations: a discussion of a relevant article in one of the topics being discussed (some suggestions will be provided) and presentations of documented code produced by the students. I encourage some of the coding work to be done jointly with fellow students, as long as individual contributions are clearly delineated. The final project will be an independent toy implementation of a discourse processing module, accompanied by a paper in the ACL format (4-8 pages, 2 column layout, see ACL proceedings), including a summary literature review, description of the approach, and evaluation on some dataset.

Absences and timely assignment submission:
Students are expected to attend all classes and to complete all assignments on time. Absences may have an adverse effect on grades in a course, up to and including failure. That said, students may excuse themselves via e-mail from up to three meetings at their discretion, provided that they make up for lost course work and submit the assignments. Any additional absences for special reasons (religious observances, athletic travel, prolonged illness etc.) may be coordinated on a case by case basis with the instructor (documentation may be required as applicable).

Participation in the 2019 shared task on discourse unit segmentation:
This semester will coincidentally overlap with the deadline for the 2019 DISRPT shared task on discourse unit segmentation. In discussion with course members, we may decide to participate in the shared task; this will then tentatively substitute our first two coding projects and alter the course schedule (optional)

Course plan: (approximate)

Approximate and tentative plan (each participant should plan to present one of the papers below or a related one). If we participate in the shared task, **weeks 5-7 may be swapped with 11-14**.

Week 1 – Introduction
Weeks 2-4 – Tiling and (Sub-)Topics

- Possible readings: Hearst (1997), Pevzner & Hearst (2002), Teufel & Moens (2002), articles from Gruber & Redeker (2014)
- Project: Tiling sub-module or discourse unit segmentation (if shared task is chosen)

Weeks 5-7 – Centering and Coherence

- Possible readings: Grosz et al. (1995), Poesio et al. (2004), Krifka (2008), Kehler & Rohde (2013), Spalek & Zeldes (2015)

- Project: Centering-based cohesion tracker or discourse unit segmentation (if shared task is chosen)

Weeks 8-10 – Entities and Coreference

- Possible readings: Recasens et al. (2010), Lee et al. (2013), Pradhan et al. (2014), Zeldes & Zhang (2016), Ma & Hovy (2016),
- Basic coreference model for language/variety using the *xrenner* coreferencer

Weeks 11-14 – Discourse Relations (RST, PDTB, SDRT)

- Possible readings: Mann & Thompson (1988), Marcu et al. (1999), Prasad et al. (2008), Surdeanu et al. (2015), Stede et al. (2016)
- Project TBD (some ideas: relation mapping [Benamara & Taboada 2015], graph simplification, modules for segmentation/classification)

Week 15 – Conclusion

Bibliography:

Braud, Chloe, Maximin Coavoux & Anders Søgaard (2017), Cross-lingual RST Discourse Parsing. In: *Proceedings of EACL 2017*. Valencia, Spain, 292–304.

Braud, Chloe, Barbara Plank & Anders Søgaard (2016), Multi-View and Multi-Task Training of RST Discourse Parsers. In: *Proceedings of COLING 2016*. Osaka, 1903–1913.

Carlson, Lynn, Daniel Marcu & Mary Ellen Okurowski (2001), Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. *Proceedings of 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001*. Aalborg, Denmark, 1–10.

Durrett, Greg & Dan Klein (2013), Easy Victories and Uphill Battles in Coreference Resolution. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*. Seattle, WA, 1971–1982.

Grosz, Barbara J., Aravind K. Joshi & Scott Weinstein (1995), Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics* 21(2), 203–225.

Gruber, Helmut & Gisela Redeker (eds.) (2014), *The Pragmatics of Discourse Coherence*. (Pragmatics and Beyond New Series 254.) Amsterdam and Philadelphia: John Benjamins.

Hayashi, Katsuhiko, Tsutomu Hirao & Masaaki Nagata (2016), Empirical Comparison of Dependency Conversions for RST Discourse Trees. In: *Proceedings of SIGDIAL 2016*. Los Angeles, CA, 128–136.

Hearst, Marti A. (1997), TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics* 23(1), 33–64.

Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw & Ralph Weischedel (2006), OntoNotes: The 90% Solution. *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. New York: ACL, 57–60.

Ji, Yangfeng & Jacob Eisenstein (2014), Representation Learning for Text-level Discourse Parsing. In: *Proceedings of ACL 2014*. Baltimore, MD, 13–24.

Kehler, Andy & Hannah Rohde (2013), *A Probabilistic Reconciliation of Coherence-driven and Centering-driven Theories of Pronoun Interpretation*. 39(1-2), 1–37.

Krifka, Manfred (2008), Basic Notions of Information Structure. *Acta Linguistica Hungarica* 55, 243–276.

Lee, Heeyoung, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu & Dan Jurafsky (2013), Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules. *Computational Linguistics* 39(4), 885–916.

Mann, William C. & Sandra A. Thompson (1988), Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text* 8(3), 243–281.

Marcu, Daniel, Estibaliz Amorrortu & Magdalena Romera (1999), Experiments in Constructing a Corpus of Discourse Trees. *Proceedings of the ACL Workshop Towards Standards and Tools for Discourse Tagging*. College Park, MD, 48–57.

Morey, Mathieu, Philippe Muller & Nicholas Asher (2017), How Much Progress have we Made on RST Discourse Parsing? A Replication Study of Recent Results on the RST-DT. In: *Proceedings of EMNLP 2017*. Copenhagen, Denmark, 1319–1324.

Pevzner, Lev & Marti A. Hearst (2002), A Critique and Improvement of an Evaluation Metric for Text Segmentation. *Computational Linguistics* 28, 1–19.

Poesio, Massimo, Rosemary Stevenson, Barbara Di Eugenio & Janet Hitzeman (2004), Centering: A Parametric Theory and Its Instantiations. *Computational Linguistics* 30(3), 309–363.

Pradhan, Sameer, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng & Michael Strube (2014), Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, MD, 30–35.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi & Bonnie Webber (2008), The Penn Discourse Treebank 2.0. *Proceedings LREC 2008*. Marrakech, Morocco.

Recasens, Marta, Eduard Hovy & M. Antònia Martí (2010), A Typology of Near-Identity Relations for Coreference (NIDENT). *Proceedings of LREC 2010*. Valletta, Malta, 149–156.

Recasens, Marta, Marie-Catherine de Marneffe & Christopher Potts (2013), The Life and Death of Discourse Entities: Identifying Singleton Mentions. *Proceedings of NAACL 2013*. Atlanta, GA, 627–633.

Renkema, Jan (2009), *The Texture of Discourse. Towards an Outline of Connectivity Theory*. Amsterdam and Philadelphia: John Benjamins.

Spalek, Katharina & Amir Zeldes (2015), Converging Evidence for the Relevance of Alternative Sets: Data from NPs with Focus Sensitive Particles in German. *Language and Cognition*.

Stede, Manfred (2012), *Discourse Processing*. (Synthesis Lectures on Human Language Technologies 4.) San Rafael, CA: Morgan & Claypool.

Stede, Manfred, Stergos Afantenos, Andreas Peldszus, Nicholas Asher & Jérémy Perret (2016), Parallel Discourse Annotations on a Corpus of Short Texts. *Proceedings of LREC 2016*. Portorož, Slovenia, 1051–1058.

Surdeanu, Mihai, Thomas Hicks & Marco A. Valenzuela-Escarcega (2015), Two Practical Rhetorical Structure Theory Parsers. *Proceedings of NAACL-HLT 2015*. Denver, CO, 1–5.

Teufel, Simone & Marc Moens (2002), Summarising Scientific Articles - Experiments with Relevance and Rhetorical Status. *Computational Linguistics* 28(4), 409–445.

Zeldes, Amir & Shuo Zhang (2016), When Annotation Schemes Change Rules Help: A Configurable Approach to Coreference Resolution beyond OntoNotes. *Proceedings of the NAACL2016 Workshop on Coreference Resolution Beyond OntoNotes (CORBON)*. San Diego, CA, 92–101.

Notice regarding sexual misconduct:
Please know that as a faculty member I am committed to supporting survivors of sexual misconduct, including relationship violence, sexual harassment and sexual assault. University policy also requires me to report any disclosures about sexual misconduct to the Title IX Coordinator, whose role is to coordinate the University's response to sexual misconduct.

Georgetown has a number of fully confidential professional resources who can provide support and assistance to survivors of sexual assault and other forms of sexual misconduct. These resources include:

Jen Schweer, MA, LPC
Associate Director of Health Education Services for Sexual Assault Response and Prevention
(202) 687-0323
jls242@georgetown.edu

Erica Shirley, Trauma Specialist
Counseling and Psychiatric Services (CAPS)
(202) 687-6985
els54@georgetown.edu

More information about campus resources and reporting sexual misconduct can be found at http://sexualassault.georgetown.edu.