Computational Discourse Modeling

(Tentative plan, subject to change)

LING-8415 (Spring 2024)
Mon+Wed, 2:00-3:15, Room TBD

Instructor:
Amir Zeldes
E-Mail: amir.zeldes@georgetown.edu
Website: https://gucorpling.org/amir/
Office: Poulton Hall 243 (office hours Wednesdays, 3:30-5:00)

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, including classic theories as Centering Theory and Alternative Semantics, and newly developing ones such as Question Under Discussion (QUD) trees. With these in mind we will look at computational implementations of systems for entity recognition, coreference resolution and discourse parsing and explore their relationship with linguistic theory and textual coherence, including symbolic, neural and prompt-based approaches using LLMs. Over the course of the semester, participants will implement their own project exploring some phenomenon within the domain of discourse processing. Intermediate programming skills (preferably in Python) are required, and a previous computational course such as Intro to NLP (LING-4400 or higher) or Computational Corpus Linguistics (LING-4427) 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, but students may suggest papers as well) 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 is usually an independent 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. Students often turn their papers into workshop or conference paper after the course (this is encouraged but not required by the class in any way).

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).

Course plan: (approximate)

Approximate and tentative plan (each participant should plan to present one of the papers below or a related one). The course plan is very flexible as this is a graduate level seminar – we can choose to tackle discourse relations before entities, skip and/or replace topics etc.

Week 1 – Introduction
Weeks 2-4 – Tiling, topics and hierarchical document models

- Possible readings: Hearst (1997), Pevzner & Hearst (2002), Teufel & Moens (2002), articles from Gruber & Redeker (2014), Xing & Carenini (2021), Xu et al. (2021)
- Project ideas: Tiling sub-module, discourse unit segmentation

Weeks 5-7 – Centering, Cohesion and Coherence

- Possible readings: Grosz et al. (1995), Poesio et al. (2004), Krifka (2008), Kehler & Rohde (2013), Spalek & Zeldes (2015), Friedrich & Palmer (2014)

- Projects: Centering-based cohesion tracking, recursive topic splitting, subsequent mention modeling

Weeks 8-10 – Entities, Anaphora and Coreference

- Possible readings: Recasens et al. (2010), Lee et al. (2013), Pradhan et al. (2014), Zeldes & Zhang (2016), Ma & Hovy (2016), Yu et al. (2020), Wu et al. (2020), Yuan et al. (2021), Zeldes (2022)…
- Projects: referring expression generation, basic coreference model for an interesting language, bridging resolution…

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

- Possible readings: Mann & Thompson (1988), Marcu et al. (1999), Prasad et al. (2008), Surdeanu et al. (2015), Stede et al. (2016), Gessler et al. (2021), Ko et al. (2021), Wu et al. (2023)…
- Projects: shallow or deep discourse parsing, relation mapping, dependency vs. constituent tree conversions, 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.

Friedrich, Annemarie & Alexis Palmer (2014), Centering Theory in natural text: a large-scale corpus study. *Proceedings of KONVENS 2014*.

Gessler, Luke, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes (2021). DisCoDisCo at the DISRPT2021 Shared Task: A System for Discourse Segmentation, Classification, and Connective Detection. In *Proceedings of Discourse Relation Parsing and Treebanking 2021 (DISRPT 2021)*, Punta Cana, Dominican Republic.

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.

Kenton Lee, Luheng He, Mike Lewis & Luke Zettlemoyer (2017). End-to-end Neural Coreference Resolution. *Proceedings of EMNLP 2017*, Copenhagen, 188–197.

Ko, Wei-Jen, Cutter Dalton, Mark Simmons, Eliza Fisher, Greg Durrett, Junyi Jessy Li (2021), Discourse Comprehension: A Question Answering Framework to Represent Sentence Connections. arXiv:2111.00701.

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.

Wu, Wei, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li (2020) CorefQA: Coreference Resolution as Query-based Span Prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6953–6963.

Wu, Yating, Ritika Mangla, Greg Durrett and Junyi Jessy Li (2023) QUDeval: The Evaluation of Questions Under Discussion Discourse Parsing . In: *Proceedings of EMNLP 2023*.

Xing, Linzi & Giuseppe Carenini (2021), Improving Unsupervised Dialogue Topic Segmentation with Utterance-Pair Coherence Scoring. *Proceedings of SIGDIAL 2021*.

Xu, Yi, Hai Zhao & Zhuosheng Zhang (2021), Topic-Aware Multi-turn Dialogue Modeling. *Proceedings of AAAI 2021*.

Yu, Juntao, Bernd Bohnet & Massimo Poesio (2020). Named Entity Recognition as Dependency Parsing. *Proceedings of ACL 2020*.

Yuan, Zheng, Chuanqi Tan, Songfang Huang & Fei Huang (2021). Fusing Heterogeneous Factors with Triaffine Mechanism for Nested Named Entity Recognition. arXiv:2110.07480v1.

Zeldes, Amir (2022), Can We Fix the Scope for Coreference Resolution? *Dialogue and Discourse*.

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 Title IX/Sexual Misconduct:

Georgetown University and its faculty are committed to supporting survivors and those impacted by sexual misconduct, which includes sexual assault, sexual harassment, relationship violence, and stalking. Georgetown requires faculty members, unless otherwise designated as confidential, to report all disclosures of sexual misconduct to the University Title IX Coordinator or a Deputy Title IX Coordinator. If you disclose an incident of sexual misconduct to a professor in or outside of the classroom (with the exception of disclosures in papers), that faculty member must report the incident to the Title IX Coordinator, or Deputy Title IX Coordinator. The coordinator will, in turn, reach out to the student to provide support, resources, and the option to meet. [Please note that the student is not required to meet with the Title IX coordinator.]. More information about reporting options and resources can be found on the Sexual Misconduct Website: https://sexualassault.georgetown.edu/resourcecenter.

If you would prefer to speak to someone confidentially, Georgetown has a number of fully confidential professional resources that can provide support and assistance. These resources include:
Health Education Services for Sexual Assault Response and Prevention: confidential email sarp@georgetown.edu

Counseling and Psychiatric Services (CAPS): 202.687.6985 or after hours, call (833) 960-3006 to reach Fonemed, a telehealth service; individuals may ask for the on-call CAPS clinician

More information about reporting options and resources can be found on the Sexual Misconduct Website.