Discourse Analysis and Its Applications

Shafiq Joty*, Giuseppe Carenini†, Raymond T. Ng†, and Gabriel Murray§
*Nanyang Technological University, Salesforce Research Asia, Singapore
†University of British Columbia, Vancouver, Canada
§University of the Fraser Valley, Abbotsford, Canada
*srjoty@ntu.edu.sg, §gabriel.murray@ufv.ca
†{carenini,rng}@cs.ubc.ca

Abstract

Discourse processing is a suite of Natural Language Processing (NLP) tasks to uncover linguistic structures from texts at several levels, which can support many downstream applications. This involves identifying the topic structure, the coherence structure, the coreference structure, and the conversation structure for conversational discourse. Taken together, these structures can inform text summarization, machine translation, essay scoring, sentiment analysis, information extraction, question answering, and thread recovery. The tutorial starts with an overview of basic concepts in discourse analysis – monologue vs. conversation, synchronous vs. asynchronous conversation, and key linguistic structures in discourse analysis. We also give an overview of linguistic structures and corresponding discourse analysis tasks that discourse researchers are generally interested in, as well as key applications on which these discourse structures have an impact.

1 Motivation

Discourse analysis has been a fundamental problem in the ACL community, where the focus is to develop tools to automatically model language phenomena that go beyond the individual sentences. With the ongoing neural revolution, as the methods become more effective and flexible, analysis and interpretability beyond the sentence-level is of particular interests for many core language processing tasks like language modeling (Ji et al., 2016) and applications such as machine translation and its evaluation (Sennrich, 2018; Läubli et al., 2018; Joty et al., 2017), text categorization (Ji and Smith, 2017), and sentiment analysis (Nejat et al., 2017). With the advent of Internet technologies, new forms of discourse are emerging (e.g., emails and discussion forums) with novel set of challenges for the computational models. Furthermore, most computational models for discourse analysis are also going through a paradigm shift from traditional statistical models to deep neural models. Considering all these novel aspects at once, this tutorial is quite timely for the community, by providing the attendees with an up-to-date, critical overview of existing approaches and their evaluations, applications, and future challenges.

2 Tutorial Outline

We start with an overview of basic concepts in discourse analysis – monologue vs. conversation, synchronous vs. asynchronous conversation, and key linguistic structures in discourse analysis. Attendees then get to learn about coherence structure and discourse parsers. We give a critical overview of different discourse theories, and available datasets annotated according to these formalisms. We cover methods for RST- and PDTB-style discourse parsing. We cover traditional methods along with the most recent works using deep neural networks, interpret them and compare their performances on benchmark datasets.

Next, we discuss coherence models to evaluate monologues and conversations based on their coherence. We then show applications (evaluation tasks) of coherence models and discourse parsers. Special attention is paid to the new emerging applications of discourse analysis such as machine translation and its evaluation, sentiment analysis, and abstractive summarization.

In the final part of the tutorial, we cover conversational structures (*e.g.,* speech acts, thread structure), computational methods to extract such structures, and their utility in downstream applications (*e.g.,* conversation summarization). Again, evaluation metrics and approaches will be discussed and compared. We conclude with an interactive discussion of future challenges for discourse analysis and its applications. In the following, we give a detailed breakdown of the tutorial content.

A. Introduction [25 mins]

1. Discourse & its different forms
   - (a) Monologue
   - (b) Synchronous & asynchronous conversations
   - (c) Modalities: written & spoken
2. Two discourse phenomena
   - (a) Coherence
   - (b) Cohesion
3. Linguistic structures in discourse & discourse analysis tasks
   - (a) Coherence structure ⇒ Discourse segmentation & parsing
   - (b) Coherence models ⇒ Coherence evaluation
   - (c) Topic structure ⇒ Topic segmentation & labeling [not covered in this tutorial]
   - (d) Coreference structure ⇒ Coreference resolution [not covered in this tutorial]
   - (e) Conversational structure ⇒ Disentanglement & reply-to structure, speech act recognition
4. Applications of discourse analysis

B. Coherence Structure, Corpora & Discourse Parsing [45 mins]

1. Discourse theories & coherence relations
   - (a) Rhetorical Structure Theory (RST) & RST Treebank (Carlson et al., 2002) & Instructional domain (Subba and Di Eugenio, 2009)
   - (b) Discourse Lexicalized Tree Adjoining Grammar (D-LTAG) & Penn Discourse Treebank (PDTB) (Prasad et al., 2005)
2. Discourse connectives & unsupervised relation identification
   - (a) Role of connectives in RST & PDTB
   - (b) Identifying discourse connectives
   - (c) Implicit and explicit relations
3. Discourse parsing in RST
   - (a) The tasks: discourse segmentation and parsing
   - (b) Role of syntax
   - (c) Traditional models – SPADE (Soricut and Marcu, 2003), HILDA (duVerle and Prendinger, 2009), CODRA (Joty et al., 2015), CRF-based model (Feng and Hirst, 2014).
   - (d) Neural models (Ji and Eisenstein, 2014; Li et al., 2014, 2016; Morey et al., 2017)
   - (e) State-of-the-Art (Wang et al., 2017; Lin et al., 2019)
   - (f) Evaluation & Discussion
4. Discourse parsing in PDTB
   - (a) The tasks: relation sense identification and scope disambiguation
   - (b) Statistical models (Pitler and Nenkova, 2009; Ziheng et al., 2014)
   - (c) Neural models (Ji and Eisenstein, 2015; Lan et al., 2017)
   - (d) Evaluation & Discussion
5. Final remarks
   - (a) Tree vs. graph structure
   - (b) Discourse Graphbank

C. Coffee Break [15 mins]

D. Coherence Models & Applications of Discourse [45 mins]

1. Overview of coherence models
   - (a) Entity grid and its extensions (Barzilay and Lapata, 2008; Elsner and Charniak, 2011b; Guinaudeau and Strube, 2013)
   - (b) Discourse relation based model (Lin et al., 2011; Pitler and Nenkova, 2008)
   - (c) Neural coherence models (Mohiuddin et al., 2018; Li and Jurafsky, 2017; Mesgar and Strube, 2018)
   - (d) Coherence models for conversations (Elsner and Charniak, 2011a; Mohiuddin et al., 2018)
2. Evaluation tasks
   - (a) Sentence ordering (Discrimination, Insertion)
   - (b) Summary coherence rating
   - (c) Readability assessment

    (d) Chat disentanglement
    (e) Thread reconstruction

3. Applications of discourse
    (a) Summarization
    (b) Generation
    (c) Sentiment analysis
    (d) Machine translation

E. Conversational Structure [35 mins]

1. Conversational structures
    (a) Speech (or dialog) acts in synchronous and asynchronous conversations
    (b) Reply-to (thread) structure in asynchronous conversations (Carenini et al., 2007)
    (c) Conversation disentanglement in synchronous conversations
2. Computational models
    (a) Speech act recognition models (Stolcke et al., 2000; Cohen et al., 2004; Ritter et al., 2010; Joty et al., 2011; Paul, 2012; Joty and Hoque, 2016; Mohiuddin et al., 2019)
    (b) Thread reconstruction models (Shen et al., 2006; Wang et al., 2008, 2011a,b)
    (c) Conversation disentanglement models (Elsner and Charniak, 2008, 2011a)
3. Evaluation & Summary of results

F. Future Challenges [15 mins]

1. Learning from limited annotated data
2. Language & domain transfer
3. Discourse generation
4. New emerging applications

Link to the Slides Our tutorial slides will be made available at https://ntunlpsg.github.io/project/acl19tutorial/

2.1 Prerequisites

Prior knowledge in basic machine learning, NLP (*e.g.,* parsing methods, machine translation), and deep learning models is essential to understand the content of this tutorial.

2.2 Similar Tutorial

We gave a similar tutorial (shorter version) at the 2018 IEEE International Conference on Data Mining (ICDM-2018), a top conference in data mining. The slides of that tutorial can be found at https://ntunlpsg.github.io/project/icdmtutorial/.

3 Instructors

Dr. Shafiq Joty[1] is an Assistant Professor at the School of Computer Science and Engineering, NTU. He is also a senior research manager at the Salesforce AI Research lab. He holds a PhD in Computer Science from the University of British Columbia. His work has primarily focused on developing discourse analysis tools (e.g., discourse parser, coherence model, topic model, dialogue act recognizer), and exploiting these tools effectively in downstream applications like machine translation, summarization, and sentiment analysis. Apart from discourse and its applications, he has also developed novel machine learning models for question answering, machine translation, image/video captioning, visual question answering, and opinion analysis. His work has appeared in major journals and conferences such as CL, JAIR, CSL, ACL, EMNLP, NAACL, IJCAI, CVPR, ECCV, and ICWSM. He served as an area chair for ACL-2019 (QA track) and EMNLP-2019 (Discourse track) and a senior program committee member for IJCAI 2019. Shafiq is a recipient of NSERC CGS-D scholarship and Microsoft Research Excellent Intern award.

Dr. Giuseppe Carenini[2] is a Professor in Computer Science at UBC. Giuseppe has broad interdisciplinary interests. His work on NLP and information visualization to support decision making has been published in over 100 peer-reviewed papers (including best paper at UMAP-14 and ACM-TiiS-14). He was the area chair for ACL'09 "Sentiment Analysis, Opinion Mining, and Text Classification" , NAACL'12 and EMNLP'19 for "Summarization and Generation", ACL'19 for Discourse; the Program Co-Chair for IUI 2015, and the Program Co-Chair for SigDial 2016. He has also co-edited an ACM-TIST Special Issue on "Intelligent Visual Interfaces for Text Analysis". In 2011, he published a co-authored book on "Meth-

[1] https://raihanjoty.github.io/
[2] https://www.cs.ubc.ca/∼carenini/

ods for Mining and Summarizing Text Conversations". He has also extensively collaborated with industrial partners, including Microsoft and IBM. He was awarded a Google Research Award, an IBM CASCON Best Exhibit Award, and a Yahoo Faculty Research Award in 2007, 2010 and 2016 respectively.

Dr. Raymond T. Ng[3] is a Professor in Computer Science and the Director of the Data Science Institute at UBC. His main research area for the past two decades is on data mining, with a specific focus on health informatics and text mining. He has published over 180 peer-reviewed publications on data clustering, outlier detection, OLAP processing, health informatics and text mining. He is the recipient of two best paper awards from the 2001 ACM SIGKDD conference, the premier data mining conference in the world, and the 2005 ACM SIGMOD conference, one of the top database conferences worldwide. For the past decade, he has co-led several large-scale genomic projects funded by Genome Canada, Genome BC and industrial collaborators. Since the inception of the PROOF Centre of Excellence, which focuses on biomarker development for end-stage organ failures, he has held the position of the Chief Informatics Officer of the Centre. From 2009 to 2014, he was the associate director of the NSERC-funded strategic network on business intelligence. Since 2016, he has been the holder of the Canadian Research Chair on Data Science and Analytics.

Dr. Gabriel Murray[4] is an Associate Professor in Computer Information Systems at the University of the Fraser Valley (UFV). His background is in computational linguistics and multimodal speech and language processing. He holds a PhD in Informatics from the University of Edinburgh, completed under the supervision of Drs. Steve Renals and Johanna Moore. His research has focused on various aspects of multimodal conversational data, including automatic summarization and sentiment detection for group discussions. Recent research also focuses on predicting group performance and participant affect in conversational data. In 2011, Dr. Murray co-authored the book "Methods for Mining and Summarizing Text Conversations".

[3] https://www.cs.ubc.ca/∼rng

[4] https://www.ufv.ca/cis/faculty-and-staff/murray-gabriel.htm

References

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

G. Carenini, R. T. Ng, and X. Zhou. 2007. Summarizing Email Conversations with Clue Words. In *Proceedings of the 16th international conference on World Wide Web*, WWW'07, pages 91–100, Banff, Canada. ACM.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. RST Discourse Treebank (RST–DT) LDC2002T07. *Linguistic Data Consortium, Philadelphia*.

William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. 2004. Learning to Classify Email into "Speech Acts". In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, EMNLP'04, pages 309–316.

David duVerle and Helmut Prendinger. 2009. A Novel Discourse Parser based on Support Vector Machine Classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 665–673, Suntec, Singapore. ACL.

Micha Elsner and Eugene Charniak. 2008. Coreference-inspired coherence modeling. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, HLT-Short '08, pages 41–44, Columbus, Ohio. Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2011a. Disentangling chat with local coherence models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1179–1189, Stroudsburg, PA, USA. Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2011b. Extending the entity grid with entity-specific features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 125–129, Portland, Oregon. Association for Computational Linguistics.

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–521. Association for Computational Linguistics.

Camille Guinaudeau and Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association* *for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 93–103.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24, Baltimore, Maryland. ACL.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association for Computational Linguistics*, 3:329–344.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse-driven language models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 332–342. Association for Computational Linguistics.

Yangfeng Ji and Noah A. Smith. 2017. Neural discourse structure for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1005. Association for Computational Linguistics.

Shafiq Joty, Giuseppe Carenini, and Chin-Yew Lin. 2011. Unsupervised Modeling of Dialog Acts in Asynchronous Conversations. In *Proceedings of the twenty second International Joint Conference on Artificial Intelligence*, IJCAI'11, Barcelona, Spain.

Shafiq Joty, Giuseppe Carenini, and Raymond T Ng. 2015. CODRA: A Novel Discriminative Framework for Rhetorical Analysis. *Computational Linguistics*, 41:3:385–435.

Shafiq Joty, Francisco Guzmán, Llus Mrquez, and Preslav Nakov. 2017. Discourse structure in machine translation evaluation. *Computational Linguistics*, 43:4:683–722.

Shafiq Joty and Enamul Hoque. 2016. Speech act modeling of written asynchronous conversations with task-specific embeddings and conditional structured models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1746–1756, Berlin, Germany. Association for Computational Linguistics.

Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, and Haifeng Wang. 2017. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1299–1308. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796. Association for Computational Linguistics.

Jiwei Li and Dan Jurafsky. 2017. Neural net models of open-domain discourse coherence. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 198–209, Copenhagen, Denmark. Association for Computational Linguistics.

Jiwei Li, Rumeng Li, and Eduard Hovy. 2014. Recursive deep models for discourse parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2061–2069, Doha, Qatar. ACL.

Qi Li, Tianshi Li, and Baobao Chang. 2016. Discourse parsing with attention-based hierarchical neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 362–371. Association for Computational Linguistics.

Xiang Lin, Shafiq Joty, Prathyusha Jwalapuram, and Saiful Bari. 2019. A unified linear-time framework for sentence-level discourse parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL'19, pages xx–xx, Florence, Italy. ACL.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 997–1006, Portland, Oregon. Association for Computational Linguistics.

Mohsen Mesgar and Michael Strube. 2018. A neural local coherence model for text quality assessment. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4328–4339, Brussels, Belgium. Association for Computational Linguistics.

Tasnim Mohiuddin, Shafiq Joty, and Dat Nguyen. 2018. Coherence Modeling of Asynchronous Conversations: A Neural Entity Grid Approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL'18, pages xx–xx, Melbourne, Australia. Association for Computational Linguistics.

Tasnim Mohiuddin, Thanh-Tung Nguyen, and Shafiq Joty. 2019. Adaptation of hierarchical structured models for speech act recognition in asynchronous conversation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL'19, pages xx–xx, Minneapolis, USA. ACL.

Mathieu Morey, Philippe Muller, and 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 the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1319–1324, Copenhagen, Denmark. Association for Computational Linguistics.

Bita Nejat, Giuseppe Carenini, and Raymond Ng. 2017. Exploring joint neural model for sentence level discourse parsing and sentiment analysis. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 289–298. Association for Computational Linguistics.

Michael J. Paul. 2012. Mixed Membership Markov Models for Unsupervised Conversation Modeling. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 94–104, Stroudsburg, PA, USA. ACL.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii. Association for Computational Linguistics.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 13–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, and Bonnie Webber. 2005. The Penn Discourse TreeBank as a Resource for Natural Language Generation. In *Proceedings of the Corpus Linguistics Workshop on Using Corpora for Natural Language Generation*, pages 25–32, Birmingham, U.K.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised Modeling of Twitter Conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 172–180, LA, California. ACL.

Rico Sennrich. 2018. Why the time is ripe for discourse in machine translation. Invited talk at the 2nd Workshop on Neural Machine Translation and Generation.

D. Shen, Q. Yang, J-T. Sun, and Z. Chen. 2006. Thread detection in dynamic text message streams. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 35–42, Seattle, Washington, USA. ACM.

Radu Soricut and Daniel Marcu. 2003. Sentence Level Discourse Parsing Using Syntactic and Lexical Information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL'03, pages 149–156, Edmonton, Canada. ACL.

A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. Van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26:339–373.

Rajen Subba and Barbara Di Eugenio. 2009. An effective discourse parser that uses rich linguistic information. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 566–574. Association for Computational Linguistics.

Hongning Wang, Chi Wang, ChengXiang Zhai, and Jiawei Han. 2011a. Learning online discussion structures by conditional random fields. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 435–444, New York, NY, USA. ACM.

Li Wang, Marco Lui, Su Nam Kim, Joakim Nivre, and Timothy Baldwin. 2011b. Predicting thread discourse structure over technical web forums. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 13–25, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yi-Chia Wang, Mahesh Joshi, William W. Cohen, and Carolyn Penstein Ros. 2008. Recovering implicit thread structure in newsgroup style conversations. In *ICWSM*. The AAAI Press.

Yizhong Wang, Sujian Li, and Houfeng Wang. 2017. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–188, Vancouver, Canada. Association for Computational Linguistics.

Lin Ziheng, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151184.