# Discourse Processing and Its Applications in Text Mining

Shafiq Joty
*Nanyang Technological University*
Singapore
srjoty@ntu.edu.sg

Giuseppe Carenini and Raymond T. Ng
*University of British Columbia*
Vancouver, Canada
{carenini,rng}@cs.ubc.ca

Gabriel Murray
*University of the Fraser Valley*
Abbotsford, Canada
gabriel.murray@ufv.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 text mining 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, essay scoring, sentiment analysis, machine translation, 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. It then covers traditional machine learning methods along with the most recent works using deep learning, and compare their performances on benchmark datasets. For each discourse structure we describe, we show its applications in downstream text mining tasks.**

***Index Terms*—discourse, conversation, coherence, dialogue acts**

## I. Rationale

With the emergence of Internet technologies, unstructured and semi-structured text data are plentiful, and discourse processing methods can help uncover structures and coherence within the texts that can enable further detailed analysis. More importantly, the uncovered structures can support a number of end-user text mining applications including text summarization, topic mining, information extraction, question answering, sentiment analysis, and machine translation.

## II. Tutorial Outline

### *A. Introduction [20 mins]*

(i) Discourse & its different forms
(a) Monologue; (b) Synchronous & asynchronous conversations; (c) Modalities: written & spoken

(ii) Linguistic structures in discourse & analysis tasks
(a) Coherence structure ⇒ Discourse segmentation & parsing; (b) Coherence models ⇒ Coherence evaluation; (c) Topic structure ⇒ Topic segmentation & labeling; (d) Coreference structure ⇒ Coreference resolution; (e) Conversational structure ⇒ Disentanglement & reply-to structure, speech act recognition

(iii) Applications of discourse analysis
(a) Text summarization & generation; (b) Machine translation; (c) Essay scoring; (d) Sentiment analysis

### *B. Discourse Parsing & Its Applications [40 mins]*

(i) Discourse annotations
(a) Rhetorical Structure Theory (RST) Treebank; (b) Penn Discourse Treebank (PDTB)

(ii) Discourse parsing with RST
(a) The tasks: discourse segmentation and parsing; (b) Traditional models – SPADE, HILDA, CODRA; (c) Neural models; (d) Evaluation & Discussion

(iii) Discourse parsing in PDTB
(a) The tasks: relation sense identification and scope disambiguation; (b) Statistical models; (c) Neural models; (d) Evaluation & Discussion

(iv) Applications of Discourse Parsing
(a) Summarization & compression; (b) Generation; (c) Coherence modeling; (d) Sentiment analysis; (e) Information extraction & QA; (f) Machine translation;

### *C. Coherence Models & Its Applications [30 mins]*

(i) Coherence models
(a) Entity grid and its extensions; (b) Other existing models; (c) Neural coherence models; (d) Coherence models for conversations

(ii) Applications (Evaluation tasks)
(a) Sentence ordering (Discrimination, Insertion); (b) Summary coherence rating; (c) Readability assessment; (d) Chat disentanglement; (e) Thread reconstruction

### *D. Conversational Structures [30 mins]*

(i) Discourse Structures in Conversations
(a) Thread identification in synchronous (e.g., multiparty chat) and asynchronous conversations (e.g., forums, emails); (b) Speech acts in synchronous & asynchronous conversations

(ii) Thread identification models for synchronous and asynchronous conversations

(iii) Speech act recognition models for synchronous and asynchronous conversations
(a) Traditional ML models; (b) Neural models

(iv) Evaluation & Applications

### *E. Conclusions & Future Challenges [15 mins]*

(i) Learning from limited annotated data; (ii) Language & domain transfer; (iii) New emerging applications