LING 520: Computational Analysis of English

Semester: FALL '16

Instructor: Sowmya Vajjala

Iowa State University, USA

10 November 2016

Class Outline

- Overview of computational modeling of discourse
- Referring expressions and anaphora resolution
- Modeling coherence
- Practice exercises (from Problem Set 7)
- Assignment 5 submission - 15th November.

Why model discourse? - 1

- Let us say there is a paragraph of text:
  Donald John Trump (born June 14, 1946) is the President-elect of the United States and a businessman. He is scheduled to take office as the 45th President on January 20, 2017. As the Republican Party's nominee for president in the 2016 election, he defeated Hillary Clinton in the general election on November 8, 2016. - Who is he in second and third sentences? (referring expressions)

Why model discourse? - 1

- Let us say there is a paragraph of text:
  Donald John Trump (born June 14, 1946) is the President-elect of the United States and a businessman. He is scheduled to take office as the 45th President on January 20, 2017. As the Republican Party's nominee for president in the 2016 election, he defeated Hillary Clinton in the general election on November 8, 2016. - Who is he in second and third sentences? (referring expressions)
- This sort of things can happen in a single sentence as well.
  Mary said she wanted to tell her friend she should have voted.
- If all this is some sort of textual discourse, we need a model to identify the antecedents to such references (she, he etc here)

Why model discourse? - 2

- Let us say there is a small paragraph of text:
  I like language. I like studying language. I like programming. I wanted to know how to combine both interests together and that is how I learnt about NLP.
- There is now another one:
  I like language. I don't like pets. I wanted to know how to reach a compromise between these two contrasting issues and ended up enrolling in a psychology class.
- What makes more sense?

Why model discourse? - 2

- Let us say there is a small paragraph of text:
  I like language. I like studying language. I like programming. I wanted to know how to combine both interests together and that is how I learnt about NLP.
- There is now another one:
  I like language. I don't like pets. I wanted to know how to reach a compromise between these two contrasting issues and ended up enrolling in a psychology class.
- What makes more sense?
- If all this is some sort of textual discourse, we need to create a model of "coherent text" to make a computer understand what is a text that is meaningful.
- Note: Semantic representations we discussed yesterday still talk at the level of sentence. What we just saw goes beyond that and looks at full discourse of the text.

Modeling discourse

Two primary problems:

- Coreference resolution
- Modeling text coherence

Coreference Resolution

- Task: identify all expressions that refer to some entity in a text. In the previous example.
- Use: very important for almost all applied NLP tasks that go beyond a sentence: summarizing a text, question answering, extracting relations between entities in a text, machine translation (if you want to cover discourse aspects) etc.

Modeling Reference Resolution

- Generally involves either rule based or statistical methods that look at the previous context of the pronouns or other entities, and other aspects like agreement of gender to decide if there is a antecedent for it.
- One ready to use approach: Stanford CoreNLP has a coreference resolution system in place.
- Demo: http://nlp.stanford.edu:8080/corenlp/
- You should be able to use it from Python through NLTK or some other library. Figure out how.

Text Coherence: RST

- Rhetorical Structure Theory is one of the theories of discourse in NLP.
- It describes relations between sentences interms of rhetorical relations.
- Some example relations: antithesis, background, concession, elaboration, purpose etc.
- RST based parsers are available (one in Python by ETS: https://github.com/EducationalTestingService/discourse-parsing)
- Note: See Radev's slides from discourse lecture.

Modeling Text Coherence: Some methods

- Coh-Metrix (http://141.225.42.86/CohMetrixHome/documentation_indices.html) does a shallow discourse modeling, but very popular.
- "Modeling coherence in ESOL learner texts" by Yannakoudakis and Briscoe, 2012. http://www.aclweb.org/anthology/W12-2004
- "Modeling Local Coherence: An Entity based approach" Barzilay and Lapata, 2007. Computational Linguistics 34 (1). https://people.csail.mit.edu/regina/my_papers/coherence.pdf
- publications related to Brown coherence toolkit (C++ code, shared freely online. https://bitbucket.org/melsner/browncoherence)

Modeling discourse: local example

- RWT project can be seen as one approach to model the discourse of research articles.

Modeling discourse at different levels of linguistic representation

- word level overlap between sentences in a text (word, stem, lemma etc)
- Looking at specific POS tags (Nouns, Pronoun overlap)
- Looking at the usage of connective words (and usage of connective words as connectives)
- Looking at how a named entity transitions from one sentence to another (e.g., subject to object)
- Looking at the length of coreference chains, or connected lexical chains in the text
- ... .

Next Week

- NLP applications: overview of machine translation, question answering, topic modeling
- NLP for CALL - Introduction
- Assignment 5 (submission and discussion)
- Final projects - status and discussion

Exercise - 1

Problem Set 7, Problem 5 and 6: Go through the description of CohMetrix text analysis tool (http://141.225.42.86/CohMetrixHome/documentation_indices.html). Read the documentation of Referential Cohesion features there, and write code to calculate adjacent and global Noun overlap. Write a program to calculate adjacent and global stem overlap.

Exercise - 2

Problem Set 7, Problem 8-9: Figure out where one can find lists of English connectives, and write a program to count all connective words in a text. Read Coh-Metrix documentation, and figure out how to modify the previous program to print the number of occurrences of different connective types in text.

Exercise - 3

Problem Set 7, Problem 7: Write a program to get groups of sentences connected by a coreference chain using Stanford corenlp.

Exercise - 4

Problem Set 7, Problem 10: Not all connective words are actually used as connectives in all con- texts. Figure out how to use "Discourse Connectives Tagger" tool (http://www.cis.upenn.edu/~nlp/software/discourse.html) to get a better estimate of connective usage in texts. Python programmers: figure out how to use the perl code in this tagger inside your Python code or somehow use the output of this code in your Python code.

Exercise - 5

Figure out how to setup and use the RST parser from ETS. If you manage to do that, think about how this is useful (if at all!) for your research interests.