Shallow Discourse Parsing

Jimmy Callin

Abstract

TODO

Contents

Preface

1 Background

1.1 Discourse

1.1.1 Parsing

1.1.2 Shallow Discourse Parsing

[Continue explaination of why D-LTAG is better, refer to Webber (2004) and Prasad et al (2008).]

A subset of Penn Treebank has been annotated with a set of discourse connectives, giving us a resource to work with to learn more about the computability of these connectives. This is a good start.

Penn Discourse Treebank follows the lexically grounded predicate-argument approach as proposed in Webber (2004). It covers the subset containing Wall Street Journal articles from the Penn Treebank, making up approximately one million tokens. When a connective explicitly appears, it will be syntactically connected to the *Arg2* argument of the discourse structure. *Arg1* is the other one. Due to *Arg2* being syntactically bounded to connective, it is easy to automatically classify *Arg2*. *Arg1* is more difficult, and Lin et al (2012) solves this by

PDTB annotates each structure with types of discourse relations according to a three level hierarchy, where the first level is made up of four classes: temporal, contingency, comparison, expansion. Each class has a second level of in total 16 types to provide a more fine-grained classification. Due to the third level being considered too fine-grained, it is ignored in this work.

1.2 Explicit vs implicit discourse connectives

I want to go to New York, but I already booked a flight. (–> I am not going to New York.)

I want to go to New York, so I already booked a flight. (–> I am going to New York.)

- Implicit
- EntRel, AltRel?

Discourse parsing will find its usefulness when current methods reach their potential given the current method of only looking within the immediate local context. When structure starts to become more important, as is already the case for e.g. QA systems but also in machine translation where pronouns may depend upon antecedents beyond the local context. Furthermore, experiments have been made where discourse

- TEMPORAL
  - Asynchronous
  - Synchronous

- CONTINGENCY
  - Cause
  - Pragmatic Cause
  - Condition
  - Pragmatic Condition

- COMPARISON
  - Contrast
  - Pragmatic Contrast
  - Concession
  - Pragmatic Concession

- EXPANSION
  - Conjunction
  - Instantiation
  - Restatement
  - Alternative
  - Exception
  - List

Figure 1.1: Two first levels of sense hierarchy.

structure measures coherence of text. See Lin, Ng, Kan (2011) Recognizing implicit discourse relations in the penn discourse treebank. Also summarization.

1.3 The CoNLL-2016 Shared Task on Shallow Discourse Parsing

The shared task for the Twentieth Conference on Computational Natural Language Learning (CoNLL-2016) continues upon the work of last year's shared task on shallow discourse parsing. It is as a part of this task we worked on the thesis.

2 Data

The first and only work to date which has used the D-LTAG for annotation is the work of *Penn Discourse Treebank* (PDTB), ...

senses	connective_token	ratio
(Expansion.Conjunction,)	7355	0.226064
(Contingency.Cause,)	4969	0.152728
(Comparison.Contrast,)	4521	0.138958
(EntRel,)	4133	0.127032
(Expansion.Restatement,)	2640	0.081143
(Temporal.Asynchronous,)	2044	0.062825
(Expansion.Instantiation,)	1392	0.042785
(Comparison.Concession,)	1268	0.038973
(Contingency.Condition,)	1114	0.034240
(Temporal.Synchrony,)	985	0.030275
(Comparison,)	484	0.014876
(Expansion.Alternative,)	435	0.013370
(Temporal.Asynchronous, Contingency.Cause)	148	0.004549
(Temporal.Synchrony, Contingency.Cause)	113	0.003473
(Temporal.Synchrony, Expansion.Conjunction)	109	0.003350
(Expansion.Conjunction, Contingency.Cause)	106	0.003258
(Expansion,)	96	0.002951
(Expansion.Conjunction, Temporal.Synchrony)	92	0.002828
(Temporal.Synchrony, Comparison.Contrast)	51	0.001568
(Expansion.Conjunction, Comparison.Contrast)	47	0.001445