A Pilot Annotation to Investigate Discourse Connectivity in Biomedical Text

Hong Yu, Nadya Frid, Susan McRoy
University of Wisconsin-Milwaukee
P.O.Box 413
Milwaukee, WI 53201
Hongyu,frid,mcroy@uwm.edu

Rashmi Prasad, Alan Lee, Aravind Joshi
University of Pennsylvania
3401 Walnut Street
Philadelphia, PA 19104, USA
Rjprasad,aleewk,joshi@seas.upenn.edu

Abstract

The goal of the Penn Discourse Treebank (PDTB) project is to develop a large-scale corpus, annotated with coherence relations marked by discourse connectives. Currently, the primary application of the PDTB annotation has been to news articles. In this study, we tested whether the PDTB guidelines can be adapted to a different genre. We annotated discourse connectives and their arguments in one 4,937-token full-text biomedical article. Two linguist annotators showed an agreement of 85% after simple conventions were added. For the remaining 15% cases, we found that biomedical domain-specific knowledge is needed to capture the linguistic cues that can be used to resolve inter-annotator disagreement. We found that the two annotators were able to reach an agreement after discussion. Thus our experiments suggest that the PDTB annotation can be adapted to new domains by minimally adjusting the guidelines and by adding some further domain-specific linguistic cues.

1 Introduction

Large scale annotated corpora, e.g., the Penn TreeBank (PTB) project (Marcus et al. 1993), have played an important role in text-mining. The Penn Discourse Treebank (PDTB) (http://www.seas.upenn.edu/~pdtb) (Prasad et al. 2008a) annotates the *argument structure*, *semantics*, and *attribution* of discourse connectives and their arguments. The current release of PDTB-2.0 contains the annotations of 1,808 Wall Street Journal articles (~1 million words) from the Penn TreeBank (Marcus et al. 1993) II distribution and a total of 40,600 discourse connective tokens (Prasad et al. 2008b). This work examines whether the PDTB annotation guidelines can be adapted to a different genre, the biomedical literature.

2 Notation

A discourse connective can be defined as a word or multiword expression that signals a discourse relation. Discourse connectives can be subordinating conjunctions (e.g., *because, when, although*), coordinating conjunctions (e.g., *but, or, nor*) and adverbials (e.g., *however, as a result, for example*). A discourse connective takes in two arguments, *Arg1* and *Arg2*. Arg2 is the argument that appears in the clause that is syntactically bound to the connective and Arg1 is the other argument. In the sentence "*John failed the exam* because **he was lazy**" the discourse connective is underlined, Arg1 appears in italics and Arg2 appears in bold.

3 A Pilot Annotation

Following the PDTB annotation manual (Prasad et al. 2008b), we conducted a pilot annotation of discourse connectivity in biomedical text. As an initial step, we only annotated the three most

BioNLP 2008: Current Trends in Biomedical Natural Language Processing, pages 92–93,
Columbus, Ohio, USA, June 2008.

important components of a discourse relation; namely, a discourse connective and its two arguments; we did not annotate attribution. Two linguist annotators independently annotated one full-text biomedical article (Verpy et al. 1999) that we randomly selected. The article is 4,937 tokens long. When the annotation work was completed, we measured the inter-annotator agreement, following the PDTB exact match criterion (Miltsakaki et al. 2004). According to this criterion, a discourse relation is in disagreement if there is disagreement on any text-span (i.e., the discourse connective or any of its two arguments). In addition, we also measured the agreement in the components (i.e., discourse connectives and the arguments). We discussed the annotation results and made suggestions to adapt the PDTB guidelines to biomedical text.

4 Results and Discussion

The first annotator identified 74 discourse connectives, and the second annotator identified 75, 68 of which were the same as those identified by the first annotator. The combined total number of discourse connectives was 81. The overall agreement in discourse connective identification was 68/81=84%.

Of the 68 discourse connectives that were annotated by both annotators, 31 were an exact match, 31 had an exact match for Arg1, and 54 had an exact match for Arg2. The overall agreement for the 68 discourse relations is 45.6% for exact match, 45.6% for Arg1, and 79.4% for Arg2. The PDTB also reported a higher level of agreement in annotating Arg2 than in annotating Arg1 (Miltsakaki et al. 2004). We manually analyzed the cases with disagreement. We found the disagreements are nearly all related to the annotation of citation references, supplementary clauses, and other conventions. When a few conventions for these cases were added, the inter-annotator agreement went up to 85%. We also found that different interpretation of a relation and its arguments by annotators plays an important role for the remaining 15% inconsistency, and domain-specific knowledge is necessary to resolve such cases.

5 New Conventions

After the completion of the pilot annotation and the discussion, we decided to add the following conventions to the PDTB annotation guidelines to address the characteristics of biomedical text:

i. Citation references are to be annotated as a part of an argument because the inclusion will benefit many text-mining tasks including identifying the semantic relations among citations.
ii. Clausal supplements (e.g., relative or parenthetical constructions) that modify arguments but are not minimally necessary for the interpretation of the relation, are annotated as part of the arguments.
iii. We will annotate a wider variety of nominalizations as arguments than allowed by the PDTB guidelines.

We anticipate that these changes will both decrease the amount of effort required for annotation and increase the reliability of the annotation.

6 References

Marcus M, Santorini B, Marcinkiewicz M (1993) Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics 19

Miltsakaki E, Prasad R, Joshi A, Webber B (2004) Annotating discourse connectives and their arguments. Paper presented at Proceedings of the NAACL/HLT Workshop: Frontiers in Corpus Annotation

Prasad R, Dinesh N, Lee A, Miltsakaki E, Robaldo L, Joshi A, Webber B (2008a) The Penn Discourse Treebank 2.0. Paper presented at The 6th International Conference on Language Resources and Evaluation (LREC). Marrakech, Morroco

Prasad R, Miltsakaki E, Dinesh N, Lee A, Joshi A, Robaldo L, Webber B (2008b) The Penn Discourse TreeBank 2.0 Annotation Manual. Technical Report: IRCS-08-01

Verpy E, Leibovici M, Petit C (1999) Characterization of otoconin-95, the major protein of murine otoconia, provides insights into the formation of these inner ear biominerals. Proc Natl Acad Sci U S A 96:529-534