Sense annotation in the Penn Discourse Treebank

Eleni Miltsakaki*, Livio Robaldo+, Alan Lee*, Aravind Joshi*

*Institute for Research in Cognitive Science, University of Pennsylvania
{elenimi, aleewk, joshi}@linc.cis.upenn.edu
+Department of Computer Science, University of Turin
robaldo@di.unito.it

Abstract. An important aspect of discourse understanding and generation involves the recognition and processing of discourse relations. These are conveyed by discourse connectives, i.e., lexical items like *because* and *as a result* or implicit connectives expressing an inferred discourse relation. The Penn Discourse TreeBank (PDTB) provides annotations of the argument structure, attribution and semantics of discourse connectives. In this paper, we provide the rationale of the tagset, detailed descriptions of the senses with corpus examples, simple semantic definitions of each type of sense tags as well as informal descriptions of the inferences allowed at each level.

1 Introduction

Large scale annotated corpora have played and continue to play a critical role in natural language processing. The continuously growing demand for more powerful and sophisticated NLP applications is evident in recent efforts to produce corpora with richer annotations [6], including annotations at the discourse level[2], [8], [4]. The Penn Discourse Treebank is, to date, the largest annotation effort at the discourse level, providing annotations of explicit and implicit connectives. The design of this annotation effort is based on the view that discource connectives are predicates taking clausal arguments. In Spring 2006, the first version of the Penn Discourse Treebank was released, making availalble thousands annotations of discourse connectives and the textual spans that they relate.

Discourse connectives, however, like verbs, can have more than one meaning. Being able to correctly identify the intended sense of connectives is crucial for every natural language task which relies on understanding relationships between events or situations in the discourse. The accuracy of information retrieval from text can be significantly impaired if, for example, a temporal relation anchored on the connective *since* is interpreted as causal.

A well-known issue in sense annotations is identifying the appropriate level of granularity and meaning refinement as well as identifying consistent criteria for making sense distinctions. Even if an 'appropriate' level of granularity can be identified responding to the demands of a specific application, creating a flat set of sense tag is limiting in many ways. Our approach to the annotation of sense