John Kraemer, Mark Finlayson, Denise Ichinco & Edward Gibson

Advances in Discourse: Theory and Annotation

Wolf & Gibson's GraphBank (2005) represented a significant advance in corpus-based investigation of discourse coherence structure; we show evidence in GraphBank for differing constraints on different classes of discourse relations. However, significant research questions cannot be answered using GraphBank, so we also introduce Story Workbench, a platform- and theory-independent extensible annotation tool.

We introduce new results showing that different classes of discourse relation obey distinct structural constraints. Wolf and Gibson used GraphBank to show that tree structures are inadequate to represent discourse coherence structure. Expanding on these results, we show that causal relations obey strong locality constraints, as do temporal relations when considered over a more limited domain of discourse segments. Elaborative relations by comparison often act non-locally via co-reference mechanisms, while respecting constituency boundaries established by other relations. Intentional relations (including attributions) establish separate modal contexts that sometimes lead to apparent exceptions to the above principles. Co-reference to entities and events, which is not explicitly represented in GraphBank, was nonetheless key to the intuitive discourse representations of the naive subjects employed in GraphBank's construction, as shown by their consistent misuse of the resemblance relation 'similarity' to signify co-reference.

Although GraphBank was adequate to establish these results, it is not well suited to anwering significant questions related to the role of lexical discourse markers, discourse segments, coreference, and entities and events. To further investigate these phenomena, we introduce a new text annotation tool, the Story Workbench, that will enable the collection of data that can address these questions. The Story Workbench implements an extremely flexible annotation scheme that makes the following four advances. First, the representation of each coherence relation can include its lexical discourse marker, if any, as well as arguments and adjuncts composed of freely-selected text-spans. Second, it represents the text's significant discourse referents (such as entities and events) along with the parts of the text that refer to them. Third, it provides the capacity to represent the syntactic structure of the text. Lastly, it organizes all of its representations using an indexing scheme anchored to the text's surface token structure. This system provides for the use of a wide variety of text annotation systems, including discourse annotation methods such as RST, GraphBank, and the Penn Discourse TreeBank system.