Situation entities and genre distinctions in the Penn Discourse TreeBank

The genre of a text or a text passage often reflects a particular manner of conveying a particular sort of information, and for a particular purpose. For example, most human readers can easily distinguish a news article from an opinion piece. Such genre distinctions are intuitively clear but theoretically only partially understood. This paper contributes to recent work on automated processing and understanding of texts by undertaking an analysis of correlations between text-level **genre characterizations** and clause-level classification of the **situation entity types** introduced to the discourse.

Our work draws on both Smith [2003] and Webber [2009]. Smith [2003] identifies linguistic correlates to the discourse mode of a text passage; distribution of situation entity types is one of the two key factors.[1] Situation entities are the events, states, generic statements, and embedded facts and propositions which clauses of text introduce to a discourse. These entities, or abstract objects, are well-studied in the linguistic literature (see Asher 1993, Vendler 1967, Dowty 1979, Smith 1991, and many others). Webber [2009] finds correlations between genre and the frequency and senses of discourse connectives found in a text. These connectives express discourse relationships between the abstract objects introduced by clauses or other units of text.

These two approaches use different levels of genre/mode characterization—Smith for passages of text and Webber for full texts. Each also focuses on a different type of linguistic correlate—where Smith looks at linguistic attributes of clauses, Webber looks at the relations *between* clauses. What they have in common is the notion that clauses introduce references to situations, whether we call those situation entities or abstract objects. This insight motivates the current work.

The analysis in Webber [2009] is based on the Penn Discourse TreeBank (PDTB), a collection of texts which have been carefully manually annotated with discourse relations and their arguments. Webber's analysis considers four genre categories: ESSAYS, NEWS, SUMMARIES, LETTERS. We select texts from the first two categories and annotate the arguments of the discourse relations with situation entity type labels. To do this, we modify the situation entity classifier of Palmer et al. [2007], use it to label the clauses, and perform manual post-correction. For example, the two clauses below are adjacent clauses from a PDTB text.[2] The PDTB annotations mark the discourse connective between the two clauses (**when**) and the sense of the discourse relation it indicates (**Temporal.Asynchronous.Succession**). To this we add annotations (**E** and **AE**) indicating the type of situation entity introduced by each clause (**event** and **abstract entity**, respectively).

(1) **(*E*)** [**When**] the dust and dirt settle in an extra-nasty mayoral race,
. [**Temporal.Asynchronous.Succession**]

(2) **(*AE*)** the man most likely to gain custody of all this is a career politician named David Dinkins.

[1] The second is the manner of progression (either temporal or metaphorical) through the text.

[2] `wsj_0765.pdtb`

Our analysis of the annotated texts considers two levels. First, we look at the distribution of situation entity types as arguments to particular classes of discourse connective. Second, we look for correlations between situation entity distribution and the genre distinctions outlined in Webber 2009. This allows us (i) to identify the presence or absence of characteristic distributions of situation entity types for the genres identified by Webber, and (ii) to investigate the relationship between discourse relations (as signalled by classes of discourse connectives) and situation entity types. In addition, the annotations and analysis are useful resources for future work in automated discourse processing and analysis.

References

Nicholas Asher. *Reference to Abstract Objects in Discourse*. Kluwer, 1993.

David Dowty. *Word Meaning and Montague Grammar*. Springer Classic Titles in Linguistics, 1979.

Alexis Palmer, Elias Ponvert, Jason Baldridge, and Carlota Smith. A sequencing model for situation entity classification. In *Proceedings for the conference of the Association for Computational Lingustics*, 2007.

Carlota S. Smith. *The Parameter of Aspect*. Kluwer, 1991.

Carlota S. Smith. *Modes of Discourse*. Cambridge University Press, 2003.

Zeno Vendler. *Linguistics in Philosophy*, chapter Verbs and Times, pages 97–121. Cornell University Press, 1967.

Bonnie Webber. Genre distinctions for Discourse in the Penn TreeBank. In *Proceedings for the conference of the Association of Computational Lingustics*, 2009.