Automatic Analysis of Discourse Structure

With the amount of information available as electronic text increasing at a staggering pace, machine understanding of natural language stands as a key barrier in the creation of computer systems that can take advantage of the knowledge encoded in written documents. Current approaches to automatic extraction of facts and other information from text focus mostly on what can be learned from single sentences, one sentence at a time, examining word-to-word relationships. This often corresponds to the notion of *who* did *what* to *whom*.

Constraining computation to single sentences has allowed researchers to explore increasingly sophisticated machine learning models that capture complex structure. In contrast, when going beyond sentence boundaries, the field is dominated by simpler models that largely ignore language structure.

This research project will explore models that target relationships that hold among larger phrases and sentences, capturing for example statements of cause and effect that involve multiple sentences, addressing the additional questions of *why*, *how* and *when* in broader contexts than what can be extracted from individual sentences, and accounting for the overall rhetorical structure of texts. This approach involves new strategies for dealing with the exponentially large search space of this task in a way that both addresses fundamental barriers to dealing with language structure in a principled way, and opens new possibilities for future research directions involving rhetorical relations such as causality, attribution, elaboration and comparison.

The proposed work aims, on one hand, to improve our understanding of efficient and effective computational modeling of natural language structure and, on the other hand, to enable deeper natural language understanding in applications such as automatic knowledge-base construction and conversational virtual human systems, allowing for the creation of smarter artificial agents with more robust language capabilities.

12015 Waterfront Drive // Playa Vista, CA 90094-2536 // 310.574.5700 tel // 310.574.5725 fax // info@ict.usc.edu
ict.usc.edu // facebook.com/USCICT // Twitter: @USC_ICT // youtube.com/USCICT

9/2013