Using Web Corpus Statistics to Infer Conceptual Structure

Brandon M. Lock
Emory University

Eugene Agichtein
Emory University

Kevin J. Holmes
Emory University

Phillip Wolff
Emory University

Abstract: The basic level is the level of conceptual structure at which categories are maximally informative. In this research, we investigated whether the privileged status of the basic level might be captured by the statistical properties of the Web. Using Google's Web search programming interface, we found that frequency ratios for terms across three levels of abstraction (superordinate, basic, and subordinate) significantly predicted human participants' spontaneous labeling of images obtained via Mechanical Turk. Specifically, the Web statistics paralleled participants' preference for superordinate labels for natural kinds (e.g., trees, fish) and basic-level labels for other categories. Further, analyses of genre-specific text from the Corpus of Contemporary American English revealed that children's texts were significantly more predictive than academic texts. Our findings suggest that distributional statistics from subsets of the Web can be used to infer properties of conceptual structure, potentially offering a powerful, high-resolution, yet low-cost tool for empirically testing theoretical predictions.