Low-Pass Semantics

Fernando Pereira (Google Inc.)

Abstract

Advances in statistical and machine learning approaches to natural language processing have yielded a wealth of methods and applications in information retrieval, speech recognition, machine translation, and information extraction. Yet, even as we enjoy these advances, we recognize that our successes are to a large extent the result of clever exploitation of redundancy in language structure and use, allowing our algorithms to eke out a few useful bits that we can put to work in applications. By focusing on applications that extract a limited amount of information from the text, finer structures such as word order or syntactic structure could be largely ignored in information retrieval or speech recognition. However, by ignoring those finer details, our language-processing systems have been stuck in an "idiot savant" stage where they can find everything but cannot understand anything. The main language processing challenge of the coming decade is to create robust, accurate, efficient methods that learn to understand the main entities and concepts discussed in any text, and the main claims made. By advancing in that direction, our systems will provide more precise answers to questions, they will verify and update knowledge bases, and they will trace arguments for and against claims throughout the written record. I will argue with examples from our recent research that we need deeper levels of linguistic analysis to do this. But I will also argue that it is possible to do much that is useful even with our very partial understanding of linguistic and computational semantics, by taking (again) advantage of distributional regularities and redundancy in large text collections to learn effective analysis and understanding rules. Thus low-pass semantics: our scientific knowledge is very far from being able to map the full spectrum of meaning, but by combining signals from the whole Web, we are starting to hear some interesting tunes.