The Evolution of an NLP System

Stephen D. Richardson

Natural Language Processing Group, Microsoft Research
One Microsoft Way, Redmond, WA 98053 USA
steveri@microsoft.com

Abstract

This talk will examine the evolution, organization, and future directions of large-scale natural language processing (NLP) systems and the components that comprise them. Microsoft's NLP system, completing its ninth year of research and development, will be used to provide concrete examples for the discussion.

Components often included in large NLP systems and reviewed in this talk will include sentence and word breaking, inflectional and derivational morphology, shallow and deep structural analysis, and generation. The creation and maintenance of lexicons containing morphological, syntactic, and semantic information, including extensive lexical knowledge bases, will be reviewed. Components specific to certain applications will also be examined, including grammar and style checking rules, named entity identification, linguistic filters for information retrieval, and transfer mappings for translation. The combination of statistical methods with traditional linguistic processing will be discussed.

The talk will cover important considerations for the development of truly usable NLP applications, including linguist-friendly development environments and automated processes for building NL systems, testing them, and evaluating progress. Microsoft's experience has confirmed the value of incrementally focusing on specific NLP applications while creating generic components and resources that are intended for use by multiple applications and users. In particular, our initial emphasis on grammar and style checking has fostered the development of very robust, broad-coverage parsers in seven languages that may be readily reused for other tasks. The benefits and challenges of reusing these components for our current work in machine translation will be examined to evaluate the effectiveness of this approach.

The concluding discussion will focus on the use (and reuse) of NLP system components and resources in computing environments of the future. In particular, Microsoft's Next Generation Windows Services (NGWS) strategy, which was recently announced for creating and employing software services across the Internet, will be examined as a fertile context for nurturing the explosive growth of linguistic services from a variety of sources.