Exploiting Hidden Meanings: Using Bilingual Text for Monolingual Annotation

Philip Resnik

Department of Linguistics
and Institute for Advanced Computer Studies
University of Maryland
College Park, Maryland 20742 USA
resnik@umd.edu

Abstract. The last decade has taught computational linguists that high performance on broad-coverage natural language processing tasks is best obtained using supervised learning techniques, which require annotation of large quantities of training data. But annotated text is hard to obtain. Some have emphasized making the most out of limited amounts of annotation. Others have argued that we should focus on simpler learning algorithms and find ways to exploit much larger quantities of text, though those efforts have tended to focus on linguistically shallow problems. In this paper, I describe my efforts to exploit larger quantities of data while still focusing on linguistically deeper problems such as parsing and word sense disambiguation. The trick, I argue, is to take advantage of the shared meaning hidden between the lines of sentences in parallel translation.

1 The Problem of Resources

1.1 Knowledge versus Data

Success in natural language processing depends crucially on good resources. In the early days, knowledge-based approaches depended heavily on good knowledge resources — grammars, lexicons, and the like. Consider LUNAR [1], which permitted users to ask questions about moon rocks using natural language sentences. As an early question answering system, LUNAR was successful not just because of a clever formalism, but also largely because of the human effort that went into a detailed characterization of linguistic alternatives, expressed as an augmented transition network grammar and lexical entries associated with that grammar.

In the late 1980s, natural language processing began to change dramatically as the result of an influsion of ideas and techniques from the speech recognition, information retrieval, and machine learning communities. Ten years ago, the "balancing act" between symbolic and statistical methods was an exciting topic for a computational linguistics workshop [2]; today it's an apt description of the