The Role of Linguistic Models and Language Annotation in Feature Selection for Machine Learning

James Pustejovsky
Department of Computer Science
Brandeis University
Waltham, MA 02454, USA
jamesp@cs.brandeis.edu

Abstract

As NLP confronts the challenge of Big Data for natural language text, the role played by linguistically annotated data in training machine learning algorithms is reaching a critical question. Namely, what role can annotated corpora play for supervised learning algorithms when the datasets become significantly outsized, compared to the gold standards used for training? The use of semi-supervised learning techniques to help solve this problem is a good next step, one that requires not less adherence to annotated data, but an even stricter adherence to linguistic models and the features that are derived from these models for subsequent annotation.

Proceedings of the 6th Linguistic Annotation Workshop, page 1,
Jeju, Republic of Korea, 12-13 July 2012.