Using a Goodness Measurement for Domain Adaptation: A Case Study on Chinese Word Segmentation

Yan Song*† and Fei Xia*
*University of Washington
Seattle, WA 98195, USA
†City University of Hong Kong
83, Tat Chee Ave., Kowloon, Hong Kong
E-mail: {yansong, fxia}@uw.edu

Abstract

Domain adaptation is an important topic for natural language processing. There has been extensive research on the topic and various methods have been explored, including training data selection, model combination, semi-supervised learning. In this study, we propose to use a goodness measure, namely, description length gain (DLG), for domain adaptation for Chinese word segmentation. We demonstrate that DLG can help domain adaptation in two ways: as additional features for supervised segmenters to improve system performance, and also as a similarity measure for selecting training data to better match a test set. We evaluated our systems on the Chinese Penn Treebank version 7.0, which has 1.2 million words from five different genres, and the Chinese Word Segmentation Bakeoff-3 data.

Keywords: word segmentation, domain adaptation, description length gain

This paper will appear in the proceedings of LREC 2012.