Abstract

This thesis proposes an approach to generating n-gram features for Conditional Random Fields (CRFs) based Chinese word segmentation (CWS) systems.

Current systems typically adopt only unigrams or bigrams of either Chinese character themselves or certain traits of characters (e.g. tones, whether it is a digit) as features, though longer n-grams may be useful to describe long-distance dependencies and other phenomena. Inspired by works in error mining, a framework of n-gram expansion has been proposed to generate n-gram of arbitrary length as features for CRF based CWS. The expansion is only triggered when certain criteria are satisfied. Under this framework, we have proposed three specific expansion methods, based on mutual information, label-entropy and label distribution bin, respectively.

Expanded n-grams can be further fed to an iterative process that ranks these candidates and selects most promising ones. Those n-gram features are added into the feature set of CRFs algorithm, and the usefulness of selected features is evaluated by measuring overall performance gain of the CWS system. Using the UPUC corpus in SIGHAN Bakeoff 2006, experiments show that error reductions between 5%~13% are achieved by adding these generated features to a CRFs based CWS system that adopts standard features. Besides, the overall system performance is comparable to top results in Bakeoff 2006.