Computational Linguistics and Chinese Language Processing
Vol. 16, No. 3-4, September/December 2011, pp. 41-60

Histogram Equalization on Statistical Approaches for Chinese Unknown Word Extraction

Bor-Shen Lin* **and Yi-Cong Chen***

Abstract

With the evolution of human lives and the spread of information, new things emerge quickly and new terms are created every day. Therefore, it is important for natural language processing systems to extract new words in progression with time. Due to the broad areas of applications, however, there might exist the mismatch of statistical characteristics between the training domain and the testing domain, which inevitably degrades the performance of word extraction. This paper proposes a scheme of word extraction in which histogram equalization for feature normalization is used. Through this scheme, the mismatch of the feature distributions due to different corpus sizes or changes of domain can be compensated for appropriately such that unknown word extraction becomes more reliable and applicable to novice domains.

The scheme was initially evaluated on the corpora announced in SIGHAN2. 68.43% and 71.40% F-measures for word identification, which correspond to 66.72%/32.94% and 75.99%/58.39% recall rates for IV/OOV, respectively, were achieved for the CKIP and the CUHK test sets, respectively, using four combined features with equalization. When applied to unknown word extraction for a novice domain, this scheme can identify such pronouns as “海角七號” (Cape No. 7, the name of a film), “蠟筆小新” (Crayon Shinchan, the name of a cartoon figure), “金融海嘯” (Financial Tsunami) and so on, which cannot be extracted reliably with rule-based approaches, although the approach appears not so good at identifying such terms as the names of humans, places, or organizations, for which the semantic structure is prominent. This scheme is complementary with the outcomes of two word segmentation systems, and is promising if other rule-based approaches could be further integrated.

* Department of Information Management, National Taiwan University of Science and Technology,
Tel: (886)-2-2703-1225 Fax: (886)-2-2737-6777
E-mail: bslin@cs.ntust.edu.tw; m9709104@mail.ntust.edu.tw