A LEXICON-CORPUS-BASED UNSUPERVISED CHINESE WORD SEGMENTATION APPROACH

Lu Pengyu, Pu Jingchuan, Du Mingming, Lou Xiaojuan and Jin Lijun
School of Management, Harbin Institute of Technology, Harbin, China
Email: lupengyu@hit.edu.cn

Submitted: Oct. 16, 2013 ***Accepted: Feb. 7, 2014*** ***Published: Mar. 10, 2014***

Abstract- This paper presents a Lexicon-Corpus-based Unsupervised (LCU) Chinese word segmentation approach to improve the Chinese word segmentation result. Specifically, it combines advantages of lexicon-based approach and Corpus-based approach to identify out-of-vocabulary (OOV) words and guarantee segmentation consistency of the actual words in texts as well. In addition, a Forward Maximum Fixed-count Segmentation (FMFS) algorithm is developed to identify phrases in texts at first. Detailed rules and experiment results of LCU are presented, too. Compared with lexicon-based approach or corpus-based approach, LCU approach makes a great improvement in Chinese word segmentation, especially for identifying n-char words. And also, two evaluation indexes are proposed to describe the effectiveness in extracting phrases, one is segmentation rate (S), and the other is segmentation consistency degree (D).

Index terms*:* Chinese word segmentation, lexicon-based, Corpus-based, word frequency, natural language processing.