Computational Linguistics and Chinese Language Processing
vol. 3, no. 2, August 1998, pp. 79-92
©Computational Linguistics Society of R.O.C.

Noisy Channel Models for Corrupted Chinese Text Restoration and GB-to-Big5 Conversion

Chao-Huang Chang*

Abstract

In this article, we propose a noisy channel/information restoration model for error recovery problems in Chinese natural language processing. A language processing system is considered as an information restoration process executed through a noisy channel. By feeding a large-scale standard corpus C into a simulated noisy channel, we can obtain a noisy version of the corpus N. Using N as the input to the language processing system (i.e., the information restoration process), we can obtain the output results C'. After that, the automatic evaluation module compares the original corpus C and the output results C', and computes the performance index (i.e., accuracy) automatically. The proposed model has been applied to two common and important problems related to Chinese NLP for the Internet: corrupted Chinese text restoration and GB-to-BIG5 conversion. Sinica Corpora version 1.0 and 2.0 are used in the experiment. The results show that the proposed model is useful and practical.

1. Introduction

In this article, we present a noisy channel (Kernighan *et al.* 1990, Chen 1996) / information restoration model for automatic evaluation of error recovery systems in Chinese natural language processing. The proposed model has been applied to two common and important problems related to Chinese NLP for the Internet: corrupted Chinese text restoration (i.e., 8-th bit restoration of BIG-5 code through a non-8-bit-clean channel), and GB-BIG5 code conversion. The concept follows our previous work on bidirectional conversion (Chang 1992) and corpus-based adaptation for Chinese homophone disambiguation (Chang 1993, Chen and Lee 1995). Several standard Chinese corpora are available to the public, such as NUS's PH corpus (Guo and Lui 1992) and Academia Sinica's sinica corpus (Huang *et al.* 1995). These corpora can be used for objective evaluation of NLP systems. Sinica Corpora version 1.0 and 2.0 were used in the

*E000/CCL, Building 51, Industrial Technology Research Institute, Chutung, Hsinchu 31015, Taiwan, R.O.C. E-mail: changch@e0sun3.ccl.itri.org.tw