Computational Linguistics and Chinese Language Processing
Vol. 11, No. 3, September 2006, pp. 183-200

An Approach to Using the Web as a Live Corpus for Spoken Transliteration Name Access

Ming-Shun Lin*, Chia-Ping Chen+ and Hsin-Hsi Chen*

Abstract

Recognizing transliteration names is challenging due to their flexible formulation and lexical coverage. In our approach, we employ the Web as a giant corpus. The patterns extracted from the Web are used as a live dictionary to correct speech recognition errors. The plausible character strings recognized by an Automated Speech Recognition (ASR) system are regarded as query terms and submitted to Google. The top *N* snippets are entered into PAT trees. The terms of the highest scores are selected. Our experiments show that the ASR model with a recovery mechanism can achieve 21.54% performance improvement compared with the ASR only model on the character level. The recall rate is improved from 0.20 to 0.42, and the MRR from 0.07 to 0.31. For collecting transliteration names, we propose a named entity (NE) ontology generation engine, called the X_{NE}-*Tree* engine, which produces relational named entities by a given seed. The engine incrementally extracts high co-occurring named entities with the seed. A total of 7,642 named entities in the ontology were initiated by 100 seeds. When the bi-character language model is combined with the NE ontology, the ASR recall rate and MRR are improved to 0.48 and 0.38, respectively.

1. Introduction

Named entities [MUC 1998], which denote persons, locations, organizations, etc., are common foci of searchers. Thompson and Dozier [1997] showed that named entity recognition (NER) could improve the performance of information retrieval systems. Capturing named entities is challenging due to their flexible formulation and novelty. The issues behind speech recognition make named entity recognition more challenging on the

* Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 106 Taiwan
E-mail: {d91022 , hhchen}@csie.ntu.edu.tw
+ Department of Computer Science and Engineering, National Sun Yat-Sen University, 70, Lien-Hai Road, Kaohsiung, 804 Taiwan
E-mail: cpchen@cse.nsysu.edu.tw

[Received February 16, 2006; Revised June 23, 2006; Accepted July 10, 2006]