MicroChiP
- Microsoft Chinese Processor

Zixin Jiang and Andi Wu

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-6399

Jiangz@microsoft.com
Andiwu@microsoft.com

(425) 703-6847
(425) 936-0985

Abstract

MicroChiP is a natural language processing system for the Chinese language. It is implemented in the general-purpose natural language processing system developed at Microsoft Research. The Chinese system has been under development for more than two years and it is now capable of analyzing a wide range of sentences with reasonably fast speed (about 2.5K bytes per second). In this demo, we will present the general analysis components as well as some specific capabilities of the system.

General Analysis

(1) Word Segmentation
We will demo our sentence-based word segmentation. Unlike most existing word breakers, our system bases its final segmentation decisions on the syntactic analysis of the sentence, thus surpassing other systems in terms of accuracy.

(2) Parsing
We will show the results of our syntactic analysis. We will also show how the parser can gracefully handle sentences that are ungrammatical or beyond the current coverage of the grammar.

(3) Logical Form
We will present the logical form produced by the system. This will be a showcase of the advantage of code sharing among different languages. Relatively little development time was taken in adding this component to the system because the mechanism for constructing logical forms is largely language-independent.

Specific Capabilities

(1) Customizability of word segmentation
The results of word segmentation produced by MicroChiP can be customized to meet the

requirements of different standards and different needs of the user. This capability is necessary because of the disagreements on segmentation standards among Chinese linguists and the diverse needs of various NLP applications. We can now let the users themselves choose the standard by setting a number of parameters.

(2) Customizability of part-of-speech tagging

In part-of-speech tagging, MicroChiP can produce different labels according to the tag set provided by the user. This capability is necessary because of the disagreement on syntactic categories among Chinese linguists and the diverse needs of NLP applications.

(3) Automatic typo detection and correction

MicroChiP has a mechanism to automatically detect the wrong character(s) in a sentence and suggest the correct character(s) to use. The correction is based on the syntactic analysis of the sentence.