Resource-based Natural Language Processing

Hitoshi ISAHARA
Computational Linguistics Group
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
isahara@nict.go.jp

Abstract

Research on natural language processing (NLP) started with so-called rule-based methodology, however, compilation of huge amount of grammar rules and dictionary entries are too difficult to develop practical systems. Then, trend of NLP research shifted to corpus-based, or statistical, systems. Thanks to the rapid improvement of computer power and data storage, nowadays we can utilize huge amount of actual linguistic data. Combining such linguistic resources and high quality language analyzer, we can extract useful linguistic information and develop practical systems for specific domain.

However, the future direction of NLP is still not obvious. Fusion of knowledge and example, or knowledge processing using linguistic resources, is one of the possibilities to develop high-performance NLP systems. As for the research target, machine translation with new paradigm and information retrieval as practical tasks are promising. To realize the fusion of knowledge and example, we try to make a computer system that utilizes linguistic knowledge of different degrees of abstraction as humans do, to make a model of human language function based on the system, and to acquire knowledge on how do humans store and use this kind of knowledge in their minds.

Based on this consideration, we are developing widely applicable and high-performance NLP technologies and linguistic resources and will open them to the public. As for linguistic resources, we already compiled and published several huge resources and aim to be one of the world-biggest NLP resource centers. We are working on the development of linguistic resources (corpora, dictionaries and other tools) as a basis of resource-based NLP. Some of them are the NICT Multilingual Corpus, the NICT Japanese Learner English (JLE) corpus, Japanese-English News Article Alignment Data, 2-million parallel sentences between Japanese and English patent documents, The Corpus of Spontaneous Japanese (CSJ), and The EDR Electronic Dictionary.

We are working on R&D into NLP using language resources. This work involves the development of fundamental NLP technologies to be utilized in machine translation systems, and the development of technologies to support the creation of large-scale corpora. To achieve efficient compilation of linguistic resources, we have developed the method to extract parallel sentences automatically from Japanese and English comparable texts. The tools which has been developed and published include parallel text alignment software and NLP software using the maximum entropy method.

We are forging ahead with practical research on augmented example-based machine translation as a verification of high-performance resource-based NLP, which is based on leading-edge research on computational linguistics and natural language processing. We launched five

978-1-4244-1610-3/07/$25.00©2007IEEE

year national project on the development of Chinese-Japanese machine translation system for scientific document in 2006.

To attain a profound understanding of language, we are working on the automatic acquisition of lexical knowledge from large-scale corpora. This result is being applied to support system for web retrieval by showing related word list to users.

We are compiling and publishing large-scale linguistic resources and incorporating knowledge on computational linguistics in order to overcome language barriers. By these means it aims to establish multilingual information processing technologies with which to harvest useful information from large quantities of web documents and multilingual texts.