Multiple-level knowledge discovery from corpus

Wang JianDe, Chen ZhaoXiong, Huang HeYan
Institute of Computing Technology
Chinese Academic of Science
PO BOX 2704 , Beijing China
100081
wangjiande@sina.com

Multiple-Level Knowledge includes knowledge of morphology, syntax, semantics, and style. These four levels of knowledge correspond to the levels of natural language, and thus can be of use in language learning, language engineering, language processing and others areas related to linguistics.

Corpora contain large amounts of language data, marked up with a set of attributes; we can derive knowledge of linguistic rules from the annotated corpus.

1) Lexical knowledge can be obtained from statistics derived from the corpus.
2) Knowledge of syntax also can be obtained from corpus data, by means of the Machine Learning algorithm.
3) Semantic rules can be found in the classified document of different semantic type through the statistic model and semantic learning function.
4) A style model can also be derived from these documents. According the arrangement of feature words and syntactic structures in the documents and the learning algorithm of Style Model.

The four levels of knowledge are not independent; they interconnect, and can convert to each other in some time.

This system is intelligent. The information of user operation and the corpus are stored as the translation resource. With the statistical model and Machine Learning algorithm, some knowledge can be extracted from the resource. The system combines the different methods of machine translation that abstract the advantage of RBMT and EBMT, and use the database to store the sentence human translated. The knowledge of translation is divided into three kinds: public, protected and private, that can be used to different translators. So when MT uses this knowledge, the system is more and more intelligent.

The above discovery can be automatic, but it needs much more corpus data, especially annotated data. So we are designing the computer-aided Multiple-Level Knowledge Discovery tool.