12 Conclusion

The ability to extract desired pieces of information from natural language texts is an important task with a growing number of potential applications. Tasks requiring locating specific data in newsgroup messages or web pages are particularly promising applications. As the amount of textual information available on-line grows, information extraction systems will become more and more important as a method of making this information available and manageable. Manually constructing such information extraction systems is a laborious task; however, learning methods have the potential to help automate the development process.

This dissertation has presented a complete information extraction system not only for research projects but also for practical application domains. This dissertation presented two viewpoints for information extraction system: a technical view and a system view.

In technical view, novel rule representation and a relational learning algorithm for learning information extraction rules were introduced. Various machine learning algorithms are used to generate different rules, patterns and also relationships. CapturePlus uses a supervised bottom-up algorithm to construct XSL-Patterns, statistical word occurrence analysis generates keywords for extracting part of text, hierarchical iterative learning to generate complicated concepts and also statistical spatial relationship analysis to contribute spatial model for template filling. The techniques are evaluated with different document types and various application domains. Evaluations showed that these learning algorithms are effective and efficient. Additionally, lots of techniques for pre-processing (such as document segmentation), result merging, extracting using constraints and combination different extracting techniques are developed. Such techniques enable CapturePlus to become an information extraction system for practical real applications. Last but not least, a new XML-Schema enabled template model is introduced for an advanced information extracting system with sophisticated template filling, taking into the structured constraints of a document more rigorously.

In the system view, a complete information extraction system CapturePlus is designed and implemented. The system distinguishes the difference between NE-tasks and TE-tasks clearly. The work showed that the task for template filling is not only a part of name extracting. Actually, template filling is more sophisticated than simple assignment of instance to template element. CapturePlus contains various Pre-processing component, NE-components and TE-models. Furthermore, search profile is used in the system. With help of search profile, extracting rules can be combined with extracting constraints
Information Extraction

and also with each other. The system provides portability, trainability, and configurable. CapturePlus is already used in several application domains such as email filtering, document management and information gathering from the Internet for various industrial clients.