11 Future Directions

The field of information extraction is fairly young and the application of machine learning to IE is an even more recent development. Simultaneously new technologies such as Semantic Web or XML, which will simplify the extraction of structured information from heterogeneous sources, are coming in the real applications. With the new technologies IE, techniques will become more robust and more useable in the real-life applications, rather than only in research areas. In following, we distinguish the possible future directions in technology and in system.

11.1 Future Directions in IE Technologies

Although there are many possibilities to do local optimisations, the performance of IE systems will not be enhanced considerably by such local optimisations. Individual solutions show good results for defined, isolated IE problems. One of the barriers to making information extraction a practical technology is the cost of adapting an extraction system to a new scenario. In general, each application of extraction will involve a different type of scenario. If implementing this scenario requires a few months of effort and the skills of the extraction system designers, the market for extraction systems will remain limited indeed. We need to have tools which will allow potential users to adapt such a system, and create an initial system in days or weeks, not months.

The basic question in developing such a customization tool is the form and level of the information to be obtained from the user (assuming that the goal is to have the customization performed directly by the user rather than by a expert system developer). If we are using a “pattern matching” system, most work will probably be focused on the development of the set of patterns. However, changes will also be needed to the template filling level, to the set of inference rules, and to the rules for creating the output templates. Unless users have considerable experience with writing patterns (regular expressions with associated actions) and some familiarity with formal syntactic structure, they may find it difficult and inconvenient to operate directly on the patterns. One possibility is to provide a graphical representation of the patterns (i.e., a finite state network), but this still exposes many of the details of the patterns.

Various Machine Learning algorithms provide simple way to create extraction rules. However, the automatically generated rules are often not satisfactory. They are either too strict (poor recall) or too rough (poor recall). Developing sophisticated and adaptive learning algorithm is a solution to the problem. Another solution could be a compensation mechanism, so that the error taken
by bad rules can be reduced with post-processing (such as TE-Profile in CapturePlus).

With the developing of new document description and presentation technologies, such as XML, documents can be written with different format than free text or HTML. IE should take more interest in such formats. Semantic Web and RDF, which is introduced by W3C, aims to represent Machine-Understandable information in Web pages (Berners 1998). Such documents are written in XML (actually in RDF, which is more formal than XML) or an ontology language such as Daml-Oil. The structured XML documents can be extracted by some standard query techniques (such as XSL-patterns or X-Query). How to create such queries, especially automatically, is another challenge.

For the task of template filling, template model can be extended from simple flat tree structure to sophisticated XML-Schema enabled model. A new TE-Agent should be designed and implemented with cardinality supporting and conflicts solving as well as preserving more probabilistic spatial constraints (see also Section 10.5).

11.2 Future Directions in IE Systems

Various systems are developed for defined IE problems and tasks. However, most of such systems are either lab products that can only be evaluated in pre-defined couples, or they are hard coded for isolated practical problems. To bring the IE systems to the real application world, systems must perform good precision and recall. This requirement can be achieved, by enhancing of IE techniques, but also by designing a good system framework. The fact that each IE component and algorithm can only solve defined, isolated tasks forces a good corporation and combination of such single techniques. Hence, the system must be able to integrate new IE components without much difficulty. The approach of an open system is introduced in order to perform a flexible and adaptive information extraction system. GATE (Cunningham 1997) is a good example for this approach.

The new version of CapturePlus is designed as distribute multi-Agent open IE system. This approach is based on the newest .NET technology and XML enabled SOAP. Figure 52 shows a blueprint of new CapturePlus.
Figure 52: Construction of a multi-Agent open IE system

The IE system consists of various IE agents, which carry out defined IE tasks. Agents are configured as autonomy server. The communication between agents and agent scheduler is taken place by CORBA, COM or XML-enable SOAP technology. Agents could be in one computer, or be distributed in different places. For the later case, a distribute data management is required in order to offer a consistent data storage.

Each agent has his own extraction rules format, which is defined with an XML file. Each search profile consists of a set of search queries. Each search query contains some query terms and a logical operation option (AND or OR) to migrate the search results extracted by the search query terms. The search query term defines search method and query information (such as keywords or rules) with respect of search options (such as match case, whole word searching etc.). For each agent there is a defined XML format to describe search query.

At first the NE Scheduler gets a search query described in XML, from the NE Profile set. Next, the Scheduler passes out query terms to corresponded NE Agents. The selection of agents is defined in the XML query file. Each NE Agent receives its own search query term described also in XML and processes the extraction. The transformation from search query XML in
search profile to various specific search query terms XML in different NE Agent takes place using XSLT. The extraction sequence is controlled by the NE Scheduler, which can also be configured by the user in a GUI.

Another direction of future systems are integration of current IE systems. The output of one IE system could be the input of another, if various IE systems focus on different tasks. Such combination could be implemented easily if there is uniform exchange formats. One possible solution is using XML to defined the data formats and also control routine.