3 Approaching the Problem

3.1 Types of Information in the Natural Language Texts

Deciding how to approach the IE problem the question stated above has to be thoroughly considered: "How does the information manifests itself in the natural language, what components of the language bear the information?". I concluded that three kinds of information are communicated by a natural language text. One of them, let us call it *primary* information, is explicitly expressed by main clauses and except for coreferences does not depend on other text parts. That is, the expressions comprising primary information are self-contained, they can be evaluated fully out of context. Primary information must be recognized before the *intersentential* information can be comprehended. It expresses the causal, chronological, conditional, etc. dependencies between main and subordinate clauses, different sentences, paragraphs, chapters, any text parts. The intersentential information turns a collection of sentences into text and allows in the first place a connected description of complex events, stories, opinions etc.

The third *implicit* kind of information can be derived from the first two kinds using an external information source, generally background or world knowledge. Implicit information cannot therefore be deduced from the text alone and requires the interrelation of text content with other knowledge acquired earlier. The majority of texts require external knowledge to be understood. However, implicitness of information can be of different origin too. Saying: *The furious battle between pirates and government troops ended in the crack of dawn as the pirates raised the white flag*, we have to consult our world knowledge and retrieve the fact that white flag is a symbol for a defeat to conclude, who won the battle. Consider on the other hand: *She lived in Berlin. James was born in the small homonymous town in Texas*, where it is sufficient to know the semantics of the word "homonymous" to infer James' place of birth.

The core of the problem stated above is extraction of the first two types of information. The attribute values of the target structure can be identified by recognizing and evaluating the primary information. The intersentential information helps to resolve ambiguities and localize the attribute values and discloses their connections to each other. Therefore capturing intersentential information is the most important prerequisite for relation extraction. The implicit information in general cannot be extracted without extensive knowledge bases and logical infer-

ence and is therefore out of the scope of IE. However, it can be gained by human analysis of the extracted data or processing it by special knowledge-based algorithms. In special cases (e.g. when the implicit fact occurs somewhere in the text explicitly, cf. the second example) even the extraction of implicit information is possible.

Examining how the primary and intersentential information is expressed, one realizes that besides the semantics of words and phrases the context and the domain play an important role in unambiguously determining the contained information. Consider the fragments *Karl fell in love with Sofia* and *Athens defeated Sparta*. Regarding their context *Karl fell in love with Sofia because of untypical for big cities jovial flair*, *Athens defeated Sparta 2 - 0* resolves the uncertainty about possible interpretations. The same conclusion would be suggested considering text domains "Capitals of the world" and "Soccer". Thus to even cover the primary information it is not sufficient to subsequently capture the semantics of words or phrases. The semantics of the whole sentence as the fundamental element of the language has to be captured that expresses completed thoughts and combines contents belonging together.

3.2 Choosing an Appropriate Approach

The above considerations serve as the main criterion choosing an appropriate approach to the problem. Statistical approaches hardly can capture complex semantics of a sentence because it cannot be "divided" in primitive features (see also the argumentation in sec. 2.4). Another serious drawback is the black-box principle of internal statistical models where you can evaluate and correct the behavior of the system only based on final results. As the intermediate state of the system can be hardly interpreted, the explanation of errors is significantly complicated and sometimes impossible.

Knowledge-based approaches make a thorough modeling of the domain and explicit formulation of human knowledge possible. However the effort of creating the knowledge resources is disproportionate to the size of the domain. Considering also the adaptation costs for new domains, this approach does not scale.

Rule-based approach combined with the a priori defined target structure offers powerful mechanism for handling the syntactic and semantic complexity of natural language without having to abandon the advantages of trainability and adaptivity. Complex semantics of a sentence can be captured in the rule patterns developing the semantic interpretation of a sentence by assigning certain roles – namely, the types of the target structure – to the relevant sentence parts. In contrast to statistical techniques rule-based approach guarantees the transparency of the internal system state. When developing such a complex system it is important to know how the learned "knowledge" of the system evolves to be able to explain unexpected or dissatisfactory results and improve the algorithms. The knowledge of the system is declaratively specified in form of rules and can be inspected at every moment during runtime. We therefore consider the rule-based method to be the most appropriate approach to the problem of information acquisition from natural language texts.

Differences to other rule-based approaches

The range of rule-based approaches is quite wide (cf. sec. 2.2), there are significant differences in the rule concept and learning algorithms. Because of the

new algorithm for rule induction and considerably different notion of linguistic patterns our approach can hardly be assigned to a certain subclass of rule-based approaches identified in 2.2. In the following major differences to other rule-based approaches are motivated and explained.

While the general notion of extraction rule specifying what has to be extracted on its left-hand side and how the extraction takes place on its right-hand side is pretty common for all approaches, the concept of rule pattern (the left-hand side of the rule) varies. Most of the approaches use simple, hardly structured patterns as rule components. Riloff restricts the patterns to be a phrase with a keyword and distinguished syntactic constituent, Crystal uses sequences of lexical items and semantic classes, $(LP)^2$ utilizes six fix features obtained from annotations made in the source documents, Nobata even gets by with syntactically enriched text sentences as patterns. Few systems (FASTUS, Gate/Annie, Whisk) encode the patterns as regular expressions using FSMs for parsing and matching. None of them uses higher-order pattern language, which makes the representation of such crucial patterns as negation or positional permutation not possible. Simple pattern language models cannot reflect the complex syntactic and semantic structure of the natural language. Therefore we propose a pattern language that is capable of representing context-free structures, negation and permutation, but also semantic and syntactic features. Resulting patterns in contrast to the mentioned are powerful and expressive enough to capture non-trivial kinds of phrases and sentences containing relevant information.

Moreover, patterns are not restricted to the plain text, they can relate to and include arbitrary XML elements. Many texts e.g. in the Internet contain additional markup for layout, structuring purposes, semantic annotations etc., that can serve as additional features for identification of relevant facts. Incorporating existing markup and linguistic annotations in the patterns increases their coverage and is especially effective for the semistructured texts. After integrating existing markup and linguistic annotations in a single XML document the rule patterns can serve as XML queries unprecedentedly reducing the problem of IE to query evaluation.

A major difference to other approaches is the algorithm for acquisition of extraction rules. Some systems acquire rules by generating huge amounts of candidate rules and selecting those with the best rankings [Ril98, Col96]. The rules are not refined because of limitations of provided semantic information or dynamic target structure. Covering algorithms improve rules iterating over the training corpus and generalizing acquired rule patterns. Often the generalization is achieved by abstraction of patterns (e.g. in $(LP)^2$) while no merging is performed. Merging patterns allows to exploit similarities of different extracted instances to achieve more effective generalization than mere abstraction of initial rules. Our approach tries to exploit all kinds of rule improvement including correction of rules to obtain an optimal rule set for extraction. Beginning with the rules generated from fact instances, which were extracted by the human, rules are corrected and generalized based on their extractions. The valuable human knowledge in form of training examples is embedded in the crucial component improving the rule quality. The inductive learning is based on the inductive definition of pattern language. The generalization of rules involves not only merging and abstraction of single patterns, but learning from the negative examples and syntactic generalization. All generalization heuristics are formally specified using the formalisms of pattern language. Thus the generalization component is more variable and therefore powerful and supported by a formal base.

The learning algorithm is aided by automatically created semantic resource revealing semantic relations such as synonymy and hypernymy between words. This component plays an important role for semantic abstraction of patterns. Other systems employ handcrafted thesauri except for Riloff's system, which automatically establishes a semantic lexicon. However, this lexicon contains only nouns with information about their semantic category.

In contrast to the systems of Riloff and Nobata it is not intended to achieve the biggest possible automation of the system reducing the human participation in the training of the system to all costs. Both authors tried to replace the human supervision by semantic resources (such as partially filled semantic lexicon, mapping of template roles to semantic domain categories). Humans controlled the intermediate results without actively intervening the learning process. The reasonable human effort for creation of training corpus is absolutely justified to avoid making simplifying assumptions (e.g. ignoring unknown words, presuming the existence of certain keywords in any text etc.) and to prevent the domain restrictions and loss of extraction quality.

Some rule-based systems do not presuppose a fix, a priori specified target structure (Crystal, $(LP)^2$). Riloff and Schmelzenbach propose a dynamic construction of target structure from predefined attributes based on the text content. A target structure adequately reflecting the domain used in our approach fulfills the double purpose. On the one hand, it directs the focus when localizing relevant content in the text, on the other hand, it provides semantic description of the domain, making the relations and dependencies between the templates and attributes of the target structure apparent. A priori specification makes the learning of extraction rules possible, since the desired information is characterized in advance and the target structure does not alter. From the theoretical point of view the target structure provides a noteworthy interpretation of information extraction as nor surjective neither injective function mapping the text content to the attributes of target structure.

Statistical and some rule-based systems reduce the identification of relevant content to detection of its start and end token. The inner structure of the extractions (in case that extractions consist of several tokens) is partially captured by the context models, but the semantic and syntactic integrity expressed by the interdependence of all tokens gets lost. Moreover, the correctness of an extraction depends on two decisions, an error in determination of one edge leads to an incorrect extraction. In fact, identification of relevant fragments localizing their borders has the advantage of easier and faster learning (it is easier to learn the properties of an edge token than those of complete extractions, fewer training instances are required) and the majority of attributes (especially with numeric and formatted values) can be adequately characterized by this model. For attributes that lack such distinguishing features capturing the attribute values as a whole provides a much more expressive and characteristic description. The extraction rules contain therefore the generalized description of complete extractions to adequately cover a wide range of potential attribute types.

Another important difference concerns the context model. The majority of rule-based and statistical approaches regards the context in an immediate vicinity of extracted fragments setting the size of the context window to a fix value. As opposed to fix context window we capture the context in surrounding contextual constituents with variable length. An important contextual feature providing evidence of relevant text fragment may be in a considerable textual distance to this fragment due to grammatical specialties of a natural language (cf. placement

of the main verbs at the end of subordinate clauses in German) and in this case would not be captured by a fix context window. While immediate context is regarded by context windows even though it is not relevant, our context representation includes only context indicated as relevant by at least two extracted instances independently from its distance to the actual extractions. This more general and flexible context model allows for complex syntax of a complete sentence abstracting from a local view of a fix context window.

3.3 Goals and Requirements

To compensate the insufficiencies of the classical rule-based approach human effort should be adequately replaced enhancing the rule-based method by learning component. The main goal of my dissertation is the development of an adaptive algorithm for inductive learning of extraction rules. Given a set of training examples in form of annotated extractions from a training corpus and a specified target structure it learns the rules for localization and extraction of relevant information, improves and generalizes them so that they can be applied to any text from the specified domain to perform actual information extraction. The algorithm will not depend on any domain and should be universally applicable. The amount of human supervision and training effort should be reduced as much as possible. Therefore creation of rules will be performed automatically by deriving them from instances of facts found in training texts.

The developed algorithm shall be thoroughly evaluated with the goal to demonstrate its effectiveness in comparison with other state of the art approaches and to assess the potential of the system comparing its results with human performance.

The development of the adaptive IE algorithm and its quantitative evaluation, though, are not the sole aspect of this work. Another important aim is to scrutinize the purpose, practical usefulness and advantage of IE technology in general using experimental results of our system GROPUS and other considered systems. It is to evaluate in what conditions application of IE is expedient, how the task should be, what kinds of texts can be managed. For this purpose an investigation of performance on text corpora with different text styles is envisioned. Important conclusions can be drawn from the analysis what factors influence the extraction quality, what are the sources of errors etc. Finally, a characterization of the range of applications, environments can be achieved where the presented approach may be usefully utilized.

Requirements

Based on foregoing problem analysis and differentiation between our and related approaches to IE we can formulate additional expectations on our system:

- ▷ The amount of training data should be reduced remaining in a reasonable relation to the size and complexity of application domain. Human effort that is necessary for annotation of training corpora is one of the reasons for the quite low utilization of IE technology. A considerable amount of human effort is replaced by the learning component. Additional improvement can be achieved making the learning procedure efficient, that is, reducing the number of training examples without significant loss of extraction quality.
- ▷ It has to be ensured that the algorithm achieves good performance without additional manually prepared semantic sources. The success of information

extraction can be notably increased using domain-specific semantic information, such as gazetteers, semantic dictionaries, type taxonomies etc. In the most cases such resources are produced by researchers to tune the system performance to the particular domain. This implies additional human effort that increases according to the domain size. We consciously do not embed any additional semantic sources in our system to examine the capabilities of our approach without fine-tuning and any human aid since the existence of such resources cannot be presupposed.

- ▷ Internal state of the system and committed errors must be explainable at any time during learning and application. This presupposes a transparent extraction model that can be realized declaratively specifying extraction rules. The advantage is that the learning steps leading to erroneous extractions can be exactly retraced. The system component responsible for the errors can be identified and eventually improved.

- ▷ A confidence measure for the correctness of extractions should be provided. The quality of extractions of the state of the art IE systems is still far from perfect. Therefore the extraction results can hardly be used for further processing without subsequent human quality control. Introducing a confidence measure would distinguish doubtful extractions from those the system was confident in. Besides, this additional qualifying parameter can be used for optimization of extraction results by selecting extractions with a certain minimum confidence measure.

- ▷ In order to make the learning algorithm more explicit and put it on a solid theoretical base it will be formalized. Rule patterns as well as operations on rules such as generalization, abstraction and correction will follow a formal specification that allows efficient implementation and later optimization of the approach.

- ▷ The algorithm has to be able to process free, semistructured and structured texts. The whole spectrum of WWW and XML documents should be covered. There are numerous approaches that are specialized in a certain kinds of text like structured or free texts. Our approach will handle all kinds of texts with a single algorithm. The evaluation will include the investigation on both free and semistructured texts.

- ▷ The learning has to be fast. Long training times are often justified by the argument that training is performed offline before the actual application, and the extraction is carried out significantly faster. This camouflages the fact that also during application a system has sometimes to be retrained due to the alternation of text input. Besides, extensive and thorough experiments cannot often be conducted because of lengthy training times.

- ▷ The algorithm should not be adapted to a specific language[1]. Provided that linguistic tools for syntactic and morphological analysis are available for a language, the algorithm can abstract from the peculiarities of the language and operate on the output of linguistic tools. Grammatical and lexical information do not have to be encoded since they can be inferred during the learning procedure. To demonstrate the language comprehensiveness of the approach it will be applied to German and English text corpora.

[1] At least Indo-European languages that share similar syntactic and morphological base are supposed to be adequately covered by the algorithm. No predication about its effectiveness for other language families can be made