13 Conclusion

Information Extraction aims at extraction of facts expressed in natural language that are underlain by a common predefined schema. The spectrum of extracted information may be defined quite widely reaching from identification of named entities to extraction of complex relation tuples. Current approaches to IE solve the basic task of identification and extraction of single attribute values. GROPUS is able to identify partial relation tuples that consist of attribute values occurring in one sentence. The identification of complete n-ary relations can be accomplished only constraining the number of relation tuples per text.

The primary objective behind IE is the idea of querying a collection of texts in a classical database fashion. The accomplishment of this goal is complicated because the results of information extraction will always be afflicted with uncertainty. However, this does not rule out IE for many practical applications that can abandon the requirement for absolute correctness and completeness. In particular, the integration of IE in the process of knowledge management promises a big gain.

13.1 Summary of our Approach and Contributions

We approached the problem of IE building on the assumption that despite the huge diversity there are certain linguistic patterns in the natural language that are used by humans for expression of certain information. In contrast to the popular statistical approaches we capture these patterns explicitly by formally encoding them and maintaining so the semantic interconnection between different linguistic elements. To live up to the complexity of natural language we introduced an expressive context-free language that maps natural language patterns to their formal representation. The formal patterns are not restricted to lexical elements. We tried to exploit all available and inducible information including available markup, data obtained by linguistic analysis and implicit text layout in order to achieve expressive and characteristic patterns.

To cope with the multifariousness of the natural language on the word level we proposed a new approach to synonym recognition based on the lexical graph constructed by reflecting the syntactic dependencies of words in a sentence in the graph structure. The synonyms are determined identifying special vertices providing evidence of the sense similarity and calculating the synonymy distance

based on the lengths of paths through these vertices that in turn depend on the statistical frequency of word co-occurrence.

Interpreting linguistic patterns as XML queries we developed a new notion of Information Extraction. The pattern-query dualism implies that pattern specification language is at the same time XML query language. We established its formal semantics introducing a novel extended sequence semantics that is especially useful for NL texts wrapped in XML. We presented an efficient query processing algorithm that is polynomial in the number of nodes in the subtree of the context node for a fix number of backtracking patterns.

The core of our approach is the inductive learning of extraction rules from a set of training examples provided by a human annotator. We introduced a novel learning algorithm featuring three central components for generalization, validation and correction of rules. In the context of rule generalization we had to solve the difficult problem of efficient determination of rule similarity that implied the comparison of hierarchical structures. Transforming hierarchical rule structures into sequences and reducing the similarity assessment to sequence alignment we could propose an algorithm that required the minimum runtime for an optimal alignment of two sequences. Rule generalization significantly benefits from the formal specification of patterns that allows to specify and implement generalizing heuristics as mathematical functions. Rule validation optimizes the set of induced rules for maximum extraction quality expressed by F-measure. Three different validation strategies have been presented making an accent on the overall rule reliability, extraction goodness of single attributes and the diversity of rules. Their effectiveness as well as the potential of other components of the learning algorithm were investigated empirically.

We extensively evaluated our approach trying to answer several questions concerning the potential of GROPUS as well as IE in general. We examined its efficiency in different environments that significantly varied in the properties of texts and the target structure. To assess the capabilities of GROPUS with respect to the current state of the art a comparison with other approaches took place in every test domain. In order to establish an objective and reasonable yardstick for interpreting results achieved by IE approaches we conducted an extraction experiment with a human using the same experimental setup. Another series of experiments served for identification of external factors and inner components that have a decisive impact on the extraction quality. In the following section we are going to discuss the evaluation results and draw possible conclusions.

13.2 Discussion of Results

13.2.1 Performance of GROPUS and other IE approaches

Performance on the Test Corpora

GROPUS has been evaluated in three different application domains. Its performance has considerably varied depending on the domain as well as on single attributes. On the Bosnian corpus that features quite restricted language style and a heterogeneous target structure GROPUS could handle well the extraction of regular attributes having more difficulties with heterogeneous attributes and failing to extract attributes with a very small number of training instances. The experiments with the size of the training corpus indicated a continuous rising tendency for bigger training corpora and confirmed that the size of Bosnian cor-

pus is too small to estimate the potential of GROPUS. The overall extraction quality was identical with the statistical system TIE.

The seminar announcement corpus features many partially structured texts and a simple target structure, so that the number of training instances per attribute is much higher than in Bosnian corpus. Besides, the applied *one answer per document* evaluation mode simplifies the extraction task requiring only one among several possible attribute values to be extracted. With a recall rate around 80% and precision around 90% GROPUS achieves a satisfactory performance outperforming the majority of state of the art rule-based systems, but being inferior to most statistical systems. Further improvement can be achieved including additional character features in the pattern specification language that play a major role in the identification of attribute values in the structured parts of seminar announcements.

The error analysis exemplarily performed on the SA corpus revealed that a big portion of errors could be obviated by a better selection of attribute values. Missed extractions constitute approximately one third of all errors, while almost one third happens due to attribute confusion and incorrect choice of the extracted value among identified candidates. An important consequence is that GROPUS identifies more correct attribute values than reflected by the recall value loosing some true positives by the imperfect selection in the *one answer per document* mode.

MUC corpus has the most challenging language style and target structure. It is therefore not surprising that the overall extraction quality was lower than on other two corpora achieving F-measure values under 30%. GROPUS obtained the best results significantly outperforming TIE and ELIE L1 (that have been among the best three systems on the SA corpus) and surpassing also the results of ELIE L2 (the best system on SA) by a couple of percent in the F-measure value. The performance for different attributes varied widely. The extraction of semantically close attributes turned out to be most challenging.

Beside the complexity of attributes the main reason for unsatisfactory results is the content of the texts that comprised a lot of ambiguity and redundant information that severely misled GROPUS and other IE systems. Relevant texts contained in average more than three relation tuples (i.e. descriptions of three different terrorist acts) and numerous coreferences of extracted attribute values. Since the coreferences have not been annotated, the systems had to guess the extracted occurrence of the attribute value, while extractions of other occurrences were counted as incorrect. In contrast to other two corpora the recall has been bigger than precision on the MUC corpus, which underlines that the selection of reliable extractions has been more difficult than the actual identification of relevant content. The recall value could have been even considerably bigger if all reasonable and semantically correct extractions of unannotated coreferences had been regarded as correct and the extraction rules had not been compromised by ostensibly wrong (but semantically correct) extractions at the training stage.

Interpretation of Human Results

The conclusions about automatic information extraction from the MUC corpus are also supported by human results. Less than the half of all attribute values have been correctly identified and only one third of all human extractions has been correct. Because of his text comprehension the identification of relevant content can be well handled even by a human non-expert, which is contradictory to the reached recall value below 50%. The even lower precision confirms the dif-

ficulties in selection of relevant content. The qualitative investigation disclosed similar error origins as in case of GROPUS including extraction of unannotated coreferences, attribute confusions and extractions of content related to the attributes of the target structure but regarded as negligible by the creator of the corpus. Even though we cannot regard the results of an experiment with a single human as representative, their very low range together with the low extraction quality achieved by IE systems are the evidence of the high inherent complexity of the extraction task on the MUC corpus.

GROPUS vs. Statistical Systems

While on the Bosnian corpus similar results have been achieved, statistical systems have been superior to GROPUS on the seminar announcement corpus, but have been outperformed by GROPUS on the MUC corpus. Considering the characteristics of the application domains and the substantial differences in the approaches these results are not surprising.

Seminar announcement corpus consists of partially structured documents featuring a form-like summary and the free text part. Many attribute values are located in the form-like part of the text and therefore not in a sentence. The text fragments in the free text part are in the most cases extracted from different sentences, which results in low information density. Therefore the more relevant characteristics of attribute values are features of single tokens like capitalization, case sensitivity, punctuation, character type (e.g. only numeric or mixed characters), certain formatting. These features can be very well captured and learned by statistical systems because their approach is based on classification of tokens. On the contrary, GROPUS cannot take advantage of its sentence-based context model, and its instruments for description of character and token features are by far not as rich as the feature vectors of statistical systems. Statistical systems are therefore generally more suitable for the extraction from non-grammatical, telegraphic in style texts.

The strength of GROPUS is its ability to explicitly capture semantic and syntactic interdependencies of attribute values and their context in a sentence. Its extraction rules can fully unfold their effectiveness on grammatically correct free texts that feature high information density. The major criteria for identification of relevant fragments in such texts are their semantic and syntactic interactions with the context so that the adequate context models are essential for the extraction success. In such environments GROPUS can leverage its sentence-based pattern model while statistical systems have difficulties to capture complex context because of their token-based view. In contrast to semistructured texts relevant fragments in the free texts are hardly characterized by simple statistical properties (such as word lengths, character properties) offering significantly less discriminating features for classification of tokens. GROPUS relies more on semantic and syntactic relations within a sentence expressed in its patterns by lexical and positional interdependencies and is therefore more robust against the absence of superficial features such as token features than statistical systems. The results on MUC corpus indicate that our rule-based approach can better handle challenging free texts than statistical systems.

Runtime

The asymptotic runtime for complete processing of a text corpus comprises the training and application times. The training time depends on the size of the training corpus, the number of induced rules and the number of iterations of the

learning cycle. Assuming a fix time for processing of a text page (i.e. applying all correct extraction rules to it) the training and extraction times are linear in the length of documents (in case of training the runtime is additionally weighted by the number of iterations). Compared to the human effort for annotating training data the computational resources are of minor importance. The turn around time for training and testing of one shuffle of MUC corpus (650 documents) is 7 min, for one shuffle of Bosnian corpus (162 documents) – 20 sec, resulting respectively in 0.6 sec and 0.03 sec extraction time per document using a standard high end PC.

Statistical systems cannot compete with GROPUS featuring much slower runtimes. The average extraction time per document of the MUC corpus is 9.97 sec for TIE and circa 102 sec for ELIE. The explicit modeling of natural language by declarative extraction rules and their induction based on well-defined formalisms proves to be the much faster and more efficient learning model than internal parameter optimization of statistical systems.

13.2.2 Influencing Factors for the Success of IE

We identified three external factors[1] having a significant impact on the success of information extraction.

Size of the Training Corpus

The size of the training corpus and connected amount of training instances play a major role for extraction quality. Increasing small training corpora results in a steep ascend of evaluation metrics. Although there is no fixed value how large a corpus should be, the minimum corpus size should at least be sufficient to complete the phase of the rapid growth of extraction quality observed on every test domain. The expedience of further annotation depends on the relation of manual effort to achieved improvement.

Bosnian corpus was too small to estimate the potential of GROPUS in this test domain since the corpus size could not be further increased even though the graphs of recall and precision featured steep slopes for maximum corpus size. No stable behavior could be reached on the seminar announcement corpus too, even though no significant improvement for bigger training corpora can be expected. Our initial expectation that the size of the training corpus should be adequate to the complexity of documents has been contradicted by the experiment on the MUC corpus. The evaluation metrics converged after the training corpus reached the half of the whole text corpus size. This experiment also emphasized that the limits of an IE system are exclusively determined by the application domain.

External Homogeneity

The experimental study clearly pointed out the influence of external homogeneity (i.e. the ratio of relevant texts in the text corpus). In particular, the precision value benefits from exclusion of irrelevant documents from the text corpus. However, an optimal filtering of texts that do not contain desired information is very hard to achieve, since the differences to relevant texts in the corpus are often very subtle. The state of the art classifiers are not able to reliably distinguish relevant texts so that utilization of text classification before the extraction process does not prove effective.

[1] The influence of the kind of texts omitted here is discussed in the next section

Since a realistic text corpus can contain a lot of irrelevant texts, the IE systems have to handle the potential incorrect extractions. GROPUS demonstrated more robust behavior against irrelevant texts than statistical systems. Comparing the results on preclassified and complete corpora the precision values of GROPUS notably differ, but the recall change is not significant while in case of statistical systems both metrics severely dropped.

Complexity of Attributes

Independent of the application domain the considerable variance of precision and recall of different attributes can be observed. The complexity of an attribute for the IE task depends on several factors. Beside the number of training instances an attribute is characterized by its semantic, structural and morphological properties and degree of locality in the text (i.e. occurrence in a regular context or at a roughly predefined position in the text). For example, date and time attributes were always among the best extracted attributes in every test domain because their values are expressed very concisely and have characteristic numeric formats.

The dependence on the number of training instances has been confirmed by the experiments with size of the training corpus. To examine the influence of other attribute properties we regarded one of the most important quantitative parameters – the expected average number of words in its values (EAL). EAL is a measure for the structural complexity of an attribute that is also affected by its semantic complexity. The experiment on the MUC corpus with a complex attribute LOCATION and its subtype attributes obtained by splitting the values of LOCATION in the subtype values gives a clear indication of the impact of the structural and semantic complexity of an attribute on its extraction quality.

Another interesting observation is the correspondence of the extraction quality of an attribute achieved by different systems relatively to other attributes. Despite the partially big differences of the absolute F-measure values attributes that were best extracted by the rule-based approach were also the easiest to process for statistical systems and the same applies for the most complex attributes. This fact leads to the conclusion that the attribute complexity is inherent to attributes and plays a major role for the issue of suitability of an application domain for IE task.

13.2.3 Utility of Internal Components

Investigating the proposed approach an important question is how the single components of GROPUS contribute to the extraction quality. The most valuable component is the rule similarity metric employing the algorithm for the optimal sequence alignment. Providing important information for rule merging rule similarity has a crucial influence on the induction step and the reliability of extraction rules in general. Rule correction proved to be another reliable instrument for improvement of the rule quality contributing to better precision and recall. The effectiveness of synonymy recognition and substitution heuristic depends on the application domain and is restricted to the enhancement of recall.

In comparison of different proposed validation strategies *RPT+APT* showed the most balanced performance achieving maximum or being close to the maximum achieved F-measure values. Because of its constantly good performance *RPT+APT* is universally applicable and integrated in GROPUS as default validation method. *Local APTs* is especially appropriate for domains with a big

training corpus and loosely interconnected attributes. *Covering setup* proves its strength in environments with low recall and complex target structures.

13.2.4 In what Environments can IE be usefully employed?

The different overall performance of examined systems on different corpora illustrates that the application domain has the crucial impact on the success of IE setting certain performance limits for adaptive approaches. Our initial conjecture was that the complexity of IE task is predominantly determined by the kind of texts in a particular domain. During our research we gained the important insight that in context of IE properties of texts cannot be regarded without considering the target structure. It is possible that the same IE system achieves different performance for two different target structures on the same text corpus. Some characteristics of the texts may or may not be relevant depending on whether they are helpful, harmful or insignificant for identification of attribute values.

Because of this tight interconnection between the texts and the target structure it is difficult to investigate the impact of the kind of text alone on information extraction. Since an application domain is also characterized by other parameters (such as the size of the text corpus, s. the previous section) that also affect the extraction quality, one has to fix several parameters in two different test domains varying only the kind of text. The establishment of realistic text corpora with some predefined characteristics (e.g. given amounts of instances of certain attributes) is very difficult so that a systematic evaluation of the impact of the kind of texts could not yet be conducted. However, based on our empirical investigations on three different test domains several conclusions are possible.

Semistructured texts feature high locality and can be reliably processed by today IE systems. The results on Bosnian corpus demonstrate that attributes with high locality can be adequately extracted even with a very small set of training items.

On all three corpora information was better extracted from the texts that were devoted to a single event (i.e. containing a single relation tuple) regardless of used evaluation mode. Generally, the concise presentation of information with a strong focus on the content captured by the target structure considerably benefits the system performance. The mention of secondary details and redundant text passages turn out to be very misleading.

The high density of information in a sentence that also implies a non-trivial sentence structure is not always a handicap. While statistical systems can better cope with simple main clauses that contain one attribute value, our rule-based approach can take advantage of strong interconnection between different attribute values.

Both classes of approaches have big difficulties with semantically ambiguous target structures. Attributes subsuming semantically heterogeneous values, too general attributes and semantically close attributes involve a significant performance drop because the respective learning models are heavily compromised by incorrect extractions and attribute confusions, which also negatively affects the extraction of other attributes.

13.3 Open Problems

Even though the remarkable progress has been achieved in the field of IE in recent time especially making the systems more autonomous and universally applicable, many problems remain hardly or not yet tackled.

The majority of approaches is not able to recognize and extract multiple occurrences of relation tuples. The systems either cannot handle relations at all assuming the target structure to be a flat slot sequence or it is presupposed that every document contains at most one relation tuple. The major difficulty arises when fragments of information belonging together are scattered in different sometimes distant parts of the text and it has to be determined whether and which fragment belongs to what complex entity. Even the most advanced approaches can hardly handle unification of partial facts beyond the sentence boundaries. The research challenge will be to find a way of reassembling composite information without establishing a complete logical representation of the text content.

A similar problem with totally different background is when the same instance of a fact appears in different forms repeatedly. Due to the richness of natural language we can refer to an entity in very many ways. However, only one occurrence containing the most complete information should be extracted. Current systems often do not recognize that text fragments refer to the same fact and extract them as a new found instance. The most straightforward solution not yet practiced would be to embed the mechanisms of coreference resolution in the extraction algorithm and develop strategies for selecting the most appropriate occurrence. A severe handicap is the unreliable coreference resolution methods, which is also an active research field.

Information extraction suffers from uncertainty of the natural language. Often facts are expressed with a certain degree of tentativeness (e.g. indirect speech: "someone reported that…", "s. o. assumed that…"). In such cases even for humans it is difficult to decide whether the information is factual and hence relevant. This may be regarded as a special case of the general issue how credible the extracted data generally is. An important task will be to determine the degree of reliability of information, which is possible deploying fuzzy methods. We made an important step towards the solution of this problem providing a confidence measure for every extraction that can be interpreted as the probability that the extracted attribute value is correct.

In this context a very important and interesting question is whether the more profound embedding of semantic analysis will contribute to the advance in IE. Recent successes of trainable systems in single-slot information extraction that almost completely forgo semantic resources and analysis suggest that it may be dispensable. But does the good performance for the simplest of IE tasks open optimistic perspectives for much more difficult problems or have the adaptive approaches to IE already reached their limit?

Future research will have to face these problems and questions because in the present state the practical usefulness of IE systems will be restricted to a quite narrow range of selected applications.

13.4 IE in Context of Knowledge Management Systems

Until this point we have regarded the application of IE in autonomous systems, which is complicated by the non-optimal results achieved by IE systems. Alternatively IE can be integrated in a broader data management environment. IE can serve as the valuable component of knowledge management systems that support the analysis, processing and systematizing of large collections of different types of data (e.g. textual, numeric or typed in a database fashion). Being integrated in a comprehensive approach that aims at the eventual presentation of all data in relational form, IE methods will provide the important link from natural language texts to typed data items that can be integrated in the database as corresponding attribute values.

Starting from a heterogeneous collection of data, different data sources have to be classified according to their origin, type, content etc. In a following step a content filter based on IR methods can establish a collection of textual sources distinguished by a certain topic that will function as the text corpus for information extraction. After identifying relevant fragments by IE methods in this corpus the extracted content can be prepared for insertion into the database. Instance unification techniques (also known as record linkage) can be used to reveal contradictory data items and identify and merge complementary information. Value normalization can transform the extracted fragments in the appropriate attribute-specific format.

A similar comprehensive approach has been pursued by known frameworks ATLAS [Lap02], GATE [Cun01] and UIMA [Fer04] that embed IE methods as a module in a pipeline of processing steps.

13.5 Final Remarks

IE is a viable, interdisciplinary technology that utilizes machine learning techniques for automatic processing of natural language. The continuing progress in this field has been leading to a significant reduction of human resources, knowledge and effort necessary to perform the extraction task and to the development of trainable approaches that can easily be adapted to different application domains and even languages. However, the step from a research field to an established technology has not yet been made.

The demand for commercial applications that are based on analysis and processing of unstructured data grows constantly. The impact of IE will increase in the next years. At the current state of development IE technology is not mature enough to be employed as a standard module in business software. The application of IE as an autonomous system is also problematic because no absolute correctness can be expected and the results are afflicted with uncertainty. However, as a part of comprehensive knowledge management process IE methods will play an important role. The utilization of information extraction in interactive environments supported by the human quality control is practicable already at the current level of development.

Information Extraction applying machine learning methods to natural language is an important and promising technique that will find a widespread deployment in the future applications. Even though no perfect solution can be expected, creating advanced learning models for natural language and extending IE beyond the basic fragment extraction will be the subject of future research.