Part III

Evaluation

In this part of the dissertation we empirically investigate the potential of GROPUS and Information Extraction based on machine learning. The quantitative evaluation of the performance of GROPUS and other approaches on different text corpora accomplishes several goals stated in the sec. 3.3. It demonstrates the effectiveness of GROPUS in different environments and sets its performance in relation to other state of the art approaches. To estimate the performance range that can be expected from GROPUS and other IE systems human extraction performance is analyzed and opposed to the results of GROPUS. The quantitative evaluation is supplemented by the qualitative analysis of the extractions that allows to interpret the results and explain the errors made by GROPUS.

Beside comparison with other systems we evaluate the usefulness of proposed system features for the IE task. The influence of semantic preprocessing, rule similarity metric, substitution heuristic, different validation strategies on the quality of extraction is investigated in the component relevance study in the twelfth chapter.

Prior to the discussion of results the tenth chapter states the questions investigated in the experiments, describes the test application domains and presents the evaluation methodology used in the experiments. The stated questions are treated in the twelfth chapter where different factors influencing the extraction quality are analyzed. In the eleventh chapter the comparison with other systems and human results is made on single corpora.

The superordinate analysis of practical usefulness of current IE systems and identification of the environments in what the application of IE is expedient take place in the last chapter. Based on the experimental results deficiencies of current and perspectives of future IE systems are outlined.

10 Introduction to the Empirical Investigation

This chapter answers three basic questions in connection with the evaluation:

▷ In what environments does the evaluation take place?

▷ What is investigated?

▷ How is the performance of the systems evaluated?

According to the evaluation goal to scrutiny, what texts can be processed sufficiently well by current IE systems, the evaluation is performed on three text corpora that significantly differ in size, language style and the complexity of the target structure. At the beginning of this chapter the properties of the text corpora relevant for IE are outlined.

One of the central objectives of the evaluation is to identify the factors that have a significant influence on the extraction quality. In this chapter we formulate several hypotheses about the dependency of extraction results on certain characteristics of the text corpus and the target structure establishing the primary focus of evaluation. The experiments described in the twelfth chapter are designed to investigate these hypotheses.

Evaluation of IE systems is not straightforward because standard correctness criteria cannot be applied in a natural language context. Since the result of an application of an IE system is many extractions of different attribute values from different texts, one can differentiate between the correctness of single extractions, correctness of extractions from a text, correctness of attribute extractions etc. To consolidate these different levels of correctness into a quantitative measure IE borrows *precision* and *recall* metrics from Information Retrieval. This chapter defines both metrics in the context of IE and describes the experimental setup.

10.1 Test Corpora

GROPUS has been conceived to extract information from any natural language text imposing no requirements on its kind or structure. As we argued in the sec. 1.4, we expect GROPUS to perform best on homogeneous, stylized texts written in a technical language. An interesting evaluation aspect is, how well can GROPUS handle other texts that significantly deviate from this optimal text style. We chose three text corpora that represent different distinct areas in the spectrum of natural language texts.

In the analysis of the text corpora we will particularly pay attention to the text properties that significantly affect the extraction quality :

▷ **Number of texts**. One of the most critical parameters for adaptive IE systems is the number of training examples that indirectly depends on the number of training texts. Beside the absolute amount the number of texts must be in a reasonable relation to the size of the target structure.

▷ **Structure of the text**. The range reaches from the form-like texts that feature a very rigid structure to fully unstructured press articles.

▷ **Text style and used language**. It is, for instance, relevant whether a text complies with a grammar of a natural language since the success of linguistic preprocessing depends on grammatical correctness of a text. A technical language with a definite terminology is easier to handle than journalistic text because of restricted diversity of the language.

▷ **Length of texts**. Long texts may contain description of several relevant events or many redundant, marginal details. Both factors increase the possibility of erroneous extractions.

▷ **Information density**. The reliability of extraction rules depends amongst others on the interdependence of attribute values in a sentence. The stronger the connection between different attributes is, the more evidence of relevant information can be assumed when a linguistic pattern matches. The context of isolated attribute values is therefore usually less reliable. The information density can be defined as the average number of extractions per sentence.

Beside the texts the properties of attributes and their relation in the target structure play an important role for the goodness of extraction:

▷ **Complexity of single attributes** is reflected in the homogeneity and structure of attribute values and in their positions in the texts. The identification of extraction boundaries is complicated if an attribute value consists of several tokens.

▷ **Functional dependencies of attributes** (disregarding the attribute values)[1] are connected with the information density of a sentence. Functional dependencies on the attribute level are reflected by the extraction patterns and contribute to the more confident identification of relevant information.

▷ **Semantic similarity of attributes** has a negative impact on extraction goodness making it more difficult to establish a correct assignment of an attribute to an extraction.

[1] Attribute B depends on A if there is an extraction of A in the text whenever there is an extraction of B independent of their values

10.1.1 Seminar Announcement Corpus

The CMU seminar announcement corpus consists of 485 newsgroup messages about forthcoming lectures, talks, presentations etc. The texts contained in the Seminar corpus can be considered as *semi-structured*, since the majority of texts has a loosely structured header where the information is presented in a table-like form. Every text contains a free text part representing more detailed information. The used language is even in the free text part quite informal and therefore not always grammatically correct.

Every text describes exactly one event. The subject of interest in this corpus is the beginning and ending time, location and speaker of the announced seminar. The information usually occurs multiple times as well in the structured header as in the free text parts. The single references to the same entity may vary throughout the document (e.g. speaker can be referred to as *Prof. Irfan Ali* or *Mr. Ali*). Attribute values comprised by the free text part are usually isolated from other extractions, that is, a sentence contains only one attribute value. The information density is therefore very low.

The missing interdependence of attribute values is also reflected by the target structure that features no functional dependencies on the attribute level. The values of two time attributes are short and feature a quite regular structure while the values of speaker and especially location significantly vary in their structure and number of tokens. Values of all attributes can be found in both structured header and the free text part and occur except for beginning time at different positions in text.

10.1.2 Bosnian Corpus

Text Documents

Bosnian corpus comprises 162 fictitious reports of military observers written in German about dislocation and movements of troops and events on the battlefield. Texts vary significantly in their completeness and length: some of them are messages consisting of a few lines, while others contain an extensive (up to several pages) description of the events over a longer period of time. The average length is however approximately the half of the page. Therefore the information is presented in a pretty condensed form. A sentence contain usually several attributes of the target structure resulting in pretty high information density. Used language is also diverse: it reaches from fully grammatical texts to sentences in the telegraphic style. Texts contain many words that are not in the dictionary such as abbreviations and named entities designating military objects. Most texts have a similar structure beginning with the date and time of the report and reporting institution, proceeding with the actual report. The final parts often contains descriptions of further actions of the observer.

The point of interest in this domain are all kinds of movements of troops, weapons and military equipment. 83 texts from 162 contain information about such movements and are regarded as relevant texts. Only factual movements are considered relevant. Intentions such as *Die ArtBrig in SRBRENICA wird mobil gemacht und wird in Kürze nach Westen aufbrechen* and assumptions of a movement are not extracted. Most relevant texts describe only one relevant event, some contain however several facts of movement, that share details about the observer.

Target Structure

The chosen target structure reflects the common attributes of the military movement (starting point, destination, the actual people or military objects which move or are moved, direction of the movement and the currently passed location). Furthermore attributes characterizing the observation of the movement (observer, date, time of observation and location where the observer was situated) are included. The complexity of attributes can be estimated by the location of the attribute values in the text and the regularity of the structure of extracted text fragments. The attribute BEOBACHTUNGSDATUM has a low complexity because its values have a similar structure (number and the name of the month separated by a point) and it almost always occurs in the same context at the beginning of the text. Although the attribute BEOBACHTUNGSZEIT has also a rather fixed structure its position in the text and context vary widely. Therefore this attribute is more complex than BEOBACHTUNGSDATUM, but less complex than OBJEKT that occurs in different contexts and has very irregular structure.

Not only the complexity of an attribute but also the number of training samples has a serious impact on the extraction quality. 740 attribute values were totally extracted. While the attribute OBJEKT (118 instances), the observation attributes (> 100) and the attribute BEWEGUNGSRICHTUNG (70) are adequately represented in the corpus, the frequency of attributes URSPRUNGSORT(45), DURCHFAHRTSORT(42) and ZIELORT(31) is rather low.

Different numbers of training instances of single attributes are not random and partially caused by the functional dependencies on the attribute level underlying the target structure. Two main dependencies can be recognized: all attributes functionally depend on the attribute OBJEKT that plays the role of the virtual primary key and constitutes the fact of movement. Therefore information about observation that occurs in almost every document is considered only if the text contains values of OBJEKT. Observation attributes in turn depend on BEOBACHTUNGSDATUM.

10.1.3 MUC corpus

The MUC corpus has been used as the test corpus for comparison of system performance in MUC3 and MUC4. The texts are about confrontations of different military and terrorist groups and the government in several Latin-American countries. 650 documents were analyzed, 388 contained relevant information. As opposed to the Bosnian corpus the texts feature the style of press messages written by the news agencies. The sentences are comparatively long and describe either already occurred events or contain indirect speech (mostly citations of officials). Texts comply fully with the English grammar and use a relatively broad variety of expressions. The average text size is about 2 pages and is considerably larger than that in the Bosnian corpus. In spite of a relatively high information density in single sentences the information is sometimes quite scattered in a document because of a big number of extracted attributes. The texts are characterized by a big number of coreferences to extracted entities. Only one of coreferences to a certain object is annotated as the actual attribute value. The identification of relevant information is complicated by a lot of linguistic fuzziness since many facts are reported in interviews, threats or assumptions of officials. Many redundant details that do not contribute to the description of the events the texts is dedicated to (e.g. mentioning details of older terrorist acts) may also confuse an IE system.

Natural language processing of the texts is complicated by the capitalization of all words. Determination of the sentence borders proceeds with a considerably high error rate. This has a negative influence on the extraction results, since important context may remain outside of the sentence borders.

Target Structure

The target structure captures the information related to terrorist acts in South America and is partially based on the extraction template provided by the MUC. The big difference is that the extracted fragments are not normalized (e.g. replaced by given values from possible value list) and are transferred without changes into the target structure. Trying to manage the tradeoff between the compactness of the target structure for the evaluation purposes and establishing a precise model for the extracted information some attributes had to be modeled on a fine-grained level to avoid ambiguity and for some the detailed model had to be sacrificed. The attributes VICTIM_TARGET, ANIMATE_VICTIMS and INANIMATE_VICTIMS, for instance, reflect the fine differentiation between the object (person or physical object) that the terrorist assault has been dedicated to and people or buildings that were close to the target and were killed, hurt or damaged. Integrating them in a single attribute (e.g. VICTIM) would lead to the semantic, contextual and syntactic heterogeneity. In contrast, the attribute TIME can include every information about the year, month, date and the daytime of the terrorist act. The relative statements such as *next week* or *yesterday* are also considered. However, values of time are more uniform and easier to capture.

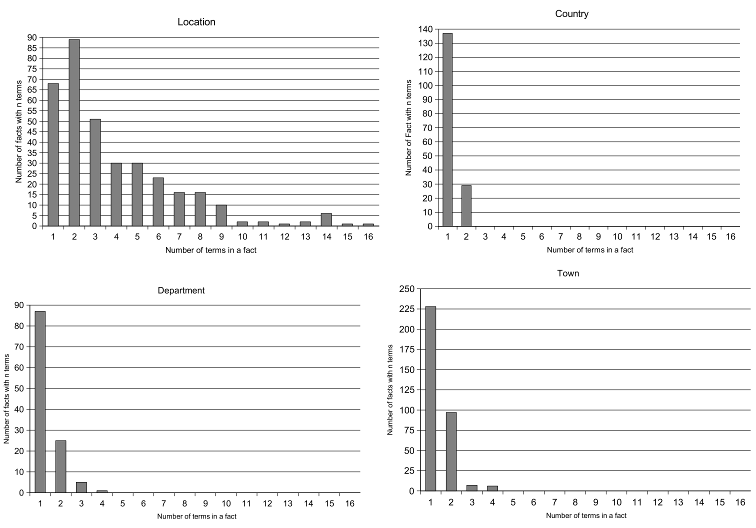

Figure 10.1: Frequency distribution of number of words in extractions of the attributes country, department, town *and* location

To demonstrate the dependency of extraction results on the complexity of attribute structure the attribute LOCATION has been examined. Extractions of these attribute feature a large structural and semantic heterogeneity since they contain all available information concerning the geographical situation of the terrorist act. Thus less complex attributes can be established by splitting the extractions of LOCATION into independent semantic units: TOWN, DEPARTMENT, COUNTRY. Corresponding attributes have been included in the target structure and the extractions of LOCATION have been split into values of the three attributes. Fig. 10.1 visualizes the frequencies of the amount of tokens in the attribute values of LOCATION and the three fine-grained attributes. LOCATION values often consist of several tokens reaching even 16 tokens in the worst case, while the values of TOWN, for instance, comprise at most two tokens. Tests on the corpus can be performed with the complex attribute LOCATION and the

Corpus	Number of texts	Relevant Texts	Attributes	Attribute values	Relations per text
Seminar	485	485	4	2841	1
Bosnian	162	83	9	740	1-5
MUC	650	388	13	3725	1-9

Table 10.1: Quantitative characteristics of investigated corpora

three less complex attributes that contain almost the same extracted fragments as LOCATION. The underlying assumption is that the extraction results can be improved without loss of information by the more proper modeling of the target structure.

Not only the system, but also a human extractor who prepares the training corpus is confronted with the difficult problem of consistent assigning of attributes to the text fragments. One serious problem is connected with the ambiguity of attributes because sometimes despite the thorough modeling it is difficult to decide which attribute is represented by the text fragment (e.g. VICTIM_TARGET or ANIMATE_VICTIMS. Another difficulty is to systematically determine the borders of extractions, i.e. to determine whether the edge words are a part of the attribute value.

10.1.4 Comparison of Corpora

Table 10.1 summarizes the quantitative properties of the texts and attributes of the target structure. As we can see, the corpora differ a lot in the relation of complexity of texts and the target structure and the number of training instances. Seminar announcement corpus contains only four attributes with totally 2841 values. Two of these attributes are quite regular. Bosnian corpus features 740 extractions distributed on 9 partially very complex attributes. Due to not very challenging texts and especially simple target structure combined with an adequate amount of training examples seminar announcement corpus is considered to be the easiest corpus for IE task. MUC corpus, on the other hand, features the most complex texts and target structure and represents the opposite edge in the range of NL texts. While the texts of Bosnian corpus are less variable than those of MUC corpus, it features a multi-faceted target structure and does not provide sufficient number of training instances. While in the seminar announcement corpus every text contains exactly one relation tuple, in two other corpora the IE systems face the difficulty to decide whether a text contains relevant data. The extraction in Bosnian and MUC corpora is additionally complicated by the fact that texts sometimes contain several relation tuples with interleaved attribute values.

Seminar announcement corpus has become the standard corpus for evaluation of IE systems and is therefore very suitable for the comparison of GROPUS with other state of the art systems. One goal of evaluation is to demonstrate that the extraction algorithm underlying GROPUS does not depend on particular language. For this purpose we demonstrate its effectiveness on English and German corpora. Bosnian corpus comprises German texts and has the advantage that another statistical IE system TIE [Sie05] has been evaluated on it. MUC corpus has been used because the performance of GROPUS on challenging texts has to be examined and the results of ELIE [Fin04a] and TIE are also available for this corpus offering the base for comparison.

10.2 Investigated Questions

Size of the training corpus

One of the dominant parameters of supervised learning approaches is the size of the training set. It has a considerable influence on the applicability of an algorithm. The creation of training corpora requires human domain knowledge and manual effort that depends on the amount of the annotated documents. A consistent extraction of facts from training texts is a time consuming and error-prone process.

On the other hand, more training examples result in a bigger diversity of extraction rules, which in turn increases their reliability and coverage. The expectation that a larger training corpus contributes to better extraction results is therefore very plausible. However, we cannot expect a continuous direct proportionality between the size of the training corpus and the extraction goodness. Resolving the tradeoff between the human effort and the quality of extraction results is one important aspect investigated experimentally.

Another related interesting issue is what size of text corpus is necessary to reach the maximum performance. Obviously, GROPUS will not extract all items correctly and achieve perfect results converging to some suboptimal level. At what level the maximum is reached and by what factors this maximum performance is determined is another subject of empirical investigation in connection with the corpus size experiments.

Linguistic complexity of attribute values

Attribute values may be expressed in the texts in a simple, straightforward way like proper names or numbers or feature a quite complex linguistic and syntactic structure. The notion of attribute complexity can be based on the location (context and position in text) and the structure of attribute values. However, since the location of most attributes is very heterogeneous, we consider the structure of extractions in order to assess the complexity of an attribute. We try to quantify the complexity of attribute structure considering the expected average number of words in the extractions we call *expected average length*:
Let N be the total number of attribute values, N_i – number of values with i words and M – maximum number of words in a value. The expected average length can be defined as $\sum_{i=1}^{M} \frac{N_i}{N} * i$. The bigger its average length is, the more complex an attribute is. It is plausible that extraction of an attribute value that consists of several words is more difficult than of those with a single word. The hypothesis in this investigation is that a attribute complexity measured by the expected average length of an attribute value is inversely proportional to the extraction goodness of an attribute.

Attributes may subsume more specific types, as the example of LOCATION and TOWN, DEPARTMENT and COUNTRY in the MUC corpus illustrates. Thus the complexity of an attribute is not necessarily its inherent property, but often the matter of the proper definition of the target structure. Thus the hypothesis mentioned above can be extended saying that the overall extraction quality may be improved by a proper modeling of the target structure. This hypothesis is going to be verified considering results for LOCATION and three derived attributes on the MUC corpus.

Homogeneity of document set

The notion of text homogeneity involves several nuances. On the one hand, it concerns formal characteristics of the documents such as their length. Other significant criteria might be the average length of sentences containing relevant information or their syntactic complexity expressed, for example, as the number of syntactic constituents in a sentence. In our analysis we focus on the *external homogeneity* of the document set, which is defined as the ratio $\frac{df}{d}$ where df is the number of documents containing extracted facts and d is the total amount of documents.

As we have seen in the previous section, only some texts in Bosnian and MUC corpora comprise relevant information, while the external homogeneity of seminar announcement corpus is 1. Regarding this problem from the perspective of machine learning the texts that do not contain relevant information can be viewed as noise in the data. One goal of the experiment will be to examine how well GROPUS can handle such noisy data. Since we know what texts in the training corpus contain relevant data, we can filter texts without extracted facts and compare the performance of GROPUS on the original "noisy" corpus and the filtered, preclassified corpus with the external homogeneity 1.

Besides, this experiment can serve to evaluate the usefulness of text classification as a preparatory step before the information extraction. If the influence of external homogeneity is significant and the text classification achieves sufficient accuracy. The system architecture may be extended by a text classification or an IR engine in the preprocessing.

Interdependence of different factors

Unfortunately, factors discussed above are far from being orthogonal. In order to evaluate the influence of a certain factor all other factors must be fixed. When, for instance, investigating the effect of attribute complexity by considering different attributes we have to assure that all attributes have approximately equal number of training instances. Since we use the same training corpus for the experiments and values of some attribute occur more often than others, it is difficult or sometimes impossible to find a shuffle of texts that satisfies the requirement for balanced number of training instances. Hence because of interdependence of single factors and missing atomicity of many parameters (since the elementary unit we can vary in the experiments is a text which consolidates many attribute values) the set of factors can often not be fixed. When analyzing the influence of a concrete factor on the extraction quality one has always to take into account possible influence of other factors.

10.3 Evaluation Methodology

10.3.1 Quantitative Metrics

Evaluating the extraction quality central questions are, how many of the expected extractions have been correctly identified and how many of made extractions are correct. These questions are reflected by the *recall* and *precision* metrics known from Information Retrieval. We have already referred to these metrics to explain certain features and effects of the learning algorithm. Below we define them formally considering first the evaluation of single extractions.

An elementary result of the application of GROPUS is a contiguous text fragment extracted as a certain attribute value. To determine whether an extraction is correct it is compared with the extractions of the human expert. The result of the comparison concerning a single extraction may be

- ▷ *true positive*, if it exactly corresponds with a human extraction, that is, the positions of the extractions conform and the correct attribute is assigned to the extraction made by IE system.
- ▷ *false positive*, if the fragment extracted by the IE system has not been extracted by the human and is therefore not among expected attribute values.
- ▷ *false negative*, if an expected extraction has not been made by the IE system.
- ▷ *partial*, if at least one of the borders of a made extraction lies within the borders of the expected extraction or vice versa. A partial extraction implies always an increment of false positives and false negatives downgrading precision and recall.

Single extractions of every attribute value from the test corpus are evaluated and true positives (*tp*), false positives (*fp*) and false negatives (*fn*) are counted. Based on the counts precision (P) and recall (R) can be determined as follows:

$$R = \frac{tp}{tp + fn}; \qquad P = \frac{tp}{tp + fp}$$

For a better comparability of the approaches it is useful to have a single measure of extraction quality. *F-Measure(F)* incorporates precision and recall being their harmonic mean:

$$F = \frac{2 \times P \times R}{P + R}$$

F-Measure rewards the balance between precision and recall, while deviations of both metrics are punished: the less balanced the values of both metrics are, the more F-Measure decreases.

10.3.2 Experimental Setup and Determination of Total Values

In all experiments except for those where the corpus size is varied the text corpora are split into training and test corpora with the ratio 80/20 in case of Bosnian and 50/50 in case of other corpora. To minimize the influence of randomness every experiment has been conducted on ten random partitions of the text corpus, we also call *shuffles*. Values of precision, recall and F-measure have been measured for every attribute of the target structure and in total.

Metric values of attributes obtained on different shuffles are combined in single experimental results calculating their means over the number of shuffles. To determine the total precision, recall and F-measure different options are available. *Macroaverage* is computed as the mean of all attribute-specific P,R and F values. In this case all attributes are considered equally important independent of how often they occur in the text corpus. For the determination of the *microaverage* all tp, fp and fn of single attributes are summed and P, R and F values are calculated with the aggregate counts. In contrast to macroaverage microaverage takes into account the frequency of occurrence of single attributes. Basically, it even abstracts from the attribute level counting all extractions. Macroaverage assigns every, even underrepresented attribute an equal weight so that the resulting value may mislead the interpretation of results.

Microaverage is therefore generally preferable for computing of total metric values. Since microaverage depends on the original "raw counts" of tp, fp and fn, which are seldom published for many approaches, a related metric is often used to compare the results between the approaches. *Weighted average* circumvents the usage of raw counts weighting the metric values of an attribute by the ratio of its extractions among all extractions in the corpus. Let P_i denote the precision and N_i – the number of values of attribute i and N – the total number of extracted fragments. Weighted average precision is $\sum_{i=1}^{n} P_1 \times \frac{N_i}{N}$, where n is the amount of attributes in the target structure.

The disadvantage of weighted average is that it does not consider the real distribution of attribute values in the training and test corpus in a concrete shuffle assuming the same ratio as in the whole text corpus. Microaverage provides, by contrast, the most precise measure reflecting the real distribution of attribute values in the test corpus. Whenever the raw counts of other approaches are available, we will therefore use microaverage for their comparison. Weighted average will otherwise be applied. After the total metric values for each shuffle are calculated as microaverage or weighted average, the total values over ten shuffles is determined as their mean.

10.3.3 Evaluation Modes

Beside structure and language style natural language texts significantly vary in the amount and diversity of comprised information. With regard to the target structure a text can contain one or more relevant events corresponding to relation tuples. Moreover, the text may contain multiple mentions of the same information, i.e. of the same attribute value we call *coreferences* to this attribute value. The question how to compare the extractions of an IE system with the expected results and decide about their correctness is therefore not trivial.

If a text contains only one relation tuple, all coreferences of attribute values are annotated by the human expert. An IE system is supposed to extract only one of the coreferences of each attribute. Therefore the extraction result for a text will comprise at most one extraction of each attribute of the target structure. This evaluation mode is referred to as *one answer per document* and applies when the target structure comprises one relation (s. sec. 4.2.4). Seminar announcement corpus is annotated and the performance of IE systems is evaluated in the *one answer per document* mode.

In case that several relation tuples occur in the text, an IE system is expected to find all occurrences of attribute values, which corresponds to *one answer per occurrence*. The major difference to *one answer per document* is also that the coreferences of attribute values are not annotated so that an IE system has to identify exactly the same occurrence of the attribute which has been annotated by the human expert. If it extracts a non-annotated occurrence, the extraction will be evaluated as wrong even though it is semantically correct. One answer per occurrence is used for Bosnian and MUC corpora.