11 Extraction Results and Comparison with Other IE Approaches and Human Performance

The "global" part of the evaluation covered by this chapter serves to assess the potential of GROPUS and to investigate its adequateness for IE task in general. For this purpose we first present and analyze the results achieved by GROPUS on the three text corpora. We compare the results with those of other approaches obtained at the equivalent conditions and finally confront them with the human performance.

Since seminar announcement corpus is de facto a standard text corpus for evaluation of IE systems, numerous results are available allowing an extensive comparison with both rule-based and statistical approaches. On the other corpora results of *TIE* and both versions of *ELIE* serve as the reference values restricting the comparison to statistical systems. Experimental results and comparison are presented separately for each corpus. The comparison based on the quantitative metrics introduced in the previous chapter is complemented by the qualitative analysis explaining different performance of GROPUS and its typical errors.

The evaluation based on comparison of precision and recall values insufficiently answers the question, how well a system accomplishes the IE task, allowing only conclusions relative to other systems. The best way to comprehend the real value of results achieved by an IE system from the human perspective is to compare them with their own performance. Human performance can be regarded as the upper limit for the results of current IE systems. We present the results of the corresponding experiment conducted on the MUC corpus at the end of this chapter.

11.1 Experiments with Bosnian Corpus

Bosnian corpus contains by far the smallest number of training examples and features a quite challenging target structure. Because of its small size the training/test split has been 80/20 to provide more training examples for the systems. Since almost the half of the texts do not contain extracted information, experiments have been conducted on both preclassified corpus from that the irrelevant texts have been removed and the original complete corpus.

11.1.1 Behavior of GROPUS for single attributes

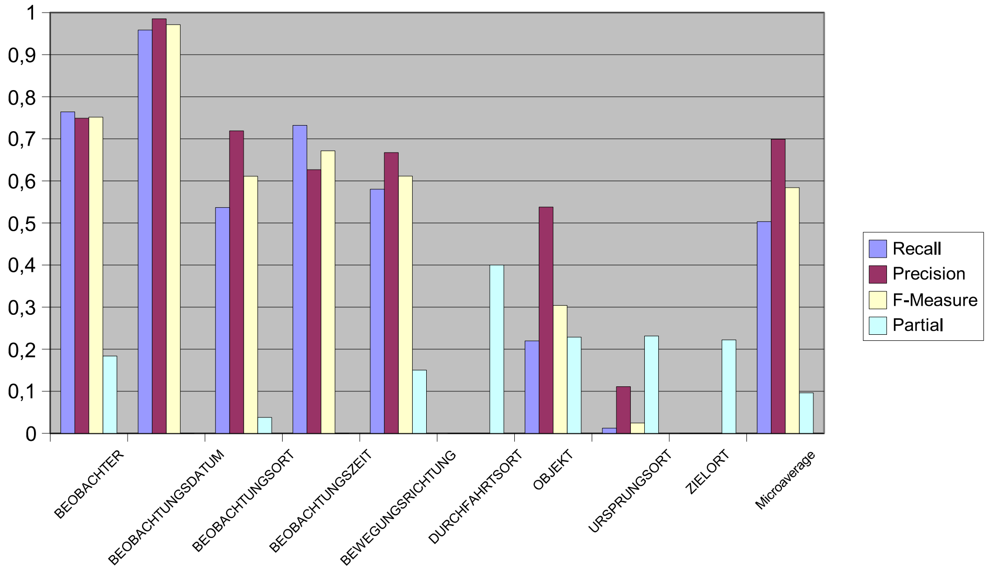

Figure 11.1: Extraction quality of single attributes on preclassified Bosnian corpus

Figure 11.1 depicts precision, recall and F-measure values on the preclassified corpus grouped by attributes. GROPUS achieved a total F-measure of 0.5839 with a precision 0.6991 significantly exceeding recall of 0.5033. The reason for this difference is that the precision of the most frequently occurring attribute OBJEKT is more than twice as high as its recall. The values of this attribute are extremely heterogeneous as demonstrated by the examples: *5 RakW APR - 40* and
1 ArtBtl mit 6 Führungsfahrzeugen, 4 Funk Kfz, 1 ZIL - Kofferfahrzeug, 2 weitere geschlossene Lkw (möglicherweise GefStd Kfz), 8 Lkw, davon 1 mit angehängtem Artillerieradargerät, 18 FAP 2026 mit gezogenen Artilleriegeschützen, davon 15 H 122 mm D - 30. Since they often contain enumerations of single military objects, the number of tokens in their values varies widely. The extraction rules that extracted shorter values proved much more reliable than those that tried to encode long values. Even though OBJEKT has the most extractions in the corpus, there were not enough training examples to establish generalized patterns for long extractions. Therefore mostly rules extracting short values could pass the validation step. They extracted OBJEKT value in the test corpus quite reliably (comparatively high precision), but since the majority of long fragments have not been extracted, the recall is quite low.

The best results have been reached by the observation attributes. Attribute BEOBACHTUNGSDATUM that is distinguished by simple structure and context achieves the best precision and recall values. The F-measure of BEOBACHTUNGSZEIT and BEOBACHTUNGSORT is lower, because in spite of simple structure they have a more variable context.

The results of BEOBACHTER and BEWEGUNGSRICHTUNG could range 20% higher by eliminating partial extractions. The extractions of BEWEGUNGSRICHTUNG beside actual direction sometimes contain geographic locations, e.g. *in den Süden* *nach Teplice*. Many extraction rules have been confused including or missing wrong or expected geographic entities in their extractions.

The extraction of attributes representing locations failed almost completely. Attributes DURCHFAHRTSORT and ZIELORT have an evidently insufficient number of training examples (46 an 31 respectively, whereas there have been even less training instances because of splitting in test and training corpus). Consequently, only few initial rules could be generated so that just a couple of similar rule pairs could be merged. It is therefore not surprising that ZIELORT and DURCHFAHRTSORT feature only partial extractions.

11.1.2 Comparison with TIE on Original and Preclassified Corpus

Unfortunately, there are only few systems that can process German texts. The results of GROPUS are compared with the performance of the statistical system TIE, which has been evaluated in the same fashion (10 random 80/20 splits). Figure 11.2 opposes the F-measure values of GROPUS and TIE obtained on the preclassified corpus. A quite remarkable result is the correspondence between the performance of both approaches. GROPUS and TIE achieved almost identical total F-measure value and the behavior for all attributes is also very similar. Both systems achieved the best results for observation attributes with BEOBACHTUNGSDATUM being the best extracted attribute and both had the most difficulties with the location attributes like ZIELORT and DURCHFAHRTSORT. For many attributes differences in the F-measure are marginal. These facts indicate that the complexity of single attributes has a bigger influence on the success of extraction from Bosnian corpus than the pursued approach.

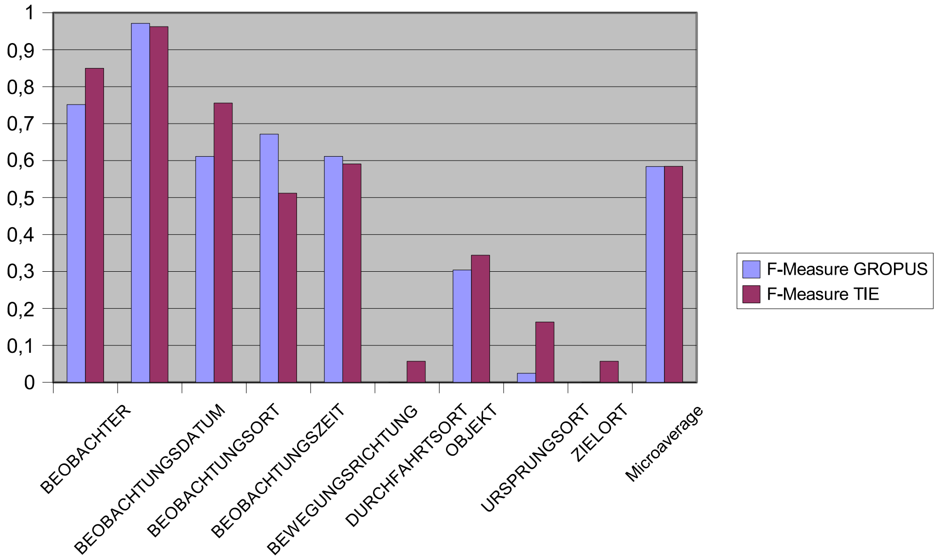

Figure 11.2: Comparison of F-measure values of GROPUS and TIE on preclassified Bosnian corpus

For some attributes the performance of both systems varies. TIE succeeded better in extraction of underrepresented geographic attributes URSPRUNGSORT, DURCHFAHRTSORT und ZIELORT and achieved also a notably higher F-measure for BEOBACHTUNGSORT. Its better performance can be explained by the additional semantic information TIE uses in contrast to GROPUS for identification of locations (e.g. address suffix identifiers). This allows a more reliable prediction of geographic entities even if not sufficient training instances are provided.

While GROPUS significantly surpassed TIE in extraction of BEOBACHTUNGSZEIT, TIE could better handle the extraction of BEOBACHTER values. The position of values of BEOBACHTUNGSZEIT in the text is quite variable. Often it is mentioned in one of the first rather rigidly formatted sentences together with other observation values. Sometimes it is mentioned in the connection with a concrete movement of military objects in the free text. In contrast to GROPUS that could cope well with both types of occurrences TIE had difficulties extracting the time values from the free text. The performance of GROPUS for BEOBACHTER seriously suffered from the very big rate of partial extractions. Here classification of single tokens combined with a resembling strategy of TIE could utilize the small number of training samples more effectively than not sufficiently specialized extraction patterns of GROPUS.

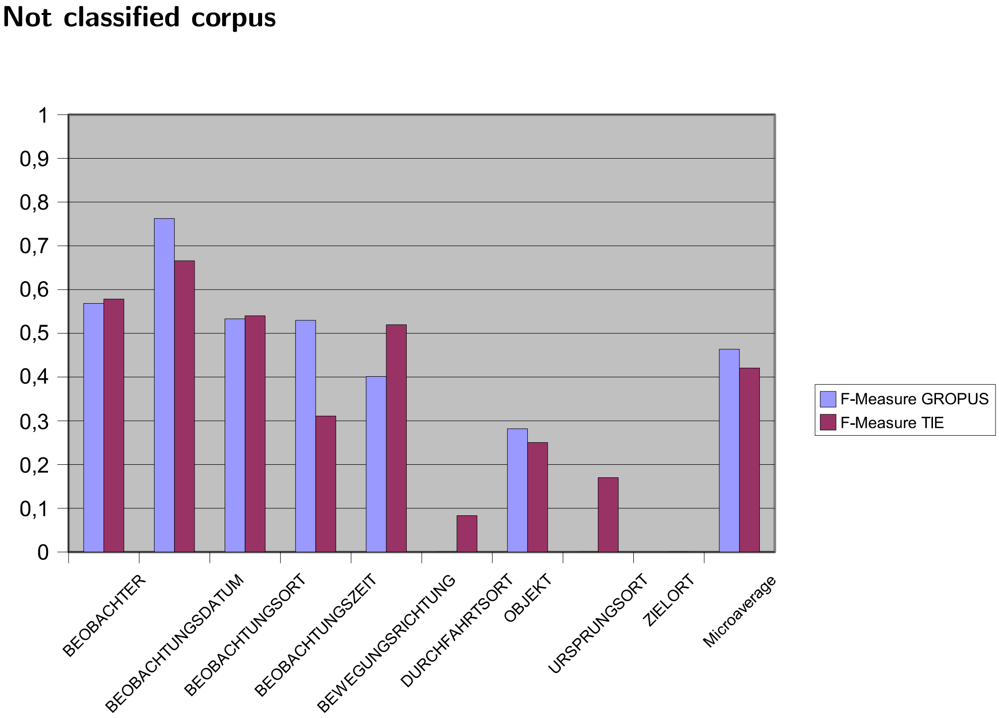

Figure 11.3: Comparison of F-measure values of GROPUS and TIE on non-classified Bosnian corpus

The conclusions derived from the figure 11.2 can only partially be transferred to the results on the non-classified Bosnian corpus displayed in figure 11.3. While there is still an extensive conformance in the behavior of the systems for most attributes, the experiment indicates that GROPUS is more robust against irrelevant texts. In addition to 4% bigger microaverage the F-measure values of BEOBACHTUNGSDATUM, -ORT and -ZEIT do not show such a considerable drop as in case of TIE. The single exception is the attribute BEWEGUNGSRICHTUNG, many values of which have been extracted from irrelevant texts (where directions also are often mentioned) that severely affected precision and involved a nearly 20% decrease of the F-measure value.

Figure 11.4 resumes the total values of recall, precision and F-measure for both variants of the Bosnian corpus. The corresponding behavior of GROPUS and TIE especially on the preclassified corpus is illustrated by nearly equal precision and recall values. On the non-classified corpus GROPUS is more confident in extraction loosing less precision and recall than TIE. Since GROPUS induces extraction rules from the training samples, their quality on nonclassified corpus is comparable with that on the preclassified corpus. The difference is that the confidence of rules is compromised by the incorrect extractions that are made from the irrelevant texts. To keep the precision at an acceptable level, some of

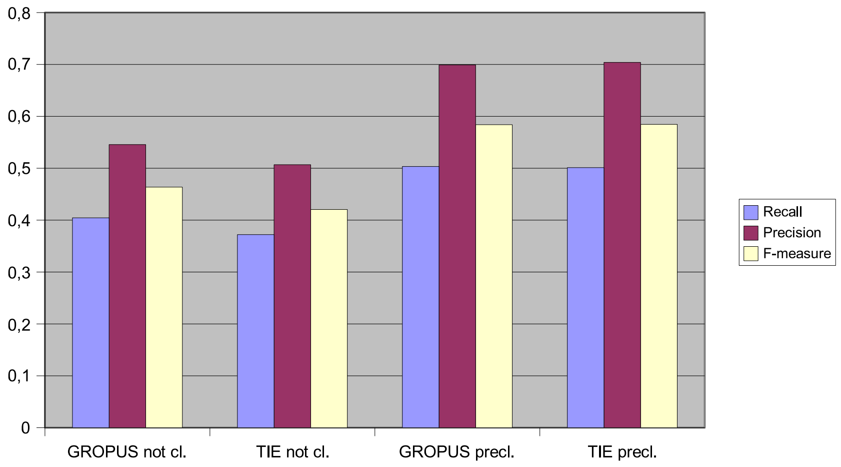

Figure 11.4: Microaverage of R, P and F on original and preclassified Bosnian corpus

them are not validated negatively influencing the recall value. However, the quality of most rules allows them to pass the validation, which results in a relatively moderate decrease of precision and recall.

The rather significant difference between precision and recall in all four experiments is an evidence that the corpus does not provide enough training examples. In case of GROPUS only limited set of initial rules is available for generalization, which results in a small number of reliable extraction rules that cannot adequately cover the large spectrum of values of the complex attributes. Because of insufficiency of similar rules at the early stages of the induction algorithm, many induced rules are rejected during the validation. The remaining validated rules extract providing an acceptable precision, are though not able to establish a satisfactory recall level. We are going to verify this explanation in the experiments with the corpus size in the next chapter.

11.2 Experiments with Seminar Announcement Corpus

In comparison to Bosnian corpus seminar announcement corpus offers significantly more training data allowing to use half of the texts for training and testing respectively. This split ratio has also been used by all other considered systems. Since every text contains extracted information there is no need for preliminary classification so that all experiments are conducted on the complete original corpus.

11.2.1 Analysis of the Quality of Extractions from Semistructured Texts

Establishing a sentence-based context model GROPUS is primarily targeted at the extraction of information from the free texts. The versatile pattern language allows however to learn extraction rules for basically any kind of texts including semistructured documents. Even though the ability to capture semantic dependencies between the attribute values cannot be fully leveraged on such texts, the structured text parts can still contribute to the successful learning providing steady, significant features of extracted information that can be incorporated by linguistic patterns.

As the figure 11.5 demonstrates, GROPUS achieved satisfactory extraction quality on the seminar announcement corpus. Comparing the results to those obtained on Bosnian corpus (cf. fig. 11.1) the performance is significantly improved.

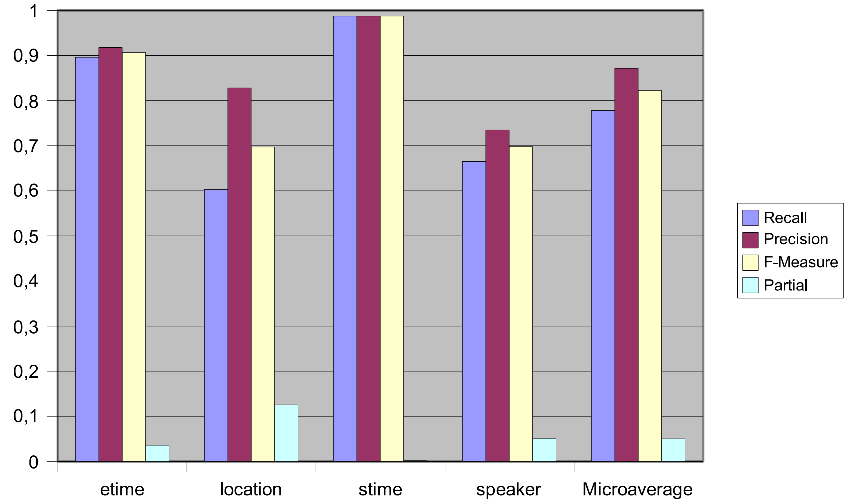

Figure 11.5: Performance of GROPUS on the seminar announcement corpus

Microaverage F-Measure increased by 24% and the extraction goodness of single attributes does not fall below the critical level.

GROPUS had no difficulties extracting regular attributes STIME and ETIME for that very good results have been achieved. Because of the relatively simple structure of their values GROPUS was able to build qualitative patterns mainly by correctly encoding their inner structure, while the context played a subordinate role in identification of their values. A big number of training examples of STIME allowed to cover almost all occurring pattern types (recall 99%), whereas significantly lower amount of training examples for ETIME prevented GROPUS from finding all types of occurrences so that the recall value is somewhat lower (circa 90%). Some partial extractions of ETIME also diminished its recall value. They have been caused by inconsistent extractions of time values that sometimes included an sometimes omitted the extensions *p.m.* and *a.m.*

As opposed to the time attributes it is much harder to infer the borders of instances of two other more complex attributes in the text only by looking on the structure of the extractions itself; their context plays a much more important role than for the time attributes. It is clearly reflected by the location in text where the extractions have been made by GROPUS: a very large portion of ETIME and STIME extractions has been made from the form-like part of the text, while LOCATION and SPEAKER have been predominantly extracted from the free text. The fact, that most extractions of SPEAKER and LOCATION occur isolated in a sentence and information density is very low, negatively influenced the extraction rules for these two attributes. It is more difficult to build reliable context models without additional evidence that relevant information is comprised by a sentence. Precision and recall values of SPEAKER and especially LOCATION also suffer from partial extractions, tagging errors etc. We provide a detailed analysis of incorrect extractions and their reasons below in sec. 11.2.2.

The overall precision is notably higher than the total recall value, which in contrast to Bosnian corpus has its reason in the validation algorithm of extraction rules (cf. 9.2.2). Recall that the algorithm tries to optimize the rule and attribute precision thresholds to select the set of validated rules that achieves the maximum F-measure. The increase of recall can be reached only at the expense of accepting overgeneralized rules that make many incorrect extractions. Since validating such rules involves a severe loss of precision RPT and APT are chosen so that these rules are excluded from the validated set to keep precision and recall more balanced and achieve a higher F-measure value.

11.2.2 Discussion of Common Errors

The precision and recall values of single attributes can be better interpreted looking closer at the extraction errors committed by GROPUS. We have already considered four basic error types in sec. 8.4.1. The "one answer per document" evaluation mode brings about another essential error type. During the application of learned extraction rules several rules may extract different values of the same attribute. Since at most one value is expected, the most reliable extraction is chosen taking the attribute and rule precisions (AP and RP) of the extracting rules into account. It is possible that a correct extraction will erroneously be omitted in favor of a wrong or partial extraction. We abbreviate this type of errors as *incorrect choice*.

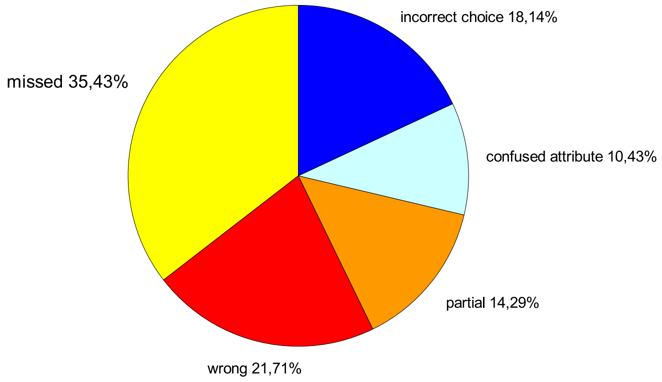

Figure 11.6: Proportions of different error types on the seminar announcement corpus

Different error types are not independent. As we already mentioned, partial extractions imply simultaneously a missed and a wrong extraction. Extraction resulting from the incorrect choice involves a partial or wrong and a missed extraction too. Confused attribute is simultaneously a wrong extraction and has a missed extraction as a consequence in the one answer per document mode. And in this mode any wrong extraction implies a missed one. To avoid double counts we assign only one category to an error according to the following preference: *confused attribute*, *incorrect choice*, *partial*, *wrong* and *missed*. Wrong and missed extractions are counted separately since they both have a negative effect on precision and recall respectively. Figure 11.6 shows the distribution of errors committed by GROPUS in ten shuffles of seminar announcement corpus.

Confused attribute errors constitute 10,43% of all errors and include three types of confusion: LOCATION→SPEAKER, SPEAKER→LOCATION and STIME→ETIME The confusion between LOCATION and SPEAKER is primarily caused by extraction patterns generated from the structured part where the extractions are not embedded in sentences but in some list or form-like structure. Because of scarce context information patterns like ** \dt[(NP|NN)] \dd[[NC: Mr.? NP*]] ** match also many locations. If there are no occurrences of LOCATION values in the free text part where more reliable extraction patterns are generated, GROPUS has to extract the fragment from the structured part where the probability of attribute confusion is higher.

Incorrect choice errors concerned primarily SPEAKER attribute and to a smaller degree LOCATION and ETIMEattributes. Some rules are misled by the fact that seminar announcements contain beside the actual speaker many different people that are mentioned in context of the seminar, and extract erroneously their names. ETIME extraction patterns sometimes match an occurrence of some time value, which does not represent the ending time of a seminar in a document where no ending time occurred. These errors are mainly responsible for the 10% precision loss of ETIME attribute.

The decision, which rule should be used for extraction, is guided amongst others by the respective attribute and rule precisions. Often the difference between the precision values of two or even more rules differ only marginally. The correct extraction of a good, reliable rule is sometimes rejected because its precision is slightly inferior to that of some other rule. Among incorrectly selected attribute values many extractions have been partial, which we are going to analyze next.

Partial errors: Almost all partial extractions are caused by overgeneralized extraction patterns. Overgeneralization is the typical effect of overfitting in the training phase. Overfitting implies that the extraction patterns are optimized for the expressions occurring in the training corpus so that a single rule achieves a very large coverage producing comparatively few wrong extraction on the training corpus. The resulting expression pattern contains though too many general elements, for instance, options and unions so that the matching precision outside of the training corpus significantly suffers. Let us consider the example where GROPUS extracted inter alia the fragment *3313 Doherty* instead of expected *3313 Doherty Hall* as the LOCATION value matching it with the pattern \dl[* \dt[PF: "time"] (\dd[PF: "@card@"] [NC: PF: "PM"])=:stime PF: "-" * ([NC: PF: "@card@" PF: "PM"])=:etime * \dt [PF: "Place"] **(\dd[(POS: "CD")? POS: "NP" ((POS: "NN" POS: "NP")? | (POS: "NP")?)] (\dd [POS: "IN" [NC: POS: "NP" POS: "NP"]])?)=:location** *]

The example is also interesting because the overgeneralized pattern part matching LOCATION value and emphasized in bold has the unexpected effect that the extraction is incomplete rather than extracting too much (*3313 Doherty* matches the red part of the pattern). The reason is that the pattern has not been created for this type of value (since it expects either another proper noun or a sequence of a noun and a proper noun or element *dd* after *Doherty*) and therefore should not be applied at this fragment. Because of optional patterns collectively matching an empty fragment it matches and causes a partial incomplete extraction.

The extraction of attribute values is complicated by the fact that LOCATION consists of many words and has a very variable structure (i.e. room number at the beginning, at the end, in the middle of the extraction, sometimes separated by commas or parentheses). Obviously 50/50 split yields too few good rules that have sufficient coverage and are specific enough to produce exact and not partial extractions. Because of insufficient specific rules GROPUS tries to cover various types of location by additional generalization of patterns (s. the example above) that leads to overgeneralization. More training examples allow GROPUS to build patterns that are responsible for extractions of one certain type of location values so that there is no need for additional generalization and overfitting is avoided. In the experiment with the variable training corpus size (cf. sec. 12.1) we expect therefore a decrease of the proportion of partial errors for larger training corpora.

Table 11.1: F-measure values of different IE systems on the seminar announcement corpus

	Rule-based Systems						Statistical Systems					
Approach	**GROPUS**	**Whisk**	**SRV**	**Rapier**	**BWI**	**(LP)^2**	**ELIE/L2**	**HMM**	**MaxEnt**	**MBL**	**SNoW-IE**	**TIE**
Reference	[Sin06a]	[Sod99]	[Fre98]	[Cal98]	[Fre00a]	[Cir01b]	[Fin04a]	[Fre99]	[Chi02]	[Zav03]	[Rot01]	[Sie05]
etime	90.6	86.0	77.9	96.2	93.9	95.5	96.4	59.5	94.2	96	96.3	**97.5**
location	69.7	66.4	72.3	72.7	76.7	75.0	86.5	83.9	82.6	**87**	75.2	80.6
speaker	69.8	18.3	56.3	53.0	67.7	77.6	**88.5**	71.1	72.6	71	73.8	85.2
stime	98.8	92.6	98.5	93.4	**99.6**	99.0	98.5	99.1	**99.6**	95	**99.6**	99.3
Weighted Avg	81.6	64.9	77.1	77.3	83.9	86.0	**92.1**	81.7	86.9	86.6	85.3	89.9
Macroaverage	82.2	65.8	76.3	78.8	84.5	86.8	**92.5**	78.4	87.3	87.3	86.2	90.7

Wrong extractions have the second largest error ratio among all errors. Many of them are also caused by overgeneralized extraction patterns, especially the underspecified context part. As we already mentioned, we expect the reduction of context overgeneralization increasing the corpus size. A considerable part of wrong extractions happens due to tagging errors. E.g. extracting ETIME the term *PostedBy* is tagged as a proper noun (NP) so that the pattern *CD NP* matches a fragment like
Time: 8:00 - 9:30
PostedBy: Neil Briscombe
producing too long extraction.

Missed extractions are in approximately one third of all cases the source of errors. It is difficult to find one predominant explanation for all missed attribute values. In many cases the encoding of the values itself has been correct while the patterns for context differed slightly with the context in the test corpus. In the opposite case the encoding of context has been absolutely conform, but the patterns for some attribute values have not matched. In very few cases there has been no adequate rule pattern for an attribute value comprised by a test text. The recall value on the seminar announcement corpus depends therefore mainly on the quality of extraction patterns.

11.2.3 Comparison with Other State of the Art IE Systems

Seminar announcement corpus has been used by many approaches as the evaluation basis. Table 11.1 presents the F-measure values of six rule-based and six statistical state of the art approaches. Since the most authors did not publish the raw counts of *tp, fp* and *fn* for single attributes, the weighted average F-measure visualized in fig. 11.7 is considered as the criterion for comparison of the overall performance.

A quite striking observation that can be made comparing the results of different systems is that the statistical systems perform notably better than the rule-based systems. GROPUS is among best rule-based systems but its results are inferior to TIE and most other statistical systems albeit the difference ranges in the 10% interval. Obviously, seminar announcement corpus with its in part form-like texts can be better handled by statistical models. GROPUS is primarily targeted at extraction of information from grammatically correct free texts. Because of its comprehensive pattern model and multi-stage rule induction its strength is the ability to capture the syntactic and semantic interdependencies between extracted fragments within a sentence. Therefore high information density benefits its performance. Many extractions in the seminar announcement corpus are made in the form-like part of the text an therefore not in a sentence. The extractions in the free text part are rather isolated from each other, which results in low information density. Instead features of single tokens like capitalization/case sensitivity, punctuation, character type (e.g. only numeric or mixed characters), certain formatting play an important role in identification of relevant information. These features can be very well captured and learned by statistical systems because their approach is generally token-based.

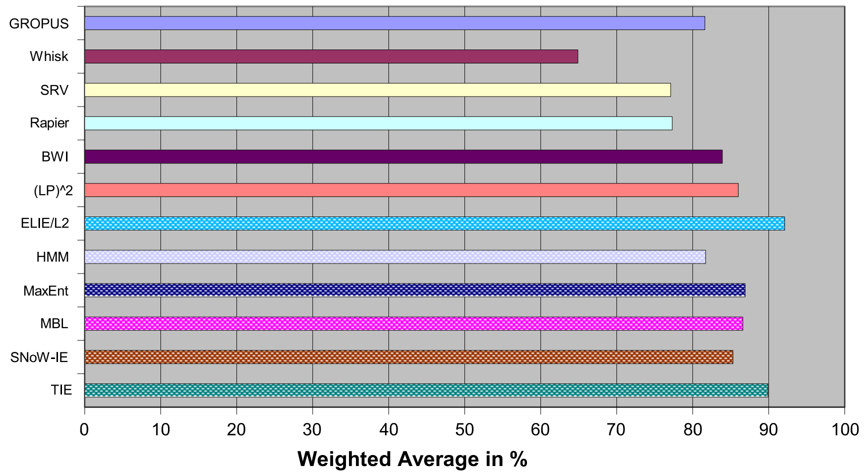

Figure 11.7: Weighted average F-measure of rule-based (plain-colored) and statistical (hatched) systems

Extending the pattern specification language of GROPUS by elements that capture the features of characters would certainly benefit its performance on this corpus. Besides, as the results of other systems demonstrate, utilization of additional semantic information such as gazetteers, ontologies or semantic dictionaries contributes to notably better performance. Without the semantic component (LP)^2 [Cir01b] reports a performance drop by 23% (from 86% to 63.1%), for BIEN [Pes03] it is 11% (from 88.9% to 77.8%) and TIE achieves a 1.3% lower weighted average. Especially the LOCATION and SPEAKER attribute take advantage from the semantic component recognizing, for example, person names, titles etc. As we already mentioned, we intentionally abstained from tuning GROPUS to a specific application domain and did not consider any sources of semantic information that might be helpful in identification of certain attributes.

11.3 Experiments with MUC Corpus

MUC corpus is the most difficult corpus because of the length of documents, high information density in the relevant documents on the one (documents contain up to nine relation tuples) and a very big ratio of irrelevant texts (almost the half of the corpus) on the other hand. Often the same relevant information is addressed by several text fragments. Since one answer per occurrence evaluation strategy is applied, the extraction task is additionally complicated by the choice of the coreference to an attribute value Besides, very often information related to the same event is presented by citing different governmental sources and press, which is not regarded as factual information by the human extractor, but can mislead an IE system to extract it. Often few details of terrorist acts are mentioned that took place in the past and are related to the current terrorist act. As often only e.g. the kind of terrorist act and its date (attributes ATTACK and ATTACKDATE) are mentioned and the text is actually devoted to another terrorist act, the human extractor does not regard this information worth to be extracted, which leads to potential erroneous extraction of an automatic IE system.

11.3.1 Discussion of Extraction Results Achieved by GROPUS

In the face of mentioned complicated factors it is not surprising that the results achieved by GROPUS on the not classified MUC corpus (s. figure 11.8) are worse in comparison with other two corpora. The recall, precision and F-measure values remain below 30% and the extraction quality of many attributes is not satisfactory. The results for single attributes demonstrate that the number of training instances is a crucial factor for the extraction quality, but it cannot be regarded independent of other factors, particularly the attribute complexity. The underrepresented attributes PERPETRATOR_NUMBER (72 instances[1]) ,INANIMATE_VICTIMS (84), COUNTRY (153) achieved low precision and recall values while ACTION(650), TOWN(308) and WEAPON(248) range among the best extracted attributes. On the other hand the attribute TIME reaches in spite of only 72 training instances the best F-measure value whereas VICTIM_TARGET with 649 attribute values is among the worst extracted attributes. The TIME values are very regular, short (EAL 1.35) and easier to identify in contrast to VICTIM_TARGET, which is the most complex attribute in the MUC target structure with EAL equal to 3.88 and highly heterogeneous values occurring in very diverse context. As we expected, GROPUS had big difficulties extracting all three victim attributes because of their complexity (reflected by a very high ratio of partial extractions) and also their semantic closeness. Their values occurred in similar context and many rules extracting the victim attributes were rejected during the validation because of confused attribute extractions, which explains their very low recall and higher precision values.

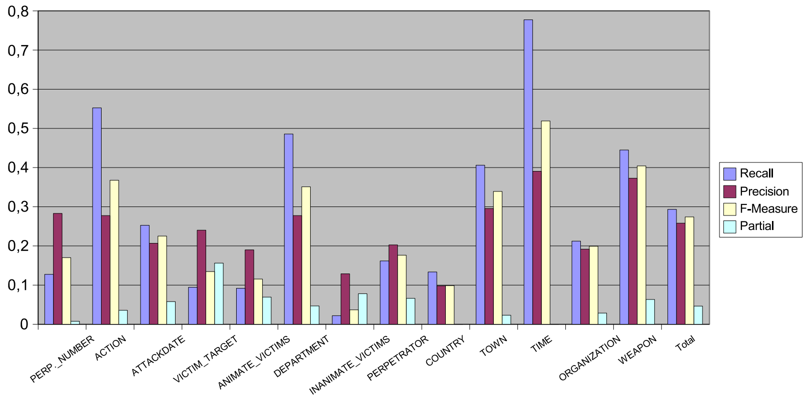

Figure 11.8: Precision, recall and F-measure results on the original MUC corpus (650 documents)

As opposed to victim attributes the recall of almost all other attributes is notably higher than their precision. Obviously, the rule and attribute precision thresholds have been set in favor of recall because it increased faster than the precision decreased (otherwise the more balanced relation between recall and precision, which is stronger rewarded by the F-measure, would be chosen). To examine this phenomenon we fixed the value of attribute precision threshold and varied the RPT in the range [0 ... 1] (s. fig. 11.9). According to the chart the F-measure value achieves its highest level and does not notably alter in the RPT range [0.16 ... 0.4] kept on this level by rapidly growing recall for the decreasing and precision for the increasing RPT. Due to faster growing recall the maximum is reached before the balance point of recall and precision for RPT equal to 0.21.

[1] the numbers in parentheses denote the total number of attribute values in the whole training corpus; a training corpus with 50/50 split contains correspondingly approximately the half of training instances

The chart also allows an estimation of maximum reachable recall. For RPT equal to 0.0, which basically means that all rules are validated, the recall is 63.4%. Thus even omitting the validation step and allowing all induced extraction rules to participate in extraction (taking in account a severe precision drop) the recall does not reach the level of seminar announcement corpus.

Another interesting observation is that the percentage of partial extraction grows almost continuously with rising RPT parallelly to the precision curve. These means that the rules featuring the highest precision on the training corpus – the most reliable rules – are responsible for the most partial extractions. This is connected with the fact that the best rules contains many general elements that sometimes cause too long or too short extractions, as we argued below analyzing the partial errors on the seminar announcement corpus (s. p. 134).

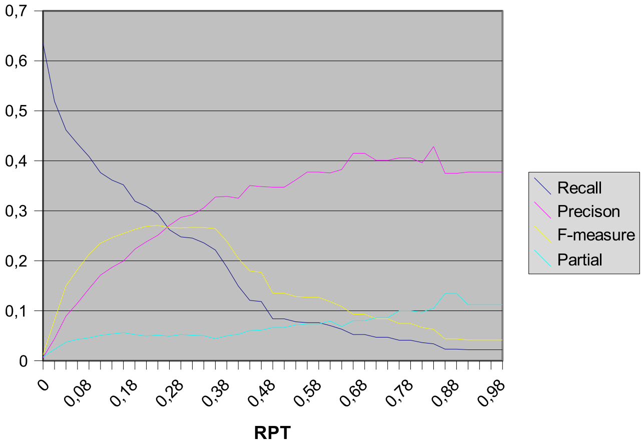

Figure 11.9: Behavior of recall and precision in dependence on RPT

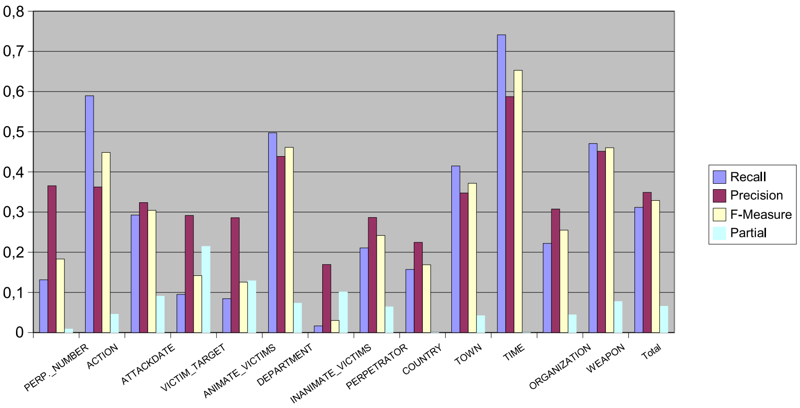

Figure 11.10: Extraction results on the preclassified MUC corpus (388 documents)

Filtering the irrelevant texts significantly changes the relation between the recall and precision as displayed in fig. 11.10. While the recall value grows rather moderately, the precision value of all attributes considerably benefits from the preliminary classification of relevant texts. The overall precision surpasses recall by about 4%. Except for this difference the shapes of figures 11.8 and 11.10 are very similar, i.e. the extraction goodness of single attributes in relation to each other is quite constant. The increase of precision is connected with the fact that irrelevant documents contain information movements of guerilla troops, victims or perpetrators of former terrorist acts without mentioning the terrorist act itself. GROPUS finds and extracts therefore many attribute values from the irrelevant texts, which are semantically correct, cannot, however, be assigned to a particular relation tuple (because of absence of other attributes) and are therefore not extracted by the human annotator. Removing irrelevant texts eliminates the possibility of such incorrect extractions, while the overall goodness of extraction rules does not change considerably in comparison to the not classified corpus.

11.3.2 Comparison with Statistical Systems TIE and ELIE

Besides GROPUS the two best statistical systems on the seminar announcement corpus – TIE and both versions of ELIE [Fin04a] – participated in the evaluation on the MUC corpus. Figure 11.11 displays the detailed comparison of weighted average F-measure for single attributes between GROPUS, TIE and ELIE. As opposed to seminar announcement corpus GROPUS outperforms TIE for every attribute considering the F-measure value. Especially remarkable is the gap in the recall values. GROPUS achieved significantly higher recall keeping the precision values in balanced proportion while TIE has a strong bias towards precision making very few extractions. Even though the absolute values of the systems considerably differ, the relative extraction quality with respect to single attributes is rather similar. So action, time and weapon are the best extracted attributes whereas victim attributes feature lowest F-measure values for both systems. Such a similar behavior of different extraction systems is an evidence that the extraction goodness depends on the complexity of single attributes (cf. sec. 12.3).

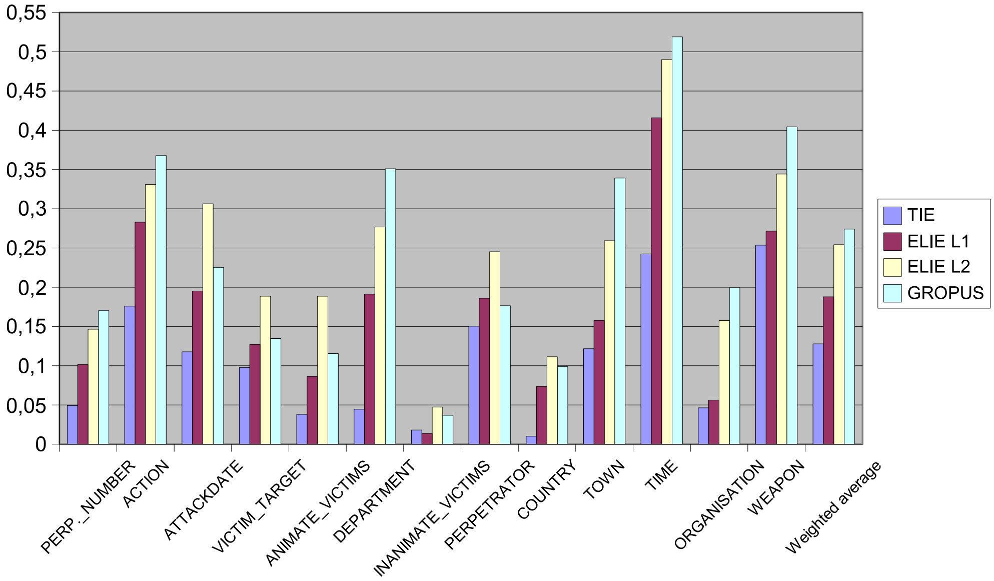

Figure 11.11: Results of GROPUS, TIE and both ELIE versions on the MUC corpus (650 documents)

The best result among statistical systems is achieved by the second level of ELIE that also showed the best performance on the seminar announcement corpus (cf. table 11.1). GROPUS achieved the best F-measure values for the majority and the best recall values for almost all attributes. The second version of ELIE succeeded better in extraction of victim attributes and perpetrator. ELIE L2 is the only statistical system that could achieve results comparable with GROPUS. In summary, featuring highest recall and F-measure results of GROPUS are superior to those of statistical systems.

MUC corpus does not include any structured or semi-structured texts or text parts, but consists completely of grammatically correct free texts. The corresponding target schema comprises 13 attributes so that the texts contain much relevant content and the information density is high. The semantics of big sentence parts or whole sentences plays an important role in the search for desired information. The main criteria for identification of relevant fragments are therefore their semantic and syntactic interaction with the context so that adequate and proper context models are essential for the success of extraction. These factors allow GROPUS to leverage its sentence-based pattern concept for establishing of complex context models and building efficient extraction rules. The token-based view of the text makes it more difficult for the statistical systems to capture the complex context. Besides, relevant text fragments are hardly characterized by simple “statistical” properties, which are numerous in the seminar announcement corpus, so that it is more difficult to find discriminating features for classification of tokens. GROPUS is more robust against the absence of superficial features such as token features (in the MUC corpus all tokens are, for example, capitalized) than statistical systems because it relies more on semantic and syntactic relations within a sentence expressed in its patterns by lexical and positional interdependencies. The better performance of GROPUS suggests that its rule-based approach can better cope with demanding free texts than statistical systems.

With the F-measure values smaller than 30% the extraction quality achieved by all systems for MUC corpus is considerably lower than for other two examined corpora. We already mentioned the characteristics of MUC corpus that complicate the extraction task: irrelevant texts negatively affect precision since the system can make more errors extracting fragments from irrelevant texts. Besides, the confidence of extraction model (e.g. extraction rules in case of GROPUS) is compromised by additional wrong extractions from these texts during training. In case of TIE this leads to very low recall and much higher precision values because the system extracts fragments only if is absolutely confident about their relevance. Big number of coreferences in the texts results in many reasonable extractions that are counted as false positives. A very detailed description of events may mislead a system that would extract some secondary detail omitting the relevant one. Big text length, mentioned linguistic impreciseness, semantically close attributes are additional factors complicating the extraction task.

GROPUS achieves already a big gain because it is in many cases able to capture the complex semantics in a sentence. The identification of relevant content has therefore already reached a significantly higher level than the actual extraction success. Therefore issues primarily related to the selection of identified content as actual extraction (especially identification of coreferences and partial relation tuples) have to be addressed to raise the performance on the MUC corpus.

11.4 Interpretation of Results in the Face of Human Performance

IE is often criticized for lacking objective criteria for evaluation of results. The F-measure of the systems falls below 30% on the MUC corpus, but does it mean that their performance is poor? To set the measured values of evaluation metrics in some reasonable relation we let a human extractor perform the same IE task the IE systems did. Human results serve as a comprehensive measure for the performance of GROPUS and other systems setting an upper limit for the extraction task on a concrete corpus.

Another important question is what prevents GROPUS and other IE systems from performing better. We already discussed many plausible factors that are intrinsic to text corpora and corresponding target structures. In this section we also regard external problems that concern the preparation of training data for supervised learning from annotated natural language texts. While text-immanent factors can be partially neutralized by the human text understanding, a human extractor faces the same external problems as an IE system during the supervised learning. Thus the human results will reveal how much influence these factors have on the extraction task.

11.4.1 Peculiarities of the Training Data for IE Based on Supervised Learning

One of the serious problems for supervised learning from natural language texts is the training examples themselves prepared by human. Beside obvious errors (such as oversights, omissions, incomplete extractions), the human extraction is usually prone to, the quality of human extraction sometimes suffers from inconsistency and bias of the extractor. Even though the obvious errors can be nearly eliminated through a thorough control by other human experts, the problem of bias still has a significant negative influence on the quality of the training corpus and the performance of machine extraction. This problem can be illustrated comparing the extractions of two human extractors on the same sample corpus with a given target structure. Without agreement on certain conventions for the extraction the text fragments extracted by the two experts will differ significantly. One of the reasons is that the experts define the relevant content differently because they differently interpret the target structure or differently evaluate the semantics of certain text passages. Some discrepancies can be caused by ambiguous definition of boundaries of compound attribute values, e.g. extracting the value of OBJEKT in Bosnian corpus one can include more or less details in the extracted fragment: *Kolonne, bestehend aus 3 LKWs, 5 Art.gesch. . . .* or *3 LKWs, 5 Art.gesch.* The experts can disagree whether a certain text fragment can be regarded as a relevant content and should be extracted or what fragment constitutes the attribute value, e.g. extracting the attribute COUNTRY from the sentence: *The bomb detonated on the border between Nicaragua and Honduras.*

The disagreement and the different interpretation have their roots in the ambiguity and complex semantics of the natural language (both in interpretation of the target structure and texts). Therefore even for simple target structures it will be very difficult for human extractors to achieve full compliance, if no conventions for the extraction are presupposed such as metadata for our target structure. We did not experimentally investigate this problem (similar experiments on agreement between domain experts usually yielded an agreement ratio around 66% [Wil93]) because much more realistic scenario is that some “gold standard” for the extraction is presupposed. In our case the gold standard is given by the actual extractions in the training corpus accomplished by a human annotator. That is, these extractions reflect what GROPUS is expected to extract from texts of a particular domain. Instead of assessing the goodness of our training examples by another human expert we focus on analysis of the strength and weakness of pursued supervised approach putting a human extractor in the place of GROPUS and establishing the same initial conditions as for Information Extraction based on supervised learning.

11.4.2 Evaluation of Human Results

Humans have a great advantage that they understand the text. Therefore the identification of desired information is a minor problem while the challenge is to learn the way the creator of the corpus defines relevant content, that is to comply with his interpretation of target structure and text semantics. Human

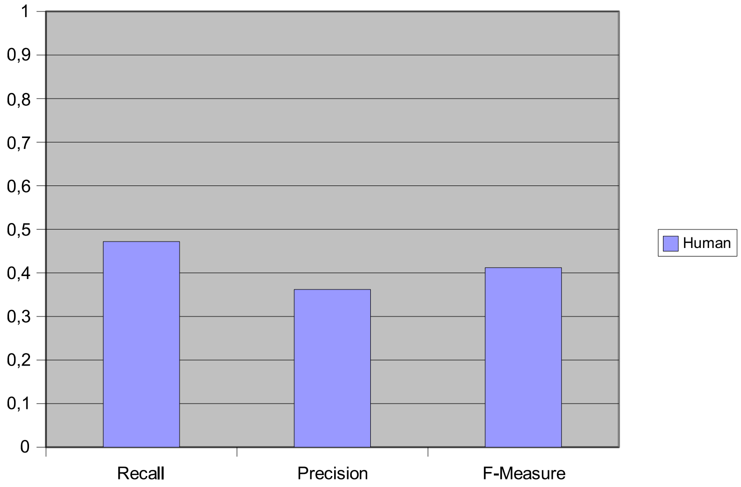

Figure 11.12: Average recall, precision and F-measure on the MUC corpus

results depend therefore to a large degree on the consistency of the training data and the ambiguity of the extraction task.

The human test extractor has been placed in the same conditions as the systems having the half of the training corpus for learning and performing extractions on the other half. No instructions about the properties of extracted attribute values or domain knowledge were provided. Because of the big effort only one random partition of MUC corpus has been processed. The results are displayed in the figure 11.12. Even though human results clearly surpassed the results of GROPUS by 17% in recall and 14% in F-measure, it is surprising that the human, who currently sets the upper limit for machine performance, did not find more than a half of relevant facts and accomplished correctly somewhat more than one third of extractions. Admittedly, this experiment is prone to criticism that it is not representative and the results may depend on personal skills etc. Besides, even without any domain knowledge the human is biased by its background and world knowledge, which makes it difficult to fully comply with the extractions of the creator of the corpus. Therefore the absolute figures are of less scientific interest than the fact that the results of human and machine extractions are comparable lying in the same quite low range and underlining the complexity of IE task in this domain.

The qualitative analysis of human performance reveals that the main sources for errors are analogous to those identified for GROPUS. A big proportion of missed extractions is caused by the extraction of a wrong coreference to an attribute value. 8.6% of all extractions are partial, which indicates that the amount of information that constitutes the attribute value is variable, depending on its interpretation. For example, test extractor identified *social-democratic leader Hector Oqueli Colindres* as VICTIM_TARGET while only the full name *Hector Oqueli Colindres* was expected. Quite surprisingly there have been many confused attribute errors. The mistakes concerned not only the semantically close victim attributes, e.g. in *Assassins paid by the cocaine cartels killed five Colombian personalities, including presidential hopeful Luis Carlos Galan* the test person extracted the red fragment as VICTIM_TARGET while the expected VICTIM_TARGET included only *five Colombian personalities* and *Luis Carlos Galan* is considered a value of ANIMATE_VICTIMS. In the sentence *the ARCE battalions indiscriminately bombed areas near Perkin* the tester put *Perkin* in the TOWN slot, but the gold standard lists *areas near Perkin* as the VICTIM_TARGET.

Even more remarkable is that the tester extracted facts from 58 documents marked as irrelevant by the gold standard and omitted 11 documents that contained relevant facts. Similarly as in the case of GROPUS the precision severely suffered from irrelevant texts. Some false negatives were caused by omitting documents that did not describe a specific incident, but summarized several acts over a certain period of time, e.g. *More than 200 officials have been murdered since the Colombian justice system began its struggle against the drug lords in 1981.*

Human results demonstrate that the extraction task on the MUC corpus is very difficult because of its challenging texts and complex, not unambiguous target structure. Looking at the examples of mistakes one can recognize that many reasonable extractions are counted as incorrect by the performance measures precision and recall. Therefore these metrics do not always fully reflect the quality of extractions, especially their semantic integrity. The reasons for the low performance of adaptive IE approaches and the human extractor are similar so that many conclusions related to the interpretation of the quantitative results can be transferred to investigated IE systems.

11.5 Runtime Comparison

Focusing on the extraction quality runtime aspects are often completely neglected in the context of IE. If an IE algorithm is executed in an offline mode, that is the training phase can be temporally separated from the application phase, only the time for the processing of a single text by already trained extraction model plays an important role. In many real-time applications the systems may be confronted with new training data (generated, for example, by the human revision of the extractions made by the system). Since many IE systems must be completely retrained given new training examples, the total time comprising training and application time is essential.

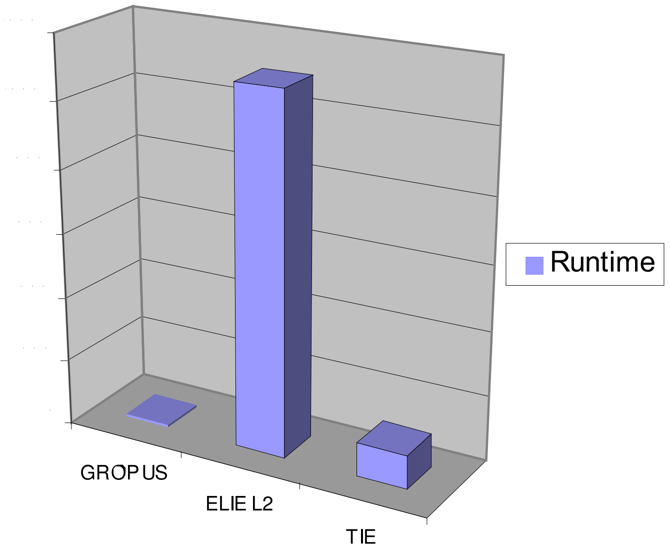

Figure 11.13: Average total runtime per shuffle of MUC corpus (650 documents)

Since absolute runtimes significantly depend on the experimental settings and hardware, we evaluated runtime on the 10 shuffles of MUC corpus measuring the total absolute runtime for training and extraction from the test corpus (as the relation of runtimes on the other corpora has been similar, we conducted the

Table 11.2: Average runtime and effectiveness measure on the MUC corpus

System	Runtime in min	$\frac{F-measure}{Runtime}$
GROPUS	7	$\frac{3.92\%}{min}$
TIE	108	$\frac{0.118\%}{min}$
ELIE	1102	$\frac{0.023\%}{min}$

experiment exemplarily on the MUC corpus).[2] Table 11.2 contains the runtimes of GROPUS, TIE and the second version of ELIE. GROPUS requires only 7 minutes to completely process one shuffle - a time that is by magnitudes shorter than the runtime of statistical systems (the relation of runtimes is visualized in fig. 11.13). TIE needs approximately the 15 times as much and ELIE L2 circa 150 times as much time for accomplishing the same task. The huge difference underlines the efficiency of the learning algorithm with optimized subalgorithms for time-consuming steps such as determination of similar rules, rule merging and validation. Beside the optimized implementation of the algorithm another serious advantage of the rule-based approach utilized in GROPUS is emphasized by the runtime comparison: statistical systems are forced to operate with huge feature vectors including several millions of features, that are indispensable for a good classification of tokens. These feature vectors imply a vast number of floating point calculations in every iteration during the training of the classifier. By contrast, GROPUS gets by with several hundred extraction rules that are derived by the well-defined formalisms based on the pattern specification language and functions operating on its elements.

The unsatisfactory runtime of ELIE is also connected with the fact that the system can only classify one attribute at a time requiring an extra training and extraction run for every attribute of the target structure. However, with more than 18 hours total time the practical applicability of the system for larger application domain is quite restricted. Trying to establish a measure that combines the extraction quality with the required runtime we regarded an *effectiveness measure* $\frac{F-measure}{Runtime}$ (cf. table 11.2). It basically expresses how many minutes a system "needs" to achieve one percent of F-measure. GROPUS is by far the most effective system achieving about $\frac{4\%}{min}$ while TIE clearly surpasses ELIE in spite of worse extraction quality.

[2] All experiments were conducted on computers with 3.0 GHz CPU and 1GB RAM.