12 What Influences Extraction Quality?

In contrast to the previous chapter, that mainly focused on the evaluation of capabilities and efficiency of GROPUS and other state of the art IE systems, the first part of this chapter identifies the characteristics of application domains that the success of information extraction depends on. In the sec. 10.2 we formulated hypotheses and questions about the role of the size of the training corpus, external homogeneity of the training set and attribute complexity for the extraction task. In the previous chapter some qualitative observations concerning their influence have been made. In this chapter we investigate these questions conducting systematic experiments on different corpora.

Beside the external factors that influence the performance of GROPUS we are interested in the evaluation of the efficiency of its internal components. The question how significant the single parts of the learning algorithm such as correction of rules are for the extraction quality is examined measuring the performance penalty when the respective component is omitted. The second part of this chapter is devoted to the study of the relevance of different validation strategies, recognition of synonymy, determination of rule similarity, rule correction and substitution heuristic.

12.1 Dependency of Extraction Goodness on the Size of the Training Corpus

Bosnian Corpus

For the experiment with the corpus size the preclassified Bosnian corpus was chosen. Since almost the half of the complete Bosnian corpus is irrelevant texts, taking a random shuffle of texts does not guarantee that e.g. 60% of the texts in the shuffle in fact contain 60% of relevant texts (since the ratios of relevant and irrelevant texts are distributed randomly as well). On the contrary, taking a preclassified shuffle ensures that a specified ratio of relevant texts is used for training. Therefore we conducted the experiment on the 10 random shuffles, which we used for evaluation of the preclassified Bosnian corpus in the previous chapter. The size of the corpus was systematically increased by 5% beginning by 5% of the total preclassified corpus (83 documents). GROPUS was trained anew

for every training set. For every nth run the first $n*5\%$ files from a shuffle were taken for training so that the training corpora with $n*5\%$ and $(n+1)*5\%$ differ only by the last 5% of documents. The total results of evaluation metrics were obtained calculating the arithmetical mean over 10 shuffles. Figure 12.1 depicts the results for recall, precision, F-measure and partial extractions.

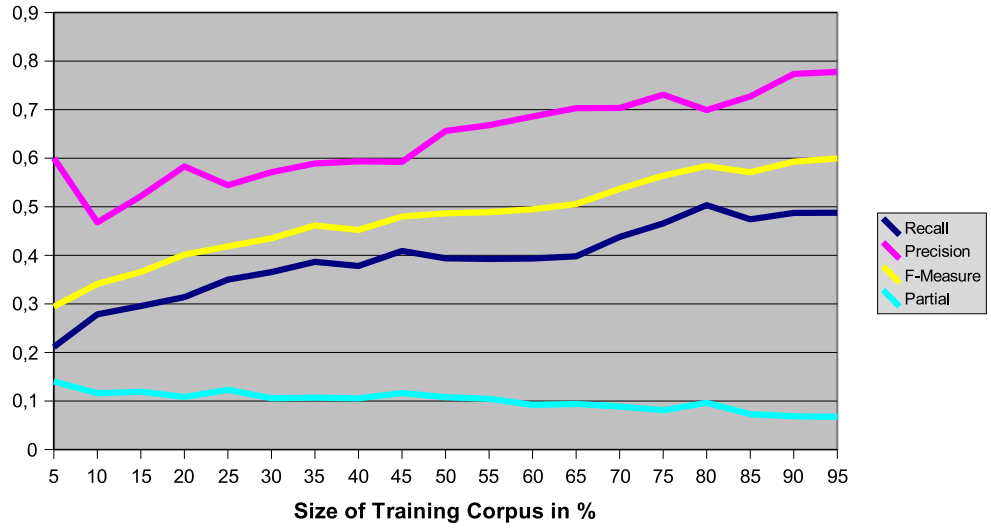

Figure 12.1: Behavior of the evaluation metrics for variable training corpus size

The recall curve features a constant growth until he training corpus reaches 45%. The following phase of till 65% coincides with the sustainable growth of the precision value being subject to fluctuations before. However the growth and stagnation phases are not random, but caused by the rule validation selecting the extraction rules that optimize the overall F-measure. After processing of 45% of training texts many general extraction patterns are already generated so that is not so easy to increase recall, while the precision can be more easily enhanced raising the precision thresholds. The effect is a continuous increase of the F-measure curve that is quite linear in the beginning and middle part. 70% offer enough potential to induce more general reliable extraction patterns so that more rules can pass the raised precision thresholds and establish a new recall boost.

The comparatively high initial recall value and its rapid growth at the beginning can be explained by the fact that reliable rules for simple observer attributes can be derived already from the couple of examples. These rules are the only few rules that are validated and enable therefore a quite high precision and a recall over 20%.

Disregarding some negligible descents of the precision at the beginning both F-measure and precision grow with the bigger size of the training corpus so that no convergent behavior can be recognized. Therefore higher precision and F-measure values can be expected for bigger training corpora. The stagnation of recall around 50% at the last four measurements may be though an evidence of convergence, but it is more likely to assume that it is again the effect of validation (similar as the segment from 45 to 65%), especially in the view of strong precision increase. Partial extractions feature almost linear decrease due to the continuous improvement of extraction patterns for the bigger corpora.

The experiment clearly shows that the size of the training corpus has a significant impact on both precision and recall values. Moreover, the curves of evaluation metrics suggest that the total corpus size is too small in order to achieve optimal

results for this application domain and the capabilities of GROPUS were not exhausted.

Seminar Announcement Corpus

The experiment on the seminar announcement corpus has been conducted analogously to that on Bosnian corpus using the same 10 shuffles . Figure 12.2[1] displays the graphs of functions of evaluation metrics in dependency on the size of the training corpus.

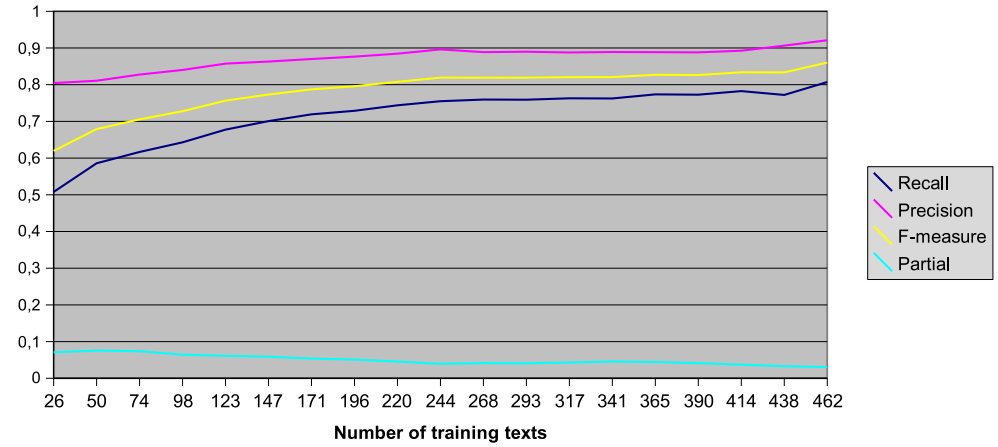

Figure 12.2: Microaverage P, R, F and partial extractions in dependency on the training corpus size of SA corpus

All evaluation metrics start in spite of very small training corpus at a much higher level in comparison to the Bosnian corpus. Since the SA corpus is much bigger, the start value of 26 training texts would correspond to 30% split in the Bosnian corpus, so that the direct comparison is not reasonable. Already a small training corpus is sufficient to learn reliable and exact extraction patterns for the regular time attributes, which explains the very high initial precision value over 80%.

Until the training corpus reaches the half of the complete corpus, the precision and recall continuously increase. While the initially steep gradient of the recall curve scales down with the increasing training/test split, the precision graph grows with almost constant, though not as high ascent. At the beginning the fast growth of recall is caused by a big amount of new rules that are derived using initial rules generated from the new texts in the increasing training corpus. After the corpus has reached its "first critical point" (here at 244 texts), it is more difficult to raise recall, because the coverage of the already generated rule set is quite high, the ratio of the new initial rules in the total amount of rules is rather low and therefore their contribution to the generalization of rules decreases. After this point the system needs much more training examples to generate enough new initial rules that can serve as the source for induction of new reliable extraction rules that are able to extract uncovered facts. To nevertheless establish an increase of F-measure function, validation heuristics enhance the recall adjusting the precision thresholds and taking in account that this may involve an interim stagnation (a phenomenon we also observed on Bosnian corpus above) or even slight decline of the precision value, which we notice in the fig. 12.2. This phase ends at 414 texts involving that the precision, F-measure as well as recall grow in the last segment of the chart.

[1] As opposed to the fig. 12.1 the X axis is labeled by the absolute numbers of the training texts

The level of partial extractions continuously decreases until the half of the total size. A slight rise for the size of 365 texts is connected with the phase of stronger growing recall in that more general rule patterns are validated taking in account more occasional partial extractions. However, from this point the percentage of partial errors sustainably decreases to the absolute minimum of 3.06% for the maximum size of the training corpus.

Even though no rapid growth of the evaluation metrics can be expected, the slow, but sustainably ascending recall and F-measure graphs do not seem to approximate a limit. The quite notable slope of precision at the last four measurements is rather an evidence of a steady increase than a random oscillation on the small test corpora. In spite of the already high level of precision and recall a better performance for bigger training corpora can therefore be expected.

MUC Corpus

Since the MUC corpus also contains a considerable portion of texts that nothing has been extracted from, analogously to the Bosnian corpus we used the preclassified corpus to exclude any randomness in the distribution of texts containing relevant data between the training and test corpora. Microaverage values of R, P and F have been measured and plotted as graphs presented in the fig. 12.3.

Figure 12.3: Influence of the training corpus size of the MUC corpus

Starting from a quite low values compared with the other two corpora GROPUS doubles the recall value and significantly increases F-measure and precision value for the 0.25 split. Between 0.25 an 0.5 splits the growth of all metrics continues with a constant, though not as steep gradient as for smaller training corpora. Until the training corpus has reached the half of the total corpus its increase has a crucial impact on th extraction quality. The numerous and complex attributes require many training examples to capture their diverse context and heterogeneous structure in adequate extraction patterns. More training texts contain more variants of expression especially providing more possibilities for rule merging. The bigger amount of merged rule pairs allows to select from a bigger pool of general rules yielding eventually a set of rules with a better coverage, which contributes to the higher recall. Beside the improved rule quality the precision benefits because the testing of derived extraction rules on the increasing training set becomes more accurate.

As opposed to the other two corpora for the splits greater than 0.5 the evaluation metric behave differently. Even though the recall graph features an ascending

graph, its growth is prone to strong oscillations. After the 0.75 split the oscillations and a very slight increase of the respective maximum values can be viewed as indications for convergence. The precision reaches its maximum at 0.6 split featuring a slight descend for bigger training corpora. The F-measure value remains at almost the same level deviating by less than 1%. On the one hand, this behavior of recall and precision for splits greater than 0.6 can be regarded as the phase in that GROPUS is forced to collect many training examples in order to generate enough new different initial rules and to increase the coverage of the rule set. We observed this behavior on the other two corpora where the system operates with precision threshold to establish a recall growth at the expense of precision. Therefore, analogously to the other corpora, the recall boost may be achieved for the bigger training corpora.

However, the negligible recall increase and the prolonged stagnation of F-Measure value can also be regarded as the strong evidence for convergent behavior. Considering the high complexity of the extraction task on the MUC corpus this implies that the performance cannot be improved for bigger training corpora and the achieved values represent the limits of our approach in this application domain. In summary, the experiment clearly demonstrates the influence of the training corpus on the extraction quality. The size of the training corpus should be chosen so that the phase of rapid growth of the evaluation metrics (circa the half of the text corpus for SA and MUC corpora) can be completed. Since the values of evaluation metrics grow more slowly after a certain point the improvement of extraction quality has to be weighed against manual annotation effort. In contrast to the other two corpora the size of Bosnian corpus does not allow to complete the phase of rapid growth, so that significantly better results can be expected for the increased training corpus. Although the results on the SA corpus are already on a quite high level, a moderate improvement still can be expected for bigger training corpora. On the contrary on the MUC corpus GROPUS seems to approximate its performance limit.

12.2 Text Classification as a Preparatory Step for IE

Texts that belong to a certain domain do not necessary contain the information specified by the target structure and are therefore "irrelevant" in the sense of IE. They can be easily removed from the corpus after it has been annotated by a human excluding all documents without any extractions. Such irrelevant texts in the training corpus certainly diminish the extraction quality negatively influencing the training and test stage. During the training the negative examples extracted from the irrelevant texts compromise the reliability of extraction rules so that rules achieving a good performance on relevant texts are nevertheless discarded because of incorrect extractions from irrelevant texts, which has a negative impact on recall. At the test stage precision suffers from extractions made in the irrelevant documents of the test corpus. However, in a realistic application domain a trained IE system may face texts that do not contain desired information. An evaluation on a not *classified* corpus (i.e. a corpus with irrelevant texts) is therefore interesting from the practical point of view.

In the previous chapter we have already considered the performance of GROPUS on preclassified and not classified corpora. In this section we summarize the results exemplarily on the MUC corpus and investigate the usefulness of the preliminary text classification based on the standard classification algorithms.

12.2.1 Performance on the Non-classified and Manually Classified Corpus

Irrelevant texts can be regarded as "noise" from the point of view of machine learning. A plausible expectation is that removing such documents will benefit the IE process is confirmed by the figure 12.4. It compares the microaverage values of precision, recall and F-measure obtained on the complete and manually classified MUC corpora (for the values of single attributes refer to fig. 11.8, 11.10).

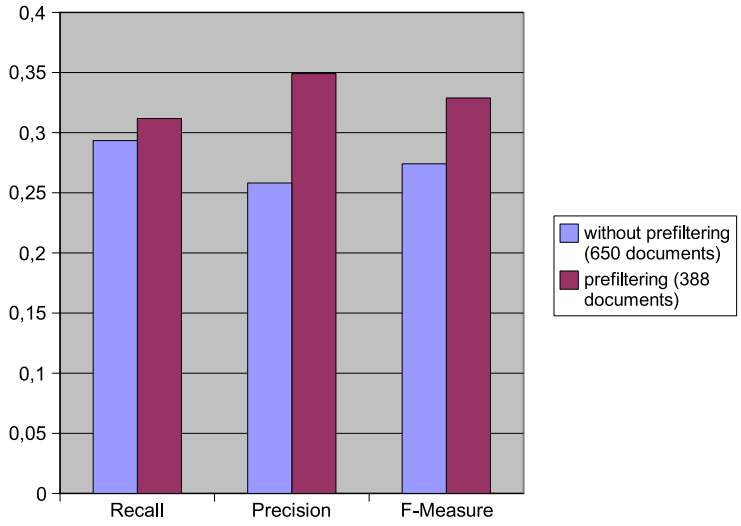

Figure 12.4: Influence of preliminary classification

The comparatively small recall difference of less than 2% is an evidence that extraction rules are only marginally affected by additional incorrect extractions during the training stage on the unfiltered corpus. However, GROPUS makes notably more (9%) incorrect extractions on the unfiltered test corpus, which also explains 5% difference in F-Measure. Even though GROPUS is quite robust against irrelevant texts, the preliminary classification can contribute to the better quality of extractions, especially to the better precision. The achieved experimental results presuppose an optimal text classifier that can correctly decide whether any domain texts contains information that should be extracted.

12.2.2 Impact of Automatic Text Classification

In this connection the utilization of text classification as a preparatory step for IE shall be investigated. Documents can be classified whether they contain relevant information before being processed by IE system. Since the task of text classification is less complex and well studied the classifier may help the IE system achieving a smaller error rate in classification than the IE system due to irrelevant documents. On the other hand there is a tradeoff between elimination of possible errors of IE system by removing irrelevant documents and impeding correct extractions by misclassifying relevant texts.

We evaluated four state of the art classification algorithms: decision rules[Wei05], RIPPER[Coh95], SVM[Joa98] and k nearest neighbors[Wei05] with different options (shared word count (with bonus) (SWC(B)) or cosine similarity) classifying the documents of Bosnian and MUC corpora (s. table 12.1). The best F-measure of 75,5% on the MUC corpus has been achieved by SVM classifier featuring a recall of 94,4% and a precision of 62,9% (incorrectly classifying $0.944 * 388(0.629^{-1} - 1) \approx 216$ from 262 irrelevant texts as relevant). Applying

corpus / classifier	*Bosnian* precision	recall	F-measure	*MUC* precision	recall	F-measure
Decision Rules	77.39%	32.07%	45.35%	62.71%	43.63%	51.46%
***k*NN SWC**	70.61%	82.28%	76.00%	66.96%	85.83%	75.23%
***k*NN SWCB**	71.30%	81.87%	76.22%	67.22%	84.50%	74.87%
***k*NN Cosine**	57.01%	78.55%	66.07%	59.59%	74.03%	66.03%
RIPPER	66.89%	73.17%	69.89%	53.48%	30.17%	38.58%
SVM	71.37%	73.48%	72.41%	62.87%	94.36%	75.46%

Table 12.1: Classification Results on Bosnian and MUC Corpora

this classifier before the actual extraction process GROPUS would in average achieve 5,6% smaller recall than in case of optimal classifier while the precision would in average decrease by $\frac{216}{262} * 9\% \approx 7,42\%$ (taking in account that 262 irrelevant texts were responsible for the 9% precision drop, cf. fig. 12.4). This would approximately correspond to an F-measure of 26,46%, which is smaller than GROPUS achieved on unfiltered corpus still ignoring the fact that recall and precision will be negatively affected by the presence of irrelevant documents in the training set. Other classification algorithms (cf. RIPPER, decision rules) achieved even recall value under 50% (worse than random classification), which indicates that they are not able to capture the subtle features distinguishing texts with relevant information.

Since the current text classification algorithm are not able to reliably identify documents containing relevant information, the utilization of text classification before the actual extraction process does not prove to be useful.

12.3 Influence of the Complexity of Attribute Values on the Extraction Quality

The results of GROPUS and other IE systems on the three text corpora point out that the quality of extraction significantly varies for different attributes. Values of some attributes can be easier identified by an IE system because they occur in a uniform context, have strong syntactic or morphological characteristics and a simpler structure. In the sec. 10.2 we introduced the expected average length as a quantitative measure characterizing the structure of an attribute.

Beside this parameters the complexity of an attribute manifests itself in its semantic properties, for example, how semantically ambiguous and how close to other attributes it is. In this section we want to examine the influence of attribute complexity regarding the semantically complex attribute LOCATION and the subtype attributes TOWN, DEPARTMENT and COUNTRY resulting from more finely granulated specification of the target structure. (cf. p. 121).

The attribute LOCATION and its subattributes are especially appropriate for the demonstration of the impact of attribute complexity because the subattributes are subordinated to LOCATION both semantically because of the subtype relation and syntactically, since LOCATION values comprise the values of the three other attributes. The expected average length and the variance of length distribution are visualized by the histograms in the fig. 10.1 that depict the frequencies of certain word numbers among all extractions of an attribute. The expected average length of LOCATION is 3.73 and the variance of the word numbers is much higher than that of the fine-grained attributes. The difference in the expected average length (TOWN - 1.43, DEPARTMENT - 1.35, COUNTRY - 1.21) emphasizes their lower attribute complexity.

To examine the influence of attribute complexity experimental results obtained under identical conditions for LOCATION and its three subattributes are compared

(fig. 12.5).[2]

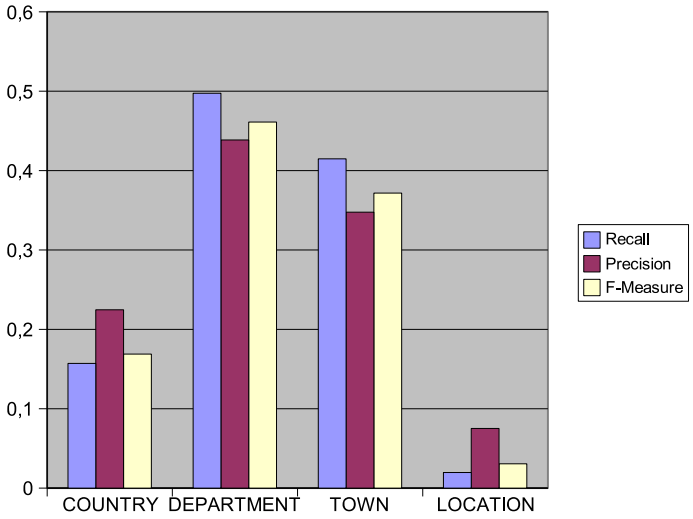

Figure 12.5: Comparison of results for attributes location, town, department *and* country

The high complexity of LOCATION is reflected in the unsatisfactory F-measure value of circa 0.031. This result demonstrates that GROPUS was not able to handle very irregular structure of LOCATION values that makes it very difficult to derive some common features and criteria for the extraction. The results of three less complex attributes are more encouraging. The worst results are achieved for COUNTRY (F-measure 0.17, expected average length - 1.21), while DEPARTMENT (expected average length - 1.35) even reaches F-measure of 0.46. Although the absolute recall and precision values are still far from optimal, the relative improvement compared with the values of LOCATION is considerable. Precision values are at least three times as large and recall values lie in a much higher range. The expected average length proves to be an adequate measure of the complexity of attribute structure.

Besides the attribute complexity the number of training instances is an important factor and has to be considered while interpreting the results. The fine-grained attributes achieve significantly better results in spite of considerably fewer training instances (in case of COUNTRY and DEPARTMENT, cf. fig. 10.1). The experiment allows two conclusions to be drawn:

- ▷ The quality of extractions depends to a high degree on the complexity of an attribute
- ▷ The goodness of extraction can be improved without semantic loss of information by proper modeling the target schema

12.4 Component Relevance Study

Beside the core components such as the pattern specification language, pattern unification algorithm, generation of initial rules, rule merging and validation the learning algorithm for extraction rules includes a number of components that are added to enhance the rule quality, but are not vital for their induction. In

[2] This experiment was conducted on the preclassified corpus to exclude the influence of irrelevant texts

the following study we want to figure out how important these components are and what effect they have on the extraction success. In our study we focus on the following components: semantic preprocessing (synonym recognition), rule similarity metric, substitution heuristic for rule generalization, rule correction and exclusion of irrelevant texts at the training stage. In the second part of the study different validation strategies presented in sec. 9.2 are investigated.

12.4.1 Effect of Single Components on the Extraction Quality

All experiments mentioned in the previous sections have been performed utilizing all components of the learning algorithm. To investigate the contribution of a single component we conduct the experiments omitting this component and compare the results with those achieved by the standard setup. In case of semantic preprocessing we exclude the synonym pattern from the pattern language (*without synonyms* denotes this configuration of GROPUS in the diagrams). When omitting rule similarity metric we use a *random similarity* instead. Substitution heuristic and rule correction are simply not applied so that a rule that has not been validated is immediately discarded. Omitting rule correction is equivalent with the exclusion of negation pattern from the pattern specification language.

Beside these four components we want to evaluate the variation of the learning algorithm ignoring the irrelevant documents at the training stage. When training GROPUS on a corpus that consists of a lot of texts without extracted information we can reduce the negative influence of irrelevant texts (cf. sec. 12.2) excluding them from the training corpus. Many rules that could not pass the validation step because of erroneous extractions from irrelevant texts achieve a satisfying rule precision and can be better tuned for the recognition of relevant content. This variation of learning algorithm has been examined on the Bosnian and MUC corpora and abbreviated as *without irrelevant texts*.

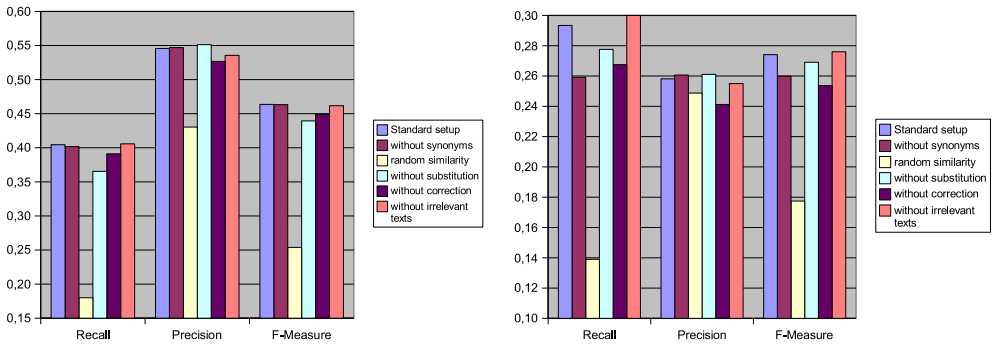

Figure 12.6: Significance of single components for P, R and F on Bosnian (left chart) and MUC corpus

The experiments in that respectively one of the components has been omitted have been performed on all three corpora[3]. We analyze the influence of every component based on the results displayed in fig. 12.6 and 12.7.

Synonymy relation recognized during semantic preprocessing should contribute to the increase of the coverage of different expression forms by extraction rules replacing concrete lexical expressions by the synonym pattern in the generalization step. The most noticeable effect has been achieved on the MUC corpus where the recall value dropped by more than four percent without using the synonym pattern while the precision increased negligibly. The same behavior of

[3] The non-classified Bosnian and MUC corpora have been used because of the experiments with filtered training corpora

precision can be observed on the other two corpora. This means that the generalization using the synonym pattern does not lead to overgeneralized extraction patterns that match irrelevant content. On the one hand, it argues for the high accuracy of our algorithm for detection of synonymy. At the same time the almost constant precision value is an evidence that the incorrectly determined synonyms are not harmful, because the context they are embedded in is only appropriate for the real synonyms, yielding eventually a semantically incorrect pattern that cannot match a real expression in the natural language (see also p. 58).

The benefit of synonym pattern for other two corpora is smaller than for MUC corpus. Especially on Bosnian corpus the improvement is negligible (recall increases by 0,23%, F-measure – by 0,04%). This fact most probably has to do with the kind of language used in the corpora: while MUC corpus features a very rich, manifold language with a big diversity of expressions, the language of Bosnian corpus is quite restricted due to the military origin. The attribute values consist of quite discrete, unambiguous terms. Th experiment confirms therefore a plausible conjecture that synonyms are more helpful for heterogeneous free texts than for restricted, quite unambiguous language.

The omission of the rule similarity metric 7.2 causes the by far biggest impact on all three corpora. To enable a reasonable rule merging the trivial similarity metric we used instead randomly selected rules that extract at least one common attribute value. However, the recall drop to less and F-measure decline to a little bit more than the half of the values achieved by the standard setup the MUC and Bosnian corpus demonstrate that the rule similarity based on structural analogy and cooccurrence of attribute values is vital for the successful rule merging. The precision fell not as drastically because the most induced incapable rules failed to pass the validation step. The comparatively better performance on the seminar announcement corpus can be explained by the big number of training examples and thus initial rules and the small size of its target structure. Given a big number of initial rules the probability is higher that among randomly selected pairs merging of some pairs yields reliable, general extraction rules. Especially for the regular time attributes several good rules are sufficient to achieve high precision and recall values. Taking a more general view the experiment confirms that the successful generalization of extraction rules is crucial for our approach.

Substitution heuristic is conceived to increase recall when the other generalizing heuristics are not able to improve the coverage of the set of correct rules. It fulfills its purpose quite successfully achieving a rise of recall by at least circa 2% and even around 4% on the Bosnian corpus. The relatively small percental contribution to the recall on seminar announcement corpus is probably connected with the already big number of good general rules, which establish already a high coverage of relevant content. Since the precision thresholds are correspondingly high, for many of substituted rules it is difficult to pass the validation step. The Bosnian corpus, on the contrary, features quite few good extraction rules due to its small size so that the potential of substitution heuristic can be better exploited.

Analogously to the synonym patterns and as we expected (s. p. 97), substitution heuristic does not have a notable effect on precision. Leaving substitution out only on the Bosnian corpus it increases by the noteworthy 0,5%. This indicates

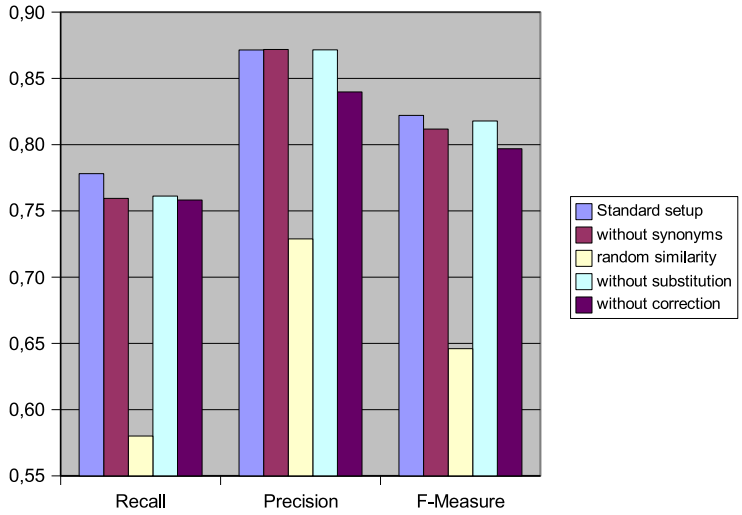

Figure 12.7: Component relevance study on the seminar announcement corpus

that the malformed patterns among those obtained by the substitution do not reflect and therefore practically do not match natural language.

Rule correction is beside the rule similarity the only component that has a measurable positive effects on both precision and recall. Our initial expectation was that recall will benefit more from the correction (as it does on the MUC corpus) because due to correction more induced rules can be validated covering more potential attribute values. Rule correction certainly also increases the precision of corrected rules eliminating false positive extractions. Obviously, on the Bosnian and SA corpora the precision of corrected rules has even been notably higher than of those that have been validated without correction so that the overall precision grew by about 2% and 3% respectively. Since the rule correction is restricted only to improvement of pattern parts that encode extracted fragments, we cannot expect a large recall or precision boost. However, notedly improving the precision and recall values on all examined corpora, rule correction proved valuable as an expedient complement to rule generalization.

As opposed to other components, **exclusion of irrelevant texts from the training** did not bring considerable improvement of the extraction quality. While on the MUC corpus the performance has been better, the F-measure value on the Bosnian corpus has been slightly worse. The differences are though so small that no conclusions about the influence of this learning variant on the overall extraction quality for single corpora can be drawn. The recall is slightly bigger and precision smaller than the corresponding values of the standard setup, since the extraction rules are not discredited by incorrect extractions from irrelevant texts, so that more less precise rules can pass the validation step (reducing the precision value on the test corpus). The low improvement of recall did not, however, fulfil our expectations. The positive effect of removing sources of incorrect extractions and increasing rule precision is obviously almost equalized by the adjustment of precision thresholds during the validation. Because of higher thresholds only a couple of new rules can be added to the final set of correct rules, which does not have a significant impact on its coverage.

The component relevance study confirms that all components except for the exclusion of irrelevant documents from the training corpus benefit the performance of GROPUS. Since most of them contribute to the generalization of rules (such as substitution heuristic or synonym pattern), a particular improvement of recall can

Validation strategy	*Bosnian* R	P	F	*SA* R	P	F	*MUC* R	P	F
RPT+APT	40.45	**54.56**	**46.36**	77.81	87.14	82.21	29.34	25.82	27.41
optimal	+2.05	−0.96	+1.56	−0.22	+1.02	+0,38	−0.58	+3.01	+1.39
local APTs	33.61	51.47	40.55	**83.23**	85.79	**84.40**	30.93	20.29	24.46
optimal	+3.73	−0.12	+2.92	+0.3	−0.06	+0.14	−2.06	+3.15	+1.32
Covering	**43.50**	42.09	42.68	56.96	**87.36**	68.92	**31.50**	**26.32**	**28.70**
optimal	+1.91	+7.29	+4.44	+6,35	−1,76	+3,01	+0.5	+0.87	+0.7

Table 12.2: Results employing different validation strategies on Bosnian, SA and MUC Corpora

be observed. Rule similarity metric proved to be the most valuable component, which had the largest impact on the quality and amount of derived extraction rules. The effectiveness of some components (like synonym pattern) depends on the environment as the varying results on different corpora demonstrate.

The study could be extended to other components and parameters of our approach, e.g. investigating the contributions of single elements of the pattern specification language such as further backtracking patterns or the role of linguistic preprocessing. However, such investigation would require a severe intervention in the learning algorithm. Exclusion of backtracking patterns such as wildcard or Kleene closure involves the revision of several core components, because rule merging, for instance, would be trivialized without these patterns. Omission of linguistic preprocessing would supersede the view of patterns as XML queries. Adjusting finer parameters such as varying the values of merging and abstraction functions for different low-level patterns might be helpful for tuning the system to a particular application domain, is though of little scientific value.

12.4.2 Evaluation of Validation Strategies

As we already have seen in the previous experiments, rule validation is one of the crucial components that determine the size and the quality of extraction rules. Rule validation has a significant impact on the evaluation metrics determining the relation between recall and precision, and it always has to be considered when interpreting evaluation results. In sec. 9.2 we presented three validation strategies that alternatively can be used in GROPUS. In all described experiments the validation strategy based on RPT and APT was employed. In the following we compare the performance of GROPUS employing each validation strategy on the three evaluation corpora. The experimental settings fully correspond to those in the previous experiment (s. 12.4.1).

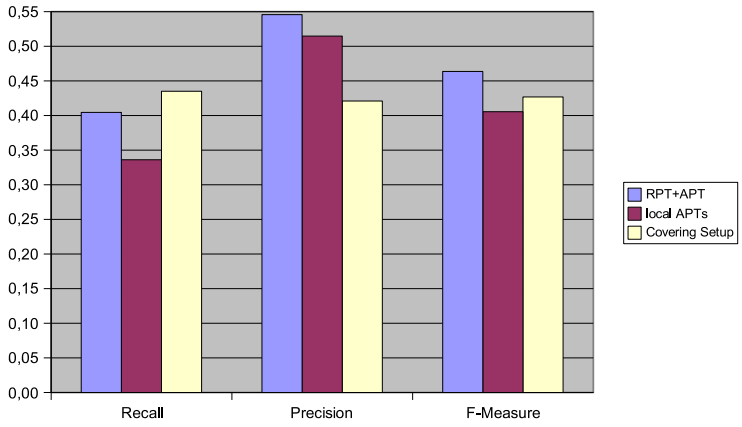

Figure 12.8: Comparison of validation strategies on the Bosnian corpus

Table 12.2 comprises the results obtained on the evaluation corpora using different validation strategies. The line *optimal* below each validation strategy

contains the percental change of the metric values employing the optimal assessment of thresholds on the test corpus. In other words, the smaller the delta to the optimal values the better the approximation of optimal threshold values of a validation strategy functions. Good approximation of maximum achievable F-measure is beside its absolute values an important characteristic of a validation method. Since the overall F-measure is optimized, the recall or precision values in optimal case can be smaller than those really achieved in favor of the bigger F-measure.

Fig. 12.8 compares the values of evaluation metrics achieved on Bosnian corpus. *RPT+APT* is clearly the most efficient validation strategy achieving highest microaverage precision and F-measure values and a good approximation of optimal values (only 1.56% difference to the optimal F-measure). The results of *local APTs* are notably inferior to those of *RPT+APT*. Here the weakness of *local APTs* becomes apparent: local maximum values for single attributes do not always contribute to and are sometimes even harmful for the overall performance. For attributes with few training instances, e.g. ZIELORT, increasing F-Measure causes huge number of false positives since the APT has to be decreased to a very low value to enable extractions of at least some attribute values. Since the false positives of ZIELORT are a part of the total amount of false positives, the total precision value considerably suffers. The same effect can be observed for attributes with high extraction quality: in this case the APT is set so high that the rules are prevented from extraction that achieve results lying above the total level, but below the level of best rules extracting this attribute. The extractions of the rejected rules would, however, improve the overall performance. Therefore validation strategies optimized for increasing the F-measure of single attributes are prone to decreasing the overall extraction quality especially in case of target structures with extremely different complexity of attributes.

Covering validation succeeds in obtaining a set of more diverse rules and achieves the highest recall value lying even 3% above the recall of *RPT+APT*. However, the precision decline is so severe (also due to the bad approximation loosing 7.29%, cf. table 12.2) that the overall performance expressed by the F-measure value cannot compete with that of *RPT+APT*. Despite the heterogeneous attributes complicated by the small size of Bosnian corpus the latter validation strategy achieves a balanced performance keeping both precision and recall on a high level compared with other validation strategies.

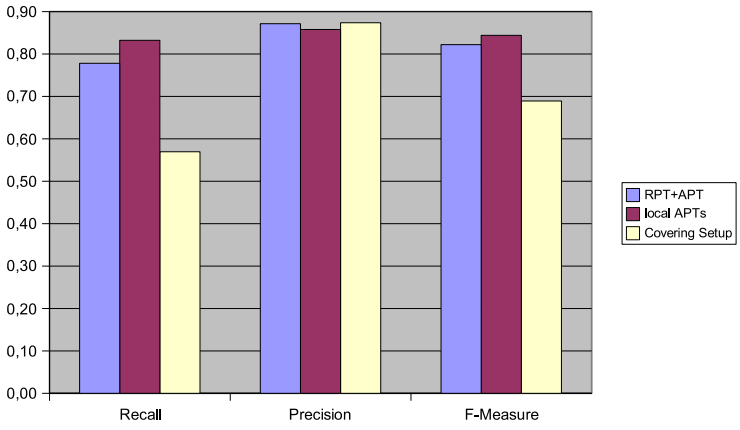

Figure 12.9: Performance of validation strategies on the SA corpus

A different picture is presented in the fig. 12.9 where the result on the SA corpus are depicted. The best total performance is shown by *local APTs* that reaches highest recall (83.23%) and F-measure (84.4%) values resulting not least from

an excellent assessment of optimal thresholds deviating only by 0.14% from the maximum achievable F-measure. As opposed to the Bosnian corpus the seminar announcement corpus offers optimal prerequisites for this validation strategy. There is a big number of training samples for each attribute and the interdependence of attributes is quite low. Therefore an independent ATP for each attribute can best reflect the complexity of single attribute setting a minimum quality level of rules extracting it. In this case the attributes LOCATION and SPEAKER notably benefit from the customize thresholds since their recall could be raised without significant loss of precision.

RPT+APT validation is slightly inferior featuring the marginally higher precision but a circa 5.5% smaller recall (cf. table 12.2). Even though the customization of attribute precision thresholds allows to achieve a bigger recall because of the special properties of the SA corpus (s. above), the gap in the total F-measure is rather small (82.21% vs. 84.39%) so that the total performance of *RPT+APT* is close to the best validation strategy on this corpus.

This conclusion does not hold for *covering setup*. Its performance suffers from the fact that the target structure of the SA corpus features only four attributes that have a big number of training examples a resulting in a quite large set of good induced extraction rules. Striving for the diversity of the rule set *covering validation* rewards exotic rules extracting some special cases but featuring a very little coverage and ignores many good rules extracting items already extracted by the best rules from the training set. This involves a considerable loss of recall because the ignored rules find many relevant fragments in the test corpus that the best rules cannot identify (an effect we discussed on p. 113). The achieved highest precision is worthless in the face of 26% deficit in recall.

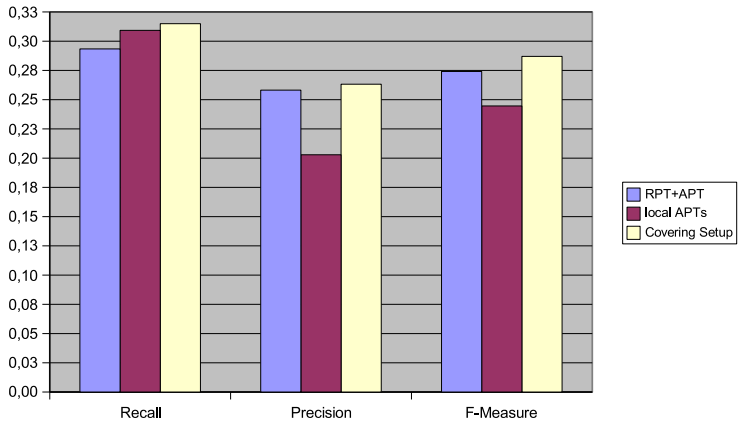

Figure 12.10: Significance of single validation methods for P, R and F on the MUC corpus

Covering setup unfolds its potential on the MUC corpus. Among all results depicted in fig. 12.10 it reaches the maximum values of all three evaluation metrics. MUC corpus features 13 attributes, many of which are underrepresented so that the most extraction rules suffer from the low precision. Selecting rules that cover as different training samples as possible the covering strategy improves the coverage on the test corpus. The expected drop of precision fails to appear. The reason is probably that covering validation strategy values inter alia quite specific rules induced in the early iterations of the learning algorithm. These rules obtain an almost constant precision on training and test corpora whereas very general rules favored by other validation strategies usually perform worse on the test corpus so that the predicted RPT and APT are more accurate in case of the *covering setup*.

RPT+APT is again the second best strategy lying around 2% behind in the F-measure value. However, the 2% difference in recall and only marginal precision lag underline the good performance of *RPT+APT* that in contrast to *local APTs* does not allow a shortfall of one of the evaluation parameters. *Local APTs* achieves a comparable recall, but a significantly smaller precision loosing also 3.15% because of bad threshold approximation. Analogously to the Bosnian corpus the precision value is again the victim of the attempt to reach better F-Measure values for the underrepresented attributes with a poor extraction quality. To increase their recall large number of incorrect extractions are taken into account.

Choice of the Validation Strategy

The notably different results on the three corpora demonstrate that, as we already pointed out in the sec. 9.2, every strategy has its advantages and drawbacks that partially compensate each other. For example, *local APTs* reveals deficiencies in handling domains with very different level of attribute complexity. Optimization of attribute-specific F-measures leads in such domains usually to the drastic decline of the overall F-measure. On the other hand, the optimization of the total F-measure with 13 local APTs (e.g. in case of MUC) is not feasible due to a huge computation in contrast to the optimization accomplished by *RPT+APT* strategy. *Covering setup* shows weakness in environments with a big number of training samples and reliable extraction rules.

Among all strategies *RPT+APT* demonstrated the most balanced performance being always among the best and achieving recall and precision values that have been at least close to the respective maximum. It also reached the best approximation of optimal F-measure value deviating by not more than 1.56%. *RPT+APT* did not reveal any particular deficits in any of the domains and performed best in the difficult conditions with an insufficient training supply on the Bosnian corpus. Controlling the overall quality of extraction rules and simultaneously customizing the validation for attributes *RPT+APT* proved to be a versatile, universally applicable validation strategy. Therefore we used it as the default validation method in all experiments and all presented experimental results are achieved with this particular setting of GROPUS (even though better results have been achieved by other strategies for particular corpora).

Tuning a system for a concrete application domain *local APTs* might be preferred in environments with a big training corpus and small interdependence between attributes while *covering setup* can be leveraged in domains with low recall and diverse, complex target structures.