18 Ablation Study and Utility of Incremental Training

This chapter will study two questions:

- How important are the various sources of information that we include in our rich context representations (cf. Sec. 12.2)? Does it make sense to include them all or would evaluation results stay the same or even improve if we omit some of them?
- How and to what degree can incremental training (cf. Sec. 3.3), one of the novelties we have introduced in the field of IE, reduce the human effort necessary to provide training data?

18.1 Ablation Study

The goal of this ablation study is to compare the relative importance of different sources of information which we include in the rich context representations (cf. Sec. 12.2) we use as input for the token classifier in our classification-based approach.

There are five major sources of information whose influence we will investigate, by removing each one of them from the generated context representations to observe how this changes results:

Markup (HTML): one of our stated novel assumptions (Sec. 8.1) was that *"Structure matters"* and that the explicit structural information contained in structured document formats such as HTML should not just be ignored by information extraction systems. However, the usual IE corpora, including the two we are using, do not contain *explicit* structural markup—the files to extract from are just plain text files (aside from the annotation of answer keys, which define the expected *output* of the extraction system and hence, obviously, cannot be used as *input*). For such cases, a weaker version of our assumption states that even the *implicit* structural information contained in plain text files ("ASCII markup") might be useful for extraction. Therefore we make this implicit markup explicit during preprocessing by using a heuristic converter (*txt2html*) that converts plain text into HTML (cf. Sec. 12.1). For the ablation study, we will skip this heuristic conversion step to find out whether the added structural markup is actually useful.

Linguistic information is added during preprocessing by invoking the *TreeTagger* to perform sentence splitting, shallow parsing and POS tagging (cf. Sec. 12.1). We will test how skipping this step affects results.

Semantic information is added to the context representations from a configurable list of dictionaries and gazetteers. By default, we use an English dictionary and a

F-measure	Default	No HTML	No Linguistic	No OSB	No Prior Ext.	No Semantic
etime	96.3	97.0	89.2	95.5	96.6	**97.2**
location	**80.1**	76.8	68.0	69.3	74.0	78.0
speaker	**81.0**	72.8	53.6	64.9	75.6	77.0
stime	99.3	**99.4**	99.1	98.7	99.3	**99.4**
Average	**88.5**	85.6	76.8	80.9	85.4	87.0

Table 18.1: Ablation Study: Seminar Announcements

few word lists related to person names and locations (listed in Sec. 12.2). We would expect this semantic information to be especially useful for the LOCATION and SPEAKER attributes in the Seminar corpus and for the ACQLOC attribute in the Acquisitions corpus, but only to a lesser degree or not at all for the other attributes, since the provided information does not cover company names, time expressions, or monetary amounts.

OSB (Orthogonal Sparse Bigrams) is a feature combination technique (cf. Sec. 11.2) we usually use to enrich the feature space, allowing the classifier to recognize and learn combinations of adjacent features occurring together. In the ablation study, we test whether this actually helps extraction performance or whether similar results can be reached without this technique.

Prior extractions: usually we include the information on the last preceding attribute values identified in the same document, to allow the classification algorithm to learn about positional relations among attributes (cf. Sec. 12.2). For the ablation study, we omit this information to check whether it is actually helpful.

Table 18.1 shows the F-measure results of performing the ablation study on the Seminar corpus (graphically represented in Fig. 18.1). The results on the Acquisitions corpus are shown in Table 18.2 and Fig. 18.2. We will only report results with incremental training, since the relative results with batch training are similar and do not offer any additional insights.

For both corpora, the *linguistic* annotations contribute most to the results—without them, the average F-measure drops by 11.7% for the Seminar corpus, by 12.5% for the Acquisitions corpus. This is not surprising—linguistic information is used by almost all IE systems and this study confirms that there are indeed good reasons for this. We would expect the relevance of linguistic information to be higher for *free texts* as contained in the Acquisitions corpus than for *semi-structured texts* as in the Seminar, since the latter are less strictly grammatical than the former and contain more non-linguistic clues. This is also confirmed by the study—the *absolute* drop in F-measure is already slightly larger for Acquisitions corpus, and the *relative* drop is much larger.

The *OSB* feature combination technique is the second most important factor for both corpora—without it, F-measure degrades by 7.6% on the Seminar and by 9.1% on the Acquisitions corpus. This confirms that it is indeed a clear benefit if the classifier is able to recognize and learn feature combinations instead of having to consider each

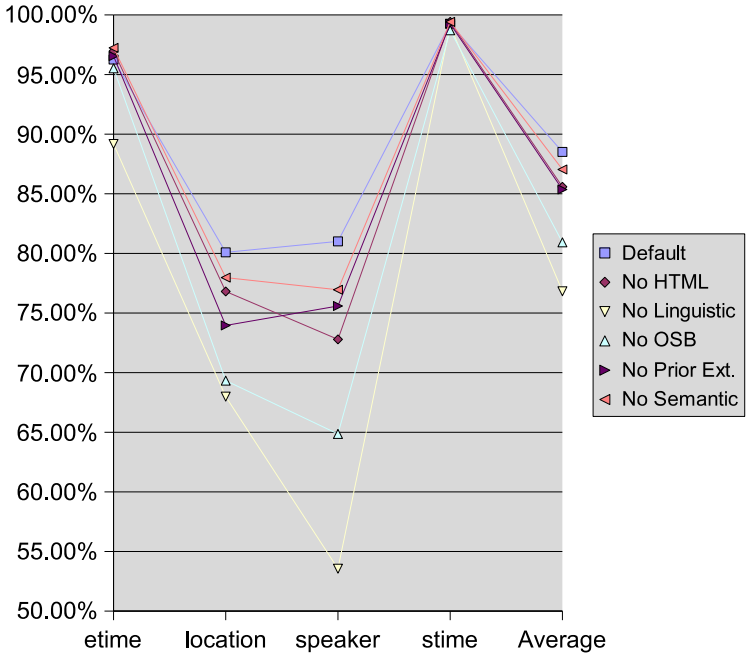

Figure 18.1: Ablation Study: Seminar Announcements

F-measure	Default	No HTML	No Linguistic	No OSB	No Prior Ext.	No Semantic
acqabr	**51.7**	49.7	33.4	43.1	50.4	51.5
acqloc	**27.3**	22.4	7.7	22.0	20.4	22.1
acquired	**49.2**	48.3	38.8	42.8	41.8	48.9
dlramt	60.9	**62.2**	55.9	54.0	59.9	61.3
purchabr	**55.3**	52.3	32.6	42.8	53.8	55.2
purchaser	**51.6**	47.4	41.4	41.5	36.0	50.1
seller	**26.0**	22.2	11.8	19.1	14.3	25.7
sellerabr	24.0	22.0	10.7	15.4	18.1	**24.2**
status	53.0	**53.2**	50.6	40.6	47.8	52.3
Average	**48.0**	45.9	35.5	38.9	41.4	47.3

Table 18.2: Ablation Study: Corporate Acquisitions

feature in isolation.

The inclusion of *prior extractions* is especially relevant for the Acquisitions corpus—here, the average F-measure drops by 6.6% without this information, while it only drops by 3.1% on the Seminar corpus. This has probably to do with the fact that there are more attributes in the Acquisitions corpus and that the (implicit) relations between them are more complicated. Interestingly, ETIME results in the Seminar corpus are slightly *improved* without this information (and STIME results are unchanged), so in this case which we had mentioned as an example, the provided information did *not* turn out to be helpful—probably because it will often be redundant since time

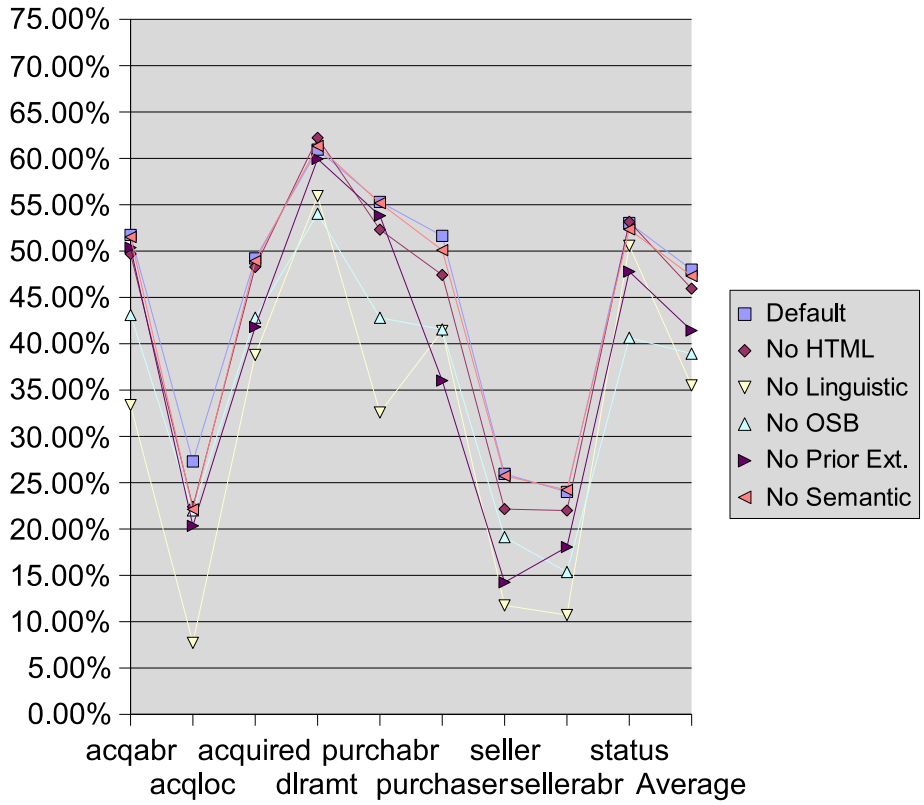

Figure 18.2: Ablation Study: Corporate Acquisitions

expressions to the left (or to the right) of a token to classify can also be recognized by their *word shapes* which we always include as features (cf. Sec.12.2).

For both corpora, the heuristically added *HTML markup* is less useful than the three kinds of information discussed so far, but omitting it still results in a drop of 2.9% for the Seminar and of 2.1% for the Acquisitions corpus. Since the original input format is plain text and the text structure is only deduced in a heuristic process, we would not have expected a much larger difference—as it is, the difference that we can observe confirms that our *"Structure matters"* assumption holds, even in these difficult circumstances (especially when considering how small the performance differences of the best systems evaluated on the Seminar corpus are, cf. Table 17.2 in Sec. 17.2).

The *semantic* information we are using turns out to be of comparatively little use—without it, average F-measure is reduced by 1.5% on the Seminar corpus and by only 0.7% on the Acquisitions corpus. This is only on average, however—as noted above, our semantic sources are targeted on person names and locations, and for the relevant attributes there is a more noticeable performance difference: 4.0% for the SPEAKER and 2.1% for the LOCATION of seminars, and 5.2% for the ACQLOC (location on acquired companies). This indicates that we might be able to improve results further by using additional gazetteers related to the attributes in a task, such as lists of company names for the Acquisitions corpus; we refrained from doing so since task-specific fine-tuning

is not our goal.

The relatively low importance of semantic sources indicates that our approach makes efficient use of syntactic and linguistic features to compensate for missing explicit semantic data. Other authors that have evaluated their approach on the Seminar corpus with and without semantic sources report a higher dependency: for the rule-learning *(LP)*^{2} system, average F-measure drops by $\approx 23\%$, from 86% to 63.1% [Cir01]; for the statistical *BIEN* system, it is $\approx 11\%$, from 88.9% to 77.8% [Pes03] (we are not aware of other such comparisons on the Seminar or of any on the Acquisitions corpus).

The ablation study confirms that all the sources of information we consider actually contribute to the good results reached by our system; none of them is generally useless (or even harmful). It also confirms, however, that noise can be a problem: while the most important sources (linguistic annotations and OSB) benefit all attributes, the less important ones tend to degrade results of a few attributes, especially for simple and regular ones such as ETIME, STIME, and DLRAMT. Hence we should also be careful not to add too much information.

We could continue performing more fine-granular studies regarding the effects of varying parameters such as the exact list of semantic sources, the number of prior extractions to consider and other parameters controlling what exactly is included in the context representation of a token. But we will not do this since such extensive parameter variation tests are not among our goals (cf. Sec. 7.4) and run a high risk of becoming corpus-specific. They are more appropriate as future work, especially when tuning the system for a specific task.

18.2 Utility of Interactive Incremental Training

In the traditional setup, as used in the preceding tests, training and test sets are clearly separated—50% of documents are used for training only (without evaluation) and the remaining 50% are used for evaluation (without any further training). In this setup, *incremental training* is just an alternative way of processing the training documents, which is faster than but not quite competitive with batch training (cf. Sec. 11.1.3 and the evaluation results in the last chapter).

However, what makes incremental training interesting is that is allows a *different* setup where the training and evaluation phases are no longer strictly separated. When incremental learning is used, it is possible to adapt the extraction model even during the evaluation phase, by allowing the classifier to train the expected attribute values (answer keys) from each document after evaluating its own predictions for this document. This corresponds to the interactive workflow described in Section 3.4, where the system proposes attribute values which are reviewed and corrected by a human supervisor. After the supervisor has corrected a document, the system updates its extraction model prior to processing the next document. With this *interactive* training and evaluation setup, the quality of the extraction proposals will continually improve, reducing the necessary human effort for providing annotated training examples and for correction.

Evaluation Set	50%		100%
Feedback	**No**	**Yes**	**Yes**
etime	96.3	97.8	94.2
location	80.1	80.2	73.2
speaker	81.0	83.9	77.0
stime	99.3	99.2	98.0
Average	88.5	89.5	84.8

Table 18.3: Results with Incremental Feedback

We have simulated and evaluated two variants of this interactive setup on the Seminar corpus—we did not repeat this test on the Acquisitions corpus, since it is meant to simulate the behavior of a real user and the bad results reached on that corpus in the standard setup (cf. last chapter) indicate that is is unlikely a user would actually consider the predictions made on that corpus to be helpful.

In the first variant, a conventional training phase of 50% is used, but in the test (evaluation) phase, for each document to test the system is in a first step asked to predict its attribute values (as usual), but then, in a second step, it is trained on the true attribute values for this document (i.e., the answer keys defined by the corpus)—this simulates a user who interactively corrects the results of the system and feeds the corrections back to the system to allow better predictions for the remaining documents. The results for this setup are shown in the medium column of Table 18.3: with this interactive feedback added, the average F-measure on the evaluation set increases to 89.5%: +1.0% compared to the results reached with incremental training in the standard setup (left column).

With this feedback mechanism it is no longer strictly necessary to start with a training-only phase; the system can be used to propose attribute values to be evaluated from the very start, using the *whole corpus* as evaluation set and no dedicated training set (0/100 split). Tested in this way, our system still reaches almost 85% F-measure over *all documents* (right column). This means the system can be beneficial to use very soon, without requiring a tedious manual annotation phase to provide initial training data.

Figure 18.3 shows the learning curve for this second variant. As can be seen, precision is high from the very start—more than 75% after the first 10 documents, more than 80% after 20. Initial recall is far lower, but it exceeds 50% after processing 50 documents and 70% after 160 documents.

An advantage of this interactive incremental setup is the reduced training burden. Figure 18.4 show the average numbers of correct predictions (true positives), missing answer keys (false negatives), and spurious predictions (false positives) measured for the conventional training set, i.e., the first 50% of documents in each test run. In the conventional setup, these documents are manually annotated, so a human user needs to perform all these extractions without any outside help. We can see that using our system to handle this task interactively can reduce this training effort enormously, since the system already proposes most of the answer keys correctly.

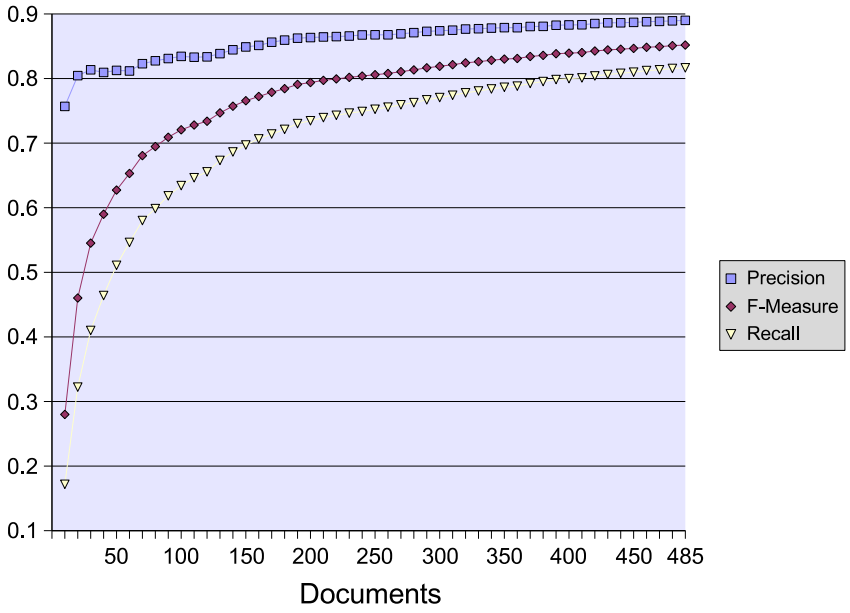

Figure 18.3: Incremental Feedback: Learning Curve (average precision, recall, and F-measure on all documents processed so far)

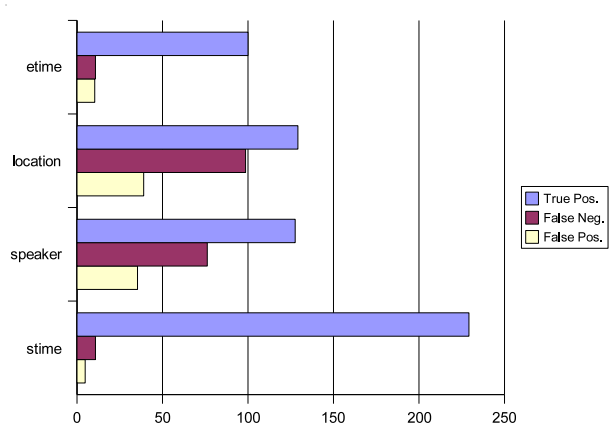

Figure 18.4: Incremental Feedback: Correct, Missing, and Spurious Predictions in the "Training Set"

	Answer Keys	**Required Corrections**	**Correction Ratio**
etime	110.8	21.2	19.1%
location	227.8	137.6	60.4%
speaker	203.8	111.6	54.8%
stime	240	15.6	6.5%
All	782.4	286	36.6%

Table 18.4: Incremental Feedback: User Effort for Correcting the "Training Set"

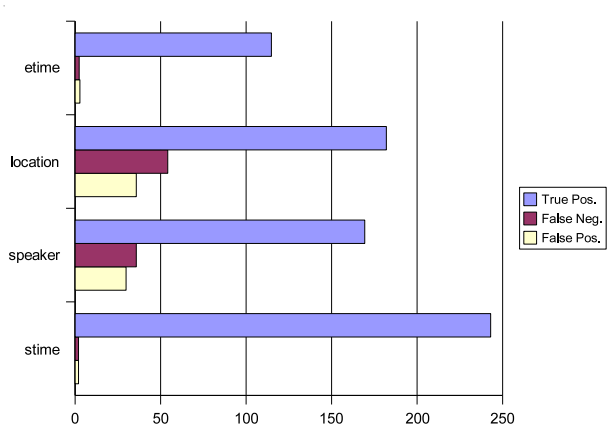

Figure 18.5: Incremental Feedback: Correct, Missing, and Spurious Predictions in the "Evaluation Set"

	Answer Keys	**Required Corrections**	**Correction Ratio**
etime	117.2	5.2	4.4%
location	236.2	90	38.1%
speaker	205.2	65.6	32.0%
stime	245	4	1.60%
All	803.6	164.8	20.5%

Table 18.5: Incremental Feedback: User Effort for Correcting the "Evaluation Set"

Table 18.4 calculates how this affects the training effort. For each attribute, it shows the number of expected answer keys (true positives + false negatives) and the number of erroneous or missing predictions that must be corrected by the human user (false positives + false negatives). The "correction ratio" is the number of required corrections divided by the number of expected answer keys. As usual, all values are averaged over the five test runs.

For the more difficult attributes SPEAKER and LOCATION, the correction ratios are about 55–60%, while for the easier time expressions they go down to 19% (ETIME) or even 6.5% (STIME). Summed over all attributes, the "correction ratio" is about 37%—the number of extractions the (simulated) user would have to perform to get a fully annotated training corpus is almost three times the number of operations required to interactively correct the predictions made by our system after it has been trained on all documents corrected so far. This shows that this interactive incremental training style we have proposed can indeed reduce the training burden in a substantial way.

Figure 18.5 and Table 18.5 show the corresponding values for the remaining 50% of documents, which are used as evaluation set in the conventional setup. For these documents, the average "correction ratio" goes down to 20%. It is noteworthy that the sum of correction operations over both halves of the corpus (≈ 450), i.e., the number of correction operations required to get a *fully* corrected corpus, is far lower (less than 60%) than the number of answer keys in the training set (≈ 780) which all need to be extracted manually in the conventional setup.