20 Weakly Hierarchical Extraction

This chapter describes our evaluation of the weakly hierarchical (WH) approach proposed in Chap. 14.

20.1 Experimental Setup

20.1.1 Named Entity Recognition

For most of the tests, we used our own system as named entity recognizer. We trained the system on the *English CoNLL-2003 Shared Task* [TKS03] data, a corpus of 1393 newswire articles from the *Reuters Corpus, Volume 1* [Reu00] annotated for NE recognition. The corpus comprises four types of named entities: PERSONs, LOCATIONs, ORGANIZATIONS and MISCELLANEOUS entities.

The *CoNLL-2003* data does not annotate temporal expressions, nor do other freely available corpora (to our knowledge). Thus we wrote and used a simple rule-based recognizer based on regular expressions to recognize TIME expressions for some additional experiments. The regularity of TIME expressions made this approach feasible.

20.1.2 Evaluation Corpora and Setup

We evaluated the weakly hierarchical approach on the two corpora used throughout this work, *CMU Seminar Announcements* and *Corporate Acquisitions*, using the evaluation setup described in Chap. 17, using evaluation setup and metrics as described before. We used batch training for all experiments reported in this chapter.

Figures 20.1 and 20.2 show the inheritance hierarchies chosen to connect the corpus-specific types with named entity types. For the *Acquisitions* corpus, both long and short names of the three kinds of organizations involved are considered subtypes of ORGANIZATION.

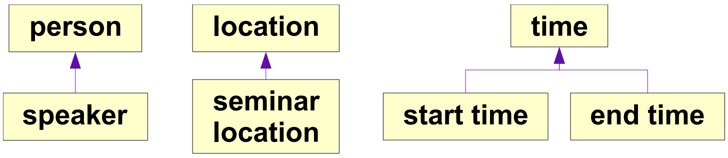


Figure 20.1: Seminar Corpus: Inheritance Hierarchy

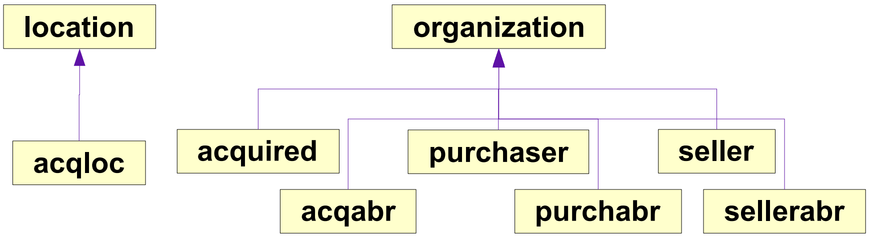


Figure 20.2: Acquisitions Corpus: Inheritance Hierarchy

Supertype	Recall
Seminar Corpus	
Speaker	40.82%
Location	0.47%
Acquisitions Corpus	
Location	27.23%
Organization	43.60%

Table 20.1: Recall Reached by Supertype Recognizers on Subtype Answer Keys

20.2 Experimental Results

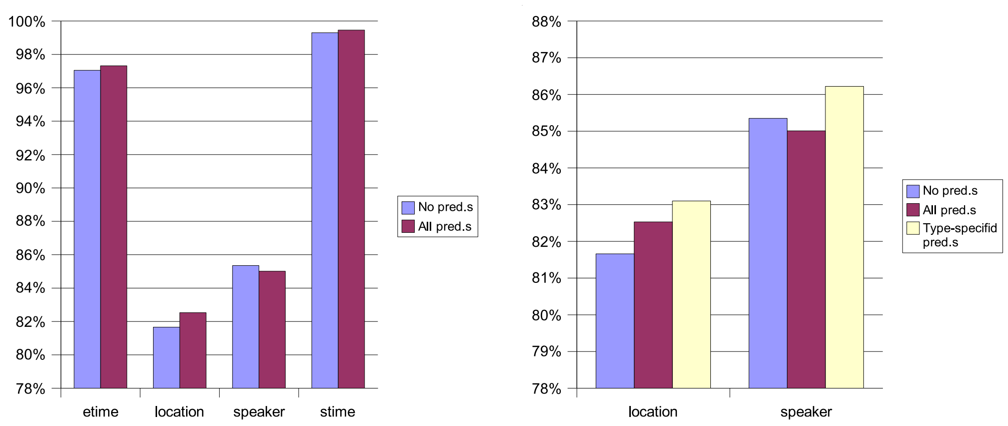
20.2.1 Strictly Hierarchical Approach

Since we did not implement the strictly hierarchical approach because of the problems discussed above (Sec. 14.3), we could not measure its performance directly. Instead we measured the recall reached by the "supertype" (NE) recognizers on the corresponding subtypes, i.e., during the first step of the SH approach (cf. Sec. 14.3), to get an upper limit of the results the SH approach would be able to reach. The low results (Table 20.1) can be regarded as a confirmation of our conjecture that the SH approach would not work since recall errors (false negatives) *cannot* be corrected in later steps. True recall would probably be lower but certainly not higher than the upper limit measured here.

The case of LOCATION (only 0.47% recall) demonstrates the *different semantics problem* discussed in Sec. 14.3.

20.2.2 Weakly Hierarchical Approach

Figure 20.3(a) shows the F-measure results reached by applying the weakly hierarchical approach indiscriminately to the Seminar corpus, by making all first-level predictions from the CoNLL corpus available to all second-level classifiers (without TIME predictions since those are not available from the CoNLL corpus). Results are mixed: LOCATION results are improved by almost 1%, but SPEAKER results degrade slightly (by 0.3%). There is also a very small positive effect on the recognition of start/end time entities, even though the "supertype" predictions do not cover temporal expressions (except by negation).



(a) All Predictions

(b) Type-specific Predictions

Figure 20.3: Seminar Corpus: F-measure Results

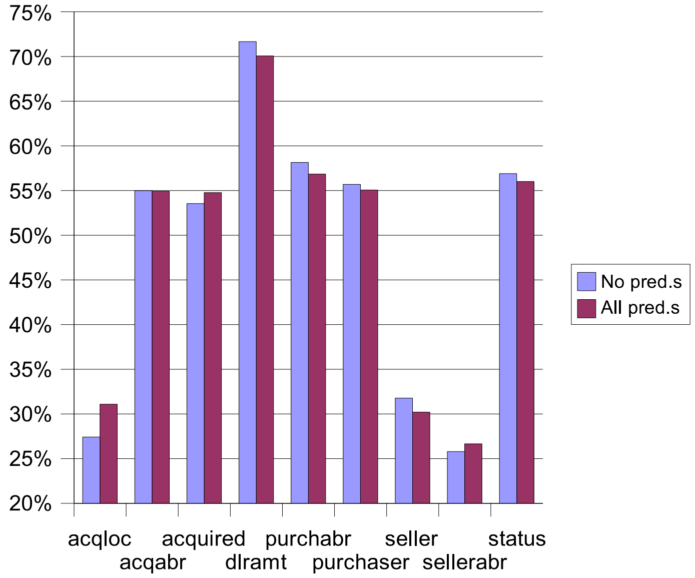
A clearer effect is reached by utilizing “supertype” predictions discriminatively (cf. Sec. 14.4), by restricting their visibility to the corresponding “subtype” classifiers (Fig. 20.3(b), right columns). In this case, the results of both LOCATION and SPEAKER are improved, by about 1.4% and 0.9%.

Figure 20.4 shows the results reached on the Acquisitions corpus. Again, the effects of indiscriminate application are very dubious (Fig. 20.4(a)): results are improved for only three fields, ACQLOC (+3.7%), ACQUIRED (+1.2%), and SELLERABR (+0.9%). For all other fields, results either stagnate or degrade, probably due to the additional noise introduced by the predictions.

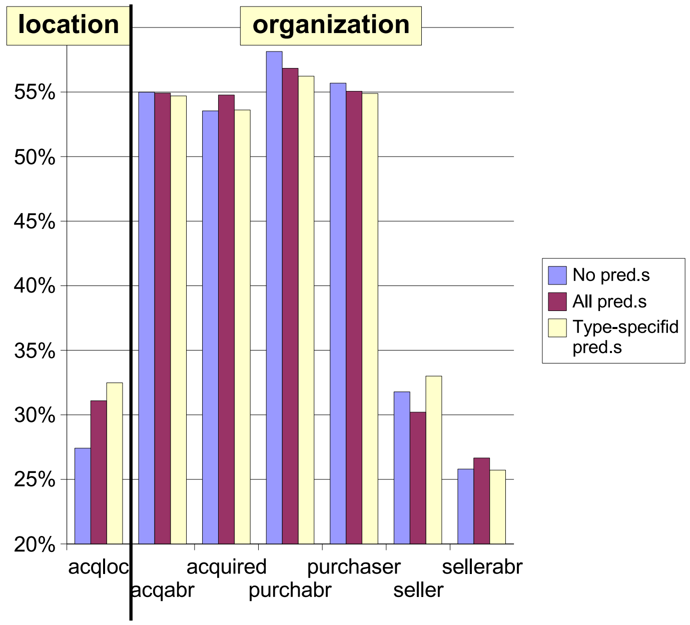
Results of discriminate application are shown in Fig. 20.4(b). As stated above, two supertypes have been used: LOCATION, which has only a single subtype (ACQLOC), and ORGANIZATION, which has six subtypes (the three kinds of companies involved and their abbreviations). ACQLOC results are clearly improved (+5.1%), but for the ORGANIZATION subtypes, this approach fails to be effective—results on most fields stagnate or degrade.

We suppose that this is at least partially caused by the fact that, for these types, the main problem is to differ between attribute values of similar types instead of locating attribute values, thus the semantic information added by the WH approach does not help much. The differentiation between long and short variants of organization names is especially tricky since they are often very similar, and the supertype information does not help at all to do this differentiation.

So check this thesis, we ran a test on a simplified variant of the Acquisitions corpus, where long and short names of each type of organization have been collapsed into a single attribute. To still require extraction of both short *and* long name, we switched from “one answer per attribute” evaluation to “one answer per different string”, i.e., all variants of each name must be found (cf. Sec. 15.2). The F-measure results are shown in Fig. 20.5. Indeed, in this setup the discriminate variant (right columns) is



(a) All Predictions



(b) Type-specific Predictions

Figure 20.4: Acquisitions Corpus: F-measure Results

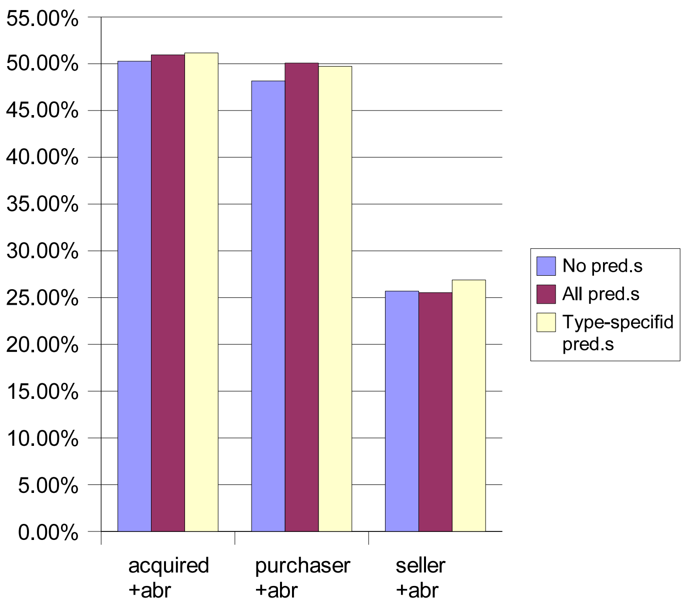


Figure 20.5: Acquisitions Corpus: Collapsing Short and Long Names

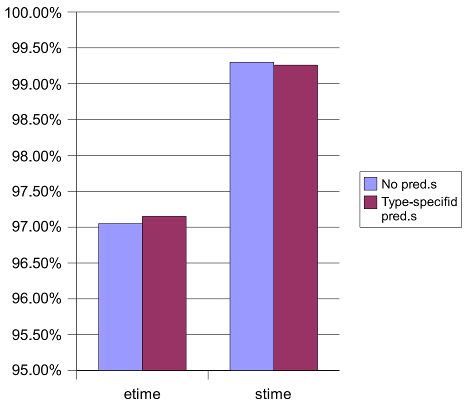


Figure 20.6: Seminar Corpus: Temporal Predictions

able to improve results on all the three types of organizations by between 0.9% (for ACQUIRED+ABR) to 1.6% (for PURCHASER+ABR).

The results of discriminatively using predictions from the TIME expression recognizer are shown in Fig. 20.6. They confirm the supposition that the WH mode is less useful for differing between subtypes of the same supertype—F-measure values are essentially unchanged, with a minimum increase in END TIME recognition and an even smaller decrease in START TIME recognition. Again, this is probably caused by the fact that the hard problem is to differ between START TIMEs, END TIMEs and time expressions that are neither; while the mere recognition of time expressions is almost trivial.

20.3 Concluding Remarks

Our results indicate that, for typical corpora, a strictly hierarchical approach would indeed not work because of the *different corpora problem* and related problems.

Results for the weakly hierarchical (WH) approach are mixed. Generally, it appears to add too much noise if applied indiscriminately. However, if applied discriminatively (for loose subtypes only), it can improve results. But it tends to fail if there are various subtypes derived from the same supertype. In such cases, the main problem is to differ between attribute values of similar types instead of locating attribute values, thus the semantic information added by the WH approach does not help much.

Based on our current results, we cannot recommend the WH approach for general application. It might be handy in some cases if information from suitable loose supertypes is available and can easily be integrated into a system, but in general there will probably be more promising and more useful ways of improving extraction quality.

Still, we believe that identifying and making accessible additional sources of information is a relevant area of research for advancing the field of IE. Finding ways to exploit sources of information in a better way and exploring further sources of information remain important topics for future work.