

A Learner with a Sense for Quality

Udo Hahn and Klemens Schnattinger

Text Knowledge Engineering Lab, Freiburg University
 Werthmannplatz 1, D-79085 Freiburg, Germany
<http://www.coling.uni-freiburg.de>

Abstract

A text understanding system with learning capabilities is presented. New concepts are acquired by incorporating two kinds of evidence – knowledge about linguistic constructions in which unknown lexical items occur and knowledge about structural patterns in ontologies such that new concept descriptions can be compared with prior knowledge. On the basis of the quality of evidence gathered this way concept hypotheses are generated, ranked according to plausibility, and the most credible ones are selected for assimilation into the domain knowledge base.

Introduction

We propose a text understanding approach in which continuous enhancements of domain knowledge bases are performed given a core ontology (such as WordNet (Fellbaum, 1998)). New concepts are acquired taking two sources of evidence into account: the prior knowledge of the domain the texts are about, and linguistic constructions in which unknown lexical items occur. Domain knowledge serves as a comparison scale for judging the plausibility of newly derived concept descriptions in the light of prior knowledge. Linguistic knowledge helps to assess the strength of the interpretative force that can be attributed to the grammatical construction in which a new lexical item occurs. Our model makes explicit the kind of quality-based reasoning that lies behind such a process.

We advocate a *knowledge-intensive* model of concept learning from sparse data that is tightly integrated with the non-learning mode of text understanding. Both learning and understanding build on a given core ontology in the format of terminological assertions, and hence make abundant use of terminological reasoning facilities. The “plain” text understanding mode can be considered as the instantiation and continuous filling of roles with respect to *single concepts* already available in the knowledge base. Under learning conditions, a *set of alternative concept hypotheses* are managed for each unknown item, with each hypothesis denoting a newly created conceptual interpretation tentatively associated with the unknown item.

A Model of Quality-Based Learning

Fig. 1 depicts how linguistic and conceptual evidence are generated and combined for continuously discriminating

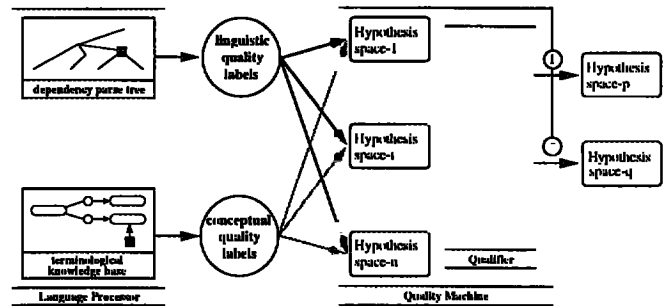


Figure 1: Architecture for Quality-Based Learning

and refining the set of concept hypotheses (the unknown item yet to be learned is characterized by the black square). The language processor yields structural dependency information from the grammatical constructions in which an unknown lexical item occurs in terms of the corresponding *parse tree*. The conceptual interpretation of parse trees involving unknown lexical items is used to derive *concept hypotheses*, which are further enriched by conceptual annotations reflecting structural patterns of consistency, mutual justification, analogy, etc. in the continuously updated *terminological knowledge base*. These kinds of initial evidence, in particular their predictive “goodness” for the learning task, are represented by corresponding sets of *linguistic* and *conceptual quality labels*. Multiple concept hypotheses for each unknown lexical item are organized in terms of a corresponding *hypothesis space*, each subspace holding different or further specialized concept hypotheses.

The *quality machine* estimates the overall credibility of single concept hypotheses by taking the available set of quality labels for each hypothesis into account. The final computation of a preference order for the entire set of competing hypotheses takes place in the *qualifier*, a terminological classifier extended by an evaluation metric for quality-based selection criteria. The output of the quality machine is a ranked list of concept hypotheses. The ranking yields, in decreasing order of significance, either the most plausible concept classes which classify the considered instance or more general concept classes subsuming the considered concept class.

Linguistic Quality Labels

Linguistic quality labels reflect structural properties of phrasal patterns or discourse contexts in which unknown lexical items occur – we assume here that the type of

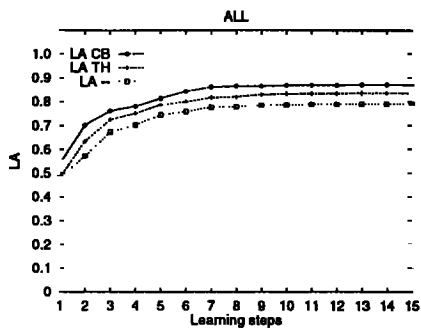


Figure 4: Learning Accuracy (LA) for the Entire Data Set

Related Work

Our approach bears a close relationship to the work of Mooney (1987), Gomez & Segami (1989), Rau et al. (1989), Hastings (1996), and Moorman & Ram (1996), who all aim at the automated learning of word meanings from context using a knowledge-intensive approach. Our work differs from theirs in that the need to cope with *several competing* concept hypotheses and to aim at a *reason-based selection* is not an issue in those studies.

The work closest to ours has been carried out by Rau et al. (1989) and Hastings (1996). Concept hypotheses are also generated from linguistic and conceptual data. Unlike our approach, the selection of hypotheses depends only on an ongoing discrimination process based on the availability of these data but does not incorporate an inferencing scheme for reasoned hypothesis selection. The difference in learning performance – in the light of our evaluation study – amounts to 8%, considering the difference between LA - (plain terminological reasoning) and LA CB values (terminological metareasoning based on the qualification calculus). Hence, our claim that we produce competitive results.

Note that the requirement to provide learning facilities for large-scale text understanders also distinguishes our approach from the currently active field of information extraction (IE) (Appelt et al., 1993). The IE task is defined in terms of a *pre-fixed* set of templates which have to be instantiated (i.e., filled with factual knowledge items) in the course of text analysis. Unlike the procedure we propose, no new templates have to be created.

Conclusion

In this paper, we have introduced a methodology for generating new knowledge items from texts and integrating them into an existing domain knowledge base. This is based on the incremental assignment and evaluation of the quality of linguistic and conceptual evidence for emerging concept hypotheses. The concept acquisition mechanism we propose is fully integrated in the text understanding mode. No specialized learning algorithm is needed, since learning is a (meta)reasoning task carried out by the classifier of a terminological reasoning system. However, heuristic guidance for selecting between plausible hypotheses comes from the different quality criteria. Our experimental data indicate that given these heuristics we achieve a high degree of pruning of the search space for hypotheses in very early phases

of the learning cycle.

Our experiments are still restricted to the case of a single unknown concept in the entire text. Generalizing to n unknown concepts can be considered from two perspectives. When hypotheses of another target item are generated and incrementally assessed relative to an already given base item, no effect occurs. When, however, two targets (i.e., two unknown items) have to be related, then the number of hypotheses that have to be taken into account is equal to the product of the number of hypothesis spaces currently associated with each of them. In the future, we intend to study such scenarios. Fortunately, our evaluation results also indicate that the number of hypothesis spaces decreases rapidly as does the learning rate, i.e., the number of concepts included in the remaining concept hypotheses. So, the learning system should remain within feasible bounds, even under these less favorable conditions.

Acknowledgements. We would like to thank our colleagues in the CLIF group for fruitful discussions and instant support. K. Schnattinger is supported by a grant from DFG (Ha 2097/3-1).

References

- Appelt, D., J. Hobbs, J. Bear, D. Israel & M. Tyson (1993). FASTUS: A finite-state processor for information extraction from real-world text. In *IJCAI'93 – Proc. 13th Intl. Conf. on Artificial Intelligence*, pp. 1172–1178.
- Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Gomez, F. & C. Segami (1989). The recognition and classification of concepts in understanding scientific texts. *J. of Exp. and Theoret. Artif. Intelligence*, 1:51–77.
- Hahn, U., S. Schacht & N. Bröker (1994). Concurrent, object-oriented natural language parsing: the PARSE-TALK model. *International Journal of Human-Computer Studies*, 41(1/2):179–222.
- Hastings, P. (1996). Implications of an automatic lexical acquisition system. In S. Wermter, E. Riloff & G. Scheler (Eds.), *Connectionist, Statistical and Symbolic Approaches to Learning in Natural Language Processing*, pp. 261–274. Berlin: Springer.
- Mooney, R. (1987). Integrated learning of words and their underlying concepts. In *CogSci'87 - Proc. 9th Annual Conf. of the Cognitive Science Society*, pp. 974–978.
- Moorman, K. & A. Ram (1996). The role of ontology in creative understanding. In *CogSci'96 – Proc. 18th Annual Conf. of the Cognitive Science Soc.*, pp. 98–103.
- Nirenburg, S. & V. Raskin (1987). The subworld concept lexicon and the lexicon management system. *Computational Linguistics*, 13(3-4):276–289.
- Rau, L., P. Jacobs & U. Zernik (1989). Information extraction and text summarization using linguistic knowledge acquisition. *Information Processing & Management*, 25(4):419–428.
- Schnattinger, K. & U. Hahn (1996). A sketch of a qualification calculus. In *FLAIRS'96 – Proc. 9th Florida AI Research Symposium*, pp. 198–203.
- Woods, W. & J. Schmolze (1992). The KL-ONE family. *Computers & Math. with Applications*, 23:133–177.