Deep Knowledge Discovery from Natural Language Texts

Udo Hahn & Klemens Schnattinger

Computational Linguistics Lab – Text Knowledge Engineering Group
Freiburg University, Werthmannplatz, D-79085 Freiburg, Germany
{hahn, schnattinger}@coling.uni-freiburg.de
http://www.coling.uni-freiburg.de

Abstract

We introduce a knowledge-based approach to deep knowledge discovery from real-world natural language texts. Data mining, data interpretation, and data cleaning are all incorporated in cycles of quality-based terminological reasoning processes. The methodology we propose identifies new knowledge items and assimilates them into a continuously updated domain knowledge base.

Introduction

The work reported in this paper is part of a large-scale project aiming at the development of SYNDIKATE, a German-language text knowledge assimilation system (Hahn, Schnattinger, & Romacker 1996). Two real-world application domains are currently under active investigation – test reports on information technology products (101 documents with 10^5 words) and, the major application, medical reports (approximately 120,000 documents with 10^7 words). The task of the system is to aggressively assimilate any facts, propositions and evaluative assertions it can glean from the source texts and feed them into a shared text knowledge pool. This goal is actually even more ambitious than the MUC task (for a survey, cf. (Grishman & Sundheim 1996)) which requires mapping natural language texts onto a highly selective and fixed set of knowledge templates in order to extract factual knowledge items only.

Given that only a few of the relevant domain concepts can be supplied in the hand-coded initial domain knowledge base, a tremendous concept learning problem arises for text knowledge acquisition systems. Any other KDD system running on natural language text input for the purpose of knowledge extraction also faces the challenge of an *open* set of knowledge templates; even more so when it is specifically targeted at *new* knowledge items. In order to break the high complexity barrier of a system integrating text understanding and concept learning under realistic conditions, we supply a natural language parser (Neuhaus & Hahn 1996) that is *inherently robust* and has various strategies to get nearly optimal results out of deficient, i.e., underspecified knowledge sources in terms of *partial, limited-depth parsing*. The price we pay for this approach is underspecification and uncertainty associated with the knowledge we extract from texts. To cope with these problems, we build on expressively rich knowledge representation models of the underlying domain (Hahn, Klenner, & Schnattinger 1996). Accordingly, we provide a start-up core ontology (such as the Penman Upper Model (Bateman *et al.* 1990)) in the format of terminological assertions. The task of the module we describe in this paper is then to position *new* knowledge items which occur in a text in that concept hierarchy and to link them with valid conceptual roles, role fillers and role filler constraints; hence, *deep* knowledge discovery.

Concept hypotheses reflecting different conceptual readings for new knowledge items emerge and are updated on the basis of two types of evidence. First, we consider the type of *linguistic* construction in which an unknown lexical item occurs in a text (since we assume that the context and type of grammatical construction forces a particular interpretation on the unknown item); second, *conceptual* criteria are accounted for which reflect structural patterns of consistency, mutual justification, analogy, etc. of concept hypotheses with concept descriptions already available in the domain knowledge base. Both kinds of evidence are represented by a set of *quality labels.* The general concept learning problem can then be viewed as a *quality-based decision task* which is decomposed into three constituent parts: the continuous generation of quality labels for single concept hypotheses (reflecting the *reasons* for their formation and their significance in the light of other hypotheses), the estimation of the overall *credibility* of single concept hypotheses (taking the available set of quality labels for each hypothesis into account), and the computation of a *preference order* for the entire set of competing hypotheses (based on these accumulated quality judgments) to *select* the most plausible ones. These phases directly correspond to the major steps underlying KDD procedures (Fayyad, Piatetsky-Shapiro, & Smyth 1996), *viz.* data mining, data interpretation, and data cleaning, respectively.

A Scenario for Deep Knowledge Discovery

In order to illustrate our problem, consider the following knowledge discovery scenario. Suppose, your knowledge of the information technology domain tells you that *Aquarius* is a company. In addition, you incidentally know that *ASI-168* is a computer system manufactured by Aquarius. By convention, you know absolutely nothing about *Megaline.* Imagine, one day your favorite computer magazine features an article starting with *"The Megaline unit by Aquarius ..".* Has your knowledge increased? If so, what did you learn already from just this phrase?

cus here on the issues of learning accuracy and the learning rate. Due to the given learning environment, the measures we apply deviate from those commonly used in the machine learning community. In concept learning algorithms like IBL (Aha, Kibler, & Albert 1991) there is no hierarchy of concepts. Hence, any prediction of the class membership of a new instance is either true or false. However, as such hierarchies naturally emerge in terminological frameworks, a prediction can be more or less precise, i.e., it may approximate the goal concept at different levels of specificity. This is captured by our measure of *learning accuracy* which takes into account the conceptual distance of a hypothesis to the goal concept of an instance rather than simply relating the number of correct and false predictions, as in IBL.

In our approach, learning is achieved by the refinement of *multiple* hypotheses about the class membership of an instance. Thus, the measure of *learning rate* we propose is concerned with the reduction of hypotheses as more and more *information* becomes available about one particular new instance. In contrast, IBL-style algorithms consider only one concept hypothesis per learning cycle and their notion of *learning rate* relates to the increase of correct predictions as more and more *instances* are being processed.

We considered a total of 101 texts taken from a corpus of information technology magazines. For each of them 5 to 15 learning steps were considered. A *learning step* is operationalized by the representation structure that results from the semantic interpretation of an utterance which contains the unknown lexical item. Since the unknown item is usually referred to several times in a text, several learning steps result. For instance, the learning steps associated with our scenario are given by: MEGALINE INST-OF UNIT and MEGALINE PRODUCT-OF AQUARIUS.

Learning Accuracy

In a first series of experiments, we investigated the *learning accuracy* of the system, i.e., the degree to which the system correctly predicts the concept class which subsumes the target concept under consideration. Learning accuracy for a single lexical item (LA) is here defined as (n being the number of concept hypotheses for that item):

$$LA := \sum_{i \in \{1 \ldots n\}} \frac{LA_i}{n} \text{ with } LA_i := \begin{cases} \frac{CP_i}{SP_i} & \text{if } FP_i = 0 \\ \frac{CP_i}{FP_i + DP_i} & \text{else} \end{cases}$$

SP_i specifies the length of the *shortest path* (in terms of the number of nodes traversed) from the TOP node of the concept hierarchy to the maximally specific concept subsuming the instance to be learned in hypothesis i; CP_i specifies the length of the path from the TOP node to that concept node in hypothesis i which is *common* both for the shortest path (as defined above) and the actual path to the predicted concept (whether correct or not); FP_i specifies the length of the path from the TOP node to the predicted (in this case *false*) concept and DP_i denotes the node *distance* between the predicted node and the most specific concept correctly subsuming the target in hypothesis i, respectively. As an example, consider Fig. 1. If we assume MEGALINE to stand for a COMPUTER, then hypothesizing HARDWARE (correct, but too general) receives an LA = .75 (note that OBJECT ISA TOP), while hypothesizing PRINTER (incorrect, but still not entirely wrong) receives an LA = .6.

Fig. 2 depicts the learning accuracy curve for the entire data set. It starts at LA values in the interval between 48% to 54% for **LA –**, **LA TH** and **LA CB** in the first learning step. **LA CB** gives the accuracy rate for the full qualification calculus including threshold and credibility criteria, **LA TH** considers linguistic criteria only, while **LA –** depicts the accuracy values without incorporating quality criteria at all, though terminological reasoning is still employed. In the final step, LA rises up to 79%, 83% and 87% for **LA –**, **LA TH** and **LA CB**, respectively. Hence, the pure terminological reasoning machinery which does not incorporate the qualification calculus always achieves an inferior level of learning accuracy than the learner equipped with it.

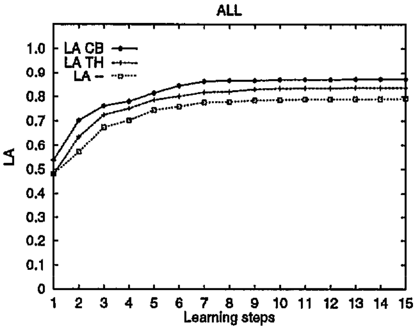

Figure 2: Learning Accuracy

Learning Rate

The learning accuracy focuses on the predictive power and validity of the learning procedure. By considering the *learning rate (LR)*, we supply data from the stepwise reduction of alternatives in the learning process. Fig. 3 depicts the mean number of transitively included concepts for all considered hypothesis spaces per learning step (each concept hypothesis relates to a concept which transitively subsumes various subconcepts; hence, pruning of concept subhierarchies reduces the number of concepts being considered as hypotheses). Note that the most general concept hypothesis in our example denotes OBJECT which currently includes 196 concepts. In general, we observed a strong negative slope of the curve for the learning rate. After the first step, slightly less than 50% of the included concepts are pruned (with 93, 94 and 97 remaining concepts for **LR CB**, **LR TH** and **LR –**, respectively). Summarizing this evaluation experiment, the system yields competitive accuracy rates (a mean of 87%), while at the same time exhibiting significant and valid reductions of the predicted concepts.

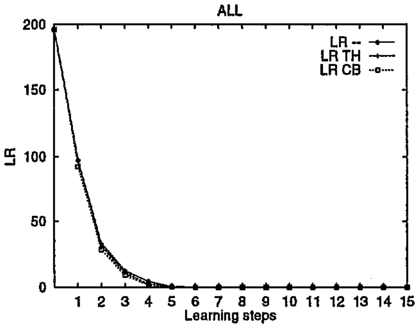

Figure 3: Learning Rate

Related Work

The issue of text analysis is only rarely dealt with in the KDD community. The reason for this should be fairly obvious. Unlike pre-structured data repositories (e.g., schemata and relations in database systems), data mining in textual sources requires to determine content-based formal structures in text strings prior to putting KDD procedures to work. Similar to approaches in the field of information retrieval (Dumais 1990), statistical methods of text structuralization are favored in the KDD framework (Feldman & Dagan 1995). While this leads to the determination of *associative relations* between lexical items, it does not allow the identification and relation of particular *facts* and *assertions* about or even *evaluations* to particular concepts. If this kind of *deep* knowledge is to be discovered, sophisticated natural language processing methodologies must come into play.

Our approach bears a close relationship to the work of, e.g., (Rau, Jacobs, & Zernik 1989) and (Hastings 1996), who aim at the automated learning of word meanings from the textual context using a knowledge-intensive approach. But 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. In the SCISOR system (Rau, Jacobs, & Zernik 1989), e.g., the selection of hypotheses depends only on an ongoing discrimination process based on the availability of linguistic and conceptual clues, but does not incorporate a dedicated inferencing scheme for reasoned hypothesis selection. The difference in learning performance – in the light of our evaluation study discussed in the previous section, at least – amounts to 8%, considering the difference between **LA –** (plain terminological reasoning) and **LA CB** values (terminological metareasoning based on the qualification calculus). Acquiring knowledge from real-world textual input usually provides the learner with only sparse, highly fragmentary clues, such that multiple concept hypotheses are likely to be derived from that input. So we stress the need for a hypothesis generation and evaluation component as an integral part of large-scale real-world text understanders operating in tandem with knowledge discovery devices.

This requirement also distinguishes our approach from the currently active field of information extraction (IE) (e.g., (Appelt *et al.* 1993)). The IE task is defined in terms of a *fixed* set of *a priori* templates which have to be instantiated (i.e., filled with factual knowledge items) in the course of text analysis. In contradistinction to our approach, no *new* templates have to be created.

Conclusion

We have introduced a new quality-based knowledge discovery methodology the constituent parts of which can be equated with the major steps underlying KDD procedures (Fayyad, Piatetsky-Shapiro, & Smyth 1996) — the generation of quality labels relates to the *data mining* (pattern extraction) phase, the estimation of the overall credibility of a single concept hypothesis refers to the *data interpretation* phase, while the selection of the most suitable concept hypothesis corresponds to the *data cleaning* mode.

Our approach is quite knowledge-intensive since knowledge discovery is fully integrated in the text understanding mode. No specialized learning algorithm is needed, since concept formation is turned into an inferencing task carried out by the classifier of a terminological reasoning system. Quality labels can be chosen from any knowledge source that seems convenient, thus ensuring easy adaptability. These labels also achieve a high degree of pruning of the search space for hypotheses in very early phases of the learning cycle. This is of major importance for considering our approach a viable contribution to KDD methodology.

Acknowledgments. We would like to thank our colleagues in the CLIF group for fruitful discussions and instant support, in particular Joe Bush who polished the text as a native speaker. K. Schnattinger is supported by a grant from DFG (Ha 2097/3-1).

References

Aha, D.; Kibler, D.; and Albert, M. 1991. Instance-based learning algorithms. *Machine Learning* 6:37–66.

Appelt, D.; Hobbs, J.; Bear, J.; Israel, D.; and Tyson, M. 1993. FASTUS: A finite-state processor for information extraction from real-world text. In *IJCAI'93 – Proc. 13th Intl. Joint Conf. on Artificial Intelligence*, 1172–1178.

Bateman, J.; Kasper, R.; Moore, J.; and Whitney, R. 1990. A general organization of knowledge for natural language processing: The PENMAN upper model. Technical report, USC/ISI.

Dumais, S. 1990. Text information retrieval. In Helander, M., ed., *Handbook of Human-Computer Interaction*, 673–700. Amsterdam: North-Holland.

Fayyad, U.; Piatetsky-Shapiro, G.; and Smyth, P. 1996. From data mining to knowledge discovery in databases. *AI Magazine* 17(3):37–54.

Feldman, R., and Dagan, I. 1995. Knowledge discovery in textual databases (KDT). In *KDD'95 – Proc. 1st Intl. Conf. on Knowledge Discovery and Data Mining.*

Grishman, R., and Sundheim, B. 1996. Message understanding conference - 6: A brief history. In *COLING'96 – Proc. 16th Intl. Conf. on Computational Linguistics*, 466–471.

Hahn, U., and Schnattinger, K. 1997. A qualitative growth model for real-world text knowledge bases. In *RIAO'97 – Proc. 5th Conf. on Computer-Assisted Information Searching on the Internet.*

Hahn, U.; Klenner, M.; and Schnattinger, K. 1996. Learning from texts: A terminological metareasoning perspective. In Wermter, S.; Riloff, E.; and Scheler, G., eds., *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, 453–468. Berlin: Springer.

Hahn, U.; Schnattinger, K.; and Romacker, M. 1996. Automatic knowledge acquisition from medical texts. In *AMIA'96 – Proc. AMIA Annual Fall Symposium. Beyond the Superhighway: Exploiting the Internet with Medical Informatics*, 383–387.

Hastings, P. 1996. Implications of an automatic lexical acquisition system. In Wermter, S.; Riloff, E.; and Scheler, G., eds., *Connectionist, Statistical and Symbolic Approaches to Learning in Natural Language Processing*, 261–274. Berlin: Springer.

Neuhaus, P., and Hahn, U. 1996. Trading off completeness for efficiency: The PARSETALK performance grammar approach to real-world text parsing. In *FLAIRS'96 – Proc. 9th Florida Artificial Intelligence Research Symposium*, 60–65.

Rau, L.; Jacobs, P.; and Zernik, U. 1989. Information extraction and text summarization using linguistic knowledge acquisition. *Information Processing & Management* 25(4):419–428.

Schnattinger, K., and Hahn, U. 1996. A terminological qualification calculus for preferential reasoning under uncertainty. In *KI'96 – Proc. 20th Annual German Conf. on Artificial Intelligence*, 349–362. Berlin: Springer.