Automatic Concept Acquisition from Real-World Texts

Udo Hahn, Manfred Klenner & Klemens Schnattinger
Freiburg University, Computational Linguistics Group
Europaplatz 1, D-79085 Freiburg, Germany
{hahn,klenner,schnattinger}@coling.uni-freiburg.de

Abstract

We introduce a concept learning methodology for text understanding systems that is based on terminological knowledge representation and reasoning. Quality-based metareasoning techniques allow for an incremental evaluation and selection of concept hypotheses. This methodology is particularly aimed at real-world text understanding environments where lexical/conceptual resources cannot be completely specified prior to text analysis and, as a consequence of partial understanding, competing concept hypotheses with different levels of credibility have to be managed.

The Concept Acquisition Problem

The work reported in this paper is part of a large-scale project aiming at the development of a German-language text knowledge extraction system for two real-world domains — information technology product reviews (approx. 100 documents with 10^5 words) and medical findings reports (approx. 120,000 documents with 10^7 words). The concept learning problem we face is two-fold. In the IT domain lexical *growth* occurs at dramatic rates – new products, technologies, companies and people continuously enter the scene such that any attempt at keeping track with these lexical innovations by hand-coding is clearly precluded. On the other hand, the medical domain is more or less lexically stable but the sheer *size* of its sublanguage (conservative estimates range about 10^6 concepts) also cannot reasonably be coded by humans in advance. Unlike Lewis (1991) we advocate a symbolically rooted learning approach in order to break the *lexical acquisition bottleneck* (Hahn, Schnattinger, & Klenner 1995; Hastings 1995). The methodology we propose is not limited to the information extraction task but can also be applied to any application framework where large volumes of texts have to be processed with incomplete lexicons and domain knowledge bases, and where the underspecification of these resources, in principle, cannot reasonably be overcome by human intervention. These conditions, obviously, apply to the whole range of information access problems, such as text filtering, routing, and classification.

We consider the problem of natural language based concept learning from a new methodological perspective, viz. one based on *metareasoning* about statements expressed in a terminological representation language. These statements either reflect *structural linguistic* properties of phrasal patterns or discourse contexts in which unknown words occur (assuming that the type of grammatical construction exercises a particular interpretative force on the unknown lexical item), or they reflect *conceptual* properties of particular concept hypotheses as they are generated and continuously refined by the ongoing text understanding process (e.g., consistency relative to already given knowledge, independent justification from several sources). Each of these grammatical, discourse or conceptual indicators is assigned a particular "quality" label.

We view the problem of choosing from among several alternatives as a *quality-based reasoning task* which can be decomposed into three constituent parts: the continuous generation of quality labels for single hypotheses (reflecting the *reasons* for their formation and their significance in the light of other hypotheses), the assessment of the *credibility* of single 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, which is based on these accumulated quality judgments.

Learning System Architecture

The concept learning methodology we propose is heavily based on the representation and reasoning facilities of terminological knowledge representation languages (Woods & Schmolze 1992). The architecture of the learner (cf. Fig. 1) provides for: *(1) quality-based asser-*

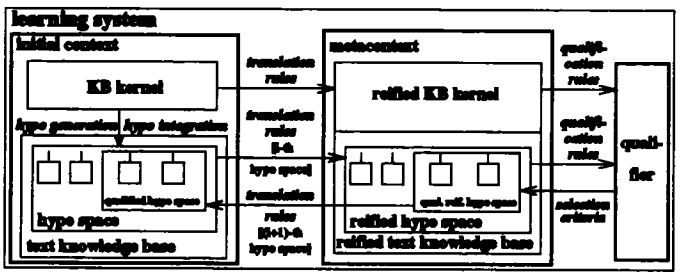

Figure 1: Architecture for Concept Acquisition

tions about propositions in a terminological language (these metastatements capture the ascription of belief to these propositions, the reasons why they came into existence, the support/weakening they receive from other propositions); *(2) metareasoning* in a terminological knowledge base about properties and relations that certain propositions may have. A learning system realizing these principles (Hahn, Schnattinger, & Klenner 1995; Schnattinger, Hahn, & Klenner 1995) has been fully implemented in LOOM (MacGregor 1994).

The formal foundations for metaknowledge and metareasoning are based on McCarthy's context model (McCarthy 1993). We here distinguish two types of contexts, viz. the initial context and the metacontext. The *initial context* contains the original terminological knowledge base (KB kernel) and the text knowledge base reflecting the knowledge acquired from the underlying text by the text parser. Knowledge in the initial context is represented without any explicit qualifications, attachments, provisos, etc. Note that in the course of text understanding – due to the working of several *hypothesis generation* rules – a hypothesis space is created which contains alternative subspaces for each concept to be learned, each one holding a different or a further specialized concept hypothesis. Various truth-preserving *translation rules* map the descriptions of the initial context to the *metacontext* which consists of the reified knowledge (Friedman & Wand 1984) of the initial context. Among the reified structures in the metacontext there is a subcontext embedded, the *reified hypothesis space*, the elements of which carry several quality annotations, e.g., reasons to believe a proposition, indications of consistency, strength of support. These quality labels are the result of the mechanisms underlying hypothesis evaluation and subsequent hypothesis selection, thus reflecting the operation of several second-order qualification rules in the *qualifier* (*quali*ty-based class*ifier*). The derived labels are the basis for the selection of those representation structures that are assigned a high degree of credibility – only those *qualified hypotheses* will be remapped to the hypothesis space of the initial context by way of (inverse) translation rules. Thus, we come full circle.

Hypothesis Generation

In the architecture we propose, text parsing and concept acquisition from texts are tightly coupled. For instance, whenever two nominals or a nominal and a verb are supposed to be syntactically related the semantic interpreter simultaneously checks whether a concept denoted by one of these objects can fill a role of the other one or is a subclass or an instance of this concept. If one of the lexical items involved is unknown (the so-called *target*), this interpretation mode generates initial concept hypotheses about the class membership of the unknown item relative to the already known item (the so-called *base* concept). We start from the assumption that the target concept fills (exactly) one of the n roles of the base concept. As further information is usually lacking, it cannot be decided on the correct role yet. Hence, n alternative hypothesis spaces are opened and the target concept is assigned as a potential filler of the i-th role in its corresponding hypothesis space. Therefore, the classifier is able to derive a suitable concept hypothesis by specializing the target concept (initial status "unknown") according to the value restriction of the base concept's i-th role.

Quality-based Reasoning

Generation of Quality Labels

In order to further constrain the hypotheses derived during the hypothesis generation phase, we supply second-order qualification rules (verbalized below). These are used to reason about the properties of (first-order) terminological descriptions available from the parser or the hypothesis generator. Their application yields quality labels for various concept hypotheses:

i: The *very positive* quality label MULTIPLY-DEDUCED is generated whenever the same role filler has been multiply derived in different hypothesis spaces.

ii: The conceptual proximity of role fillers of a (non-ACTION) concept, which share a common concept class leads to the *positive* quality label SUPPORTED.

iii: The inherent symmetry between two instances mutually related via two quasi-inverse relations (as "inverted" role fillers of each other) is expressed by the *positive* quality label CROSS-SUPPORTED.

iv: The negative assessment for any attempt to fill the same mandatory case role of an ACTION concept more than once by different role fillers is expressed by the *negative* quality label ADDITIONAL-ROLE-FILLER.

Assessment of Quality Labels

During each learning step several qualification rules may fire and thus generate various quality labels. In order to select the most credible hypotheses from each cycle, we take the direction (positive/negative) and the individual 'strength' of each label into account:

Threshold Criterion. Select all hypothesis spaces that do not contain any quality label INCONSISTENT-HYPO. Additionally, the maximum number of verb interpretations is required. If several of these hypothesis spaces exist, the one(s) with the least number of ADDITIONAL-ROLE-FILLER labels is (are) chosen.

Only those hypotheses that reach the credibility threshold after each quality assessment cycle are transferred from the metacontext back to the initial context.

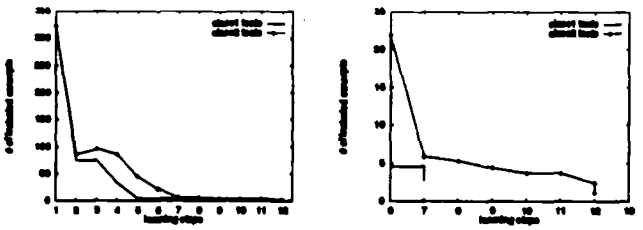

Figure 2: Included Concepts per Learning Step

Preference Ranking

After several cycles of hypothesis space pruning, a final ranking of those concept hypotheses is produced that have repeatedly passed the Threshold Criterion:

Ranked Prediction Criterion. Select all hypothesis spaces with the maximum number of MULTIPLY-DEDUCED labels. If there are more than one, rank these spaces in decreasing order according to the number of SUPPORTED labels they contain. If there are more than one with the same maximum number of MULTIPLY-DEDUCED and SUPPORTED labels, rank these spaces in decreasing order according to the number of CROSS-SUPPORTED labels.

Empirical Evaluation

A first round of experiments was run to evaluate the effectiveness of the quality-based concept learner; 18 texts which contained 6344 tokens were considered.

The mean *learning accuracy* of the system, i.e., the degree to which the system correctly predicts the concept class which subsumes the target concept[1] amounts to 91%. In 14 of the 18 cases the Threshold Criterion reduces the number of concept hypotheses to *1*. In 10 of the 18 cases the predicted concept matched the target completely, in 7 cases only "near misses" (path length penalty of *1*) occurred, while only a single prediction was considered inadequate.

Learning accuracy focuses on the final result of the learning process, while the *learning rate* addresses the step-wise development of the learning process. Fig. 2 contains the mean number of transitively included concepts for all considered hypothesis spaces per learning step (each hypothesis denotes a concept which transitively subsumes various subconcepts). We grouped the 18 texts in two classes in order to normalize the number of propositions they contained. The left side in Fig. 2 gives the overall mean learning rate for both classes. The right one (zooming at the learning steps 6 to 12) focuses on the reduction achieved by the Ranked Prediction Criterion, yielding a final drop of approximately 50% for each class (class 1: from 4.9 to 2.6 concepts, class 2: from 3.7 to 1.9 concepts).

[1]The accuracy measure is based on shortest and common path length criteria for the concept nodes involved.

Related Work

Our approach bears a close relationship to the work of Mooney (1987), Hastings (1995), Gomez (1995), and Soderland *et al.* (1995), who aim at the automated learning of *word meanings* (concepts) from textual context. However, the need to cope with *several competing* concept hypotheses is not an issue in these studies. As almost any parser available for realistic text understanding tasks currently suffers from the incapability to generate complete parses (i.e., to produce a full understanding of the underlying texts), usually more than one concept hypothesis (indicating different topics of the text) is likely to be derived from a given NL input. Therefore, we stress the need for a hypothesis generation and evaluation component as an integral part of any robust NLU system that learns in tandem with such coverage-restricted devices. Our contribution thus constitutes a part of the developing methodological infrastructure for effective information access.

Acknowledgments. This work was supported by a grant from DFG (account Ha2097/2-1). We gratefully acknowledge the provision of the LOOM system from USC/ISI.

References

Friedman, D., and Wand, M. 1984. Reification: Reflection without metaphysics. In *Proc. 1984 ACM Symposium on Lisp and Functional Programming*, 348–355.

Gomez, F. 1995. Acquiring knowledge about the habitats of animals from encyclopedic texts. In *Proc. KAW'95*.

Hahn, U.; Schnattinger, K.; and Klenner, M. 1995. Learning from texts: A terminological meta-reasoning perspective. In *IJCAI-95 Workshop "New Approaches to Learning for Natural Language Processing"*, 143–150.

Hastings, P. M. 1995. Implications of an automatic lexical acquisition system. In *Working Notes from the IJCAI-95 Workshop on "New Approaches to Learning for Natural Language Processing"*, 170–175.

Lewis, D. 1991. Learning in intelligent information retrieval. In *ML'91: Proc. 8th Intl. Workshop*, 235–239.

MacGregor, R. 1994. A description classifier for the predicate calculus. In *Proc. AAAI'94*, 213–220.

McCarthy, J. 1993. Notes on formalizing context. In *Proc. IJCAI'93*, 555–560.

Mooney, R. 1987. Integrated learning of words and their underlying concepts. In *Proc. 9th Annual Conf. of the Cognitive Science Society*, 974–978.

Schnattinger, K.; Hahn, U.; and Klenner, M. 1995. Terminological meta-reasoning by reification and multiple contexts. In *Proc. EPIA'95*, 1–16.

Soderland, S.; Fisher, D.; Aseltine, J.; and Lehnert, W. 1995. CRYSTAL: Inducing a conceptual dictionary. In *Proc. IJCAI'95*, 1314–1319.

Woods, W., and Schmolze, J. 1992. The KL-ONE family. *Computers & Mathematics with Applic.* 23(2-5):133–177.