Towards Text Knowledge Engineering

Udo Hahn & Klemens Schnattinger

LIF Computational Linguistics Group
Text Knowledge Engineering Lab
Freiburg University
Werthmannplatz 1, D-79085 Freiburg, Germany
http://www.coling.uni-freiburg.de

Abstract

We introduce a methodology for automating the maintenance of domain-specific taxonomies based on natural language text understanding. A given ontology is incrementally updated as new concepts are acquired from real-world texts. The acquisition process is centered around the linguistic and conceptual "quality" of various forms of evidence underlying the generation and refinement of concept hypotheses. On the basis of the quality of evidence, concept hypotheses are ranked according to credibility and the most credible ones are selected for assimilation into the domain knowledge base.

Introduction

Knowledge engineering is still an expert task. Though a variety of architectures have been proposed to date (Buchanan *et al.* 1983), the dominating paradigm for the process of eliciting and maintaining domain ontologies continues to focus on the interactive knowledge transfer from humans to machines (Stefik 1995). Some experimental activities tried to make use of machine learning methods in order to induce knowledge from structured data repositories (Morik *et al.* 1993), but even fewer efforts were targeted at unstructured natural language texts as a source for automating knowledge engineering processes (Gomez & Segami 1990).

We propose here such a text-based knowledge acquisition methodology, in which domain knowledge bases are continuously enhanced as a by-product of text understanding processes – hence, *text knowledge engineering*. New concepts are acquired taking two sources of evidence into account: background knowledge of the domain the texts are about, and linguistic patterns 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 assesses the strength of the interpretative force that can be attributed to the grammatical construction in which new lexical items occur. Our model makes explicit the kind of qualitative reasoning that is behind such a learning process.

This is, then, a knowledge-intensive model of concept acquisition, tightly integrated with the non-learning mode of text understanding. The "plain" text understanding mode can be considered the instantiation and continuous role filling of *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.

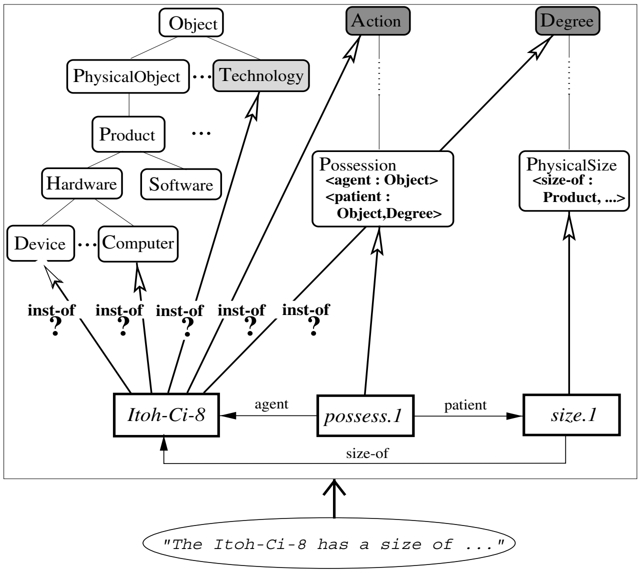

Figure 1: A Sample Scenario

For illustration purposes, consider the following scenario as depicted in Fig. 1. Suppose, your knowledge of the information technology domain tells you nothing about an *Itoh-Ci-8*. Imagine, one day your favorite technology magazine features an article starting with *"The Itoh-Ci-8 has a size of ..."*. Has your knowledge increased? If so, what did you learn from just this phrase?

The text knowledge acquisition process starts upon the reading of the unknown lexical item *"Itoh-Ci-8"*. In this initial step, the corresponding hypothesis space incorporates all the top level concepts available in the ontology for the new lexical item *"Itoh-Ci-8"*. So, the concept ITOH-CI-8 may be a kind of an OBJECT, an ACTION, a DEGREE, etc. As a consequence of processing the noun phrase *"The Itoh-Ci-8"* as the grammatical subject of the verb *"has"*, ITOH-CI-8 is related via the AGENT role to the ACTION

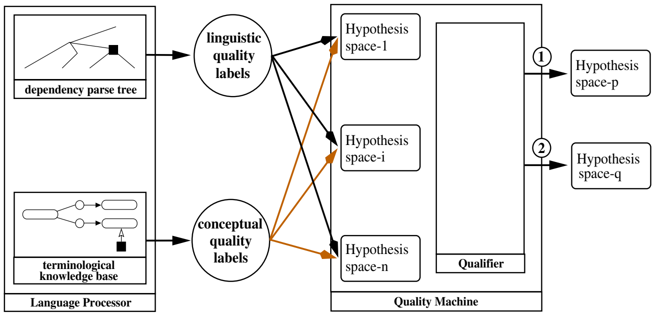

Figure 2: Architecture for Text Knowledge Engineering

concept Possession, the concept denoted by *"has"* (lexical ambiguities, e.g., for the verb *"has"*, lead to the creation of alternative hypothesis spaces). Since Possession requires its agent to be an Object, Action and Degree are no longer valid concept hypotheses. Their cancellation (cf. the darkly shaded boxes) yields significant reduction of the huge initial hypothesis space. The learner then aggressively specializes the remaining single hypothesis to the immediate subordinates of Object, *viz.* PhysicalObject and Technology, in order to test more restricted hypotheses which – according to more specific constraints – are easier falsifiable.

In addition, the *linguistic* constraints for the verb *"has"* indicate that the grammatical direct object relation is to be interpreted in terms of a conceptual patient role. Accordingly, the phrase " *. . . has a size of . . .* " is processed such that $size.1$ is the patient of the Possession relationship. Also, agent and patient are both restricted by specific *conceptual* constraints. These come into play in the subsequent semantic interpretation step, where possible conceptual relations between the agent and patient are tried.

A straightforward translation of the basic conceptual relations contained in the utterance above yields the following terminological expressions:

(P1) $size.1$: PhysicalSize
(P2) $size.1$ size-of $Itoh\text{-}Ci$-8

Assertion P1 indicates that $size.1$ is an instance of the concept class PhysicalSize and P2 relates $size.1$ and $Itoh\text{-}Ci$-8 via the binary relation size-of.

Given the conceptual roles attached to PhysicalSize, the system recognizes that all specializations of Product can be related to the concept PhysicalSize (via the role size-of), while for Technology no such relation can be established. So, we prefer the conceptual reading of Itoh-Ci-8 as a kind of a Product over the Technology hypothesis (cf. the grey-shaded boxes).

A Model of Text Knowledge Engineering

The methodology and corresponding system architecture for text knowledge elicitation is summarized in Fig. 2. It depicts how linguistic and conceptual evidence are generated and combined to continuously discriminate and refine the set of concept hypotheses (the unknown item yet to be learned is characterized by the black square).

The *language processor* (for an overview, cf. Hahn, Schacht, & Bröker (1994)) yields structural dependency information from the grammatical constructions in which an unknown lexical item occurs in terms of the corresponding *parse tree*. The kinds of syntactic constructions (e.g., genitive, apposition, comparative), in which unknown lexical items appear, are recorded and assessed later on relative to the credit they lend to a particular hypothesis. The conceptual interpretation of parse trees involving unknown lexical items in the *terminological knowledge base* (cf. Woods & Schmolze (1992) for a survey of terminological, Kl-one-style knowledge representation) is used to derive *concept hypotheses*, which are further enriched by conceptual annotations reflecting structural patterns of consistency, mutual justification, analogy, etc. This kind of initial evidence, in particular its predictive "goodness" for the learning task, is 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 (cf. Schnattinger & Hahn (1996) for details of this metareasoning process).

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 grammatical construction exercises a particular interpretative force on the unknown item and, at the same time, yields a particular level of credibility for the hypotheses being derived therefrom. As a concrete example of a high-quality label, consider the case of Apposition. This label is generated for constructions such as *".. the printer @A@ .."*, with "@..@" denoting the unknown item. The apposition almost unequivocally determines "@A@" (considered as a potential noun)[1] to denote a Printer. This assumption is justified independent of further conceptual conditions, simply due to the nature of the linguistic construction being used. Still of good quality but already less constraining are occurrences of the unknown item in a CaseFrame construction as illustrated by *".. @B@ has a size of .."*. In this example, case frame specifications of the verb *"has"* that relate to its agent role carry over to *"@B@"*. Given its final

[1] Such a part-of-speech hypothesis can be derived from the inventory of valence and word order specifications underlying the dependency grammar model we use (Hahn, Schacht, & Bröker 1994).

semantic interpretation, "@B@" may be anything that has a size. Considering an utterance like *"The Itoh-Ci-8 has a size of .."*, we may hypothesize that, in an information technology domain, at least, the concept ITOH-CI-8 can tentatively be considered a PRODUCT (which IS-A PHYSICALOBJECT and, hence, always provides a HAS-SIZE relation).

Depending on the type of the syntactic construction in which the unknown lexical item occurs, different hypothesis generation rules may fire. As in a sample phrase such as *"The switch of the Itoh-Ci-8 .."*, genitive noun phrases place only a few constraints on the item to be acquired. In the following, let *target* be the unknown item *("Itoh-Ci-8")* and *base* be the known item *("switch")*, whose conceptual relation to the target is constrained by the syntactic relation in which their lexical counterparts co-occur. The main constraint for genitives says that the target concept fills (exactly) one of the *n* roles of the base concept. Since the correct role cannot yet be decided upon, *n* alternative hypotheses have to be posited (unless additional constraints apply), and the target concept has to be assigned as a filler of the *i*-th role of base in the corresponding *i*-th hypothesis space. As a consequence, the classifier is able to derive a suitable concept hypothesis by specializing the target concept (initially TOP, by default) according to the value restriction of the base concept's *i*-th role. Additionally, this rule assigns a syntactic quality label to each *i-th* hypothesis indicating the type of syntactic construction in which target and base co-occur.

After the processing of *"The Itoh-Ci-8 has a size of .."*, the target ITOH-CI-8 is already predicted as a PRODUCT. Prior to continuing with the phrase *"The switch of the Itoh-Ci-8 .."*, consider some fragments of the conceptual representation for SWITCHes:

$$\begin{array}{ll}
\text{(P3)} & \text{SWITCH-OF} \doteq {}_{\text{SWITCH}}|\text{PART-OF}|_{\text{HARDWARE}} \\
\text{(P4)} & \text{SWITCH} \doteq \\
& \forall \text{HAS-PRICE.PRICE} \sqcap \\
& \forall \text{HAS-WEIGHT.WEIGHT} \sqcap \\
& \forall \text{SWITCH-OF.} \begin{pmatrix} \text{OUTPUTDEV} \sqcup \text{INPUTDEV} \sqcup \\ \text{STORAGEDEV} \sqcup \text{COMPUTER} \end{pmatrix}
\end{array}$$

The relation SWITCH-OF is defined by P3 as the set of all PART-OF relations which have their domain restricted to SWITCH and their range restricted to HARDWARE. In addition, (P4) reads as "all fillers of HAS-PRICE, HAS-WEIGHT, and SWITCH-OF roles must be concepts subsumed by PRICE, WEIGHT, and the disjunction of (OUTPUTDEV ⊔ INPUTDEV ⊔ STORAGEDEV ⊔ COMPUTER), respectively". So, three roles have to be considered for relating the target ITOH-CI-8, as a tentative PRODUCT, to the base concept SWITCH. Two of them, HAS-PRICE and HAS-WEIGHT, are ruled out due to the violation of a simple integrity constraint (PRODUCT does not denote a unit of measure). Therefore, only the role SWITCH-OF must be considered. Due to the definition of SWITCH-OF (cf. P3), ITOH-CI-8 is immediately specialized to HARDWARE by the classifier. Since the classifier aggressively pushes the hypothesizing to be maximally specific, four distinct hypotheses are immediately created due to the specific range restrictions of the role SWITCH-OF expressed in (P4), the definition of the concept SWITCH, *viz.* OUTPUTDEV, INPUTDEV, STORAGEDEV and COMPUTER, and they are managed in four distinct hypothesis spaces, h_1, h_2, h_3 and h_4, respectively. We sketch their contents roughly in the following concept descriptions (note that for *Itoh-Ci-8* we also include parts of the implicit IS-A hierarchy):

$(\textit{Itoh-Ci-8} : \text{OUTPUTDEV})_{h_1}, (\textit{Itoh-Ci-8} : \text{DEVICE})_{h_1}, ..,$
$(\textit{switch.1}\ \text{SWITCH-OF}\ \textit{Itoh-Ci-8})_{h_1}$

$(\textit{Itoh-Ci-8} : \text{INPUTDEV})_{h_2}, (\textit{Itoh-Ci-8} : \text{DEVICE})_{h_2}, ..,$
$(\textit{switch.1}\ \text{SWITCH-OF}\ \textit{Itoh-Ci-8})_{h_2}$

$(\textit{Itoh-Ci-8} : \text{STORAGEDEV})_{h_3}, (\textit{Itoh-Ci-8} : \text{DEVICE})_{h_3}, ..,$
$(\textit{switch.1}\ \text{SWITCH-OF}\ \textit{Itoh-Ci-8})_{h_3}$

$(\textit{Itoh-Ci-8} : \text{COMPUTER})_{h_4}, (\textit{Itoh-Ci-8} : \text{HARDWARE})_{h_4}, ..,$
$(\textit{switch.1}\ \text{SWITCH-OF}\ \textit{Itoh-Ci-8})_{h_4}$

Conceptual Quality Labels

Conceptual quality labels result from comparing the representation structures of a concept hypothesis with those of alternative concept hypotheses or already existing representation structures in the underlying domain knowledge base from the viewpoint of structural similarity, compatibility, etc. The closer the match, the more credit is lent to a hypothesis. For instance, a very positive conceptual quality label such as M-DEDUCED is assigned to multiple derivations of the same concept hypothesis in different hypothesis (sub)spaces. Positive labels are also assigned to terminological expressions which share structural similarities, though they are not identical. The label C-SUPPORTED, e.g., is assigned to any hypothesized relation $R1$ between two instances when another relation, $R2$, already exists in the knowledge base involving the same two instances, but where the role fillers occur in "inverted" order (note that $R1$ and $R2$ need not necessarily be semantically inverse relations such as with *"buy"* and *"sell"*). This rule of *"cross"* support captures the inherent symmetry between concepts related via quasi-inverse conceptual relations.

Quality annotations of the conceptual status of concept hypotheses are derived from qualification rules. For instance, the rule for the label M-DEDUCED applies to the case where the same assertion is deduced in at least two different hypothesis spaces (cf. h_1 and h_2 in the expression below). That assertion, e.g., $(\textit{Itoh-Ci-8} : \text{DEVICE})_{h_1}$ in the example below, is then annotated by a high-quality label. In technical terms, an instance of the quality label M-DEDUCED is created (for a formal specification of several qualification rules, including the representation of and metareasoning with quality assertions, cf. Hahn, Klenner, & Schnattinger (1996)). Considering our example, for ITOH-CI-8 the concept hypotheses OUTPUTDEVice, INPUTDEVice and STORAGEDEVice were derived independently of each other in different hypothesis spaces. Hence, DEVICE, as their common superconcept, has been *multiply* derived by the classifier in each of these spaces, too. Accordingly, this hypothesis is assigned a high degree of confidence by issuing the conceptual quality label M-DEDUCED:

$$\begin{array}{l}
(\textit{Itoh-Ci-8} : \text{DEVICE})_{h_1} \ \wedge \ (\textit{Itoh-Ci-8} : \text{DEVICE})_{h_2} \\
\Longrightarrow \\
(\textit{Itoh-Ci-8} : \text{DEVICE})_{h_1} : \text{M-DEDUCED} \ \ldots \ \ldots
\end{array}$$

Quality-Based Classification

Whenever new evidence for or against a concept hypothesis is brought forth in a single learning step all concept hypotheses are reevaluated. First, weak or even untenable hypotheses are discarded. A quality-based selection among the remaining hypothesis spaces is grounded in *threshold levels* (later on referred to as **TH**). Their definition takes linguistic evidence into account. At the first threshold level, all hypothesis spaces with the maximum of Apposition labels are selected. If more than one hypothesis is left to be considered, only concept hypotheses with the maximum number of CaseFrame assignments are approved at the second threshold level. Those hypothesis spaces that have fulfilled these threshold criteria will then be classified relative to two different *credibility levels* (later on referred to as **CB**). The first level of credibility contains all hypothesis spaces which have the maximum of M-Deduced labels, while at the second level (again, with more than one hypothesis left to be considered) those are chosen which are assigned the maximum of C-Supported labels. A comprehensive terminological specification of the underlying qualification calculus is given by Schnattinger & Hahn (1996).

For an illustration, consider the first utterance, once again: *"The Itoh-Ci-8 has a size of .."*. An assignment of the syntactic quality label CaseFrame is triggered only in those hypothesis spaces where the unknown item is considered a PhysicalObject (cf. Table 3, learning step 1). The remaining hypotheses (cf. Table 3, learning step 2) cannot be annotated by CaseFrame, since the concepts they represent (e.g., MentalObject, Norm) have no property such as PhysicalSize. As a consequence, their hypothesis spaces are ruled out by the criterion set up at the second threshold level, and the still valid concept hypothesis PhysicalObject is further refined as Product. As far as the sample phrase *"The switch of the Itoh-Ci-8 .."* is concerned, four more specific hypothesis spaces are generated from the Product hypothesis, three of which stipulate a Device hypothesis. Since the conceptual quality label M-Deduced has been derived by the classifier, this result yields a ranking with these three Device hypotheses preferred over the one associated with Computer (cf. Table 3, learning step 3).

Evaluation

In this section, we present some data from an empirical evaluation of the text knowledge acquisition system. We start with a consideration of canonical performance measures (such as recall, precision, etc.) and then focus on the more pertinent 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 target 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.

	Phrase	**Semantic Interpretation**
1. 2.	The *Itoh-Ci-8 has* a *size* of ..	(possess.1, agent, Itoh-Ci-8) (possess.1, patient, size.1) ↦ (size.1,size-of,Itoh-Ci-8) ↦ (Itoh-Ci-8,has-size,size.1)
3.	The *switch* of the *Itoh-Ci-8* ..	(switch.1, switch-of, Itoh-Ci-8) ↦ (Itoh-Ci-8,has-switch,switch.1)
4.	The *housing* from the *Itoh-Ci-8* ..	(housing.1, case-of, Itoh-Ci-8) ↦ (Itoh-Ci-8,has-case,housing.1)
5.	*Itoh-Ci-8* with a *main memory* ..	(memory.1, memory-of, Itoh-Ci-8) ↦ (Itoh-Ci-8,has-memory,memory.1)
6.	*Itoh-Ci-8*'s *LED lines* ..	(LED-line.1, part-of, Itoh-Ci-8) ↦ (Itoh-Ci-8,has-part,LED-line.1)
7.	*Itoh-Ci-8*'s *toner supply* ..	(tonerSupply.1, part-of, Itoh-Ci-8) ↦ (Itoh-Ci-8,has-part,tonerSupply.1)
8.	*Paper cassette* of *the Itoh-Ci-8* ..	(paperSupply.1, part-of, Itoh-Ci-8) ↦ (Itoh-Ci-8,has-part,paperSupply.1)
9.	*Itoh-Ci-8* with a *resolution rate* ..	(resolution.1, rate-of, Itoh-Ci-8) ↦ (Itoh-Ci-8,has-rate,resolution.1)

Table 1: Interpretation Results of a Text Featuring *"Itoh-Ci-8"*

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 possible 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.

The knowledge base on which we performed our experiments contained 325 concept definitions and 447 conceptual relations. The upper-level concepts of that ontology were taken from Nirenburg & Raskin (1987). We considered a total of 101 texts (= **SizeofTestSet** below) randomly selected from a corpus of information technology magazines. For each of them, 5 to 15 learning steps were considered. A *learning step* captures the final result of all semantic interpretation processes being made at the level of hypothesis spaces after new textual input has been supplied in which the item to be learned occurs. In order to clarify the input data available for the learning system, cf. Table 1. It consists of nine single learning steps for the unknown item *"Itoh-Ci-8"* that occurred while processing the entire text. Each learning step is associated with a particular natural language phrase in which the unknown lexical item occurs and the corresponding semantic interpretation data.

Canonical Performance Measures

In a first series of experiments, we neglected the incrementality of the learner and evaluated our system in terms of its *bare off-line* performance. By this we mean its potential to determine the correct concept description at the end of each text analysis considering the outcome of the final learning step only. Following previous work on evaluation measures for learning systems (Hastings 1994), we distinguish here the following parameters:

- **Hypothesis** denotes the set of concept hypotheses derived by the system as the final result of the text understanding process for each target item;

	CAMILLE	–	TH	CB
Correct	*	100	100	99
OneCorrect	*	21	26	31
ConceptSum	*	446	360	255
RECALL := $\frac{\textbf{Correct}}{\textbf{SizeofTestSet}}$	44%	99%	99%	98%
PRECISION := $\frac{\textbf{Correct}}{\textbf{ConceptSum}}$	22%	22%	28%	39%
PARSIMONY := $\frac{\textbf{OneCorrect}}{\textbf{SizeofTestSet}}$	14%	21%	26%	31%

Table 2: Performance Measures

- **Correct** denotes the number of cases in the test set in which **Hypothesis** contains the correct concept description for the target item;
- **OneCorrect** denotes the number of cases in the test set in which **Hypothesis** is a singleton set which contains the correct concept description only;
- **ConceptSum** denotes the number of concepts generated for all of the target items considering the entire test set.

Measures were taken under three experimental conditions (cf. Table 2). In the second column (indicated by **–**), we considered the contribution of only the terminological reasoning component to the concept acquisition task, the third column contains the results of incorporating the (linguistic) threshold criteria (denoted by **TH**), while the fourth one incorporates (linguistic as well as conceptual) credibility criteria (designated by **CB**). The data indicate a surprisingly high recall rate. The slight drop for **CB** (98% relative to 99%) is due to an incidental selection fault during processing. The values for precision as well as those for parsimony are consistently in favor of the full qualification calculus (**CB**).

In an attempt to relate these results of the quality-based learner to a system close in spirit to our approach, we chose CAMILLE (Hastings 1994), considering versions 1.0, 1.2, 2.0, and 2.1, and the results reported for recall, precision, and parsimony in the assembly line and the terrorism domain (cf. Table 2, column one). Not surprisingly, the precision of our terminological reasoning component, the LOOM system (MacGregor 1994), is equal to CAMILLE's,[2] but our system outperforms CAMILLE significantly on the evaluation dimensions **TH** and **CB** with respect to any of the performance measures we considered. Unlike CAMILLE, our learner also consistently improves as more and more information becomes available for an unknown target item (cf. the following section).

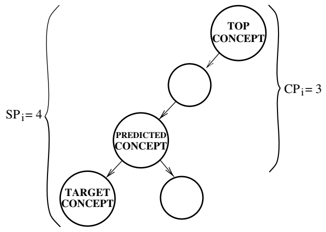

Figure 3: LA for an Underspecified Concept Hypothesis

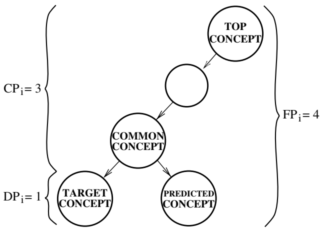

Figure 4: LA for a Slightly Incorrect Concept Hypothesis

Learning Accuracy

In a second series of experiments, we investigated the *learning accuracy* of the system, i.e., the degree to which it correctly predicts the concept class which subsumes the target concept to be learned. Learning accuracy (LA) is defined as (n being the number of concept hypotheses for the target):

$$LA := \sum_{i \in \{1 \ldots n\}} \frac{LA_i}{n} \quad \text{with}$$

$$LA_i := \begin{cases} \frac{CP_i}{SP_i} & \text{if } FP_i = 0 \\ \frac{CP_i}{FP_i + DP_i} & \text{else} \end{cases}$$

SP_i specifies the length of the *shortest path* (in terms of the number of nodes being 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 to 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 ($FP_i = 0$, if the prediction is correct), and DP_i denotes the node *distance* between the predicted false node and the most specific common concept (on the path from the TOP node to the predicted false node) still correctly subsuming the target in hypothesis i. Sample configurations for concrete LA values involving these parameters are depicted in Fig. 3, which illustrates a correct, yet too general prediction with $LA_i = .75$, while Fig. 4 contains a false prediction with $LA_i = .6$. Though the measure is sensitive to the depth of concept graphs, it produced adequate results in the technology domain we considered. As the graphs in knowledge bases for "natural" domains typically have an almost canonical depth that ranges between seven to ten nodes from the most general to the most specific concept, our measure seems to generalize to other domains as well.[3]

Given the LA measure from above, Table 3 and Table 4 illustrate how alternative concept hypotheses for ITOH-CI-8 develop in accuracy from one step to the other. The numbers in brackets in the column **Concept Hypotheses** indicate

[2] Hastings (1994, page 71) mentions that "... classifier systems [like LOOM] provide a very similar inference mechanism to CAMILLE's." This statement is backed up by our precision data which exhibit equal values for our system and CAMILLE. Hastings (ibid.) also rightly observes that "... they [the classifier systems] stop short of inferring the best hypotheses." The specialization procedure developed for CAMILLE resembles the one underlying our system. Contrary to Hasting's approach, however, we evaluate the different, more specific hypotheses with respect to linguistic and conceptual evidence and arrive at a ranked list of hypotheses based on **TH** and **CB** criteria. This way, more specific hypotheses simultaneously pass an evidential filtering mechanism that significantly increases the system's learning performance.

[3] We tested the WORDNET lexical database (Fellbaum 1998), a common-sense ontology, in order to determine concept paths of maximal length. In the computer domain, the maximum path length amounts to eight nodes. For the entire ontology, the maximum path length of eleven nodes was found in the biology domain. The data were collected by one of our colleagues, Katja Markert.

Concept Hypotheses	LA –	LA TH	LA CB
PhysicalObject(176)	0.30	0.30	0.30
MentalObject(0)	0.16	0.16	0.16
InformationObject(5)	0.16	0.16	0.16
MassObject(0)	0.16	0.16	0.16
Norm(3)	0.16	0.16	0.16
Technology(1)	0.16	0.16	0.16
Mode(5)	0.16	0.16	0.16
Feature(0)	0.16	0.16	0.16
	ø:0.18	ø:0.18	ø:0.18
Learning Step 1			
Product(136)	0.50	0.50	0.50
MentalObject(0)	0.16		
InformationObject(5)	0.16		
MassObject(0)	0.16		
Norm(3)	0.16		
Technology(1)	0.16		
Mode(5)	0.16		
Feature(0)	0.16		
	ø:0.20	ø:0.50	ø:0.50
Learning Step 2			
Computer(5)	0.50	0.50	
OutputDev(9)	0.80	0.80	0.80
StorageDev(5)	0.55	0.55	0.55
InputDev(2)	0.55	0.55	0.55
	ø:0.60	ø:0.60	ø:0.63
Learning Step 3			

Table 3: Learning Steps 1 to 3 for the Sample Text

for each hypothesized concept the number of concepts subsumed by it in the underlying knowledge base (cf. also our notion of learning rate, as introduced below); **LA CB** gives the accuracy rate for the full qualification calculus including threshold and credibility criteria, **LA TH** for threshold criteria only, while **LA –** depicts the accuracy values produced by the terminological reasoning component without incorporating any quality criteria. As can be read off from both tables, the full qualification calculus produces either the same or even more accurate results on average, equally many or fewer hypothesis spaces (indicated by the number of rows), and derives the correct prediction more rapidly (in step 6) than the less knowledgeable variants (in step 9).

The data also illustrate the continuous specialization of concept hypotheses achieved by the terminological classifier, e.g., from PhysicalObject in step 1 via Product in step 2 to OutputDEVice, Printer, and LaserPrinter in step 3, 4, and 5, respectively. The overall learning accuracy – due to the learner's aggressive specialization strategy – may even temporarily decrease in the course of hypothesizing (e.g., from step 3 to 4 or step 5 to 6 for **LA –**, as well as for **LA TH**), but the learning accuracy value for the full qualification calculus (**LA CB**) always increases.

Fig. 5 depicts the learning accuracy curve for the entire data set (101 texts). We also have included the graph depicting the growth behavior of hypothesis spaces (Fig. 6). For both data sets, we distinguish again between the measurements for **LA –**, **LA TH** and **LA CB**. In Fig. 5, we start from LA values in the interval between 48% to 54% for **LA –/LA TH** and **LA CB**, respectively, in the first learning step, whereas the number of hypothesis spaces (**NH**) range between 6.2 and 4.5 (Fig. 6). In the final step, learning accuracy rises up from 79%, 83% to 87% for **LA –**, **LA TH** and **LA CB**, respectively, and the **NH** values reduce to 4.4, 3.6 and 2.5 for each of the three criteria, respectively.

Concept Hypotheses	LA –	LA TH	LA CB
Notebook(0)	0.43	0.43	
Portable(0)	0.43	0.43	
PC(0)	0.43	0.43	
Workstation(0)	0.43	0.43	
Desktop(0)	0.43	0.43	
Printer(3)	0.90	0.90	0.90
VisualDev(2)	0.66	0.66	0.66
Loudspeaker(0)	0.66	0.66	0.66
Plotter(0)	0.66	0.66	0.66
RW-Store(2)	0.50	0.50	0.50
RO-Store(1)	0.50	0.50	0.50
Mouse(0)	0.50	0.50	
Keyboard(0)	0.50	0.50	
	ø:0.54	ø:0.54	ø:0.65
Learning Step 4			
Notebook(0)	0.43	0.43	
Portable(0)	0.43	0.43	
PC(0)	0.43	0.43	
Workstation(0)	0.43	0.43	
Desktop(0)	0.43	0.43	
LaserPrinter(0)	1.00	1.00	1.00
InkjetPrinter(0)	0.75	0.75	0.75
NeedlePrinter(0)	0.75	0.75	0.75
	ø:0.58	ø:0.58	ø:0.83
Learning Step 5			
Notebook(0)	0.43	0.43	
Portable(0)	0.43	0.43	
PC(0)	0.43	0.43	
Workstation(0)	0.43	0.43	
Desktop(0)	0.43	0.43	
LaserPrinter(0)	1.00	1.00	1.00
	ø:0.53	ø:0.53	ø:1.00
Learning Step 6,7,8			
LaserPrinter(0)	1.00	1.00	1.00
	ø:1.00	ø:1.00	ø:1.00
Learning Step 9			

Table 4: Learning Steps 4 to 9 for the Sample Text

The pure terminological reasoning machinery always achieves an inferior level of learning accuracy and generates more hypothesis spaces than the learner equipped with the qualification calculus. Also, the inclusion of conceptual criteria (**CB**) supplementing the linguistic criteria (**TH**) helps a lot to focus on the relevant hypothesis spaces and to further discriminate the valid hypotheses (on the range of 4% of precision). Note that an already significant plateau of accuracy is usually reached after the third step (*viz.* 67%, 73%, and 76% for **LA –**, **LA TH**, and **LA CB**, respectively, in Fig. 5; the corresponding numbers of hypothesis spaces being 6.1, 5.1, and 3.7 for **NH –**, **NH TH**, and **NH CB**, respectively, in Fig. 6). This indicates that our approach not only yields competitive accuracy rates (a mean of 87%) but also finds the most relevant distinctions in a very early phase of the learning process, i.e., it requires only a *few* examples.

Learning Rate

The learning accuracy focuses on the predictive power of the learning procedure. By considering a third type of measure, the *learning rate*, we supply data from the step-wise reduction of alternatives for the learning process. Fig. 7 depicts the mean number of transitively included concepts for all considered hypothesis spaces per learning step (each concept hypothesis denotes a concept which transitively subsumes various subconcepts). Note that the most general concept hypothesis, in our example, denotes Object which currently includes 196 concepts. In general, we observed a

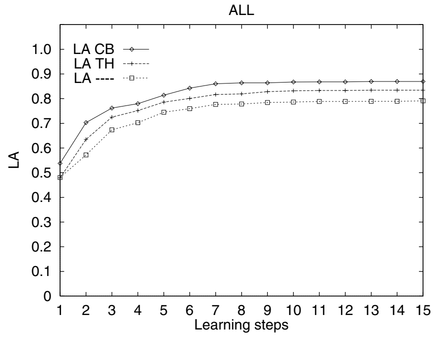

Figure 5: Learning Accuracy (LA)

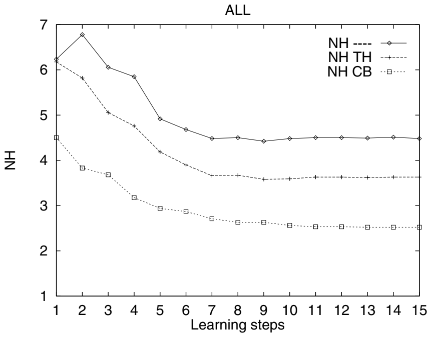

Figure 6: Number of Hypotheses (NH)

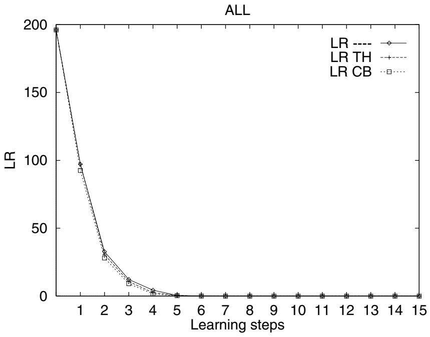

Figure 7: Learning Rate (LR)

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). Again, learning step 3 is a crucial point for the reduction of the number of included concepts (ranging from 16 to 21 concepts). Summarizing this evaluation experiment, the quality-based learning system exhibits significant and valid reductions of the predicted concepts (up to two, on the average).

Related Work

Our approach bears a close relationship to the work of Granger (1977), Mooney (1987), Berwick (1989), Rau, Jacobs, & Zernik (1989), Gomez & Segami (1990), Hastings (1994), and Moorman & Ram (1996), who all aim at the automated learning of word meanings from 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* in terms of the *quality* of arguments is not an issue in these studies. Learning from real-world texts usually provides the learner with only sparse and fragmentary evidence such that multiple hypotheses are likely to be derived and a need for a hypothesis evaluation arises.

The work closest to ours has been carried out by Rau, Jacobs, & Zernik (1989) and Hastings (1994). They also generate concept hypotheses from linguistic and conceptual evidence. 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 for Rau *et al.*'s system – in the light of our evaluation study (cf. Fig. 5, final learning step) – amounts to 8%, considering the difference between **LA -** (plain terminological reasoning) and **LA CB** values (terminological metareasoning based on the qualification calculus). Similarly strong arguments hold for a comparison of our results with Hasting's (1994) approach at the precision dimension, with an even greater advantage for the full qualification calculus (39%) over terminological-style reasoning in the CAMILLE System (22%). Hence, our claim that we produce competitive results.

Note that the requirement to provide learning facilities for real-world text knowledge engineering 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. In particular, no *new* templates have to be created. This step would correspond to the procedure we described in this contribution.

In the field of knowledge engineering from texts, our system constitutes a major achievement through the complete automatization of the knowledge elicitation process (cf. also Gomez & Segami (1990)). Previous studies mainly dealt with that problem by either hand-coding the content of the textual documents (Skuce *et al.* 1985), or providing semi-automatic, interactive devices for text knowledge acquisition (Szpakowicz 1990), or using lexically oriented statistical approaches to text analysis (Shaw & Gaines 1987).

Conclusion

Knowledge-based systems provide powerful forms of reasoning, but it takes a lot of effort to equip them with the knowledge they need by means of manual knowledge engineering. In this paper, we have introduced an alternative solution based on the fully automatic processing of expository texts. This text knowledge engineering methodology is based on the incremental assignment and evaluation of the quality of linguistic and conceptual evidence for emerging concept hypotheses. No specialized learning algorithm is needed, since learning is a (meta)reasoning task carried out by the classifier of a terminological reasoning system. However, strong 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.

The procedure for text knowledge engineering was tested on a medium-sized knowledge base for the information technology domain. The choice of a single domain reduces the number of possible conceptual ambiguities when concept hypotheses are created, in particular when compared with common-sense ontologies such as WORDNET (Fellbaum 1998). However, one might envisage partitioning mechanisms in order to control the activation of reasonable portions of a knowledge base and thus escape from a prohibitive explosion of the number of alternatives to be pursued.

Actually, we also like to contrast our text knowledge engineering approach to standard machine learning algorithms like ID3, k-nearest neighbor and Bayesian classifers. Initial evidence from current experiments indicates that either the number of hypotheses they generate become prohibitively large, even in the medium-sized knowledge base we use (especially for k-nearest neighbor), or the learning accuracy drops down very seriously (e.g., for ID3). The outcome of these experiments might clarify the usefulness of standard ML algorithms for the text knowledge engineering task.

It should also be obvious that the accuracy of our text knowledge engineering procedure is dependent on the input supplied by the parser. This is particularly true of false semantic interpretations (cf. Table 1), which directly misguide the reasoning process of the learner. Missing data, however, are far less harmful, since the knowledge acquisition procedure needs only a few examples to narrow down the search space, as has become evident from the evaluation study.

In our experiments, learning was 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 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 associated with each of them. In the future, we intend to study such test cases, too. Fortunately, the number of hypothesis spaces decreases rapidly (cf. Fig. 6) as does the learning rate (cf. Fig. 7) so that the learning system should remain within feasible bounds.

Acknowledgements. 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 – Proceedings 13th International Joint Conference on Artificial Intelligence.*, 1172–1178. San Mateo, CA: Morgan Kaufmann.

Berwick, R. 1989. Learning word meanings from examples. In Waltz, D., ed., *Semantic Structures. Advances in Natural Language Processing*. L. Erlbaum. 89–124.

Buchanan, B.; Barstow, D.; Bechtal, R.; Bennett, J.; Clancey, W.; Kulikowski, C.; Mitchell, T.; and Waterman, D. 1983. Constructing an expert system. In Hayes-Roth, F.; Waterman, D.; and Lenat, D., eds., *Building Expert Systems*. Reading, MA: Addison-Wesley. 127–167.

Fellbaum, C., ed. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Gomez, F., and Segami, C. 1990. Knowledge acquisition from natural language for expert systems based on classification problem-solving methods. *Knowledge Acquisition* 2(2):107–128.

Granger, R. 1977. FOUL-UP: a program that figures out meanings of words from context. In *IJCAI'77 – Proc. of the 5th Intl. Joint Conf. on Artificial Intelligence.*, 172–178.

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*. Springer. 453–468.

Hahn, U.; Schacht, S.; and Bröker, N. 1994. Concurrent, object-oriented natural language parsing: the PARSETALK model. *International Journal of Human-Computer Studies* 41(1/2):179–222.

Hastings, P. 1994. *Automatic Acquisition of Word Meaning from Context*. Ph.D. Dissertation, Computer Science and Engineering Department, University of Michigan.

MacGregor, R. 1994. A description classifier for the predicate calculus. In *AAAI'94 – Proceedings 12th National Conference on Artificial Intelligence.*, 213–220. Menlo Park, CA: AAAI Press & Cambridge, MA: MIT Press.

Mooney, R. 1987. Integrated learning of words and their underlying concepts. In *CogSci'87 – Proceedings of the 9th Annual Conference of the Cognitive Science Society*, 974–978. Hillsdale, NJ: L. Erlbaum.

Moorman, K., and Ram, A. 1996. The role of ontology in creative understanding. In *CogSci'96 – Proceedings of the 18th Annual Conference of the Cognitive Science Society*, 98–103. Mahwah, NJ: L. Erlbaum.

Morik, K.; Wrobel, S.; Kietz, J.-U.; and Emde, W. 1993. *Knowledge Acquisition and Machine Learning: Theory, Methods, and Applications*. London: Academic Press.

Nirenburg, S., and Raskin, V. 1987. The subworld concept lexicon and the lexicon management system. *Computational Linguistics* 13(3/4):276–289.

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 – Proceedings 20th Annual German Conference on Artificial Intelligence*, 349–362. Springer.

Shaw, M., and Gaines, B. 1987. KITTEN: knowledge initiation and transfer tools for experts and novices. *International Journal of Man-Machine Studies* 27(3):251–280.

Skuce, D.; Matwin, S.; Tauzovich, B.; Oppacher, F.; and Szpakowicz, S. 1985. A logic-based knowledge source system for natural language documents. *Data & Knowledge Engineering* 1(3):201–231.

Stefik, M. 1995. *Introduction to Knowledge Systems*. San Francisco, CA: Morgan Kaufmann.

Szpakowicz, S. 1990. Semi-automatic acquisition of conceptual structures from technical texts. *International Journal on Man-Machine Studies* 33:385–397.

Woods, W., and Schmolze, J. 1992. The KL-ONE family. *Computers & Mathematics with Applications* 23:133–177.