

Ontology and Lexicon Evolution by Text Understanding

Udo Hahn and Kornél G. Markó¹

Abstract. We introduce a methodology for automating the maintenance and extension of domain taxonomies, combined with the acquisition of lexical knowledge on the basis of a natural language understanding system. By defining the strength of various forms of evidence, conceptual as well as lexical hypotheses are ranked according to qualitative plausibility criteria and the most reasonable ones are selected for assimilation into the already given domain ontology and lexical class hierarchy.

1 Introduction

Natural language understanding systems require knowledge-rich resources to reason with: beside lexical and morphological information and syntactic rules or constraints, semantic as well as conceptual knowledge must be available. Since the coding of this information by humans is inevitably incomplete when facing real-world scenarios, specification gaps of various knowledge sources have to be filled automatically. Some emphasis has been put on providing machine learning support for single knowledge sources – morphological [9], lexical [12, 15], syntactic [1, 3], semantic [4] or conceptual knowledge [8, 18]. But only Cardie [2], up to now, has made an attempt to combine these isolated streams of linguistic knowledge acquisition.

In this paper, we also propose an integrated approach to learn *different* types of relevant knowledge for natural language processing systems *simultaneously*. New concepts are acquired and positioned in the given concept taxonomy whilst the grammatical status of their lexical correlates is identified and stored in a lexical class hierarchy.

In the next section, we will describe the architecture of SYNDIKATE, a text understanding system which learns new concepts while understanding a text. The model of the knowledge acquisition process is then introduced informally by discussing a concrete learning scenario in Section 3. Section 4 outlines the methodology of generating concept hypotheses and their selection by taking quality criteria into account. A compact survey of an evaluation of the grammar and concept acquisition task is given in Section 5. Finally, in Section 6 we discuss the advantages and drawbacks of our approach in the light of current research and our own evaluation results.

2 System Architecture of SYNDIKATE

In this section, we briefly introduce the text understanding system SYNDIKATE (SYNthesis of Distributed Knowledge Acquired from TEXTs) [5], into which the learning procedure is integrated. Two different domain knowledge bases are currently available for the system, one covering information technology with focus on the hardware domain, the other dealing mainly with the anatomy and pathol-

ogy subdomains of medicine [17]. SYNDIKATE relies on two major kinds of knowledge:

Grammatical knowledge for syntactic analysis is given as a fully lexicalized dependency grammar [7]. Such a grammar mainly consists of the specification of local valency constraints between lexical items, or more precisely, between a potential syntactic head (e.g., a noun) and a possible syntactic modifier (e.g., a determiner or an adjective). Valency constraints also include restrictions on word order, compatibility of morphosyntactic features, as well as semantic integrity conditions. In order to relate two lexical items via a dependency relation $\delta \in \mathcal{D} := \{\text{specifier, subject, dir-object, ...}\}$, all valency constraints must be fulfilled. In this approach, lexeme specifications, to which lexical items are attached, form the leaf nodes of a lexicon tree. These lexical nodes are further abstracted in terms of a hierarchy of word class specifications at different levels of generality, which reflect stronger (or weaker) constraints these classes embody as one descends (ascends) the word class hierarchy. This leads to a specification of word class names $\mathcal{W} = \{\text{VERBAL, VERBFINITE, SUBSTANTIVE, NOUN, ...}\}$ and a subsumption relation $isa_{\mathcal{W}} = \{(\text{VERBFINITE, VERBAL}), (\text{NOUN, SUBSTANTIVE}), \dots\} \subset \mathcal{W} \times \mathcal{W}$, which characterizes specialization relations between word classes.

Conceptual knowledge is expressed in terms of a KL-ONE-like knowledge representation language [21]. A domain ontology (we here consider the IT domain) consists of a set of concept names $\mathcal{F} := \{\text{COMPANY, HARDDISK, ...}\}$ and a subsumption relation $isa_{\mathcal{F}} = \{(\text{HARDDISK, STORAGEDEVICE}), (\text{IBM, COMPANY}), \dots\} \subset \mathcal{F} \times \mathcal{F}$. Concepts are linked by conceptual relations. The corresponding set of relation names $\mathcal{R} := \{\text{HAS-PART, DELIVERAGENT, ...}\}$ denotes conceptual relations which are also organized in a subsumption hierarchy $isa_{\mathcal{R}} = \{(\text{HAS-HARD-DISK, HAS-PHYS-PART}), (\text{HAS-PHYS-PART, HAS-PART}), \dots\} \subset \mathcal{R} \times \mathcal{R}$.

The result of a syntactic analysis is captured in a dependency graph, in which nodes represent words only. These nodes are connected by dependency relations taken from \mathcal{D} . The semantic interpretation rests on well-defined configurational patterns within a dependency graph, so-called *semantically interpretable subgraphs* [14]. Such a subgraph is given by a connection of two content words via a number of edges without another content word intervening on that path. Whenever during the incremental analysis process a semantically interpretable subgraph is completed, a semantic interpretation process is triggered, which consists of a search for conceptual relations in the knowledge base between the conceptual correlates of the two content words in the minimal subgraph.

3 Sample Learning Scenario

Suppose, you never heard anything about “R600MX” or “Vaio” before. Imagine, one day, your favorite computer magazine features an article starting with “The R600MX of the company Vaio costs approx-

¹ Text Knowledge Engineering Lab, Linguistische Informatik, Albert-Ludwigs-Universität Freiburg, Werthmannplatz 1, D-79085 Freiburg, Germany, <http://www.coling.uni-freiburg.de/>

imately 1600 Euros.” Has your knowledge increased? If so, what did you learn from just this phrase?

Initially, from a grammatical point of view, the lexical item “R600MX” can be regarded as an instance of one of the top-level open-class part-of-speech categories (i.e., NOMINAL, ADVERB and VERBAL)² or of their descendents, respectively (cf. Figure 1). During the processing of the phrase “of the company” as a potential attribute of the yet unknown item “R600MX”, the ADVERB hypothesis can be rejected, due to violating grammatical constraints (neither a noun (“company”), nor an article (“the”) can modify an adverb, cf. the darkly shaded box in Figure 1). Furthermore, since no valency description for a determiner is specified in the VERBAL or ADJECTIVE word class (neither for their descendents), only the SUBSTANTIVE hypothesis remains valid (cf. the grey shaded boxes in Figure 1). This hypothesis can be further refined to the class of NOUNS, because first, the PRONOUN subclass does not provide a dependency relation for an article and, second, this set describes a closed class, which is completely specified.

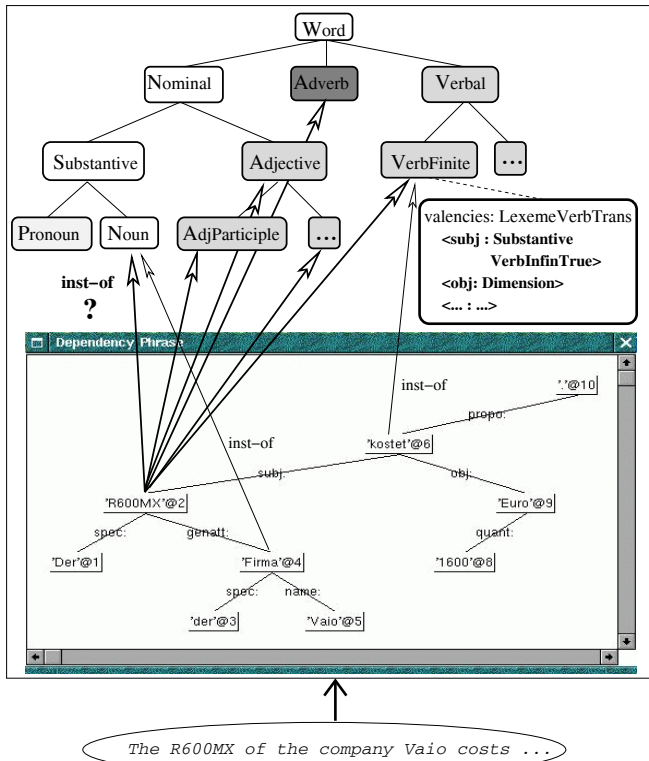


Figure 1. Sample Scenario — Grammatical Learning

From a conceptual perspective, when processing the word “R600MX”, the initial hypothesis space for the new lexical item incorporates all the top level concepts available in the given domain

² While the distinction between NOMINAL and VERBAL should be obvious, the prominent role of ADVERB at the top level of word class categories might not be. NOMINAL as well as VERBAL carry grammatical information such as case, gender, number, or tense, mood, aspect, respectively, none of which is shared by ADVERBS. As class hierarchies derive from the principle of property inheritance, and ADVERBS lack common features with other word classes, they form an independent class on their own. This explains the prominent role of ADVERB at the highest level of class abstraction (cf. Hahn *et al.* [7] for a discussion of the object-oriented design of the underlying grammar/parser).

ontology, i.e. OBJECT, ACTION, DEGREE, etc. (cf. Figure 2). At the stage after processing the phrase “of the company”, the word “R600MX” is linked with “company” via a specific dependency relation ($gen[itive]att[ribute]$), which makes a semantically interpretable subgraph. Therefore, the conceptual correlate of “R600MX” must be something that is semantically related with the concept COMPANY in the domain ontology. Now, consider a fragment of the conceptual representation for companies:

- (P1) COMPANY \sqsubseteq LEGAL-PERSON \sqcap
 $\forall HAS-OFFICE.OFFICE \sqcap$
 $\forall HAS-MEMBER.NATURAL-PERSON \sqcap$
 $\forall PRODUCES.PRODUCT \sqcap$
 ...

The concept COMPANY is defined (P1) as a subclass of LEGAL-PERSON and all the fillers of the relations HAS-OFFICE, HAS-MEMBER and PRODUCES must be concepts subsumed by OFFICE, NATURAL-PERSON and PRODUCT, respectively. Whilst HAS-MEMBER is a role inherited from a concept that subsumes COMPANY, viz. LEGAL-PERSON, the other roles are attached to the concept itself (and are inherited by all subconcepts of COMPANY). All roles from P1 have to be considered for relating the conceptual representation of “R600MX” to the semantic correlate of “company”. As a consequence, “R600MX” can be regarded as a kind of OFFICE, NATURAL-PERSON or PRODUCT, respectively (cf. Figure 2).

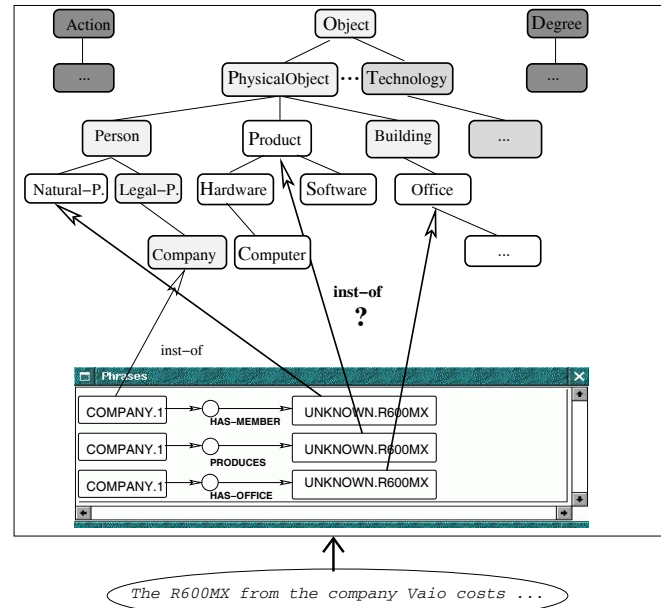


Figure 2. Sample Scenario — Conceptual Learning

Continuing our example, the (unknown) word “Vaio” has to be integrated in the existing dependency graph. On the one hand, if it is treated as a NOUN, it can be attached to the word “company” by establishing an apposition relation between these two items (“the company Vaio”). Following semantic interpretation rules, “Vaio” can immediately be classified as a kind of COMPANY, since appositions uniquely restrict the semantic interpretability. On the other hand, if we interpret “Vaio” as an ADVERB, it modifies the following verb “costs”, though a semantic correlate cannot be stated for this alternative.

In addition, after processing the verb, the NOUN “*R600MX*” is bound to “*costs*” via the `subj[ect]` dependency relation. The conceptual representation of “*costs*”,

$$(P2) \text{ COST} \sqsubseteq \text{STATE} \sqcap \\ \forall \text{COST-PATIENT.}(\text{PRODUCT} \sqcup \text{SERVICE}) \sqcap \\ \forall \text{COST-CO-PATIENT.PRICE}$$

forces the semantic interpreter to translate the given syntactic structure to the role `COST-PATIENT`, leading again to the hypothesis that “*R600MX*” is a kind of `PRODUCT` or `SERVICE`. So, `PRODUCT` as the conceptual correlate of “*R600MX*” is derived twice, and is therefore the preferred reading at this stage of text analysis. Proceeding with the rest of the sentence “*approximately 1600 Euros*”, the value and dimension statements are interpreted as `COST-CO-PATIENT` of the verb concept `COST`, viz. `PRICE`.

Summarizing, after reading the sentence “*The R600MX of the company Vaio costs approximately 1600 Euros*” one can state that “*R600MX*” is a NOUN, which is hypothesized to be a kind of `PRODUCT`, whilst “*Vaio*” can be a NOUN or `ADVERB` from a grammatical point of view, and — if one prefers the NOUN reading, e.g., by multiple derivations of this hypothesis in the sample text — its semantic correlate can be classified to the concept `COMPANY`.

When we analyze a text in this manner, more evidence for generated hypotheses or specializations of them can be collected. For example, the processing of a noun phrase as “*the touchpad of the R600MX*” results in alternative hypotheses, viz. all concepts that are related with `TOUCHPAD` in the domain ontology, e.g., `OWNER`, `PRODUCER`, `NOTEBOOK`, etc. Since only the concept `NOTEBOOK` is consistent with the multiply derived hypothesis `PRODUCT`, this specialization is preferred for the ongoing analysis. Furthermore, a case frame assignment as in “*the R600MX has a touchpad*” is even more restrictive, because in this case the new item “*R600MX*” can only be an entity that is related with the concept `TOUCHPAD` via the role `HAS-PART` in the domain ontology, as with `NOTEBOOK` only. As one can see, different syntactic structures provide different *levels of quality* of concept hypotheses.

4 Learning by Quality

As mentioned above, the generation of concept hypotheses is determined by the syntactic context in which an unknown word appears. When the syntactic analysis identifies the pattern of an apposition (“*the notebook R600MX*”) or an exemplification (“*the R600MX is a notebook*”) a single concept hypothesis can immediately be derived by considering the semantic correlate of the known lexical item in the syntactic context of the unknown word. Since appositions and exemplifications are strong indicators of the validity of a derived hypothesis, this information is attached to the corresponding hypothesis in terms of so-called *linguistic quality labels* $LQ \in \{\text{APPOSITION, EXEMPLIFICATION, CASE-FRAME-ASSIGNMENT, PP-ATTACHMENT, GENITIVE-ATTRIBUTION}\}$. When an unknown lexical item occurs in the syntactic context of a case frame assignment (“*the R600MX has a touchpad*”) role restrictions of the verb concept form the corresponding hypotheses³ and the appropriate linguistic quality label is assigned. With regard to concept hypothesis

³ Since, e.g., the role `HAS-TOUCHPAD` is defined as a specialization of the role `HAS-PART` with its domain restricted to `NOTEBOOK` and its range restricted to `TOUCHPAD`, the classifier of the underlying terminological logic system immediately specializes the conceptual correlate of a verb (e.g. “*have*”) according to the conceptual role fillers computed by the semantic interpreter (cf. [14]).

generation, the syntactic context of prepositional phrase attachments and genitive assignments are less restrictive, since all roles attached to the conceptual correlate of the dependency-related known lexical item must be taken into account (cf. the example in Section 3). The labels `PP-ATTACHMENT` or `GENITIVE-ASSIGNMENT` are then assigned to the corresponding concept hypothesis, respectively. This leads to a partially ordered quality relation $>_{LQ}$ along the linguistic dimension of quality labels `LQ`:

$$>_{LQ} = \{ \text{(APPOSITION, CASE - FRAME - ASSIGNMENT)}, \\ \text{(APPOSITION, GENITIVE-ASSIGNMENT)}, \\ \text{(APPOSITION, PP-ATTACHMENT)}, \\ \text{(EXEMPLIFICATION, CASE-FRAME-ASSIGNMENT)}, \\ \text{(EXEMPLIFICATION, GENITIVE-ASSIGNMENT)}, \\ \text{(EXEMPLIFICATION, PP-ATTACHMENT)}, \\ \text{(CASE-FRAME-ASSIGNMENT, GENITIVE-ASSIGNMENT)}, \\ \text{(CASE-FRAME-ASSIGNMENT, PP-ATTACHMENT)} \}$$

In addition to the linguistic dimension, *conceptual quality labels* $CQ \in \{\text{MULTIPLY-DEDUCED, SUPPORTED, CROSS-SUPPORTED, ADDITIONAL-ROLE-FILLER}\}$ are assigned to concept hypotheses when specific conceptual patterns arise in the text knowledge base. For example, the multiple derivation of the same concept hypothesis during text analysis is regarded as a strong indicator for its validity, and therefore the label `MULTIPLY-DEDUCED` will be attached to it. Another, slightly weaker label is `SUPPORTED`. It is assigned to a concept hypothesis, if the discourse entity in focus is already conceptually related to another entity via the same conceptual role (e.g., if “*Vaio*” is already known as being the producer of a specific notebook, and “*R600MX*” is related to “*Vaio*”, it is also hypothesized as a notebook). The label `CROSS-SUPPORTED` lends evidence to a hypothesis, when two discourse objects - one of them representing the unknown item - are related via two similar roles (e.g., the hypothesis that “*Vaio*” is the producer of “*R600MX*” is supported by the statement that “*Vaio*” is also known to be the vendor of “*R600MX*”). Negative evidence is represented by the label `ADDITIONAL-ROLE-FILLER`. It is assigned whenever an action role (`AGENT`, `PATIENT`, `CO-PATIENT`) is multiply filled (e.g., when two different companies are assumed to be the producer of the same product). As for `LQ`, the following quality order relation $>_{CQ}$ can be defined along the conceptual dimension of quality labels `CQ`:

$$>_{CQ} = \{ \text{(MULTIPLY-DEDUCED, SUPPORTED)}, \\ \text{(MULTIPLY-DEDUCED, CROSS-SUPPORTED)}, \\ \text{(MULTIPLY-DEDUCED, ADDITIONAL-ROLE-FILLER)}, \\ \text{(SUPPORTED, ADDITIONAL-ROLE-FILLER)}, \\ \text{(CROSS-SUPPORTED, ADDITIONAL-ROLE-FILLER)} \}$$

In addition, the goodness of linguistic quality labels is ranked higher than all conceptual quality labels. In order to estimate the quality of hypotheses, the learning procedure collects all hypotheses with the highest amount of `APPOSITION` labels. Based on this set, all elements with the highest amount of `CASE-FRAME-ASSIGNMENT` labels are selected for further discrimination, and so on, according to the order relations $>_{LQ}$ and $>_{CQ}$. After the processing of the whole text the highest ranked hypothesis is selected for assimilation into the concept taxonomy, for each unknown lexical item. On the other hand, from the grammatical point of view, the most frequent word class hypothesis of each unknown word is chosen for integration into `SYNDIKATE`'s lexicon.

5 Learning Performance

The domain knowledge base on which we performed our evaluation experiments contained approximately 3,000 concepts and relations from the information technology (IT) domain, the grammatical class hierarchy was composed of 80 word classes. We randomly selected 48 texts from our corpus of IT magazines. This sample contained a total amount of 75 unknown words from a wide range of word classes (excluding VERBALS), as well as 48 descriptions of new products, i.e., new concepts to be learned. In this experiment, we evaluated the learner’s 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 [8], we distinguish here the following parameters:

- **Hypothesis** denotes the set of concept or grammatical class hypotheses derived by the system as the final result of the text understanding process for each target item;
- **Correct** denotes the number of cases in the test set in which **Hypothesis** contains the correct concept or grammatical class description for the target item;
- **OneCorrect** denotes the number of cases in the test set in which **Hypothesis** is a singleton set, which contains only the correct concept or grammatical description;
- **HypoSum** denotes the number of different hypotheses generated by the system for the target item considering the entire test set.

We measure the performance of the lexicon as well as the concept learner in terms of recall and precision, with **TestSet** denoting the number of target items to be learned:

$$\text{RECALL} := \frac{\text{Correct}}{\text{TestSet}} \quad \text{PRECISION} := \frac{\text{Correct}}{\text{HypoSum}}$$

5.1 Lexicon Learning

The task of lexicon learning is to predict the most specific word class for an unknown lexical item, given a hierarchy which covers all relevant word classes for a particular natural language. The learner starts from quite general word class hypotheses which are continuously refined as more discriminatory evidence comes in.

The data in Table 1 indicates that the system dealt with 75 instances of unknown lexical items. This number includes cases of word class ambiguities, as well as instances of word classes other than SUBSTANTIVES (but excluding occurrences of VERBALS). We first discuss the results of the basic learning procedure (data in column one), and then turn to a heuristic refinement (column two). So, in the both learning modes, in 71 of the 75 cases word class hypotheses could be generated (in four cases data was so weak that no hypothesis could be created). In 67 of the 71 cases, the set of word class hypotheses for an unknown lexical item included the correct prediction, whereas in 31 cases this set contained only the correct grammatical description. Counting all word class hypotheses generated at all by the parser leads to 89% recall and 63% precision (column one).

These results can be substantially improved with respect to precision, when we add a simple heuristics (cf. column two). At the end of the full learning cycle various word class hypotheses may still remain valid for one unknown lexical item (in the test set, this happened in [71–31 =] 40 cases). Rather than considering this outcome as the final result, we resolved the indeterminacy by summing all occurrences of single word class predictions for each unknown word over all learning steps, i.e., at any point where the unknown word appeared in a new syntactic pattern. We then considered the word class

Table 1. Performance Data for Lexicon Learning

	Basic	Basic + Heuristic
TestSet	75	75
Correct	67	67
OneCorrect	31	58
HypoSum	106	78
RECALL	89.3%	89.3%
PRECISION	63.2%	85.9%

prediction with the highest number of occurrences as the preferred word class hypothesis. This heuristic leads to 89% recall and 86% precision for a fully unsupervised learning procedure. For 58 lexical items (instead of 31 in the basic procedure), there was only one and correct hypothesis, while also the diversity of hypotheses generated at all was far more restricted (78 instead of 106).

A straightforward comparison of the results achieved by the lexicon learner with to-day’s best performing part-of-speech (POS) taggers (with recognition accuracy ranging between 97-99% [20, 1]) should be carried out with caution. The reason being that the diversity and specificity of the word classes we employ is different from comparable grammars. For instance, the number of word classes we use (on the order of 80) is more than twice the number of those in Treebank-style grammars (with 36 POS tags [10]). So, incorrect guesses are more likely when more choices can be made and quite specific word classes have to be predicted. Furthermore, our grammar does not only provide POS information (i.e., lexical categories such as noun, adjective, etc.) but also comes up with rich additional grammatical knowledge. So, once a word class is hypothesized, grammatical information associated with this word class (such as valency frames, word order constraints, or morphosyntactic features) comes for free due to the organization of the grammar as a lexical class hierarchy.

5.2 Ontology Learning

The evaluation study we performed for the ontology learning task was conducted under two varying experimental conditions. On the one hand, we wanted to assess the potential of the quality calculus for ontology learning under *optimal* conditions. By this, we refer to a state of the system where the parser as well as the domain knowledge base have access to sufficiently rich specifications so that ‘complete’ (in the sense of non-corrupted) parse trees, discourse structures and semantic interpretation results can be generated from textual input. In essence, this is a framework for testing the *learning methodology* proper. These ideal conditions are relaxed under *realistic* conditions. By this, we refer to a ‘frozen’ state of the system’s knowledge sources prior to analyzing the test set. Hence, grammar specifications may be lacking, conceptual specifications may be fragmentary or missing at all so that deficient representation structures are likely to emerge depending on the breadth and depth of specification gaps.⁴ This is then a framework for testing SYNDIKATE’s current learning functionality and *system performance*. The difference between

⁴ The effect of incomplete knowledge on the quality of semantic interpretation for randomly sampled texts is assessed in Romacker & Hahn [13].

Table 2. Performance Measures for Concept Learning under Realistic and Optimal Conditions

	Realistic			Optimal		
	TR	TR+LQ	TR+LQ+CQ	TR	TR+LQ	TR+LQ+CQ
TestSet	71	71	71	48	48	48
Correct	27	27	26	34	34	33
OneCorrect	10	23	24	7	27	27
HypoSum	257	117	78	346	174	115
RECALL	38.0%	38.0%	36.6%	70.8%	70.8%	68.8%
PRECISION	10.5%	23.1%	33.3%	9.8%	19.5%	28.7%

these two measuring scenarios may elucidate, however, the potential of a knowledge-intensive approach to text analysis when it faces unfriendly real-world conditions.⁵

Both for optimal as well as realistic conditions, measures were taken under three experimental conditions (cf. Table 2). In the first and the fourth column (indicated by **TR**), we considered the contribution of a plain terminological reasoning component, the classifier, to the concept acquisition task, the second and the fifth column contain the results of incorporating linguistic quality criteria only, as a supplement to the classifier (denoted by **TR+LQ**), while the third and sixth column mirror linguistic as well as conceptual quality criteria (designated by **TR+LQ+CQ**), as a supplement to terminological reasoning.

Under realistic conditions, not only the 48 new product names were dealt with as unknown words by the system, but also 23 other lexical items from a wide range of word classes had to be taken into account. So the size of the **TestSet** varies for the realistic task (71 items)⁶ and for the optimal one (48 items). This also explains why we do not provide lexicon learning data for the optimal case. Since non-NOUN hypotheses simply do not meet the triggering condition of the learning system (e.g., an apposition involving an unknown noun is no longer an apposition when we assume the unknown item to be, e.g., an adjective), lexicon learning is (almost) trivial in the ideal case when we exclude verbal items from further consideration.

Under realistic test conditions learning without the qualification calculus, just relying on terminological reasoning, leads to particularly disastrous precision results (11%) at a recall of 38%. By incorporating all quality criteria, recall decreases slightly (37%), whereas precision increases up to 33%. It is obvious that the full calculus yields an enormous reduction of the number of hypotheses generated by the plain terminological reasoning component. At the same time, by refining the set of hypotheses, only one correct hypothesis was rejected in our test set. In 34% of all learning cases our system derives a single and valid concept hypothesis (**TR+LQ+CQ**).

Though the test set is smaller under optimal test conditions, a greater amount of correct hypotheses is generated, leading to recall values of 71% (**TR** and **TR+LQ**) and 69% (**TR+LQ+CQ**). The sur-

prisingly high numbers of hypotheses generated at all result in only slightly lower precision values (10%, 20% and 29%, respectively). Due to perfect parses in the optimal test scenario, more linguistic evidence is available and, therefore, much more concept hypotheses are collected. Nevertheless, in 56% of all learning cases there is only one and correct prediction (**TR+LQ+CQ**).

The source documents we dealt in our evaluation are test reports from the information technology domain. As it turned out, this text genre is highly suited for the concept acquisition method described in this contribution, due to following reasons: First, one can assume, that a particular target item (e.g., a product name) is usually featured in an article, and second, the unknown item often appears as part of an apposition or exemplification in the leading sentences of a test report. Obviously, authors tend to provide the proper interpretation context for a new concept in quite an early stage of text understanding ("*The notebook R600MX...*", "*The R600MX is a notebook...*"). The presented approach for concept acquisition directly takes advantage of this observation.

The SYNDIKATE system is also designed for the automatic content analysis of medical texts [6]. In this context or in related domains like pharmacology, further experiments are necessary in order to estimate the performance when it comes to the extraction of disease names, pharmaceutical product names, bio-catalysts, etc.

6 Conclusions

Knowledge-based systems provide powerful means for reasoning, but it takes a lot of effort to equip them with the knowledge they need, usually by manual knowledge engineering. In this paper, we have introduced an alternative solution. It is based on an automatic learning methodology in which concept and grammatical class hypotheses emerge as a result of the incremental assignment and evaluation of the quality of linguistic and conceptual evidence related to unknown words. No specialized learning algorithm is needed, since learning is a (meta)reasoning task carried out by the classifier of a terminological reasoning system [16].

This distinguishes our methodology from Cardie's case-based approach [2] which also combines conceptual and grammatical learning, but where the actual learning task is delegated to the C4.5 decision tree algorithm. Cardie's approach also requires some supervision (interactive grammatical encoding of the context window surrounding the unknown word), while our method operates entirely unsupervised. We share with her the view that learning should encompass several linguistic dimensions simultaneously (parts of speech, semantic and conceptual encodings) within a unified approach, and

⁵ A similar comparison of learning performance has been conducted by Cardie [2], who also distinguishes access to perfect vs. sparse dictionary information for a case-based learner.

⁶ As mentioned above, in four cases no valid word class hypotheses could be generated by the lexicon learner. Under such a circumstance, no concept learning is triggered, since linguistic quality criteria cannot be determined. In all other cases, a NOUN hypothesis was derived, even when this categorization was incorrect. On the other hand, a classification of an unknown substantive as a non-NOUN word class only did not occur in our test set.

should also avoid any explicit hand-coding heuristics to drive the acquisition process.

The work closest to ours with respect to the ontology learning problem has been carried out by Rau et al. [11] and Hastings and Lytinen [8]. They also generate concept hypotheses from linguistic and conceptual evidence. Unlike our approach, their selection of hypotheses depends only on an ongoing discrimination process based on the availability of this data but does not incorporate an inferencing scheme for reasoned hypothesis selection. The crucial role of quality considerations becomes obvious when one compares plain and quality-annotated terminological reasoning for the learning task (cf. Table 2).

As far as the qualification calculus is concerned the system of quality labels is still under investigation and needs further evaluation. We are currently working with a system of 15 linguistic and 6 conceptual quality labels the ordering of which (under preference considerations) has proved to be stable. A particularly interesting feature of our approach is that it does not require a learning mechanism on its own but is fully integrated in the terminological reasoning mode, the basis of proper text understanding.

The main disadvantage of our approach is that a profound amount of a priori knowledge is required. We provide domain knowledge bases which were specified up to the level of so-called base categories [19]. These are concepts which are needed for structuring the basic concept set of a domain (say, computers, printers, hard disks, operating systems, programming languages, etc. in the IT domain), but do not extend to more specialized concepts. We may guess that the set of base level categories which is characteristic of the IT domain amounts to 5,000 to 10,000 categories.

Extracting adequate conceptual representations of abstract terms fails in our studies, since their base categories are modeled only superficially, that means with just few conceptual roles which are indispensable for the learner. Their semantics remains unclear, often even for humans.

REFERENCES

- [1] Eric Brill, 'Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging', *Computational Linguistics*, **21**(4), 543–565, (1995).
- [2] Claire Cardie, 'A case-based approach to knowledge acquisition for domain-specific sentence analysis', in *AAAI'93 – Proceedings of the 11th National Conference on Artificial Intelligence*, pp. 798–803. Washington, D.C., July 11-15, 1993. Menlo Park, CA & Cambridge, MA: AAAI Press & MIT Press, (1993).
- [3] Eugene Charniak, 'Tree-bank grammars', in *AAAI'96/IAAI'96 – Proceedings of the 13th National Conference on Artificial Intelligence & 8th Innovative Applications of Artificial Intelligence Conference*, volume 2, pp. 1031–1036. Portland, Oregon, August 4-8, 1996. Menlo Park, CA & Cambridge, MA: AAAI Press & MIT Press, (1996).
- [4] Fernando Gomez, Carlos Segami, and Richard Hull, 'Determining prepositional attachment, prepositional meaning, verb meaning and thematic roles', *Computational Intelligence*, **13**(1), 1–31, (1997).
- [5] Udo Hahn and Martin Romacker, 'Content management in the SYNDIKATE system: How technical documents are automatically transformed to text knowledge bases', *Data & Knowledge Engineering*, **35**(2), 137–159, (2000).
- [6] Udo Hahn, Martin Romacker, and Stefan Schulz, 'How knowledge drives understanding: Matching medical ontologies with the needs of medical language processing', *Artificial Intelligence in Medicine*, **15**(1), 25–51, (1999).
- [7] Udo Hahn, Susanne Schacht, and Norbert Bröker, 'Concurrent, object-oriented natural language parsing: The PARSETALK model', *International Journal of Human-Computer Studies*, **41**(1/2), 179–222, (1994).
- [8] Peter M. Hastings and Steven L. Lytinen, 'The ups and downs of lexical acquisition', in *AAAI'94 – Proceedings of the 12th National Conference on Artificial Intelligence*, volume 1, pp. 754–759. Seattle, WA, USA, July 31 - August 4, 1994. Menlo Park, CA: AAAI Press & MIT Press, (1994).
- [9] Christian Jacquemin, 'Guessing morphology from terms and corpora', in *SIGIR'97 – Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, eds., N. J. Belkin, A. D. Narasimhalu, and P. Willett, pp. 156–165. Philadelphia, PA, USA, July 27-31, 1997. New York, NY: ACM, (1997).
- [10] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz, 'Building a large annotated corpus of English: The Penn Treebank', *Computational Linguistics*, **19**(2), 313–330, (1993).
- [11] Lisa F. Rau, Paul S. Jacobs, and Uri Zernik, 'Information extraction and text summarization using linguistic knowledge acquisition', *Information Processing & Management*, **25**(4), 419–428, (1989).
- [12] Ellen Riloff and Rosie Jones, 'Learning dictionaries for information extraction by multi-level bootstrapping', in *AAAI'99/IAAI'99 – Proceedings of the 16th National Conference on Artificial Intelligence & 11th Innovative Applications of Artificial Intelligence Conference*, pp. 474–479. Orlando, Florida, July 18-22, 1999. Menlo Park, CA: Cambridge, MA: AAAI Press & MIT Press, (1999).
- [13] Martin Romacker and Udo Hahn, 'An empirical assessment of semantic interpretation', in *NAACL 2000 – Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 327–334. Seattle, Washington, USA, April 29 - May 4, 2000. San Francisco, CA: Morgan Kaufmann, (2000).
- [14] Martin Romacker, Katja Markert, and Udo Hahn, 'Lean semantic interpretation', in *IJCAI'99 – Proceedings of the 16th International Joint Conference on Artificial Intelligence*, volume 2, pp. 868–875. Stockholm, Sweden, July 31 - August 6, 1999. San Francisco, CA: Morgan Kaufmann, (1999).
- [15] Barry Schiffman and Kathleen R. McKeown, 'Experiments in automated lexicon building for text searching', in *COLING 2000 – Proceedings of the 18th International Conference on Computational Linguistics*, volume 2, pp. 719–725. Saarbrücken, Germany, 31 July - 4 August, 2000. San Francisco, CA: Morgan Kaufmann, (2000).
- [16] Klemens Schnattinger and Udo Hahn, 'Quality-based learning', in *ECAL'98 – Proceedings of the 13th European Conference on Artificial Intelligence*, ed., W. Wahlster, pp. 160–164. Brighton, U.K., August 23-28, 1998. Chichester: John Wiley, (1998).
- [17] Stefan Schulz and Udo Hahn, 'Knowledge engineering by large-scale knowledge reuse: Experience from the medical domain', in *Principles of Knowledge Representation and Reasoning. Proceedings of the 7th International Conference – KR 2000*, eds., A. G. Cohn, F. Giunchiglia, and B. Selman, pp. 601–610. Breckenridge, CO, USA, April 12-15, 2000. San Francisco, CA: Morgan Kaufmann, (2000).
- [18] Stephen Soderland and Wendy Lehnert, 'WRAP-UP: A trainable discourse module for information extraction', *Journal of Artificial Intelligence Research*, **2**, 131–158, (1994).
- [19] Barbara Tversky, 'Where partonomies and taxonomies meet', in *Meanings and Prototypes. Studies in Linguistic Categorization*, ed., S. L. Tsohatzidis, 334–344, London, New York: Routledge, (1990).
- [20] Aro Voutilainen, 'A syntax-based part-of-speech analyser', in *EACL'95 – Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 157–164. Dublin, Ireland, March 27-31, 1995. Association for Computational Linguistics, (1995).
- [21] William A. Woods and James G. Schmolze, 'The KL-ONE family', *Computers & Mathematics with Applications*, **23**(2/5), 133–177, (1992).