Pacific Symposium on Biocomputing 7:338-349 (2002)

CREATING KNOWLEDGE REPOSITORIES FROM BIOMEDICAL REPORTS: THE MEDSYNDIKATE TEXT MINING SYSTEM

UDO HAHN MARTIN ROMACKER
Text Knowledge Engineering Lab, Freiburg University, D-79098 Freiburg, Germany
http://www.coling.uni-freiburg.de

STEFAN SCHULZ
Department of Medical Informatics, Freiburg University Hospital, D-79104 Freiburg
http://www.imbi.uni-freiburg.de/medinf

MEDSYNDIKATE is a natural language processor for automatically acquiring knowledge from medical finding reports. The content of these documents is transferred to formal representation structures which constitute a corresponding text knowledge base. The system architecture integrates requirements from the analysis of single sentences, as well as those of referentially linked sentences forming cohesive texts. The strong demands MEDSYNDIKATE poses to the availability of expressive knowledge sources are accounted for by two alternative approaches to (semi)automatic ontology engineering. We also present data for the knowledge extraction performance of MEDSYNDIKATE for three major syntactic patterns in medical documents.

1 Introduction

The application of methods from the field of natural language processing to biological data has long been restricted to the parsing of molecular structures such as DNA [1,2]. More recently, however, efforts have also been directed to capturing content from biological documents (research reports, journal articles, etc.), either dealing with restricted information extraction problems such as name recognition for proteins or gene products [3,4,5], or more sophisticated ones which aim at the acquisition of knowledge relating to protein or enzyme interactions, molecular binding behavior, etc. [6,7,8,9].

Current information extraction (IE) systems, however, suffer from various weaknesses. First, their range of understanding is bounded by rather limited domain knowledge. The templates these systems are supplied with allow only factual information about particular, a priori chosen entities (cell type, virus type, protein group, etc.) to be assembled from the analyzed documents. Also, these knowledge sources are considered to be entirely static. Accordingly, when the focus of interest of a user shifts to (facets of) a topic not considered so far, new templates must be supplied or existing ones must be updated manually. In any case, for a modified set of templates the analysis has to be rerun for the entire document collection. Templates also provide either no or severely limited inferencing capabilities to reason about the template fillers – hence, their understanding depth is low. Finally, the potential of IE systems for

dealing with textual phenomena is rather weak, if it is available at all. Reference relations spanning over several sentences, however, may cause invalid knowledge base structures to emerge so that incorrect information may be retrieved or inferred.

With the SYNDIKATE system family, we are addressing these shortcomings and aim at a more sophisticated level of knowledge acquisition from real-world texts. The source documents we deal with are currently taken from two domains, *viz.* test reports from the information technology domain (ITSYNDIKATE [10]) and medical finding reports, the framework of the MEDSYNDIKATE system [11]. MEDSYNDIKATE is designed to acquire from each input text a maximum number of simple facts ("The *findings correspond to* an *adenocarcinoma*."), complex propositions ("*All* mucosa layers show an inflammatory infiltration that mainly consists of lymphocytes."), and evaluative assertions ("The findings correspond to a *severe chronical* gastritis."). Hence, our primary goal is to extract conceptually deeper and inferentially richer forms of relational information than that found by state-of-the-art IE systems. Also, rather than restricting natural language processing intentionally to few templates, we here present an open system architecture for knowledge extraction where text understanding is constrained only by the unpredictable limits of available knowledge sources, the domain ontology, in particular.

To achieve this goal, several requirements with respect to language processing proper have to be fulfilled. As most of the IE systems, we require our parser to be robust to underspecification and ill-formed input (cf. the protocols in [12]). Unlike almost all of them, our parsing system is particularly sensitive to the treatment of textual reference relations as established by various forms of anaphora [13]. Furthermore, since SYNDIKATE systems rely on a knowledge-rich infrastructure, particular care is taken to provide expressive knowledge repositories on a larger scale. We are currently exploring two approaches. First, we automatically enhance the set of already given knowledge templates through incremental concept learning routines [14]. Our second approach makes use of the large body of knowledge that has already been assembled in medical taxonomies and terminologies (e.g., the UMLS). That knowledge is automatically transformed into a description logics format and, after interactive debugging and refinement, integrated into a comprehensive medical knowledge base [15].

2 System Architecture

In the following, major design issues for MEDSYNDIKATE are discussed, with focus on the distinction between sentence-level and text-level analysis. We will then turn to two alternative ontology engineering methodologies satisfying the need for the (semi)automatic supply of large amounts of background knowledge.

The overall architecture of SYNDIKATE is summarized in Figure 1. The general task of any SYNDIKATE system consists of mapping each incoming text, T_i, into a

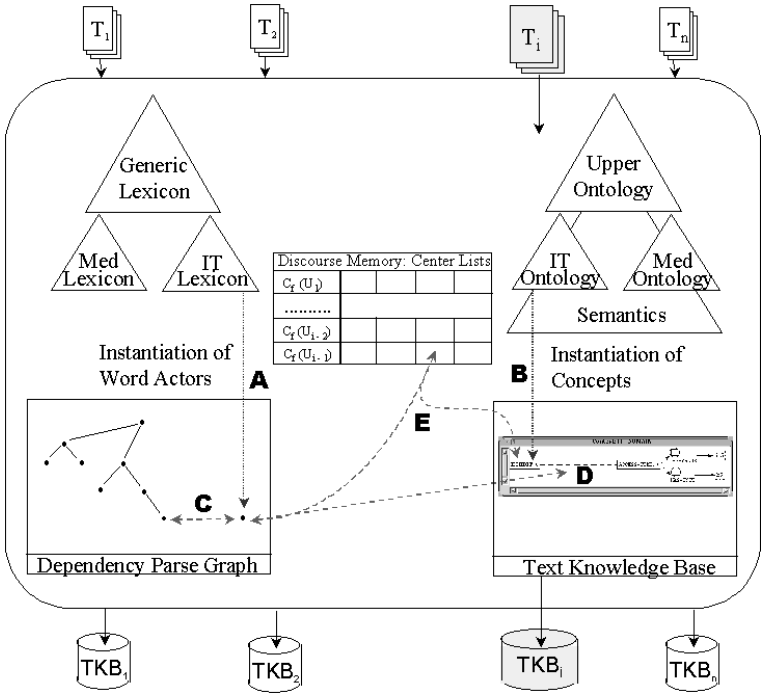

Figure 1: System Architecture of SYNDIKATE

corresponding *text knowledge base*, TKB_i, which contains a formal representation of T_i's content. This knowledge will be exploited by various information services, such as inferentially supported fact retrieval or text summarization.

2.1 Sentence-Level Understanding

Grammatical knowledge for syntactic analysis resides in a fully lexicalized dependency grammar (cf. [12] for details), we refer to as *Lexicon* in Figure 1. Basic word forms (lexemes) constitute the leaf nodes of the lexicon tree, while grammatical generalizations from lexemes appear as lexeme class specifications at different levels of abstraction. The *Generic Lexicon* in Figure 1 contains entries which are domain-independent (such as *move*, *with*, or *month*), while domain-specific extensions are kept in specialized lexicons serving the needs of particular subdomains, e.g., IT (*notebook, hard disk*, etc.) or medicine (*adenocarcinoma*, *gastric mucosa*, etc.).

Conceptual knowledge is expressed in a KL-ONE-like representation language (cf.[11] for details). These languages support the definition of complex concept descrip-

tions by means of conceptual roles and corresponding role filler constraints which introduce type restrictions on possible fillers. Taxonomic reasoning can be defined as being primitive (following explicit links), or it can be realized by letting a classifier engine compute subsumption relations between complex conceptual descriptions. A distinction is made between concept classes (types) and instances (representing concrete real-world entities). Most lexemes (except, e.g., pronouns, prepositions) are directly associated with one (or, in case of polysemy, several) concept type(s). Accordingly, when a new lexical item is read from the input text, a dedicated process (word actor) is created for lexical parsing (step A in Figure 1), together with an instance of the lexeme's concept type (step B). Each word actor then negotiates dependency relations by taking syntactic constraints from the already generated dependency tree into account (step C), as well as conceptual constraints supplied by the associated instance in the domain knowledge (step D) [12]. As with the *Lexicon*, the ontologies we provide are split up between one that serves all applications, the *Upper Ontology*, while specialized ontologies account for the conceptual structure of particular domains, e.g., information technology (NOTEBOOK, HARD-DISK, etc.), or medicine (ADENOCARCINOMA, GASTRIC-MUCOSA, etc.).

Semantic knowledge is concerned with determining relations between instances of concept classes based on the interpretation of so-called minimal *semantically interpretable subgraphs* of the dependency graph. Such a subgraph is bounded by two content words (nouns, verbs, adjectives) which may be directly linked by a single dependency relation or indirectly by a sequence of dependency relations linking non-content words only (e.g., prepositions, auxiliaries). Hence, a conceptual relation may either be constrained by dependency relations (e.g., the *subject:* relation may only be interpreted conceptually in terms of AGENT or PATIENT roles), by intervening non-content words (e.g., some prepositions impose special role constraints, such as *"with"* does in terms of HAS-PART or INSTRUMENT roles), or it may only be constrained by conceptual compatibility between the concepts involved (e.g., for genitives) [16]. The specification of semantic knowledge shares many commonalities with domain knowledge. Hence, the overlap in Figure 1.

2.2 Text-Level Understanding

The proper analysis of textual phenomena prevents inadequate text knowledge representation structures to emerge in the course of sentence-centered analysis [17]. Consider the following text fragment:

(1) Der Befund entspricht einem hochdifferenzierten *Adenokarzinom*.
(The findings correspond to a highly differentiated *adenocarcinoma*.)

(2) *Der Tumor* hat einen Durchmesser von 2 cm.
(*The tumor* has a diameter of 2 cm.)

With purely sentence-oriented analyses, *invalid* knowledge bases are likely to emerge, when each entity which has a different denotation at the text surface is treated as a formally distinct item at the symbol level of knowledge representation, although different denotations refer literally to the same conceptual entity. This is the case for *nominal anaphora*, an example of which is given by the reference relation between the noun phrase *"Der Tumor" (the tumor)* in Sentence (2) and *"Adenokarzinom" (adenocarcinoma)* in Sentence (1). A false referential description appears in Figure 2, where TUMOR.2-05 is introduced as a new representational entity, whereas Figure 3 depicts the adequate, intended meaning at the conceptual representation level, *viz.* maintaining ADENOCARCINOMA.6-04 as the proper referent.

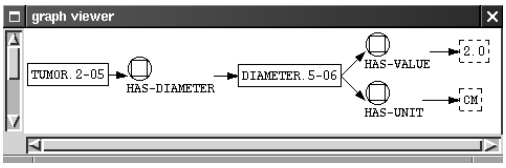

Figure 2: Unresolved Nominal Anaphora

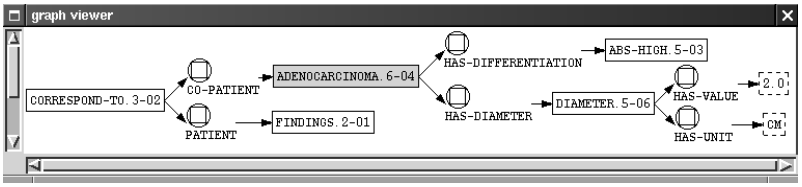

Figure 3: Resolved Nominal Anaphora

The methodological framework for tracking such reference relations at the text level is provided by *center* lists [13] (cf. step E in Figure 1). The ordering of their elements indicates that the most highly ranked element is the most likely antecedent of an anaphoric expression in the subsequent utterance, while the remaining elements are ordered according to decreasing preference for establishing referential links.

S_1	[FINDINGS.2-01: Befund, ADENOCARCINOMA.6-04: Adenokarzinom]
S_2	[ADENOCARCINOMA.6-04: Tumor, DIAMETER.5-06: Durchmesser, CM: cm]

Table 1: Center Lists for Sentences (1) and (2)

In Table 1, the tuple notation takes the conceptual correlate of each noun in the text knowledge base in the first place, while the lexical surface form appears in second place. Using the center list of Sentence (1) for the interpretation of Sentence (2) results in a series of queries whether FINDINGS is conceptually more special than TUMOR (answer: No) or ADENOCARCINOMA is more special than TUMOR (answer:

Yes). As the second center list item for S_1 fulfils all required constraints (mainly the one that ADENOCARCINOMA IS-A TUMOR), in the conceptual representation structure of Sentence (2), TUMOR.2-05, the literal instance (cf. Figure 2), is replaced by ADENOCARCINOMA.6-04, the referentially valid identifier (cf. Figure 3). As a consequence, instead of having two unlinked sentence graphs for Sentence (1) and (2) (e.g., cf. Figure 2) the reference resolution for nominal anaphora leads to joining them in a single coherent and valid text knowledge graph in Figure 3.

Given a fact retrieval application, the validity of text knowledge bases becomes a crucial issue. Disregarding textual phenomena will cause dysfunctional system behavior in terms of incorrect answers. This can be illustrated by a query `Q` such as

```
Q :  (retrieve ?x (Tumor ?x))
A-:  (Tumor.2-05, Adenocarcinoma.6-04)
A+:  (Adenocarcinoma.6-04)
```

which triggers a search for all instances in the text knowledge base that are of type TUMOR. Given an invalid knowledge base (cf. Figure 2), the incorrect answer (A-) contains two entities, *viz.* TUMOR.2-05 and ADENOCARCINOMA.6-04, since both are in the extension of the concept TUMOR. If, however, a valid text knowledge base such as the one in Figure 3 is given, only the correct answer, ADENOCARCINOMA.6-04, is inferred (A+).

2.3 Ontology Engineering

MEDSYNDIKATE requires a knowledge-rich infrastructure both in terms of grammar and domain knowledge, which can hardly be maintained by human efforts alone. Rather a significant amount of knowledge should be generated automatically. For SYNDIKATE systems, we have chosen a dual strategy. One focuses on the incremental learning of new concepts while understanding the texts, the other is based on the reuse of available comprehensive (though semantically weak) knowledge sources.

Concept Learning from Text. Extending a given core ontology by new concepts as a by-product of the text understanding process builds on two different sources of evidence — the already given domain knowledge, and the grammatical constructions in which unknown lexical items occur in the source document.

The parser yields information from the grammatical constructions in which lexical items occur in terms of the labellings in the dependency graph. The kinds of syntactic constructions in which unknown words appear are recorded and later assessed relative to the credit they lend to a particular hypothesis. Typical linguistic indicators that can be exploited for taxonomic integration are, e.g., appositions *('the symptom @A@'*, with '@A@' denoting the unknown word) or exemplification phrases *('symptoms like @A@')*, These constructions almost unequivocally determine '@A@' when considered as a medical concept to denote an instance of a SYMPTOM.

The conceptual interpretation of parse trees involving unknown words in the text knowledge base leads to the derivation of concept hypotheses, which are further enriched by conceptual annotations. These reflect structural patterns of consistency, mutual justification, analogy, etc. relative to already available concept descriptions in the ontology or other concept hypotheses. Grammatical and conceptual evidence of this kind, in particular their predictive "goodness" for the learning task, are represented by corresponding sets of linguistic and conceptual quality labels. Multiple concept hypotheses for each unknown lexical item are organized in terms of hypothesis spaces, each of which holds alternative or further specialized conceptual readings. An inference engine coupled with the classifier, the so-called quality machine, estimates the overall credibility of single concept hypotheses by taking the available set of quality labels for each hypothesis into account (cf. [14] for details).

Reengineering Medical Terminologies. The second approach makes use of the large body of knowledge that has already been assembled in comprehensive medical terminologies such as the UMLS [18]. The knowledge they contain, however, cannot be applied directly to MEDSYNDIKATE, because it is characterized by inconsistencies, circular definitions, insufficient depth, gaps, etc., and the lack of an inference engine.

The methodology for reusing weak medical knowledge consists of four steps [15]. First, we create automatically KL-ONE-style logical expressions by feeding a generator with data directly from the UMLS, i.e., the concepts and the semantic links between concept pairs. In a second step, the imported concepts, already in a logical format, are submitted to the classifier of the knowledge representation system (in our case, LOOM) in order to check whether the terminological definitions are consistent and non-circular. For those elements which are inconsistent, their validity is restituted and definitional circles are removed manually by a medical domain expert. In the final step the knowledge base which has emerged so far is manually rectified and refined (e.g., by checking the adequacy of taxonomic and partonomic hierarchies).

3 Evaluating Knowledge Extraction Performance

3.1 Evaluation Framework

In quantitative terms, SYNDIKATE is neither a toy system nor a monster. The *Generic Lexicon* currently includes 5,000 entries, while the *MED Lexicon* contributes 3,000 entries. Similarly, the *Upper Ontology* contains 1,500 concepts and roles, to which the *MED Ontology* adds 2,500 concepts and roles. However, recent experiments with reengineering the UMLS have resulted in a very large medical knowledge base with 164,000 concepts and 76,000 relations [15] that is currently under validation.

We extracted the text collection from the hospital information system of the University Hospital in Freiburg (Germany). All finding reports in histopathology

from the first quarter of 1999 were initially included, altogether 4,973 documents. However, for the time being MEDSYNDIKATE covers especially the subdomain of gastro-intestinal diseases. Thus, 752 texts out of these 4,973 were extracted semi-automatically in order to guarantee a sufficient coverage of domain knowledge. From this collection, a random sample of 90 texts was taken and divided into two sets. 60 of them served as the training set which was used for parameter tuning of the system. The remaining 30 texts were then used to measure the performance of the MEDSYNDIKATE system with unseen data. The configuration of the system was frozen prior to analyzing the test set.

In the empirical study proper, three basic settings of dependency graphs were evaluated, *viz.* ones containing genitives, prepositional phrases, as well as constructions including modal verbs or auxiliaries. Genitives and prepositional phrases relate fundamental biomedical concepts via associated roles at the conceptual level. Modal and auxiliary verbs create a complex syntactic environment for the interpretation of verbs, and, hence, the conceptual representation of medical processes and events. For each instance of these configurations semantic interpretations were automatically computed the result of which was judged for accuracy by two skilled raters.

Still, the way how a (gold) standard for semantic interpretation can be set up is an issue of hot debates [19]. In fact, conceptually annotated medical text corpora do not exist at all, at least for the German language. At this level, the ontology we have developed eases judgements, since it is based on a fine-grained relation hierarchy with clear sortal restrictions for role fillers. In anatomy, e.g., we use relations such as ANATOMICAL-PART-OF, which is itself a subrelation of PHYSICAL-PART-OF and PART-OF, and specialize it in order to account for subtle PART-OF relationships. A very specific relation such as ANATOMICAL-PART-OF-MUCOSA refers to a precise subset of entities to be related by the interpretation process. Therefore, relating BRAIN to MUCOSA by ANATOMICAL-PART-OF-MUCOSA obviously would be considered as incorrect, whereas relating LAMINA-PROPRIA-MUCOSAE would be considered a reasonable interpretation.

3.2 Quantitative Analysis

The following tables contain data for both the training and the test set indicating the quality of knowledge extraction as obtained for the three different syntactic settings. Besides providing data for recall and precision, the tables are divided into two assessment layers: *"without interpretation"* means that the system was not able to produce an interpretation because of specification gaps, i.e., at least one of the two content words in a minimal dependency graph under consideration was not specified. Note that even for the training set which was intended to generate optimal results we were unable to formulate reasonable and generally valid concept definitions for some of the

content words we encountered (e.g., for fuzzy expressions of locations: *"In der Tiefe der Schleimhaut"* (*"In the depth of the mucosa"*)). The second group *"with interpretation"* is divided into four categories. The label *correct (non-ambiguous)* qualifies, if just a single and correct conceptual relation was computed by the semantic interpretation process. However, if the result was correct but yielded more than one conceptual relation, the label *correct (ambiguous)* was assigned. An interpretation was considered *incorrect* when the conceptual relation was inappropriate. Finally, *NIL* was used to indicate that an interpretation was performed (both concepts for the content words were specified) but no conceptual relation could be computed.

Genitives. In the medical domain, as indicated by Table 2 the recall and precision values for the interpretation of genitives are very encouraging both for the training set (92% and 93%) and the test set (93% and 93%), respectively.[a] However, since genitives, in general, provide no additional constraints how the conceptual correlates of the two content words involved can be related, the number of ambiguous interpretations amounts to 13% and 36%, respectively.

	Training Set	Test Set
Recall	92%	93%
Precision	93%	93%
# occurrences ...	168	91
... **with interpretation**	158 (94%)	86 (95%)
[confidence intervals]	[89%-97%]	[90%-98%]
..... correct (non-ambiguous)	125.5 (75%)	48.5 (53%)
..... correct (ambiguous)	22 (13%)	33 (36%)
..... incorrect	6.5	3.5
..... NIL	4	1
... **without interpretation**	10 (6%)	5 (5%)

Table 2: Evaluation of Genitives

Auxiliaries and Modals. Table 3 contains the results for modal verbs or auxiliaries. A semantic interpretation of modal/auxiliary verb complexes relates a content-bearing verb with the conceptual correlate of the syntactic subject. In case of a passive construction the direct-object-to-subject normalization has to be carried out.

Recall and precision for the training set are high (94% and 98%, respectively) and, therefore, indicate that semantic interpretation can cope with almost all occurrences given an optimal degree of specification. The values for recall and precision dropped to 80% and 84%, respectively, in the test set. The increase of *NIL* results reveals that the granularity of the underlying domain model is insufficient as far as conceptual relations are concerned. Although the corresponding concepts are modelled, no conceptual relation between them could be determined.

[a] Confidence intervals for .95 probability are given in square brackets.

	Training Set	Test Set
Recall	94%	80%
Precision	98%	84%
# occurrences ...	131	55
... **with interpretation**	125 (95%)	52 (95%)
[confidence intervals]	[92%-99%]	[84%-99%]
..... correct (non-ambiguous)	122 (93%)	43,5 (79%)
..... correct (ambiguous)	1	0
..... incorrect	0	0,5
..... NIL	2	8 (15%)
... **without interpretation**	6 (5%)	3 (5%)

Table 3: Evaluation of Modal Verbs and Auxiliaries

Prepositional phrases (PPs) are crucial for the semantic interpretation of a text, since they introduce a wide variety of conceptual relations, such as spatial, temporal, causal, or instrumental ones. The importance of PPs is reflected by their relative frequency. In the training set and the test set, we encountered 1,108 prepositions, which is a little bit less than 10% of the words in both sets (approximately 11,300).[b] Provided also that the preposition's syntactic head and its modifier participate in the interpretation, at the phrase level, more than 25% of the texts' contents is encoded by PPs (certainly, this data also reflects a considerable degree of genre dependency).

	Training Set	Test Set
Recall	85%	85%
Precision	79%	81%
# occurrences ...	562	278
... **with interpretation**	548 (98%)	253 (91%)
[confidence intervals]	[96%-99%]	[86%-93%]
..... correct (non-ambiguous)	401,5 (71%)	167 (60%)
..... correct (ambiguous)	32,5 (6%)	37,5 (13%)
..... incorrect	43 (8%)	30,5 (11%)
..... NIL	71 (13%)	18 (6%)
... **without interpretation**	14 (2%)	25 (9%)

Table 4: Evaluation of Prepositional Phrases

Considering the results for semantic interpretation of PPs (cf. Table 4), the values for recall and precision are almost the same for the training set and the test set. Recall climaxed at 85% for both the training set and the test set, whereas precision reached 79% for the training set and 81% for the test set. Getting almost the same performance

[b] Only 940 of these 1,108 were included in the empirical analysis, since 168 did not form a minimal subgraph. Phrases like *"zum Teil"* (*"partly"*) map to a single meaning — as evidenced by the English translation correlate — and were therefore excluded.

for both sets also reveals a stable level of semantic interpretation of PPs.[c]

4 Conclusions

We have introduced MEDSYNDIKATE, a system for mining knowledge from biomedical reports. Emphasis was put on the role of various knowledge sources required for 'deep' text understanding. When turning from sentence-level to text-level analysis, we considered representational inadequacies when text phenomena were not properly accounted for and, hence, proposed a solution based on centering mechanisms.

The enormous knowledge requirements posed by our approach can only be reasonably met when knowledge engineering does not rely on human efforts only. Hence, a second major issue we have focused on concerns alternative ways to support knowledge acquisition and guarantee, this way, a reasonable chance for scalability of the system. We made two proposals. The first one deals with an automatic concept learning methodology that is fully embedded in the text understanding process, the other one exploits the vast amounts of medical knowledge assembled in various knowledge repositories such as the UMLS. We, finally, provided empirical data which characterizes the knowledge extraction performance of MEDSYNDIKATE in terms of three major syntactic structures, *viz.* genitives, modals and auxiliaries, and prepositional phrases. These reflect, at the linguistic level, fundamental categories of biomedical ontologies — states, processes, and actions.

References

1. D. B. Searls. String variable grammar: A logic grammar formalism for the biological language of DNA. *Journal of Logic Programming*, 24(1/2):73–102, 1995.
2. S. Leung, C. Mellish, and D. Robertson. Basic gene grammars and DNA-ChartParser for language processing of *Escherichia coli* promoter DNA sequences. *Bioinformatics*, 17(3):226–236, 2001.
3. F. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi. Toward information extraction: Identifying protein names from biological papers. In *PSB 98 – Proceedings of the 3rd Pacific Symposium on Biocomputing*, pages 705–716. Maui, Hawaii, 4-9 January, 1998.
4. N. Collier, C. Nobata, and J. Tsujii. Extracting the names of genes and gene products with a hidden Markov model. In *COLING 2000 – Proceedings of the 18th International Conference on Computational Linguistics*, pages 201–207. Saarbrücken, Germany, 31 July - 4 August, 2000.

[c] Corresponding data for an alternative test scenario, knowledge extraction from information technology (IT) test reports, is not as favorable as for the medical domain. For PPs, 70% / 64% recall and 77% / 66% precision were measured for the training set and the test set, respectively. A reasonable argument why we achieved better results in the medical domain than in the IT world might be that in the medical texts a considerably lower degree of linguistic variation is encountered.

5. T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi. Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, 17(2):155–161, 2001.
6. M. Craven and J. Kumlien. Constructing biological knowledge bases by extracting information from text sources. In *Proc. of the 7th Intl. Conference on Intelligent Systems for Molecular Biology*, pages 77–86. Heidelberg, Germany, August 6-10, 1999.
7. C. Blaschke, M. A. Andrade, C. Ouzounis, and A. Valencia. Automatic extraction of biological information from scientific text: Protein-protein interactions. *Intelligent Systems for Molecular Biology*, 7:60–67, 1999.
8. K. Humphreys, G. Demetriou, and R. Gaizauskas. Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structures. In *PSB 2000 – Proceedings of the 5th Pacific Symposium on Biocomputing*, pages 502–513. Honolulu, Hawaii, USA, 4-9 January, 2000.
9. T. Rindflesch, J. Rajan, and L. Hunter. Extracting molecular binding relationships from biomedical text. In *ANLP 2000 – Proceedings of the 6th Conference on Applied Natural Language Processing*, pages 188–195. Seattle, WA, USA, April 29 - May 4, 2000.
10. U. Hahn and M. 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.
11. U. Hahn, M. Romacker, and S. Schulz. How knowledge drives understanding: Matching medical ontologies with the needs of medical language processing. *Artificial Intelligence in Medicine*, 15(1):25–51, 1999.
12. U. Hahn, N. Bröker, and P. Neuhaus. Let's PARSETALK: Message-passing protocols for object-oriented parsing. In H. Bunt and A. Nijholt, editors, *Advances in Probabilistic and Other Parsing Technologies*, pages 177–201. Dordrecht, Boston: Kluwer, 2000.
13. M. Strube and U. Hahn. Functional centering: Grounding referential coherence in information structure. *Computational Linguistics*, 25(3):309–344, 1999.
14. U. Hahn and K. Schnattinger. Towards text knowledge engineering. In *AAAI'98 – Proceedings of the 15th National Conference on Artificial Intelligence*, pages 524–531. Madison, Wisconsin, July 26-30, 1998.
15. S. Schulz and U. Hahn. Knowledge engineering by large-scale knowledge reuse: Experience from the medical domain. In *Proc. 7th Intl. Conf. on Principles of Knowledge Representation and Reasoning*, pages 601–610. Breckenridge, CO, April 12-15, 2000.
16. M. Romacker, S. Schulz, and U. Hahn. Streamlining semantic interpretation for medical narratives. In *AMIA'99 – Proceedings of the Annual Symposium of the American Medical Informatics Association*, pages 925–929. Washington, D.C., Nov. 6-10, 1999.
17. U. Hahn, M. Romacker, and S. Schulz. Discourse structures in medical reports – watch out! The generation of referentially coherent and valid text knowledge bases in the MEDSYNDIKATE system. *International Journal of Medical Informatics*, 53:1–28, 1999.
18. *Unified Medical Language System*. Bethesda, MD: National Library of Medicine, 2001.
19. P. Zweigenbaum, J. Bouaud, B. Bachimont, J. Charlet, and J.-F. Boisvieux. Evaluating a normalized conceptual representation produced from natural language patient discharge summaries. In *AMIA'97 – Proceedings of the 1997 AMIA Annual Fall Symposium (formerly SCAMC).*, pages 590–594. Nashville, TN, October 25-29, 1997.