Automatic Knowledge Acquisition from Medical Texts

Udo Hahn, Klemens Schnattinger & Martin Romacker
Text Knowledge Engineering Lab
Freiburg University, Germany

An approach to knowledge-based understanding of realistic texts from the medical domain (viz. findings of gastro-intestinal diseases) is presented. We survey major methodologicalfeatures of an object-oriented, fully lexicalized, dependency-based grammar model which is tightly linked to domain knowledge representations based on description logics. The parser adheres to the principles of robustness, incrementality and concurrency. The substrate of automatic knowledge acquisition are text knowledge bases generated by the parser from medical narratives, which represent major portions of the content of these documents.

INTRODUCTION

Information contained in medical free texts (such as patient reports or discharge summaries) is relevant to a variety of different retrieval, coding and inference purposes. It should, for instance, provide support for medical decision making, for mapping data into medical coding systems or for quality assurance of medical treatment. With the growing availability of medical documents in machine-readable form, procedures for automatically analyzing and formatting textual data gain more and more importance, because hand-coding and manual indexing are time-consuming and usually error-prone. Therefore, automatic knowledge acquisition from medical free texts is highly desirable for hospital data management.

We currently port a German language text understanding system originally developed for the domain of information technology to the medical domain. In this paper, we will outline the basic methodological principles of our approach and illustrate them briefly with an example taken from a sample medical text.

CHALLENGES OF MEDICAL TEXT UNDERSTANDING

Acquiring knowledge from realistic, routinely produced medical texts at a reasonable level of detail and accuracy requires to meet at least two fundamental challenges – a methodological and an engineering one. The *methodological* challenge consists of supplying an almost excessive amount of different kinds of knowledge, which are needed for proper text understanding. Of primary importance are grammar specifications capturing the linguistic knowledge – not only at the lexical (morphology), phrasal, clausal and sentential level (syntax, semantics), but also at the discourse level. By this, we mean phenomena of local as well as global coherence of texts as exemplified by (pro)nominal anaphora [1] or functional anaphora [2] as well as more global patterns of thematic progression in texts [3]. Unless these *discourse coherence* phenomena are adequately dealt with, referentially invalid and incohesive text knowledge bases are likely to emerge, thus seriously deteriorating the quality and subsequent usability of the acquired knowledge. Structural (i.e., syntax-based) linguistic descriptions, however, as sophisticated as they may be, are still not expressive enough for comprehensive inferencing and, thus, have to be backed up by conceptual knowledge of the underlying domain. This requires to specify a consistent, balanced *ontology*, with descriptions pertaining to major concepts (categories), their hierarchical and aggregational interrelations, and their incorporation into rule expressions to allow for various inference modes. Both levels are linked by mediating semantic representations at the linguistic level, which relate to conceptual specifications at the ontological (sometimes also called the encyclopedic) level.

The *engineering* challenge can be characterized by the observation that, at least in realistic text understanding scenarios, sufficiently complete and in-depth specifications at the linguistic and conceptual level are the exception rather than the rule. Hence, appropriate processing strategies have to be supplied to proceed with text understanding even in the presence of lacking or underspecified knowledge. Ideally, the degradation of the text understander's performance should be proportional to the degree of linguistic and conceptual underspecification encountered. Together with the observation that realistic text understanders also have to cope with ungrammatical (ill-formed) input, an inherent potential for *robustness* has to be built into these systems. Ignoring the quest for robust processing either leads to a huge knowledge engineering overhead (*viz.* advance manual coding of all lacking pieces of information) or fatal system behavior, *viz.* blockade at every occurrence of an unknown lexical item or an extra- or ungrammatical expression. Note also that this engineering challenge can easily be transferred into a methodological one by supplying text understanders with automatic learning facilities in order to actively account for novel input.

In the remainder of this paper, we elaborate in greater detail on our approach to text knowledge acquisition

0195—4210/96/$5.00 © 1996 AMIA, Inc.

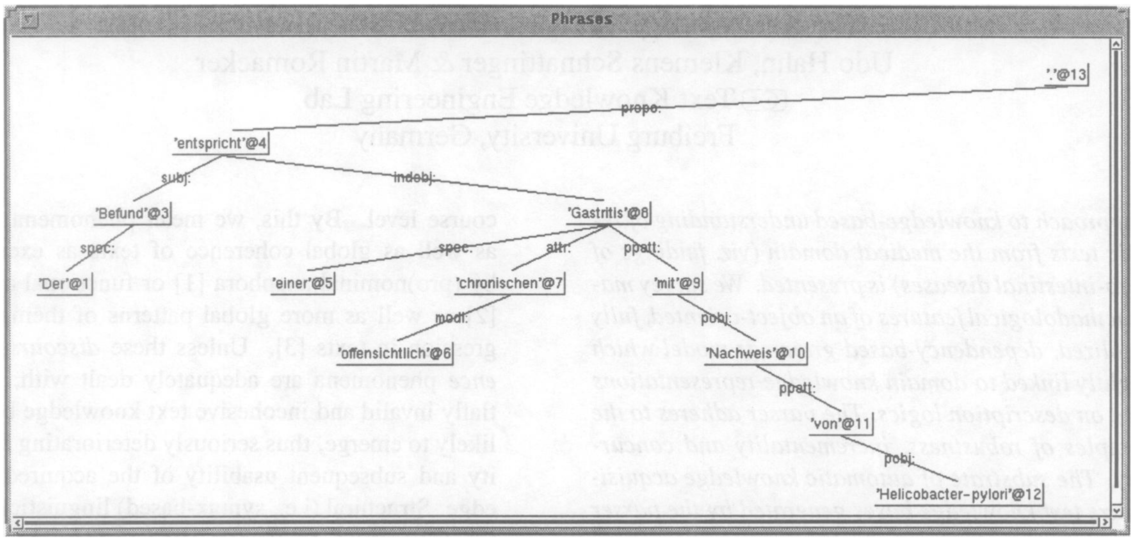

Figure 2: Sample Dependency Tree

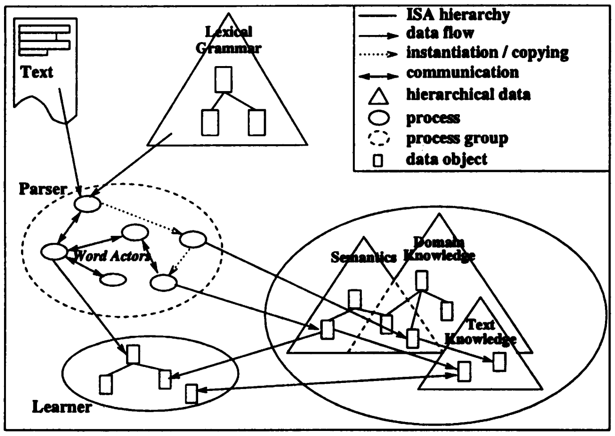

Figure 1: System Architecture

based on the above methodological principles. We present a model of automatic text understanding, with emphasis on the knowledge sources used and their underlying theoretical constructs. Procedural aspects of text parsing, e.g., robustness, preference and prediction strategies (cf. [4]), as well as issues relating to automatic concept learning (cf. [5]) are given almost no attention in this paper.

ARCHITECTURE OF THE TEXT UNDERSTANDING KERNEL SYSTEM

We here consider the basic knowledge sources and their organization in terms of the architecture (see Fig. 1) underlying the text understanding kernel system PARSETALK (for a survey, cf. [6]).

Linguistic knowledge comes with a completely lexicalized, head-oriented *grammar* system. Each lexical item is associated with specifications of its part of speech (e.g., noun, verb, preposition), morphosyntactic features (e.g., gender, case, mode, tense marking), word order constraints and a valency frame which specifies the potential modifiers the lexical item may govern as a head. Part of the valency description are semantic compatibility criteria (ako sortal constraints) between the head and each modifier.

The lexical distribution of declarative grammatical knowledge is complemented by the provision of lexicalized control knowledge. Knowledge of this sort is locally encoded in terms of scripts, i.e., collections of methods by which the behavior of a lexical parsing process is determined upon the reception of an incoming message. The object-oriented specification style of the lexical grammar accounts for the variety of control patterns between heterogeneous knowledge sources and the flexible switching of strategies to ensure efficient as well as robust system performance.

An ubiquitous feature of the PARSETALK system are various forms of inheritance hierarchies, e.g., for parts of speech, morphosyntactic features, by which the size and complexity of the grammar specifications can be kept manageable and consistent. The lexical items form the leave nodes of such hierarchies, while inner nodes contain linguistic generalizations based, e.g., on distribution patterns for parts of speech.

Viewed from the angle of the *parser*, whenever a lexical item is read from the textual input and it is identified as an entry in the lexicon, a lexical process, we refer to as *word actor*, is instantiated. Word actors combine lexicalized grammar knowledge and positional information from the text. They drive the parsing process by way of message passing using well-defined, linguistically motivated protocols (cf. [4] for a deeper account). Each of these protocols accounts for different functional relations lexical items may join, e.g.,

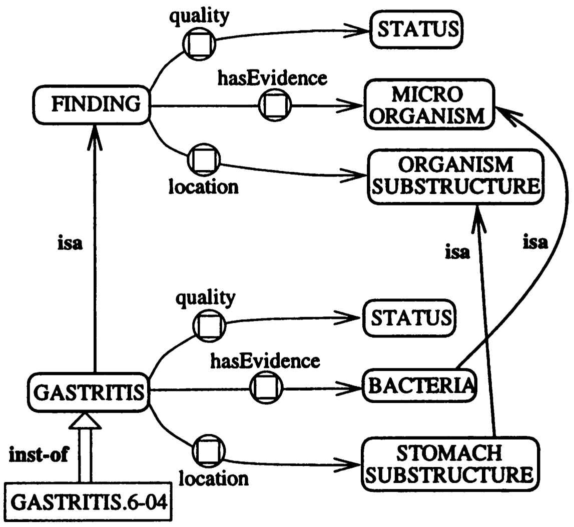

Figure 3: Fragment of the Concept Hierarchy

dependency relations [4], anaphoric relations [2], etc. The parsing process is carried out incrementally, concurrently and, due to efficiency reasons, basically incomplete (for a motivation, cf. [4]). The target structure produced by the interactions of word actors at the syntax level are dependency trees labeled with functional relations such as SUBJ (subject), INDOBJ (indirect object) (cf. Fig. 2).

The need for a *domain knowledge* base within a formally specified framework for natural language understanding is getting more and more accepted in the field of medical informatics (e.g. [7]). It requires medical knowledge to be expressed in a sound and consistent manner. As a consequence, conceptual taxonomies, for instance, can be computed from formalized concept descriptions rather than being predefined. Though knowledge engineering in an ontologically coherent (sub)domain should make use of existing knowledge compilations, e.g., ICD-10, SNOMED, MESH, UMLS, as much as possible, due to the lack of formal specification and an insufficient level of granularity only a limited amount of this knowledge can be reused even if it is judged transferable, in principle [8].

In our approach, we chose a representation formalism based on terminological logics to incorporate semantic and conceptual constraints in the parsing process. The domain knowledge base is built from concept definitions that are organized in an acyclic directed graph by subsumption (subconcept) relations. (For a survey of major properties of terminological languages, cf. [9].)

In the example given in Fig. 3, the concept FINDING subsumes the concept GASTRITIS. Both have the same set of roles, but the subsumption relation is inferred from the range restrictions for the role LOCATION (the concept ORGANISMSUBSTRUCTURE is specialized by the concept STOMACHSUBSTRUCTURE) and for HASEVIDENCE (the concept MICROORGANISM is specialized by the concept BACTERIA).

Linguistic and conceptual knowledge are linked in two ways. First, any valency specification contains a semantic compatibility check, whose evaluation is directly transferred to and executed in the domain knowledge base. Second, any semantic computation based on confirmed syntactic structures (e.g., the processing of intermediate structures for prepositional attachments, modal verbs, quantifiers or pronouns) is carried out in an autonomous context. Contexts are data structures which are fully embedded in the same terminological reasoning system we use for the representation and processing of domain knowledge. Since the coupling of the semantic and conceptual level of analysis is pretty tight, the overwhelming majority of interpretation processes can directly be executed in the domain knowledge base (more precisely, the *text knowledge* base which constitutes a text-specific instantiation of the original domain knowledge base; cf. Fig. 1). Only those semantic interpretation processes which require longer transactions are encapsulated at the semantic processing level. After termination, the results of these computations are immediately transferred to the text knowledge base.

A SAMPLE PARSE

In the following example, we focus on the interplay between syntactic analysis, semantic interpretation and conceptual reasoning. Consider the sentence: *"Der vorläufige Befund entspricht einer offensichtlich chronischen Gastritis mit Nachweis von Helicobacter-pylori." ("The preliminary findings correspond to an apparently chronic gastritis with evidence for helicobacter-pylori.")*. Syntactic analysis of the sentence produces a dependency structure as depicted in Fig. 2, where nodes correspond to word actors marked with an integer indicating their text position. Edges between nodes indicate an established dependency relation labeled by the corresponding relation tag. Note that due to a lacking lexical specification the text item *"vorläufige"* on position 2 has been skipped (cf. [4] for a corresponding protocol description), while due to lacking conceptual specifications the text item *"offensichtlich"* on position 6 though still covered by the parse will, finally, receive no conceptual interpretation. Fig. 4 contains the conceptual representation from the text knowledge base as generated for the sample sentence. Compared to the dependency tree, the concept graph looks by far "slimmer". This is due to the normalizing operation of semantic rules.

Our aim in semantic interpretation is to categorize on the epistemological status of concepts in order to avoid

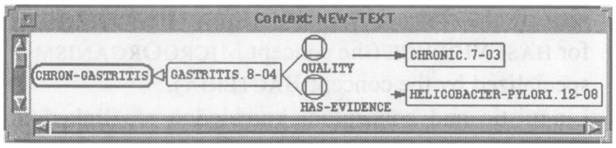

Figure 4: Sample Concept Graph

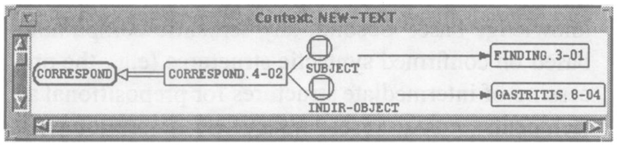

Figure 5: Sample Semantic Graph

an intractable number of interpretation rules. Therefore, all semantic interpretation rules are attached to categories which are inherited by concepts as with the conceptual counterpart CORRESPOND *("to correspond to")* of the surface lexeme *"entspricht"*. Furthermore, rules are attached to parts of speech that make exclusively part of a special semantic level (e.g., prepositions, predicative verbs). The semantic interpretation rule for CORRESPOND (cf. Fig. 6), for instance, establishes a co-reference relation between the fillers of the roles SUBJECT and INDIR-OBJECT at the conceptual level. Since (cf. Fig. 5) the filler of the role INDIR-OBJECT, *Gastritis*.8-04, specializes (is subsumed by) the filler of the role SUBJECT, *Finding*.3-01 (according to the ISA relation stated in Fig. 3), both can be merged by a recursive attribute unification algorithm (the function **subsumes+**). Furthermore, the verb instance *Correspond*.4-02 can be removed, because its "meaning" is already encoded in the structures resulting from the merge. An analogous interpretation rule handles the conceptual attachment of the instance *Helicobacter-pylori*.12-08 to the instance *Gastritis*.8-04 via the role HASEVIDENCE. Finally, the classifier of the terminological knowledge representation system assigns *Gastritis*.8-04 as an instance of the concept CHRON-GASTRITIS *("chronic gastritis")*. As a final result of the text understanding process, various linguistic surface expressions are mapped to canonical representation structures from the underlying domain knowledge base. Continuously proceeding this way, sentence by sentence, the text knowledge base gets incrementally augmented by the processes of conceptual interpretation and knowledge integration.

RELATED WORK

In the last few years, there has been a growing interest in developing text analysis system that yield semantic representations of medical narratives for further information access [10, 11, 12]. The foundations in this area were already laid in the seventies by the LINGUISTIC STRING PROJECT (LSP). LSP provides a fully developed, broad coverage medical language processing system [13]. Narrative texts are syntactically analyzed by taking semantic restrictions for possible constituent structures into consideration that are defined for the medical sublanguage. So-called *information formats* serve as templates for representing characteristic statement types for the sublanguage [14]. In the LSP system, however, no strict separation is made between syntactic knowledge and semantic/conceptual knowledge. Semantic restrictions for establishing syntactical relations are all *explicitly* precoded in the grammar rules and must therefore be entirely anticipated in the design phase of the system. By the use of terminological knowledge representation systems, these restrictions can be inferred at run-time from the concept definitions.

```
EXISTS v, s, o :
  v : CORRESPOND ⊓
  v SUBJECT s ⊓
  v INDIR-OBJECT o  ⟹
IF (subsumes+(s,o) OR subsumes+ (o,s) THEN
  TELL  s CO-REFERENT o
```

Figure 6: Sample Semantic Interpretation Rule

Baud *et al.* [15] derive a semantic representation of medical narratives by exploiting the conceptual relations of proximate words in a sentence using a simple pattern matching method for syntactic analysis. Based on the clustering of words in sentences, conceptual graphs are used as knowledge respresentation formalism [16] to determine the semantic relations. Although being very effective for the task the system is designed for, this approach is likely to run into problems when it has to deal with text phenomena and more complex sentence structures like coordination.

In the GALEN project [17] the task of knowledge acquisition from medical narratives has led to a division of the system into several logically independent components. While there are several source languages (English, German and French) to be parsed, the representational structures are considered language independent. A strict separation between the concept representation system (GRAIL) and the medical coding schemes is realized in GALEN. As in our approach, the KL-ONE-like knowledge representation language GRAIL prohibits syntactic structure building for semantically invalid statements on the basis of formally specified concept descriptions.

CONCLUSIONS

We sketched the major building blocks of a system for automatic knowledge acquisition from texts, which integrates heterogenous knowledge sources with different representation formalisms in its architecture. Emphasis was laid on the interplay between syntactic analysis, semantic interpretation and conceptual reasoning. Since we are dealing with "realistic" texts, ungrammatical and extragrammatical input must be processed. Hence, robustness turns out to be a primary issue for the design of realistic text understanding systems.

The knowledge acquisition prototype for German language medical texts we have implemented is fed with input from the medical documentation center of the Hospital of Freiburg University [18]. We have concentrated so far on the domain of gastro-intestinal diseases. The lexicon currently contains about 1,500 entries, the medical knowledge base consists of approximately 550 concepts and 350 roles. The PARSETALK system is implemented in an actor dialect of Smalltalk, while the knowledge base is implemented in LOOM [19]. The prototype so far has only been evaluated with respect to specialized functionalities, e.g., its potential for text coherence analysis [1] or parsing efficiency [4], while a comprehensive and focused investigation of its information system capabilities is still under way.

Acknowledgments

We like to thank the members of our group and Prof. Dr. Klar and Dr. Schulz from the Department of Medical Informatics at Freiburg University for fruitful discussions. We also gratefully acknowledge the provision of the LOOM system from USC/ISI.

References

1. M. Strube and U. Hahn. Functional centering. In *ACL'96 - Proc. 34th Annual Meeting of the Association for Computational Linguistics*, pages 270–277. Santa Cruz, California, June 23-28, 1996. San Francisco, CA: Morgan Kaufmann, 1996.
2. U. Hahn and M. Strube. PARSETALK about functional anaphora. In G. McCalla, (Ed.), *Advances in Artificial Intelligence. Proc. 11th Biennial Conf. of the Canadian Society for Computational Studies of Intelligence (AI'96)*, pages 133–145. Toronto, Ontario, Canada, May 21-24, 1996. Berlin: Springer, 1996. (LNAI, 1081).
3. U. Hahn. Topic parsing: accounting for text macro structures in full-text analysis. *Information Processing & Management*, 26(1):135–170, 1990.
4. P. Neuhaus and U. Hahn. Trading off completeness for efficiency: the PARSETALK performance grammar approach to real-world text parsing. In *FLAIRS'96 - Proc. 9th Florida Artificial Intelligence Research Symposium*, pages 60–65. Key West, Florida, May 20-22, 1996.
5. U. Hahn, M. Klenner, and K. Schnattinger. Learning from texts: a terminological metareasoning perspective. In S. Wermter, E. Riloff, and G. Scheler, (Eds.), *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pages 453–468. Berlin: Springer, 1996. (LNAI, 1040).
6. U. Hahn, S. Schacht, and N. Bröker. Concurrent, object-oriented dependency parsing: the PARSETALK model. *International Journal of Human-Computer Studies*, 41(1-2):179–222, 1994.
7. K. Campbell, A. Das, and Musen M. A logical foundation for representation of clinical data. *Journal of the American Medical Informatics Association*, 1(3):218–232, 1994.
8. G. Carenini and J. Moore. Using the UMLS semantic network as a basis for constructing a terminological knowledge base: a preliminary report. In *SCAMC'93 - Proc. 17th Annual Symposium on Computer Applications in Medical Care*, pages 725–729, 1993.
9. W. Woods and J. Schmolze. The KL-ONE family. *Computers & Mathematics with Applications*, 23(2-5):133–177, 1992.
10. C. Berrut. Indexing medical reports: the RIME approach. *Information Processing and Management*, 26(1):93–109, 1990.
11. F. Volot, P. Zweigenbaum, B. Bachimont, M. Ben Said, J. Bouaud, M. Fieschi, and J. Boisvieux. Structuration and acquisition of medical knowledge. Using UMLS in the conceptual graph formalism. In *SCAMC'93 - Proc. 17th Annual Symposium on Computer Applications in Medical Care*, pages 710–714, 1993.
12. P. Zweigenbaum, B. Bachimont, J. Bouaud, J. Charlet, and J. Boisvieux. A multi-lingual architecture for building a normalised conceptual representation from medical language. In *SCAMC'95 - Proc. 19th Annual Symposium on Computer Applications in Medical Care*, pages 357–361, 1995.
13. N. Sager, C. Friedman, and M. Lyman. *Medical Language Processing. Computer Management of Narrative Text.* Reading, MA: Addison-Wesley, 1987.
14. N. Sager, M. Lyman, C. Bucknall, N. Nhan, and L. Tick. Natural language processing and the representation of clinical data. *Journal of the American Medical Informatics Association*, 1(2):142–160, 1994.
15. R. Baud, A. Rassinoux, and J. Scherrer. Natural language processing and semantical representation of medical texts. *Methods of Information in Medicine*, 31(2):117–125, 1992.
16. J. Sowa. *Conceptual Structures: Information Processing in Mind and Machine.* Reading, MA: Addison-Wesley, 1984.
17. A. Rector and W. Nowlan. The GALEN project. *Computer Methods and Programs in Biomedicine*, 45(1-2):75–78, 1994.
18. R. Klar, A. Zaiß, A. Timmermann, and U. Schrade. The information system of the Freiburger University Hospital. In K. Adlassnig et al., (Ed.), *MIE - Proc. Medical Informatics Europe*, pages 46–50. Vienna, Austria, August 1991. Berlin: Springer, 1991.
19. R. MacGregor. A description classifier for the predicate calculus. In *AAAI'94 - Proc. 12th National Conf. on Artificial Intelligence.*, pages 213–220. Seattle, Wash., July 31 - August 4, 1994. Menlo Park: AAAI Press/MIT Press, 1994.