Text Structures in Medical Text Processing: Empirical Evidence and a Text Understanding Prototype

Udo Hahn & Martin Romacker
Text Knowledge Engineering Lab, Freiburg University
Werthmannplatz 1, D-79085 Freiburg, Germany

We consider the role of textual structures in medical texts. In particular, we examine the impact the lacking recognition of text phenomena has on the validity of medical knowledge bases fed by a natural language understanding front-end. First, we review the results from an empirical study on a sample of medical texts considering, in various forms of local coherence phenomena (anaphora and textual ellipses). We then discuss the representation bias emerging in the text knowledge base that is likely to occur when these phenomena are not dealt with — mainly the emergence of referentially incoherent and invalid representations. We then turn to a medical text understanding system designed to account for local text coherence.

INTRODUCTION

With the overall diffusion of electronic text processing technology in clinical offices and, more recently, the unlimited access to text resources in the Internet, a vast potential for medical information supply arises, at least in principle. The natural language processing community has responded to the urgent needs of real-world text processing in the medical domain by providing simple and robust analytical devices. These techniques usually exploit statistical methods, pattern-matching methodologies or finite-state techniques [1, 2]. An implicitly held assumption of these approaches is that medical texts (discharge summaries, findings reports, etc.) can be considered a sequence of phrases or sentences lacking any further *inter*dependencies. By the latter, we mean *local coherence* phenomena discussed in the natural language processing community under the heading of anaphora or reference relations, or, even more ambitious, *global coherence* phenomena which cover the entire macro organization of texts in terms of coherence realations and rhetorical structures (for a survey, cf. [3]). As a consequence, currently available analytic devices for medical text processing are actually sentence processors or even more restricted phrase processors which treat their input, sentences and phrases, in isolation only.

In this paper, we shall challenge this view. We stipulate that medical texts, as any other text sort, exhibit textual structures and that disregarding these structural relations will lead to underdetermined or even invalid content representations. To render support to our argument we conducted an empirical investigation of medical findings reports from a large clinical text database in order to assess whether this issue is really relevant. First, we will describe the experimental setting and elaborate on the quantitative distribution of various text phenomena in the sample. We will then turn to the consequences of not taking textual structures into account and show how referentially incoherent and referentially invalid text knowledge representation structures are likely to emerge. This is illustrated considering a small text fragment as analyzed by our system prototype, the Medical Knowledge SYNDIKATE. We shall focus on those aspects of the system design which account for the proper analysis of text phenomena. Up to now, such a functionality has to the best of our knowledge not been provided by any other system for medical text processing (for a survey, cf. [4]).

EMPIRICAL SETTING

In order to render substance to our claim that accounting for text structures is vital for medical text processing, we analyzed a randomly chosen sample of 103 reports on histological findings taken from the clinical information system of the University Hospital at Freiburg. These (German language) texts – a fragment of which appears below – deal with biopsy material for diagnosing gastro-intestinal diseases or leukemia.

(1) In einem Partikel mit 4 mm Durchmesser wurde eine Magenschleimhaut vom Antrumtyp erfaßt.
(A gastric mucous membrane of the antrum type was seized in a particle with a diameter of 4mm.)

(2) 1. *Sie* weist schwachgradig verlängerte Foveolen auf.
(*It* reveals slightly lengthened foveolas.)
2. Die intakte *Schleimhaut* weist schwachgradig verlängerte Foveolen auf.
(The intact *mucous membrane* reveals slightly lengthened foveolas.)

(3) Das ödematöse *Stroma* wird vermehrt von Lymphozyten infiltriert.
(The edematous *stroma* is increasingly infiltrated by lymphocytes.)

The total number of words amount to approximately 17,200, giving an average of 170 terms per document. Single texts range from a minimum of 25 up to a maximum of 650 words depending on the complexity of histological analyses and the severity of the findings. We considered the following types of text phenomena:

- ***Pronominal anaphors*** relate an anaphoric expression (pronouns such as *"it"* or *"they"*) to an antecedent in the preceding text by placing various morphosyntactic constraints on the antecedent

1091—8280/97/$5.00

(e.g., agreement in number or gender). Proper antecedent selection also incorporates semantic or conceptual compatibility constraints on a tentative referent given the context the anaphoric expression is embedded in. As an example, consider the relation between the anaphoric expression *"it"* and its corresponding antecedent *"gastric mucous membrane"* in sentences (2.1) and (1), respectively. Note that conceptual criteria are decisive, as they reveal that only *"gastric mucous membrane"* can be properly related to *"foveolas"*, while *"particle"*, another morphosyntactically admissible antecedent, cannot.

- ***Nominal anaphors*** relate an anaphoric expression (a noun or a noun phrase) to an antecedent in the preceding text by reducing the morphosyntactic constraints to number agreement. An additional syntactic constraint is imposed which requires the anaphoric expression to occur in a definite noun phrase. Also, a conceptual generalization constraint is added such that the anaphoric expression must be conceptually more general than the antecedent. As an example, consider the relation between the anaphoric expression *"the ... mucous membrane"* and its corresponding antecedent *"gastric mucous membrane"* in sentences (2.2) and (1), respectively.
- ***Textual ellipses*** relate a textelliptic expression (a noun or a noun phrase) to an antecedent in the preceding text by placing a conceptual constraint such that the elliptical expression is associated with its extrasentential antecedent by a suitable conceptual role. The missing conceptual link between those two discourse elements must be inferred in the domain knowledge base in order to establish the local coherence of the discourse. The only grammatical constraint left is that the textual ellipsis must occur in a definite noun phrase. As an example, consider the relation between the textelliptic expression *"the ... stroma"* and its corresponding antecedent *"gastric mucous membrane"* in sentences (3) and (2), respectively, via the conceptual role CONSISTS-OF, i.e., GASTRICMUCOUSMEMBRANE CONSISTS-OF STROMA. Note also that we implicitly assume that the (pro)nominal anaphors have already been resolved to guarantee proper reference resolution for subsequent utterances.

The results of the empirical study for medical texts are summarized in Table 1. As the Knowledge SYNDIKATE, prior to porting it to the medical domain, has originally been developed for analyzing information technology (IT) test reports, we have already gathered empirical data of the occurrence of textual phenomena in the IT test domain (details of these results are discussed in [5]). In IT texts, anaphora and textual ellipses occur at an almost balanced rate (we also gathered quantitative evidence that anaphora are the dominating textual phenomenon in newspaper and, in particular, in literary texts). The quantitative distribution of textual phenomena in the medical texts we investigated exhibits, however, an entirely surprising result. The data show that *textual ellipses* are the major glue for establishing local coherence in medical texts (two thirds of all textual phenomena), while anaphora, pronominal anaphora in particular, play a far less important role than in any other text sort. This is interesting insofar as the phenomenon of textual ellipsis, unlike the broad coverage of anaphoric phenomena, has only received marginal attention in the field of natural language processing so far (cf. [6] for a fully worked out algorithmic proposal). The immense importance of textual ellipsis and the remarkable ratio of nominal (17%) compared to extrasentential pronominal anaphora (3%) is clearly an indication of the primary orientation in medical texts to convey facts in a densely written manner presupposing a considerable degree of medical background knowledge. Stylistic criteria, mainly the source of using pronominal anaphora, have far less impact. The residual category (Rest) with about 13% of the text phenomena incorporates rather complicated cases of referential coherence phenomena such as plural anaphora, reference to (sub)sets, etc. for which no conclusive procedure has been found so far. A major portion of these phenomena, slightly more than 3% of the data, is constituted by metonymies, a special form of figurative speech, for which we have already worked out a resolution algorithm (cf. [7] for more details).[1] It is interesting to note, however, that the longer the medical texts grow, the more likely is the usage of pronominal anaphors and the incorporation of metonymies (clearly, a tribute to the increasing degree of "textuality" in these longer texts).

Summing up, local text coherence structures are frequent phenomena in medical texts. About 3% of all terms directly relate to textual phenomena and on the range of 5 to 10% of the terms in a text are involved,

[1] An expression A is considered a metonymy, if A deviates from its "standard denotation" (often causing a sortal conflict which gives rise to some kind of type coercion) in that it stands for an entity B which is not expressed explicitly but is conceptually related to A via a (usually conventionalized) conceptual relation r. As an example, consider the following succession to the text fragment (1) to (3):
Die weitere endoskopische und bioptische Kontrolle des *Patienten* ist angezeigt. (Further endoscoptic and bioptic screening of the *patient* is required.)
After resolving the whole-for-part metonymy, which holds for PATIENT ($= A$) and STOMACH ($= B$) via the role r = HASORGAN, and relating STOMACH and GASTRICMUCOUSMEMBRANE via HASPART-like roles local coherence is, finally, established.

Text	Pronominal Anaphora		Nominal Anaphora		Textual Ellipsis		Rest		Total
Gastritis	4	2.7%	46	31.3%	72	49.0%	25	17.1%	147
Leukemia	9	3.2%	29	10.1%	216	75.5%	32	11.2%	286
Total	13	3.0%	75	17.3%	288	66.5%	57	13.1%	433

Table 1: Distribution of Textual Phenomena

as textual structures are usually embodied by noun phrases.

THE SYNDIKATE SYSTEM

The basic architecture of the text understanding system that implements the local coherence recognition facilities needed to account for the text phenomena just discussed has already been described at the 1996 AMIA Fall Symposium [8]. Its natural language processing kernel, the PARSETALK system, is composed of a fully lexicalized, head-oriented dependency grammar [9] and an associated object-oriented, concurrent parser. Main features of this kernel system are robustness with respect to ungrammatical and extragrammatical input and a partial understanding performance depending on the depth and breadth of the knowledge sources made available [10]. The domain knowledge is represented using a hybrid terminological approach, the LOOM system [11], in order to provide appropriate semantic and conceptual constraints for the text understanding processes. A quite unique feature of the Medical Knowledge SYNDIKATE is constituted by its learning facilities which are tightly integrated with the parsing process [12]. Given the exorbitant size of medical sublanguages needed in a real-world text processing environment (conservative estimates range on the order of several millions of terms), hand-coding is clearly precluded. Available broad-band ontologies such as ICD-10, SNOMED, MeSH or UMLS, on the other hand, may possibly serve as a high-level ontological "grid". But given the strong requirements concerning the specificity of terminological knowledge structures required for deep text understanding these ontologies are neither detailed nor formally explicit enough to be integrated on-the-fly for the purpose of sophisticated text understanding [13]. By this, we mean, e.g., text ellipsis resolution which strongly depends on the availability of rich sets of conceptual relations for each concept. Hence, our focus on learning methodologies which allow for the acquisition of new concepts, roles, role fillers, and role filler constraints as text understanding incrementally proceeds.

The Centering Model for Local Coherence

In order to account for the text phenomena introduced in the previous section, our text analysis module is based on the *centering* model [14]. This approach offers a methodology for ranking possible antecedents on a theoretically justified preference scale. The centering model is intended to capture the local coherence of discourse by considering the coherence among the utterances in a particular discourse segment (say, a paragraph of a text). Local coherence is opposed to global coherence, i.e., coherence with other segments in the discourse. Discourse entities serving to link one utterance to other utterances in a particular discourse segment are organized in terms of centers. Each utterance U_i in a discourse segment is assigned a set of *forward-looking centers*, $C_f(U_i)$, and a unique *backward-looking center*, $C_b(U_i)$. The forward-looking centers of U_i depend only on the expressions that constitute that utterance, previous utterances provide no constraints on $C_f(U_i)$. The elements of $C_f(U_i)$ are partially ordered to reflect relative prominence in U_i. The most highly ranked element of $C_f(U_i)$ that is *realized* in U_{i+1} (i.e., is associated with an expression that has a valid interpretation in the underlying semantic/conceptual representation language) is the $C_b(U_{i+1})$. The ranking imposed on the elements of the C_f reflects the assumption that the most highly ranked element of $C_f(U_i)$ is the most preferred possible antecedent of an anaphoric expression in U_{i+1}, while the remaining elements are (partially) ordered according to decreasing preference for establishing referential links. The ordering in centers we base our analysis on incorporates *functional* information structure considerations, i.e., notions such as given *vs.* new information, theme (topic) *vs.* rheme (comment) [5], and has recently been extended to account for global reference relations between larger discourse segments [15].

SAMPLE TEXTUAL ANALYSIS

In this section we will consider a typical example of textual phenomena occurring in the medical text corpus. In particular, we will discuss a case of textual ellipsis as illustrated by the sentences (1) to (3).

Fig. 1 depicts the result of parsing sentence (1) in terms of the corresponding dependency tree which represents the syntactic structure produced by the PARSETALK parser [9, 10]. After semantic interpretation the dependency graph for sentence (1) is considerably flattened and appears in its ultimate conceptual interpretation form in Fig. 2. The corresponding centering structures for sentence (1) are provided in Table 2. Its backward-looking center is empty at the beginning of the text.

The ranking of the forward-looking centers directly reflects the linear precedence of discourse entities in the utterance U_1. The items in $C_f(U_1)$ then constitute the potential textual referents for the discourse entities in utterance U_2. Considering, e.g., utterance (2.2) as a possible succession of the text, the definite noun phrase *"the ... mucous membrane"* indicates an anaphoric reference. A conceptual generalization relation of the underlying concept, MUCOUSMEMBRANE, neither holds for PARTICLE nor for DIAMETER, but can be confirmed for the third list element, *viz.* GASTRICMUCOUSMEMBRANE. Hence, the resolution is performed with respect to this concept as indicated by the first element of $C_f(U_{2.2})$ (GASTRICMUCOUSMEMBRANE, *"mucous membrane (Schleimhaut)"*) — the first component specifies the conceptual denotation and the second one indicates the linguistic surface form. As this is the first realization of any of the elements from $C_f(U_1)$, the $C_b(U_{2.2})$ becomes GASTRICMUCOUSMEMBRANE. In the case of utterance U_3, again, a definite noun phrase, *"the ... stroma"*, occurs, signalling another case of referential relation. A nominal anaphora is precluded, since no element from $C_f(U_{2.2})$ applies for a conceptual generalization relation to STROMA. However, the criteria for textual ellipsis, the existence of a CONSISTS-OF relation to GASTRICMUCOUSMEMBRANE, are fulfilled. In cases of textual ellipsis, the proper antecedent is introduced in the centering list, though it is lexically not realized (indicated by "—" for the second tuple component in the corresponding list entry). The conceptual interpretation for sentence (3) appears in Fig. 3. Note that Fig. 2 and 3 depict the conceptual target structures that can be achieved by a parsing device that does not account for text phenomena at all. In order to assess our contribution in terms of text structure recognition, consider Fig. 4. It contains the intended conceptual interpretation for sentences (1) and (3) (we here leave out the representation structures contributed by sentence (2)). Note that the essential conceptual link to be established as a result of text structure computations — the resolution of the text ellipsis relationship between instances of the concepts GASTRICMUCOUSMEMBRANE and STROMA via the conceptual role CONSISTS-OF — is missing in Fig. 2 and 3, respectively. Hence, these concept graphs remain *unconnected*. The connectivity of the concept graph in Fig. 4 is a result mainly from accessing the centering lists for the text fragment under consideration (cf. Table 2), as it indicates which element might be related to another one given its textual context.

This illustrates our claim concerning the referential *incoherence* of text knowledge bases under the assumption that textual ellipsis relations were not resolved.

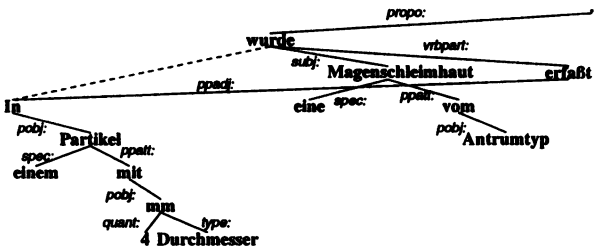

Figure 1: Dependency Graph for Utterance (1)

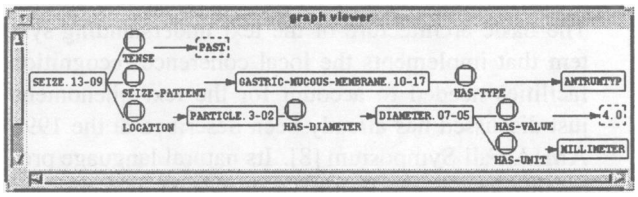

Figure 2: Concept Graph for Utterance (1)

Similar effects occur in terms of referential *invalidity* for cases of a lacking account of (pro)nominal anaphora.

(1)	Cb:	—
	Cf:	[PARTICLE: Partikel, DIAMETER: 4 mm Durchmesser, GASTRICMUCOUSMEMBRANE: Magenschleimhaut, ANTRUMTYPE: Antrumtyp]
(2.2)	Cb:	GASTRICMUCOUSMEMBRANE: Schleimhaut
	Cf:	[GASTRICMUCOUSMEMBRANE: Schleimhaut, FOVEOLA: Foveolen]
(3)	Cb:	GASTRICMUCOUSMEMBRANE: —
	Cf:	[GASTRICMUCOUSMEMBRANE: —, STROMA: Stroma, LYMPHOCYTE: Lymphozyten]

Table 2: Centering Data for Text Fragment (1) to (3)

RELATED WORK

The role of anaphora in scientific texts has already been investigated from the descriptive perspective related to their occurrence patterns (for abstract texts, cf. the study of [16]). Unfortunately, most of the computational studies focus on purely methodological issues and, in particular, do not consider real-world texts (cf., e.g., [17] for one of the rare exceptions).

In the medical domain, only [18] has dealt with anaphora in the environment of a text analysis system. More recently, [19] considered the problem of resolving metonymies based on a graph traversal approach similar in spirit to our work on metonymies [7]. Temporal reasoning which accounts for local coherence in the framework of medical text analysis has been considered by [20]. What is lacking in all these studies is a unified methodology for accounting for a broad spectrum of referential phenomena. This is where our proposal based on the centering model comes in.

CONCLUSIONS

We have outlined a unified methodology based on the centering model to deal with various forms of lo-

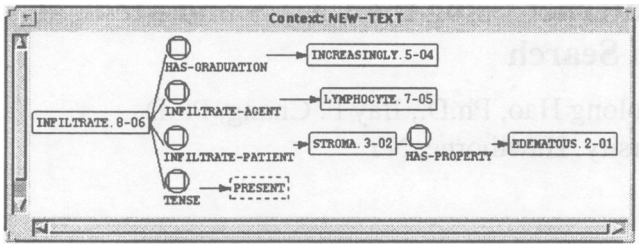

Figure 3: Concept Graph for Utterance (3)

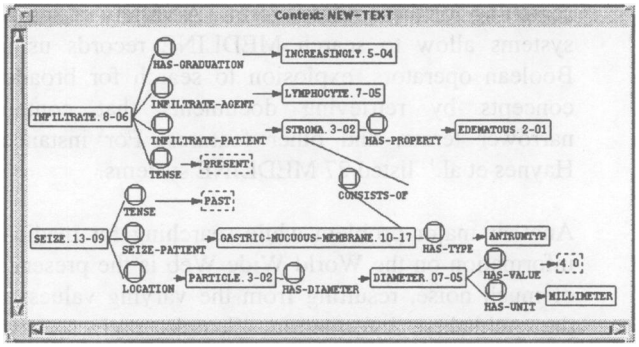

Figure 4: Combined Concept Graph for (1) and (3)

cal text coherence phenomena in medical texts, *viz.* (pro)nominal anaphora and textual ellipses. We have argued for the necessity to account for these text structures based on an empirical study of the distribution of these phenomena in a sample of medical texts.

The Medical Knowledge SYNDIKATE, a natural language text understander specifically designed to account for text structures in medical texts, has been presented as a text analysis environment capable of accounting for these text phenomena. While we have gathered substantial empirical evidence for SYNDIKATE's text analysis performance in the IT domain [5], after successful porting, we continue to run evaluation experiments in the medical domain. We have reasoned expectations that the failure rate we encountered in the IT domain (*viz.* approximately 10 to 15%) can be carried over to the medical domain.

Acknowledgements. We would like to thank our colleagues in the CLIF group and Stefan Schulz from the Department of Medical Informatics for fruitful discussions. M. Romacker is supported by a grant from DFG (Ha 2097/5-1).

References

[1] K.A. Spackman and W.R Hersh. Recognizing noun phrases in medical discharge summaries: An evaluation of two natural language parsers. In J.J. Cimino, (Ed.), *Proc. 1996 AMIA Annual Fall Symposium (formerly SCAMC)*, pages 155–158, 1996.

[2] D. Evans, N.D. Brownlow, W.R. Hersh, and E.M. Campbell. Automatic concept identification in the electronic medical record: An experiment in extracting dosage information. In J.J. Cimino, (Ed.), *AMIA'96 - Proc. 1996 AMIA Annual Fall Symposium (formerly SCAMC)*, pages 388–392, 1996.

[3] B.J. Grosz, M.E. Pollack, and C.L. Sidner. Discourse. In M.I. Posner, (Ed.), *Foundations of Cognitive Science*, pages 437–468. Cambridge: MIT Press, 1989.

[4] P. Spyns. Natural language processing in medicine: An overview. *Methods of Information in Medicine*, 35(4-5):285–301, 1996.

[5] M. Strube and U. Hahn. Functional centering. In *ACL'96 – Proc. 34th Annual Meeting of the Association for Computational Linguistics*, pages 270–277, 1996.

[6] U. Hahn, K. Markert, and M. Strube. A conceptual reasoning approach to textual ellipsis. In W. Wahlster, (Ed.), *ECAI'96 – Proc. 12th European Conference on Artifical Intelligence*, pages 572–576, 1996.

[7] K. Markert and U. Hahn. On the interaction of metonymies and anaphora. In *IJCAI'97 - Proc. 15th International Conference on Artificial Intelligence*, 1997.

[8] U. Hahn, K. Schnattinger, and M. Romacker. Automatic knowledge acquisition from medical texts. In J.J. Cimino, (Ed.), *AMIA'96 - Proc. 1996 AMIA Annual Fall Symposium (formerly SCAMC)*, pages 383–387, 1996.

[9] 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.

[10] 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, 1996.

[11] R. MacGregor. A description classifier for the predicate calculus. In *AAAI'94 – Proc. 12th National Conference on Artificial Intelligence.*, pages 213–220, 1994.

[12] U. Hahn and K. Schnattinger. Deep knowledge discovery from natural language texts. In *KDD'97 - Proceedings of the 3rd Conference on Knowledge Discovery and Data Mining*, 1997.

[13] 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.

[14] B.J. Grosz, A.K. Joshi, and S. Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995.

[15] U. Hahn and M. Strube. Centering in-the-large: Computing referential discourse segments. In *ACL '97/ EACL '97 - Proc. 35th Annual Meeting of the Association for Computational Linguistics and of the 8th Conference of the European Chapter of the Association for Computational Linguistics*, 1997.

[16] B.A. Fox. *Discourse Structure and Anaphora in Written and Conversational English.* Cambridge: Cambridge University Press, 1987.

[17] C. Kennedy and B. Boguraev. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *COLING'96 – Proc. 16th Intl. Conference on Computational Linguistics*, pages 113–118, 1996.

[18] G. Proszeky. Processing clinical narratives in Hungarian. In *COLING'86 – Proc. 11th Intl. Conference on Computational Linguistics*, pages 365–367, 1986.

[19] J. Bouaud, B. Bachimont, and P. Zweigenbaum. Processing metonymy: A domain-model heuristic graph traversal approach. In *COLING'96 - Proc. 16th Intl. Conference on Computational Linguistics*, pages 137–142, 1996.

[20] L. Hirschman and G. Story. Representing implicit and explicit time relations in narrative. In *IJCAI'81 – Proc. 7th Intl. Joint Conference on Artificial Intelligence*, pages 289–295, 1981.