Towards Temporal Segmentation of Patient History in Discharge Letters

Galia Angelova
Institute of Information and Communication Technologies, Bulgarian Academy of Sciences (IICT-BAS)
Sofia, Bulgaria
galia@lml.bas.bg

Svetla Boytcheva
IICT-BAS and
University of Library Studies and Information Technologies
Sofia, Bulgaria
svetla.boytcheva@gmail.com

Abstract

This paper reports about ongoing work in automatic identification of temporal markers and segmentation of patient histories into episodes. We discuss the discourse structure of the *Anamneses* in Bulgarian hospital discharge letters and present experiments with a corpus of 1,375 anonymised discharge letters of patients with endocrine and metabolic diseases. Our IE prototype discovers 32,445 key terms in the corpus, among them more that 7,000 occurrences of drug names and about 7,500 occurrences of diagnoses. The temporal markers occur 8,248 times usually paired with tokens pointing the direction of time "forward" or "backwards". Temporal markers are identified with precision 84%, recall 57% and f-measure 67.9%.

1 Introduction

Medical informatics has made little progress in the temporal representation and reasoning tasks [1]. This is partly due to the complexity of the free text descriptions in clinical narratives where temporal information is presented. On the other hand, there is no agreement about the essence of clinical temporal models and the concepts and relationships that have to be taken into account. Research on temporal information interpretation is still in its embryonic stage according to [2]. The progress requires theoretic models as well as large training corpora of annotated medical texts which are expensive to construct.

This paper summarises work in progress on discharge letters structuring and experiments in automatic extraction of temporal markers. We deal with discharge letters in Electronic Health Records (EHR) in Bulgarian language. These letters contain predefined sections due to the general practice of structuring clinical notes into sections which dates back to the 60's and 70's of the last century as a result of centralised regulations[1]. Our present IE system, which analyses discharge letter texts, automatically identifies the *Anamnesis* (*Patient history*). This section contains a sketchy abstract, manually prepared by medical experts, who summarise the patient history in order to communicate it to another doctor. Explicit temporal markers designate the main phases in disease development, the main interventions and their effects. The unified text format is a motivation for its inclusion in methods for automatic episode recognition. Our main idea is to select a discourse structure theory and to try extracting pieces of information that have meanings as discourse units. The paper presents our first results in this direction.

Section 2 overviews related approaches. Section 3 considers the context of our work: the discourse structure of discharge letters and existing prototypes for section splitting and information extraction. Section 4 presents the evaluation of the current component for automatic extraction of temporal markers as well as our research agenda for building timelines of clinical events. Section 5 contains the conclusion.

2 Related Work

Research on temporal information processing is a relatively recent activity in biomedical NLP. Savova et al. [4] presents an annotation schema with temporal relations based on TimeML [5] and analyses the potential of TimeML tags as annotation tool for clinical narratives. The general objective is to build a temporal relation discovery component and a reasoner to create timelines of clinically relevant concepts. The authors consider five *Event* classes including *Occurence* (events that happened) and *State* (a condition or

[1] The list of sections in Bulgarian hospital discharge letters is published as a legal Agreement in the Official State Gazette, Article 190(3) [3].

Proceedings of the Workshop on Biomedical Natural Language Processing, pages 49–54,
Hissar, Bulgaria, 15 September 2011.

state, e.g. symptoms, descriptors and chronic conditions). The paper explicates important fine-grained characteristics of events and temporal relations in clinical texts which are related to linguistic units.

Five tags for marking up temporal information are suggested in [6]: *reference point*, *direction*, *number*, *time unit*, and *pattern*. The authors identified 254 temporal expressions in 50 discharge summaries and represented them using the suggested scheme. The inter-rater agreement was 75% which shows the complexity of temporal annotations even when simple tags are used.

Harkema et al. [7] presents an algorithm called ConText which identifies clinical conditions that are described in clinical reports: they can be *negated*, *hypothetical*, *historical*, or experienced by someone *other* than the patient. ConText infers the status of a condition with regard to these properties from simple lexical clues occurring in the context of the condition. The study deals with 4,654 annotations from 240 clinical reports: 2,377 annotated conditions in the development set and 2,277 annotated conditions in the test set. The evaluation summarises results obtained in a six-token window (*stw*) and end-of-sentence (*eos*) contexts. For "*historical*" condition, ConText achieves *stw* precision 78% and recall 70% as well as *eos* precision 77% and recall 79% across all report types that contain such conditions.

Hripcsak et al. model temporal information as contraint satisfaction problem [8]. Medical events from 231 discharge summaries are represented as intervals, and assertions about events are represented as constraints. Up to 151 medical events and 388 temporal assertions were identified per complete discharge summary. Non-definitional assertions were explicit (36%) or implicit (64%) and absolute (17%), qualitative (72%), or metric (11%). Implicit assertions were based on domain knowledge and assumptions, e.g., the section of the report determined the ordering of events. The source texts contained no instances of discontinuous temporal disjunction. The authors conclude that a simple temporal constraint satisfaction problem appears sufficient to represent most temporal assertions in discharge summaries and may be useful for encoding electronic medical records.

In our present work, we are mostly influenced by [6], which is attractive because of its relative simplicity, and [8].

3 Project context

The general objectives of our project are: *(i)* to develop a system for knowledge extraction from discharge letter texts and *(ii)* to design algorithms for searching conceptual patterns in the extracted clinical facts. Recognising the events and ordering them in timelines is an essential part of the project research agenda.

3.1 Materials

The discharge letters in our corpus contain sections which can be automatically recognised with accuracy 99.99% using their headers. Sometimes the structure is not strictly kept due to section merging, changing the section headers, skipping (empty) sections and replacing the default section sequence. Table 1 shows the percentage of discharge letters with available standard sections in the corpus of 1,375 EHRs.

Section title	***Discharge letters containing the section***
Diagnoses	100%
Anamnesis	100%
Past diseases	88.52%
Allergies, risk factors	43.56%
Family Medical History	52.22%
Patient Status	100%
Lab data, clinical tests	100%
Examiners' comments	59.95%
Debate	100%
Treatment	26.70%

Table 1. Formatting discharge letters according to centralised state regulations

Having the potential to identify automatically the *Anamnesis* section, which contains the patient history, we can plan further research tasks related to its temporal segmentation.

3.2 Methods

We have developed extractors of ICD-10 codes (the International Classification of Diseases, v. 10[2]) and ATC codes (the Anatomical Therapeutic Chemical Classification System[3]) from discharge letter texts [9, 10, 11]. The tool for diagnosis extraction assigns ICD-10 codes to Bulgarian nominal phrases designating disease names with 84.5% precision [9]. The drug extractor assigns ATC codes to Bulgarian drug names with f-measure 98.42% and achieves 93.85% f-

[2] http://www.nchi.government.bg/download.html
[3] http://www.who.int/classifications/atcddd/en/

measure for the dosage recognition [10]. Automatic extraction of "current treatment" is possible with highest accuracy for the drugs discussed in the *Anamnesis* (precision 88%, recall 92.45%, f-measure 90.17%, overgeneration 6%) [11].

Figure 1 shows the connections between the text (pre-)processing components that have been developed so far in our projects. The anonymised input discharge letter is split into section; after tokenisation and sentence splitting, the temporal markers are identified and episodes are tagged. Within the episodes, diseases and drugs are recognised by the existing tools. Other entities to be recognised in the episodes are *conditions* (symptoms, complains and status) and *treatment outcome*. The system works on a resource bank of medical nomenclatures, terminologies and linguistic resources. The output consists of annotated anamneses whit tagged episodes.

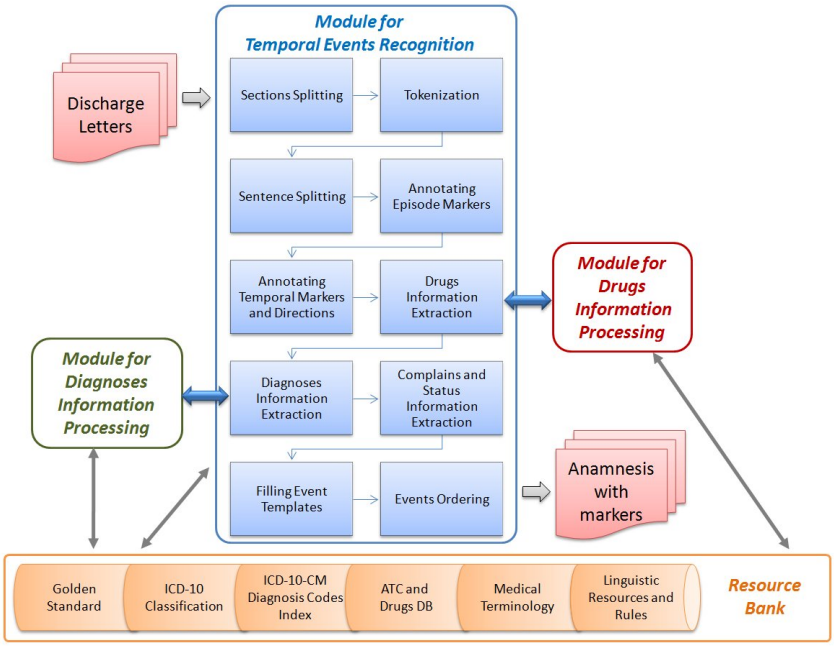

Figure 1. Text pre-processing components and extractors of diseases, drug names, and temporal markers

4 Extraction of temporal markers and ideas for structuring episodes

4.1 Evaluation results

We have performed experimental tests with 1,375 discharge letters where our IE prototype discovers 32,445 key terms or markers in the *Anamneses* (in average 23.59 per discharge letter). The distribution of these terminologies and temporal markers is presented in Table 2.

Temporal Marker	*Occurrences*	*EHRs*	*Avg*
drug names	7,108	1,213	5.86
diagnoses	7,565	1,292	5.86
complains	1,274	841	1.51
temporal	8,248	1,373	6.01
direction	8,249	1,374	6.01
Total	32,445	1,375	23.59

Table 2. Recognised entities in 1,375 discharge letters

Not surprisingly the temporal and direction markers occur as pairs, because direction markers point "forward" or "backward" from the time marker. The share of temporal and direction markers is significant (51% in total, see Fig. 2). These figures explicate the importance of temporal information for the case history description.

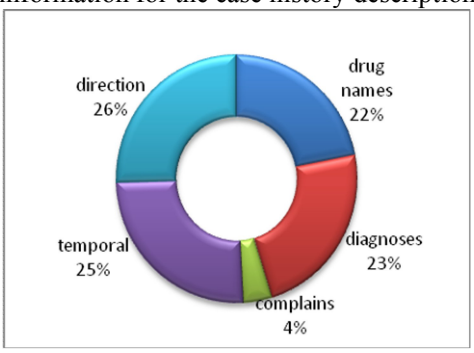

Figure 2. Percentage of temporal markers

The present recall in the recognition of the temporal markers in about 57% and the precision is 84% (f-measure 67.9%).

4.2 Segmentation into episodes

In medicine, an episode comprises all activities that are performed between the diagnosis of a disease and its cure (or stabilisation in case of chronic diseases). Studying various approaches to determine and annotate the granularity of temporal intervals, we view the patient history episodes as sets of events defined via the explicit temporal markers stated by the physicians who examine and treat the patients.

Most discharge letters in our corpus concern patients with endocrine and metabolic diseases diagnosed decades ago. In general only the major illness phases are discussed in the *Anamnesis* together with the treatment and medication changes. We consider below an example of a case history written in 2010:

Example 1. `Diabetes Mellitus diagnosed 5-6 years ago, manifested by most symptoms. At the beginning started treatment with Maninil only, afterwards in combination with Siofor. After few months the Maninil was replaced by Diaprel. Since October 2005 treated with Insulin Novomix 30 – 32E in the morning, 26E in the evening with diagnosed diabetic retinopathy. Complains of strong pains in the feet mostly at night.`

We believe that human experts declare explicitly the most important temporal markers which are sufficient (in their view) to adequately and unambiguously communicate the case history to another medical doctor. Therefore, we consider these markers as intentional signals for discourse segmentation. Our temporal model is framed using three tags suggested in [6]:

- *reference point*, *direction*, and *temporal expression*

plus additional tags needed for our project:

- *diagnoses or disorders*,
- *complains or symptoms*,
- *drugs/treatment applied* as well as
- *treatment outcome.*

There could be several diagnoses or symptoms enumerated in one episode as well as more than one drug correspondingly prescribed to the patient.

Let us consider the episodes in Example 1 which are defined by the explicit temporal expressions. Interpreting the text and ordering the temporal markers in a time progression scale according to the concrete moments, we construct the representation shown in Table 3. The conventional literal *'now'* denotes the speech/writing moment, in this case the moment of hospitalisation in 2010. Table 3 integrates the results of two extractors that were referred to in Section 3.2.

Ep1	Reference point	*Now* minus 5-6 years
	Direction	forward
	Temporal expression	5-6 years ago
	Diagnoses	Diabetes Mellitus E10
	Complains, symptoms	
	Drugs/Treatment	Maninil A10BB01
	Drugs/Treatment	Siofor 1 A10BA02
	Treatment outcome	
Ep2	Reference point	*(Now* – (5-6 years)) + few months
	Direction	forward
	Temporal expression	After few months
	Diagnoses	
	Complains, symptoms	
	Drugs/Treatment	Diaprel A10BB09
	Drugs/Treatment	Siofor 1 A10BA02
	Treatment outcome	
Ep3	Reference point	October 2005
	Direction	forward
	Temporal expression	Since October 2005
	Diagnoses	Diabetic retinopathy H36.0
	Complains, symptoms	
	Drugs/Treatment	Insulin Novomix 30 – 32E mon., 26E ev.
	Treatment outcome	
Ep4	Reference point	*Now*
	Direction	
	Temporal expression	
	Diagnoses	
	Complains, symptoms	strong pains in the feet
		…..

Table 3. Temporal segmentation with integration of automatic diagnoses and drugs extraction

Studying manually the discharge letters in our corpus, we think that the episodes resemble discourse segments as introduced in [12]. It seems reasonable to consider every temporal marker as a cue phrase signaling a new episode, because cue phrases express the intention of the writer to emphasize on major disease phases. In Example 1 and Table 3 we can also follow the *topical focus shifts* [13]: an episode might

- retain the topic of the previous one and contain references to the discourse entities in

the previous episode. For instance, *episode 2* which was uttered immediately after *episode 1* refers to the entity Maninil introduced in *episode 1*;

- shift the topic and start discussion of another entity like e.g. *episode 3*.

Practical guidance of how to recognise segments is given in [14] where a discourse segment is viewed as a sequence of clauses that display local coherence. The following properties are listed as features that should hold within a segment:

(i) Resolution of references should be possible by techniques based on recency;
(ii) The time is fixed or there is a simple progression;
(iii) A fixed set of background assumptions is relevant to all clauses in the segment;
(iv) From intentional perspective, all the sentences in the segment contribute to a common discourse purpose, i.e. the same communicative goal motivates the writer to utter all clauses in the segment;
(v) From informational perspective, all the sentences in the segment are related to each other by some temporal, causal or rhetoric relations, i.e. all sentences and phrases combine together to describe a coherent event or situation.

Following these five properties, we might view the elementary *episode 1* and *episode 2* as a single segment because they discuss the progression of the diabetes, while *episode 3* is focused on the diabetes complication *Diabetic retinopathy* diagnosed in 2005. We also note that the clauses in *episode 1* and *episode 2* form a focus space where the diabetes progress in considered; in computational linguistics the focus spaces are organised in hierarchies and the preferred option is to open each topic only once.

These structural features are seen in most discharge letters that we have studied manually but we are far from final generalisation of our empirical observations. At the end we note that most temporal markers contain explicit references to time which help to construct the "elementary episodes". Only 0.014% of the temporal markers contain vague statements like "*long-term* diabetes" and only 0.01% use expressions like "since *several* years/weeks/days". The relative temporal markers are oriented in two directions:

- To the time marker of the immediately preceding previous episode, e.g. *since then*, *before that*, *after that,* etc. and
- To the moment *Now* when the discharge letter is written: like e.g. *few months ago*. Such episodes might elaborate a past event or period no matter that they are oriented according to the present moment.

The next task in our research agenda is to develop heuristics enabling to automatically position the temporal markers on a linear time scale with respect to the actual date of hospitalisation which is known in the Hospital Information System. In this way the episodes might be ordered in a sequence using a simple procedure which tries to calculate the actual date and constructs a list of linearly-ordered reference points. Moreover, applying the discourse coherence considerations, we could aim at semantic coupling of episodes which discuss the same topic (like *episode 1* and *episode 2* that are adjacent and display local coherence). This seems to be a challenge but one can aim to achieve it because adjacent episodes should be uttered in consecutive sentences dealing with the same entities. Figure 3 is obtained in this way. At present we only extract temporal markers and build "*elementary*" episodes of patient history phases.

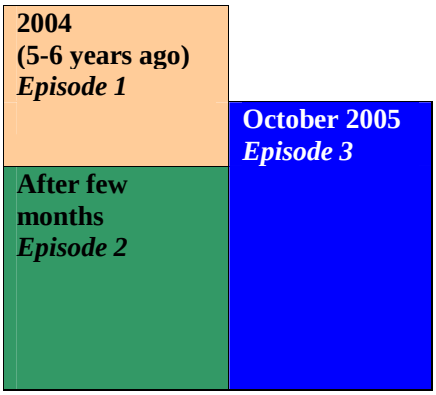

Figure 3. Grouping episodes into segments with local coherence for an *Anamnesis* written in 2010

Background medical knowledge might also help for the automatic grouping of adjacent episodes into intervals. For instance our system needs to perform semantic analysis of *episodes 1* and *2* in order to understand that Maninil A10BB01 was replaced by Diaprel A10BB09 in a few months. However, these two drugs belong to the same generic group (which is seen by the ATC code) and cannot be taken together, and this medical default might support the semantic analysis of the textual description. This fact is another hint that *episode 2* elaborates *episode 1* and presents further description of the treatment

related to diabetes. Therefore, it is easy to find informational signals that *episode 2* belongs to the local context of *episode 1*.

5 Conclusion

This paper presents work in progress aiming at the automatic segmentation of episodes in the patient history (usually events or periods) as described in discharge letters.

So far, we have found no case of temporal ambiguity which prohibits the manual annotation (although it is based on human interpretation of the text semantics). Further research needs to be done for designing a heuristic strategy of how to "glue" adjacent groups of clauses together because they display local coherence properties. To solve such a task our system needs to understand which episode elaborates the previous one. The present paper reports first findings in this respect.

Acknowledgments
The research work presented here is supported by grant DO 02-292 "Effective search of conceptual information with applications in medical informatics", funded by the Bulgarian National Science Fund in 2009-2012.

The authors are grateful to the paper reviewers for the useful comments and numerous suggestions.

References

1. Zhou L., C. Friedman, S. Parsons, and G. Hripcsak. *System architecture for temporal information extraction, representation and reasoning in clinical narrative reports*. In Proc. AMIA Ann. Symp. 2005, pp. 869–873.
2. Zhou L. and G. Hripcsak. *Temporal reasoning with medical data - a review with emphasis on medical natural language processing*. J. Biom. Informatics 2007, 40(2), pp. 183-202.
3. *National Framework Contract* between the National Health Insurance Fund, the Bulgarian Medical Association and the Bulgarian Dental Association, Official State Gazette №106 (2005), updates №68(2006) and №101(2006), Sofia, Bulgaria, http://dv.parliament.bg/.
4. Savova, G., S. Bethard, W. Styler, J. Martin, M. Palmer, J. Masanz, and W. Ward. *Towards Temporal Relation Discovery from the Clinical Narrative.* In Proc. AMIA Annual Symposium 2009, pp. 568–572.
5. Sauri R., J. Littman, B. Knippen, R. Gaizauskas, A. Setzer and J. Pustejovky. *TimeML annotation guidelines*, Version 1.2.1, 31 January 2006. Available online at http://www.timeml.org/site/publications/timeMLdocs/annguide_1.2.1.pdf.
6. Hyun S., S. Bakken and S.B. Johnson. *Markup of temporal information in electronic health records.* In Stud. Health Technologies and Informatics Vol. 122, 2006, pp. 907-908.
7. Harkema, H., J. Dowling, T. Thornblade, and W. Chapman. *Context: An Algorithm for Determining Negation, Experiencer, and Temporal Status from Clinical Reports*. J Biomed Inform. 2009 42(5): 839–851.
8. Hripcsak G., L. Zhou, S. Parsons, A. K. Das, and S.B. Johnson. *Modeling electronic discharge summaries as a simple temporal constraint satisfaction problem*. JAMIA 2005, 12(1), pp. 55-63.
9. Tcharaktchiev, D., G. Angelova, S. Boytcheva, Z. Angelov, and S. Zacharieva. *Completion of Structured Patient Descriptions by Semantic Mining*. In: Koutkias et al. (Eds.), *Patient Safety Informatics - Adverse Drug Events, Human Factors and IT Tools for Patient Medication Safety*, Stud. Health Techn. Inform. 166, 2011, pp. 260-269.
10. Boytcheva, S. Shallow Medication Extraction from Hospital Patient Records. In: *Koutkias* Koutkias et al. (Eds.), *Patient Safety Informatics – Adverse Drug Events, Human Factors and IT Tools for Patient Medication Safety*, Stud. Health Techn. Inform. 166, 2011, pp. 119-128.
11. Boytcheva, S., D. Tcharaktchiev and G. Angelova. *Contextualization in automatic extraction of drugs from Hospital Patient Records.* In A. Moen et al. (Eds.) *User Centred Networked Health Case*, Proc. of MIE-2011, the 23th Int. Conf. of the European Federation of Medical Informatics, Stud. Health Techn. Inform. 169, 2011, pp. 527-531.
12. Grosz, B. and C. Sidner. *Attention, Intention and the Structure of Discourse*. Computational Linguistics 1986, 12(3), Reprinted in RNLP.
13. Grosz, B. *The representation and use of focus in a system for understanding dialogues*. IJCAI 1977, pp. 67-76, Reprinted in RNLP.
14. Allen, J. *Natural Language Understanding*, 2nd Edition, Benjamin/Cummings, 1995.