Towards Structuring Episodes in Patient History

Galia Angelova[1], Svetla Boytcheva[1,2] and Dimitar Tcharaktchiev[3]

[1] Institute of Information and Communication Technology, Bulgarian Academy of Sciences, 25A Acad. G. Bonchev Str., 1113 Sofia, Bulgaria, galia@lml.bas.bg
[2] University for Library Studies and Information Technology, Sofia, Bulgaria
[3] University Specialised Hospital for Active Treatment of Endocrinology, Medical University Sofia, Bulgaria

Abstract. This article presents current results in automatic Information Extraction (IE) from hospital patient records. The aim is to construct a temporal sequence of important facts about phases in disease development, by recognising the main events that are described in the anamnesis (case history). Normally all important stages of illness progressing are documented with the corresponding treatment and its effect. The paper presents the conceptual structure of "episode", which is designed as a simple conceptual graph to support our automatic IE procedures. Representing patient histories as simple CGs would enable comparison of cases. The article also reports about the evaluation of the extraction procedures which discover temporal markers signaling the episodes.

Keywords: Natural Language Processing (NLP), Event recognition, Temporal markers, Episode sequencing, Capturing structured information from free text

1 Introduction

Defining events is difficult as the perspective can vary depending on people knowledge and task context. Any observable or hypothetic occurrence in certain location at a particular time is an event. A medical event is, for instance, the diagnosing of a disease accompanied by a number of examinations – which can take weeks, but swallowing a single pill is an event as well. In narrative texts the events are described with various depth and granularity; normally we assume that any verb refers to an event or state but an event instance can be also expressed by several consecutive sentences.

In Artificial Intelligence, events are treated as entities distinct from the things that participate in the happening or occurrence [1]. In linguistics, viewing events as quantifiable entities enables to consider their instances as particular individuals with specific participants, time and location. Modern approaches in software modeling treat events as basic units, fundamental for representing world semantics [2]; according to this event-driven paradigm, the situation is an event occurrence which might require reaction. From the perspective of software design and development, [2] suggests features and attributes to describe events, among them: *modal* (absence, always, sometimes, not selected); *temporal* (sequence, increasing, decreasing, non increasing, non decreasing and mixed); *spatio-temporal* (moving in consistent direction, moving in mixed direction, stationary, moving towards). The approach [2] models the events from a higher perspective, with primary focus on their tracking and planning the eventual reactions. In contrary, computational linguistics constructs fine-grained representations of event descriptions in text, viewing time as an essential aspect of the text understanding. Recently the markup language TimeML for annotation of events and temporal relations was developed [3]. Events in TimeML are situations that happen or occur, they can be punctual or last some time, and may be expressed by means of verbs, nominalisations, adjectives, predicative clauses, or prepositional phrases. In this way TimeML suggests a text-based framework for detailed event annotation which facilitates event recognition in applications which aim at the automatic analysis of free text.

Our research project aims to discover patterns in disease developments by searching similar case histories in a corpus of hospital patient records and the respective discharge letters. This includes, among others, extraction of structured event descriptions from free texts in order to explicate the temporal relations that may hold implicitly between the events and participating things. Various approaches to event modeling are possible but we select those which are closest to the semantic structure of the discharge letters. An event is, for instance, the diagnosing of a disease at some moment of time; the treatment prescribed then can be viewed as a feature. The next event might be the occurrence of certain disease complications after several years, which are treated by some drugs and so on. Thus we can consider the case history as a sequence of phases or episodes which summarise the chronology of disease-relevant events. Explicit temporal markers enable the identification of such events in the clinical records. Additionally, the narrative convention (to utter events in the sequence of appearance) helps much to capture structured temporal information. In this paper we present current research results in the automatic identification of clinical episodes and their representation as conceptual structures.

The article is structured as follows. Section 2 overviews some related work. Section 3 introduces the project context by describing the input data, the project objectives and the motivations behind our definition of episodes. Section 4 discusses the current prototype for automatic identification of temporal markers and presents its evaluation. Section 5 contains the conclusion and summarises plans for further work.

2 Related Work

Our tasks are focused on text analysis, therefore we summarise some language-related approaches to event and time recognition. At first we consider the mark-up language TimeML [3] in more depth because its tags give us useful hints about the nature of text-based event identification. The main tags that explicate links between events and entities are *Event*, *TLink*, *SLink* and *ALink* (but they all have a range of further attributes, see [3]). The *Event* types are REPORTING, PERCEPTION, ASPECTUAL, OCCURRENCE, I(ntentional)_ACTION, I(ntentional)_STATE, and STATE. Events may be expressed by means of tensed or untensed verbs, nominalisations, adjectives, predicative clauses, or prepositional phrases. *Events* participate in *TLinks*, *SLinks* and

ALinks only by means of their corresponding event instance identifiers (IDs). A *TLink* or *Temporal Link* represents the temporal relation that holds between events, times or between an event and a time. The *TLinks* shows whether the involved entities are simultaneous, before, after, immediately before, immediately after, including, being included, during, beginning, begun by, ending, identity, and set/subset. A *Tlink* is inserted in the annotation whenever a temporal relationship holding between events or an event and a time is described in the text. A *SLink* or *Subordination Link* is used for contexts introducing relations between events; *SLink* sorts are modal, factive, counter-factive, evidential, negative evidential, and conditional. These links can be expressed be lexical means or structural means, by purpose clauses and conditional constructions. An *ALink* or *Aspectual link* represents the relationship between an aspectual event and its argument event; there are five types of *ALinks*: initiation, culmination, termination, continuation, reinitiation. Another three major tags are *Timex3* (annotating temporal expressions), Signal (annotating temporal prepositions, conjunctions and modifiers) and MakeInstance (annotating the actual realisation of an event since TimeML distinguish between event tokens and event instances). Examples of fully annotated sentences are given in [3]. In 2005, a corpus called TimeBank1.2 of around 200 news report documents from various sources, annotated with TimeML temporal and event information, was used as a gold standard corpus of event recognition. The paper [4] presents Evita, an application for identification of events in natural language texts. As the authors claim, Evita is unique in that it is not limited to any pre-established list of relation types (events), nor is it restricted to a specific domain. Evita was evaluated by comparing its performance against TimeBank1.2. The system performance was 74.03% precision and 87.31% recall, which is comparable to the interannotation agreement scores for the task of tagging verbal and nominal events. In 2007, the 2007 TempEval competition was organised, with a stricter annotation interface and a simplified set of temporal relations. Systems performed well on its tense identification task, but poorly on the other tasks which often required multiple stages of implicit temporal logic.

In the area of biomedical NLP, the research on event recognition is a relatively recent activity. The article [5] summarises the trends in the general NLP community and analyses the potential of TimeML tags as annotation tool for clinical narratives. The general objective of the authors is to extend the system cTAKES with a temporal relation discovery component and a reasoner to create timelines of clinically relevant concepts. An annotation schema for temporal relations based on TimeML is presented in [5]. Some TimeML features are reduced and restricted, other features are modified or added. For instance, *Timex3* objects are definitive references to time that provide concrete temporal references (e.g. *yesterday, 4 days ago, December 2003*). States and conditions are labeled as *Events* in addition to traditional events. *Tense* is an attribute of *Events* and refers to the temporal relation of the event to the time of the patient-physician encounter (not to the grammatical verb tense). Only five *Event* classes are considered (out of the seven suggested in TimeML); *Occurence* is used for events which happened, *State* denotes a condition or state (symptoms, descriptors and chronic conditions are *States*), the other *Event* types are *Perception*, *Reporting* and *Aspectual*. *Degree* (with values *all*, *none*, *little* and *most*) is an attribute the *Events*. In this way [5] explicates useful characteristics of events in clinical narratives.

The article [6] provides important information about features of patient clinical conditions which are described in clinical reports: they can be *negated*, *hypothetical*, *historical*, or experienced by someone *other* than the patient. The authors propose an algorithm called ConText which 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 4654 annotations from 240 clinical reports: 2377 annotated conditions in the development set and 2277 annotated conditions in the test set. The evaluation summarises results obtained in a six-token window (*stw*) and end-of-sentence (*eos*) contexts. ConText shows reasonable to good performance for *negated* (*stw* precision 99% and recall 96%, *eos* precision 98% and recall 98%), *historical* (*stw* precision 78% and recall 70%, *eos* precision 77% and recall 79%), and *hypothetical* (*stw* precision 100% and recall 34%, *eos* precision 100% and recall 93%) conditions across all report types that contain such conditions. Conditions experienced by someone *other* than the patient are very rarely found in the corpus and the ConText performance is: *stw* precision 100% and recall 67%, *eos* precision 100% and recall 67%. The authors conclude that "a comprehensive solution to the problem of determining whether a clinical condition is *historical* or *recent* requires knowledge above and beyond the surface clues picked up by ConText".

An alternative representation for marking up temporal information in patient records is suggested in [7]. There are five tags for marking up temporal information: *reference point*, *direction*, *number*, *time unit*, and *pattern*. The authors identified 254 temporal expressions in 50 discharge summaries and represented them using the suggested (relatively simple) scheme.

Another interesting result in presented in [8]. Medical events from 231 discharge summaries were represented as intervals, and assertions about events were represented as constraints. Some 46-151 medical events and 118-388 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.

A detailed review of temporal reasoning with medical data is given in [9]. The authors state that "minimal work has been done in medical informatics on temporal representation and reasoning problems". Indeed, only few groups are active in this difficult domain. The article [10] discusses experiments with TimeText, a temporal reasoning system, and presents evaluation of its accuracy in answering time-oriented clinical questions. TimeText generated temporal relations about the endpoints (start or finish) of pairs of medical events. Independent human raters determined that 97% of 295 manually generated temporal relations were correct as well as 96.5% of 995 system-generated temporal relations. The system captured 79% of 307 temporal relations determined to be clinically important by the subjects and raters. TimeText answered 84% of the temporal questions correctly. One of the few systems dealing with temporal information in a language other than English is MedSyndikate, which processes medical findings reports and discovers simple facts, complex propositions and evaluative assertions. For the conceptual representation of medical processes and events, modal and auxiliary verbs are analysed, and the final conceptual representation uses an anaphora resolution component [11]. All listed approaches and systems provide useful hints and design considerations for our task. At the end we note that conceptual graphs are applied in medical document processing, see e.g. [12], so our idea to model temporally-related facts as conceptual structures in aligned to resent research in medical informatics.

3 The project context

The general objectives of our project are *(i)* to develop a system for knowledge discovery and extraction from the patient record (PR) texts and *(ii)* to investigate algorithms for searching conceptual patterns in the extracted clinical facts [13]. We deal with a corpus of anonymised hospital PRs of diabetic patients, delivered by the University Specialised Hospital for Active Treatment of Endocrinology (USHATE) which belongs to the oldest and largest Medical University in Bulgaria.

The hospital USHATE treats citizens from allover the country with specific, complex history cases. The discharge letters contain a brief summary of the most important facts related to the diseases treated in USHATE and enumeration of accompanying diseases. Some patients have up to 30 diagnoses listed in the hospital PRs (but most have up to 7 diagnoses); the case history is summarised with a specific level of granularity into several paragraphs. We have processed about 6300 PRs and all of them conform to the accepted style to overview the chronology of patient illnesses (certainly, the summary quality depends on the author, but the intention to provide it is always there). Hence, in the project we actually work with summaries written by human experts.

3.1. Input data

The input texts in our experiment are free-text paragraphs of anonymised discharge letters. In Bulgaria the discharge letter structure is mandatory for all hospitals (it is published in the Official State Gazette, as a part of a legal Agreement between the Bulgarian Medical Association and National Health Insurance Fund). The PR text should contain the following sections: *(i)* personal details; *(ii)* diagnoses of the leading and accompanying diseases; *(iii)* anamnesis (personal medical history), including current complains, past diseases, family medical history, allergies, risk factors; *(iv)* patient status, including results from physical examination; *(v)* laboratory and other tests findings; *(vi)* medical examiners comments; *(vii)* debate; *(viii)* treatment; *(ix)* recommendations.

This structure could provide appropriate context for temporal interpretation of the extracted events but in reality it is not strictly kept. Although the sections are mandatory, many PRs are structured differently due to section merging, changing the section headers, skipping (empty) sections and replacing the default section sequence. Table 1 shows some statistics about availability of the above-listed sections in a training corpus of 1300 USHATE PRs. Nevertheless the accepted style is more or less followed; for instance it is quite unusual to see patient history discussed outside the Anamnesis and Debate sections. Therefore, while designing our temporal models, we focus on the Anamnesis which is automatically discovered in almost all PRs.

Diag-Noses	Anam-nesis	Past disea-ses	Allergi-es, risk factors	Family history	Patient status	Lab tests	Exami-ners co-mments	Deba-te	Treat-ment
100	100	88,52	43,56	52,22	100	100	59,95	100	26,70

Table 1. Percentage of PRs including standard sections (which are automatically recognised)

Most of the PRs present patients with diabetes diagnosed decades ago. In general only the major illness phases are discussed together with the treatment and medication changes. There are no detailed descriptions of medical events; the PRs rather contain sketchy abstracts. We consider below 2 examples of case histories in our corpus (examples 1 is written in 2004, example 2 - in 2010):

Example 1. Diabetes Mellitus diagnosed in 2003, manifested by most symptoms - polyuria, polydipsia, lost 20 kg in 6 months with reduced appetite. Prescribed Maninil 3,5 mg 1+1 tabl. for a period of 3 months. Since then no blood sugar was tested and no further therapy was carried out. 20 years ago enlarged thyroid, sometimes the patient had suffocation and palpitation, but no examinations were made and no therapy was carried out.

Example 2. Diabetes Mellitus diagnosed 5-6 years ago, manifested by most symptoms. At the beginning started treatment with maninil only, afterward in combination with siofor. Few months ago 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 from strong pains in the feet mostly at night.

These original summaries of patient histories are presented in the Anamnesis of the respective PRs. In addition the following subsections might be included there with corresponding headers: Accompanying or Past diseases, Family history, Risk factors, and explicit statement about patient Allergies. Our studies in the last years show that in many countries the discharge letters have no structure as the one presented here.

3.2. Definition of episodes

In medicine, an episode comprises all activities that are performed between the diagnosis of disease and its cure; normally the episode is decomposed to goals and actions. The patient-related documentation is related to this default fragmentation of healthcare tasks [14]. For chronic diseases instead of cure we talk about stabilisation, e.g. the diabetes is compensated. Studying various approaches to determine and annotate the granularity of temporal intervals, when important clinical events occur, we shall consider as episodes *sets of events defined via the explicit temporal markers uttered by the physicians* who examine and treat the patients. We remind that our information extraction procedures actually work on summaries; we believe that human experts declare explicitly the most important temporal markers which are sufficient (in their view) to adequately communicate the case history to another medical doctor. Therefore, we consider these markers as primary signals for diseases progression phases. Our model is framed using three tags suggested in [7]:

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

plus additional tags needed for our project:

- *diagnoses, complains or symptoms* (i.e. what happens, occurs or is found during the episode) as well as
- *drugs*/*treatment* applied during the episode.

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 Examples 1 and 2. Interpreting the text as human beings, for Example 1 we can construct the representation shown in Table 2. We use the conventional literal '*now*' to denote the speech/writing moment. Ideally, we should be able to build correct temporal sequences of all events in clinical texts but [7] cites 75% inter-annotators agreement when 50 discharge summaries were manually annotated by 254 temporal expressions. It also remains unclear whether one should annotate periods when nothing happens, like e.g. episode 4 in example 1.

Ep1	Reference point	*Now* minus 20 years
	Direction	forward
	Temporal expression	20 years ago
	Diagnoses, complains, symptoms	enlarged thyroid, sometimes suffocation and palpitation
	Drugs/Treatment	no
Ep2	Reference point	2003 (diagnosis point)
	Direction	backward
	Temporal expression	in 6 months
	Diagnoses, complains, symptoms	lost 20 kg with reduced appetite
	Drugs/Treatment	no
Ep3	Reference point	2003 (diagnosis point)
	Direction	forward
	Temporal expression	in 2003
	Diagnoses, complains, symptoms	Diabetes Mellitus, polyuria, polydipsia
	Drugs/Treatment	Maninil 3,5 mg 1+1 tabl. (for a period of 3 months)
Ep4	Reference point	2003 (diagnosis point) plus 3 months
	Direction	forward
	Temporal expression	since then
	Diagnoses, complains, symptoms	-
	Drugs/Treatment	no
Ep5	Reference point	*Now* (moment of hospitalisation)
		…..

Table 2. Manually-constructed temporal event sequencing for the case presented in Example 1.

We are aiming at the automatic recognition of temporally-sequenced episodes; therefore as another example we show an extract produced by our present IE prototype while processing the sentences given in Example 2. The diagnoses are recognised when an ICD-10 code is juxtaposed to the respective phrases (ICD is the International classification of Diseases, version 10). The extractor encodes the drugs by their ATC code (ATC is the Anatomical Therapeutic Chemical Classification System).

Ep1	Reference point	*Now* minus 5-6 years
	Direction	forward
	Temporal expression	5-6 years ago
	Diagnoses, complains, symptoms	Diabetes Mellitus E29
	Drugs/Treatment	Maninil A10BB01
	Drugs/Treatment	Siofor 1 A10BA02
Ep2	Reference point	October 2005
	Direction	forward
	Temporal expression	since October 2005
	Diagnoses, complains, symptoms	Diabetic retinopathy H36
	Diagnoses, complains, symptoms	strong pains in the feet
	Drugs/Treatment	Insulin Novomix 30 – 32E in the morning, 26E in the evening
Ep3	Reference point	*Now* minus few months
	Direction	forward
	Temporal expression	few months ago
	Diagnoses, complains, symptoms	
	Drugs/Treatment	Diaprel A10BB09
Ep4	Reference point	*Now*
		…..

Table 3. Automatic temporal event sequencing for the case presented in example 2.

Modeling of time is important in information extraction because it would enable deeper meaning understanding. It would support the construction of timeline(s) positioning all events that are described in the text. Complex models of time have been developed, like e.g. the model of Reichenbach, which link the grammatical tense with the event time (we note that tense is a typical feature of verbs but temporal information may be also expressed by temporal adverbs, prepositions etc). However, verbs and complete sentences are rare in clinical narratives. Time is signaled by (short) temporal phrases which are relatively easy to identify. Similarly to the discourse segments considered in natural language understanding [15], time intervals do not overlap in arbitrary manner; we notice that episodes are described in a coherent way before passing to the description of a new episode. In general we find many similarities between the episodes, as defined here, and the discourse segments that are composed in a tree-like structure to build the coherent discourse.

Some temporal references are hard to recognise, for instance *'since then'* in episode 4 of example 1 could be problematic for humans too. The same holds for phrases like e.g. *'Two medication courses were made in 1990 and 1991'*. The events, happening within the episodes, are easier to position and interpret after the recognition of episode boundaries.

We note that the relative temporal clauses need to be resolved by calculation of actual dates or periods (because we want to build a chronologic model of episodes). Some temporal markers are easy to interpret, for instance Episode 1 in Table 3 starts ′5-6 years ago′ which can be interpreted as 5,5 years ago. This episode is shown at Figure 2. Having in mind that the discharge letter in Example 2 was written in February 2010, then the reference point is October 2004. The beginning of episode 3 ′few months ago′ can be considered as October or November 2009.

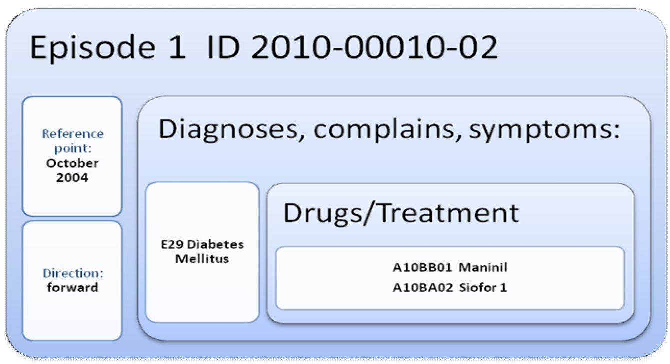

Figure 2. Calculation of dates in Episode 1 of Example 2

We develop our temporal framework by corpus-based investigations on manually annotated discharge letters and evaluation experiments with prototypes of IE components. Our present results are summarised in the next section.

4 Automatic discovery of episodes

In the present IE prototype we have integrated previously developed software components:

- module for extraction of drug names, dosage, frequency and route, which identifies[1] 1537 drug names in 6200 PRs with F-score 98.42% and dosage with F-score 93.85%, [16];

[1] The IE accuracy is measured by the *precision* (percentage of correctly extracted entities as a subset of all extracted entities), *recall* (percentage correctly extracted entities as a subset of all available entities in the corpus) and the *F-score* = 2**Precision***Recall* / (*Precision* + *Recall*).

- module for automatic assignment of ICD-10 codes to diagnoses [17], which was extended considerably to tackle the diagnoses in a large corpus of 6200 PRs [18]. Its current precision is 84.5% for the disease names occurring in the PR section Diagnoses.

In this way our efforts were directed mostly to recognition of symptoms and complain, as well as on systematic study of various temporal markers.

Regarding the automatic recognition of symptoms and complain in the free text of hospital PRs, these are described by a variety of expressions, ranging from the medical terminology of symptoms given in textbooks and encyclopedia to the free explanations and paraphrases, using the wording of the patient. There are many cue phrases, which additionally explain and emphasize some details, as well as inclusion of stories told when the patient is interviewed in the hospital admission office (the latter may also contain temporal references). In this way the automatic identification of symptoms and complain is a complicated task which requires incremental construction of lexicons and training corpora for specific diseases of interest.

At present we perform experimental tests with 1375 discharge letters where the IE prototype discovers 29178 key terms or markers (in average 21,22 key terms and markers per PR). The distribution of these terminologies and temporal markers is the following one:

- 7092 occurrences of drug names were met in 1213 discharge letters,
- 6436 diagnoses are referred to in 1292 discharge letters,
- 1274 complains are recognised in 841 discharge letters and
- 7149 temporal markers were identified in 1374 discharge letters.

It turns out that the hospital PRs contain a significant amount of temporal information, see Fig. 2.

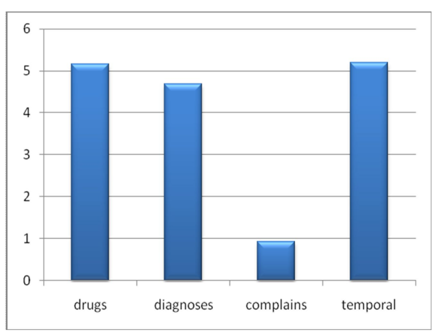

Figure 2. Average number of entities (diagnoses, drugs, symptoms and temporal markers) in 1375 discharge letters of diabetic patients

Considering the percentage of key terms and markers, we see that the share of temporal markers is significant (33%, see Figure 3). This proves the importance of temporal information for the description of the case history. Major errors in the automatic recognition of episode markers and their correct interpretation are due to the following phenomena:

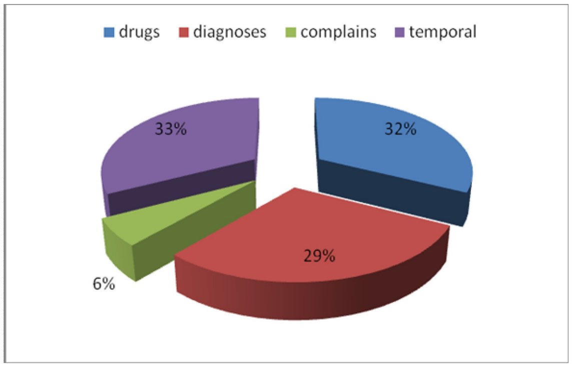

Figure 3. Percentage of temporal markers in discharge letters

- often use of *abbreviations* to denote intervals of time: *'y.'/'ye'./'yea.'* for *'year'*, followed or not by full stop or other punctuation mark (in Bulgarian the word *'year'* (*'година'*) consist of 6 symbols and there are more variants for abbreviations. The same holds for *'month'*, *'week'*, *'day'* etc.;
- sophisticated *prepositional phrases* for marking start, duration, cycle or interval: *'in 3 months'*, *'per 3 months'*, *'for 3 months'* etc. Often the understanding is possible only by interpretation in the context;
- *ambiguity* in the use of temporal phrases, e.g. *'per day'* may participate in the dosage of some drug and then *'day'* should not be treated as a temporal marker, which signals a new episode;
- *reference to multiple time moments* in one token, e.g. *'5-6 years ago'*, *'2001-02'*, etc.
- *variety of tokens denoting the same time*, e.g. *'September 2009'*, *'m09.2009'*, *'M09.2009'*, *'Sept.2009'*, *'09/2009'* and so on;
- *fuzzy* and *non-determined references*, like e.g. *'few months ago'*;
- *anaphoric references* to previously introduced moments of time, e.g. *'since then'*, *'simultaneously'*, *'before'*, *'after'*, *'after that'*, *'several months after that'*, *'about the end of the year'*, *'at the same time'* etc.;
- *spelling errors.*

The temporal markers are identified by an empirically-elaborated context-free grammar, which is run initially with simple rules and is under incremental development. We present here a fragment of few simplified rules with literals in English, which recognise markers like *'since 2 weeks'*, *'since about 3 years'* and so on:

```
<time-marker> ::= <past-definite-period> | <past-definite-month-period> | <past-definite-month-ro-period >
                  | <duration-year> | ...
<past-definite-period> ::= <past-definite-preposition> | <integer> | <temp-duration>
<past-definite-preposition> ::= "since" | "since about" | ...
<temp-duration> ::= "months" | "years" | "weeks" | "days "
```

The present recall in the recognition of the temporal markers in about 57% and the precision is 84%. There is some over-generation too, i.e. the system generates more temporal markers than appropriate. We consider our present achievements as work in progress, which has to be developed further.

Episodes are ordered in a sequence by a simple procedure which tries to calculate the actual date and constructs a list of linearly-ordered reference points. We note that complicated time reasoners are needed to cope with the interpretation of temporal information in clinical narratives but according to [19] these research tasks at in their embryonic stage. The progress requires theoretic models as well as large training corpora of annotated texts which are too expensive to construct.

At the end we would like to comment that temporal modelling of clinical texts can borrow theories from computational linguistics and contextualise them accordingly. For instance, it is interesting to note how the discourse focus is shift. Example 2 shows that 3 episodes are uttered in the following sequence: *(i)* the oldest one (point of diagnosis in 2004), *(ii)* a recent one (2009), and *(iii)* an older one (2005). The narrative convention is not kept in this case because the writer prefers to close the issue (focal space) of Maninil which is replaced by Diaprel because of its ineffectiveness, so all the information about the application of Maninil is gathered in neighboring sentences. This example shows that discussing the medications is more important for the writer than the narrative convention. In our view a deeper study of the specific clinical discourse would help to acqure empiric rules for temporal reasoning in this domain, by explicating how temporal information is communicated in clinical texts.

5 Conclusion

Extraction of time-related information is a challenging research task. It is very important in medical informatics because time in medicine is essential to assess the speed of disease manifestation and development, the progress and effectiveness of treatments and so on. Success in extraction of temporal information would improve the clinical decision support systems. Unfortunately the task is very difficult and does not become simpler when the considerations are narrowed down from general NLP to medical texts. Much work is needed to develop the necessary corpora and conceptual resources which might support the implementation of advanced prototypes.

This article presents our current results in automatic segmentation of case histories into chronology of episodes. We have achieved some progress due to already implemented components which deliver reliable information about the diagnoses and medication events. The elaboration of a time reasoner with proper complexity is an important aspect of the temporal modeling and is a target for our future work.

Acknowledgements. The research work presented in this paper 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.

References

1. Sowa, John F. (2000) *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks Cole Publishing Co., Pacific Grove, CA.
2. Etzion, O. and P. Niblett. *Event Processing in Action*. Manning Publications Co., 2010.
3. 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.
4. Saurı R., B. Knippen, M. Verhagen, and J. Pustejovsky. *Evita: A robust event recognizer for question-answering systems*. In Proc. Int. Conference on Human Language Technologies – Empirical Methods in NLP, 2005.
5. 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.
6. 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 October; 42(5): 839–851.
7. 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.
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*. Journal of American Medical Informatics Association 2005, 12(1), pp. 55-63.
9. 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 Annual Symposium 2005, pp. 869–873.
10. Zhou L., S. Parsons, and G. Hripcsak. *The evaluation of a temporal reasoning system in processing clinical discharge summaries.* Journal of American Medical Informatics Association 2008 15(1), pp. 99-106.
11. Hahn U., M. Romacker, and S. Schulz. *MedSyndikate – a natural language system for the extraction of medical information from findings reports*. International Journal of Medical Informatics 2002, 67(1–3), pp. 63–74.
12. Ruiz E., M. Chilov, S.B. Johnson, and E. Mendonça. *Developing multilevel search filters for clinical questions represented as conceptual graphs.* In Proc. AMIA Annual Symposium 2008, p. 1118.
13. Boytcheva, S. and G. Angelova. *Towards Extraction of Conceptual Structures from Electronic Health Records*. In: Rudolph, S., F. Dau, and S. O. Kuznetsov (Eds.): Proceedings of the 17th Int. Conf. on Conceptual Structures (ICCS'09), July 2009, Moscow, Russian Federation. Springer, Lecture Notes in Artificial Intelligence 5662, pp. 100–113.
14. Tcharaktchiev, D. *Hospital Information Systems*. Sofia, Kama, 2003 (in Bulgarian).
15. Allen, J. Natural Language Understanding, 2nd Edition, Benjamin/Cummings, 1995.
16. Boytcheva, S. *Shallow Medication Extraction from Hospital Patient Record*, Submitted to 2nd Int. PSIP Workshop on Patient Safely through Intelligent Procedures in Medication, Paris, May 2011.
17. Boytcheva S. *Assignment of ICD-10 Codes to Diagnoses in Hospital Patient Records in Bulgarian*. In: Alfred, R., G. Angelova and H. Pfeiffer (Eds.). Proc. of the Int. Workshop "Extraction of Structured Information from Texts in the Biomedical Domain" (ESIT-BioMed 2010), associated to ICCS-2010, Kuching, Sarawak, Malaysia, 26 July 2010, Published by MIMOS BERHAD, ISBN 978-983-41371-3-7, July 2010, pp. 56-66.
18. Boytcheva S., Zh. Angelov, G. Angelova and D. Tcharaktchiev. *Semantic Mining and Information Extraction from Bulgarian Texts of Hospital Patient Records*, Deliverable 2.4 of PSIP Project Patient Safely through Intelligent Procedures in Medication, January 2011, www.psip-project.eu

19. Zhou L. and G. Hripcsak. *Temporal reasoning with medical data - a review with emphasis on medical natural language processing*. Journal of Biomedical Informatics 2007, 40(2), pp. 183-202.