A Workbench for Temporal Event Information Extraction from Patient Records

Svetla Boytcheva[1,2], Galia Angelova[1]

[1]Institute of Information and Communication Technologies,
Bulgarian Academy of Sciences (IICT-BAS), Sofia, Bulgaria

[2]American University in Bulgaria, Blagoevgrad, Bulgaria

sboytcheva@aubg.bg, galia@lml.bas.bg

Abstract. This paper presents a research prototype for temporal event information extraction from hospital discharge letters in Bulgarian. An algorithm for extraction of primitive events automatically sets markers for patients' complaints, drug treatment and diagnoses with precision about 90%. Specific domain knowledge is further used to generate compound events and to identify some relations between event time sequences. Absolute and relative time information enables ordering the generated compound events using semi-intervals and fuzzy logic. Some negated events are analyzed as well to better structure the patient history.

Keywords: Information Extraction, Health Informatics, Temporal Events

1 Introduction

Temporal information extraction (IE) from narratives in patient records is a relatively new task in biomedical natural language processing (NLP) [19]. These IE systems rely on time-oriented patient data and time-based medical knowledge. Temporal reasoning is their key feature [9, 5, 17]. Various temporal IE systems have been demonstrated: summarizing data from temporal clinical databases, reasoning on temporal clinical data for therapeutic assessments, and modeling uncertainty in clinical knowledge and data [1]. Recent activities are often focused on annotation tasks. Initiatives like "i2b2/VA Track on Challenges in NLP for Clinical Data" [14] were conceived in the context of TimeML and TimeBank. TimeML is an ISO standard for annotation of temporal information [12]; TimeBank is a hand-annotated corpus conforming to TimeML [13]. In this approach all events (medical and others) are tagged. A clinical annotation schema based on TimeML was used to annotate corpus of more than 5000 tokens [15]. The HPI TimeML Corpus contains TimeML-annotated "History of Present Illness" sections of 44 discharge summaries [11]. Additional annotation for tense (past, present, future) corresponding to the whole event time with respect to patient hospital entrance is associated to each event. It is quite difficult to compare and assess the resulting corpora because most of them are not publicly available.

Although there are available significant research results for temporal events extraction in English, they cannot be directly applied for patient records (PRs) in Bulgarian due to the specific medical terminology and the lack of digital resources. A mixture of terminology in Bulgarian, Latin and transliterated to Cyrillic Latin terms occurs in Bulgarian medical PRs. Another specificity is that the anonymisation procedure removes admission and discharge dates from PRs, which causes difficulties in further linkage between events when the text processing is done outside the hospital system.

Our approach recognizes dates and prepositional phrases containing temporal expressions. Specific domain knowledge is used to combine them into inter-related compound events. Temporal representation and reasoning are based on fuzzy logic.

Section 2 presents the discourse structure of discharge letters and discusses time interval representations. Section 3 considers existing prototypes for section splitting and temporal IE. Section 4 presents the evaluation of the current workbench for automatic extraction of temporal markers. Section 5 sketches further work and the conclusion.

2 Materials

We process anonymised hospital Discharge Letters in Bulgarian. We identify temporal events and their inter-relations from the Anamnesis section of the Patient Record (PR) that narrates the patient's disease history with rich temporal references. Our IE experiments were performed on a training corpus of 1,300 and test corpus of 6,200 anonymised hospital PRs for patients with endocrine and metabolic diseases. We adopt the event classification [8] into primitive and composite events. Primitive events can be explicit and temporal (relative and absolute).

The PRs text is split automatically into episodes (1). We assume that the Patient history is represented as a sequence of adjacent episodes:

$$e_1, e_2, \ldots, e_n \tag{1}$$

An episode is represented as a vector (2) of primitive events in the episode:

$$e_i = < PID_{i1}, PID_{i2}, \ldots, PID_{ik} > \tag{2}$$

Different episodes can contain different numbers of primitive events.

Primitive events contain descriptions for diagnoses, complaints, drugs, procedures or complications related to them and they are associated with time markers:

$$PE(PID, PRID, TimeBegin, TimeEnd, Mode, Type, Data) \tag{3}$$

Here `PID` is an unique ID associated with the primitive event; `PRID` is patient record's ID; `TimeBegin` is the initial time when the event is initiated; `TimeEnd` is the event termination time; `Mode` can be *positive* (the event happened), *negative* (the event is negated) and *conditional* (the event is optional and there are additional conditions that identify whether the event can happen); `Type` specifies information described in this primitive event - about drug, complaints, diagnose, procedure, lab result, or status; `Data` contains structured information (e.g. diagnose/drug codes).

We use previously developed extractors of ICD-10 (the International Classification of Diseases, v. 10[1]) and ATC (the Anatomical Therapeutic Chemical Classification System[2]) codes. They assign ICD-10 codes to disease names with 84.5% precision and ATC codes to drug names with f-measure 98.42% [6,7].

Most of the time episodes in patient discharge letters contain no explicit information about beginnings and ends of intervals, which is required by Allen's theory of temporal intervals [2]. Several Allen's relations require the equality of two or more interval boundaries. However we process data about chronic deceases and in most cases only the beginning is marked. In order to deal with imperfect and incomplete information in patient records we use Freksa's theory of semi-intervals [10]. According to this theory the semi-intervals provide the following advantages: (i) they are rather natural entities both from a cognitive and from a computational point of view; (ii) coarse knowledge can be processed directly; computational effort is saved; (iii) incomplete knowledge about events can be fully exploited; (iv) incomplete inferences made on the basis of complete knowledge can be used directly for further inference steps. Freska defines eleven semi-interval relationships and depending on the types of deformation of events and their relations, he defines different neighborhood structures: A-neighbor relation - when for two events three of the four semi-intervals are fixed and allow the fourth to be moved; B-neighbor relation – in case we leave the duration of events fixed and allow complete events to be moved in time; C-neighbor relation – when we leave the 'temporal location' of an event fixed and allow the duration of the events to vary.

An interesting point in the use of time and tenses in natural language was brought out by Anscombe's investigation into the meanings of *before* and *after* [4]. For instance, from "The infection was present after the fever ended" it does not follow that the fever ended before the infection was present. Thus, *before* and *after* are not strict converses. Note that, however, from "The infection started after the fever started," we can indeed conclude that the fever started before the infection started. Therefore *before* and *after* are converses when they link instantaneous events.

Another problem occurs in reasoning, because in First Order Logic, objects exist timelessly, time being just another dimension; in tenser approaches, "now" is a point of time in a separate class. Thus Fuzzy Logic [18] is a more convenient approach for modeling temporal data and reasoning on it.

3 Methods

3.1 System Architecture

The temporal information extractor processes the input PR text as follows (Fig. 1):

- ***Sections splitting*** - based on regular expressions that recognize more than 70 keywords for sections names as well as missing or merged sections;

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

- ***Setting markers for primitive events*** (drug names, diagnoses, complaints) module - uses about 80 regular expressions. This module also uses especially developed extractors of ICD-10 and ATC codes for diagnoses and drug names [6,7];

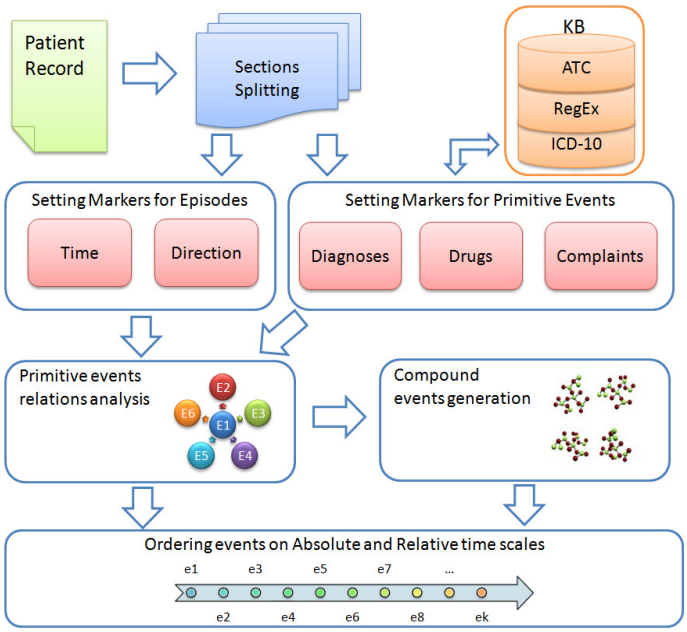

Fig. 1. System Architecture

- ***Setting markers for episodes*** (time and direction) module - uses about 50 rules defined as regular expressions based on keywords for directions, absolute and relative time (date and duration) [3,16]. Keyword lists include various types of absolute time representation like e.g. month description – by name, by Roman and Arabic numerals; dates and years for instance "since June 2009", "in 2010" etc. There are also many keywords/phrases for relative time representation like e.g. "since then", "after the puberty", "after X was stopped" etc.

- ***Primitive event relation analysis*** - this module groups primitive event for drugs, complaints and diagnoses according to episodes, time and directions markers. The module sets `TimeBegin` and `TimeEnd` values for primitive events. In principle, an episode contains no/single/many time markers. In the first case we assume that this episode continues the explanations from the last time marker t_i mentioned in some preceding $episode_i$ (4). Thus all events from e_{im} to e_{jn} in the current $episode_j$ are related to the same time t_i

$$\underbrace{\ldots \; t_i \;\; e_{im} \;\; \ldots \;\; e_{ik}}_{episode_i} \ldots \underbrace{e_{j1} \;\; e_{j2} \;\; \ldots \;\; e_{jn}}_{episode_j} \tag{4}$$

In case there is a single time marker in the episode (5) – all primitive events from e_{i1} to e_{ik} are related to this time t_i.

$$\underbrace{e_{i1} \quad \dots \quad e_{im-1} \quad t_i \quad e_{im} \quad \dots \quad e_{ik}}_{episode_i} \tag{5}$$

In case of many time markers (6) – the primitive events in the episode are split according to them, and all primitive events between two time markers are related to the time marker preceding them

$$\underbrace{\dots \quad t_{i1} \quad e_{im} \quad \dots \quad e_{ik} \quad t_{i2} \quad e_{ik+1} \quad \dots \quad e_{in}}_{episode_i} \tag{6}$$

In this example the primitive events from e_{im} to e_{ik} are related to the time t_{i1} and primitive events from e_{ik+1} to e_{in} are related to the time t_{i2}.

- ***Compound event generation*** - this module groups primitive events in compound events like e.g. treatment (all drugs prescribed for the period, all diagnoses). This allows further reasoning and finding cause-effect relations. Absence of complaints, symptoms, treatment with some drugs or diagnose is important information for further patient treatment.
- ***Ordering events on Absolute and Relative time scales*** - the algorithm for event ordering is based on directed multi-graphs representation. Where time markers are nodes (states), the edges represent primitive events, and they are incident with the beginning and end time nodes. Two graphs are generated – one for relative and one for absolute time scales. Some of the fuzzy relations are resolved.

3.2 Example

The patient history for PR ID 03211 has been automatically broken into primitive evens. Table 1 shows the extracted primitive events: 4 diagnoses (E1, E4, E11 and E12), 8 drugs (E2, E3, E5-E9, E14) and 3 complaints (E10, E13, E15). They are grouped according to the type of time markers – dates and durations. For only two of them (E1 & E2) are associated `TimeBegin` and `TimeEnd` values for absolute time scale (Fig. 2). Even in the narrative PR there is nothing mentioned about the value of `TimeBegin` for E2, therefore it is assigned to the same time as E1, because they are recognized as related primitive events grouped in compound event treatment. The absolute time scale starts at the patient's birth date and ends at the admission date. The relative time scale values later are resolved by mapping it to the absolute time scale and according to the admission date. The Compound Events generated from the primitive events pool by mapping time intervals to absolute and relative time scales and resolving their overlapping using Fuzzy logic are shown on Fig. 3. The resulting set contains 6 treatments (Diagnoses, Drugs, Complaints), but some of them are incomplete.

Table 1. Extracted events information from PR ID 03211

Event	ID	Time Begin	Time End	Mode	Type	Data
E1	03211	since 2001	N/A	pos	diagnose	E11, Non-insulin-dependent diabetes mellitus
E2	03211	N/A	until 2009	pos	drug	A10BB09, Diaprel MR, 2 tabl
E3	03211	about 1 year ago	N/A	pos	drug	A10BA02, Metformin NIHFI, 3x850mg
E4	03211	more than 15 years ago	N/A	pos	diagnose	I158, Other secondary hypertension
E5	03211	at present	N/A	pos	drug	C03BA11, Tertensif SR, 1 tabl
E6	03211	during the last week	N/A	pos	drug	C09AA02 , Renapril, 40mg
E7	03211	at present	N/A	pos	drug	C09AA02 , Renapril, 2x10 mg
E8	03211	at present	N/A	pos	drug	C07AB02, Betaloc, 50mg
E9	03211	at present	N/A	cond	drug	C02AC01, Chlophazolin, when needed
E10	03211	about 2 years ago	N/A	pos	complaints	serum creatinine, increased levels
E11	03211	about 1 year ago	N/A	pos	diagnose	O11X, Pre-exist hypertens disorder with superimposed proteinuria
E12	03211	at present	N/A	neg	diagnose	H360, Diabetic retinopathy
E13	03211	during last 10 days	N/A	pos	complaints	Hypertension
E14	03211	during last 10 days	N/A	pos	drug	C02AC01, Chlophazolin
E15	03211	at present	N/A	neg	complaints	decompensation of diabetes mellitus

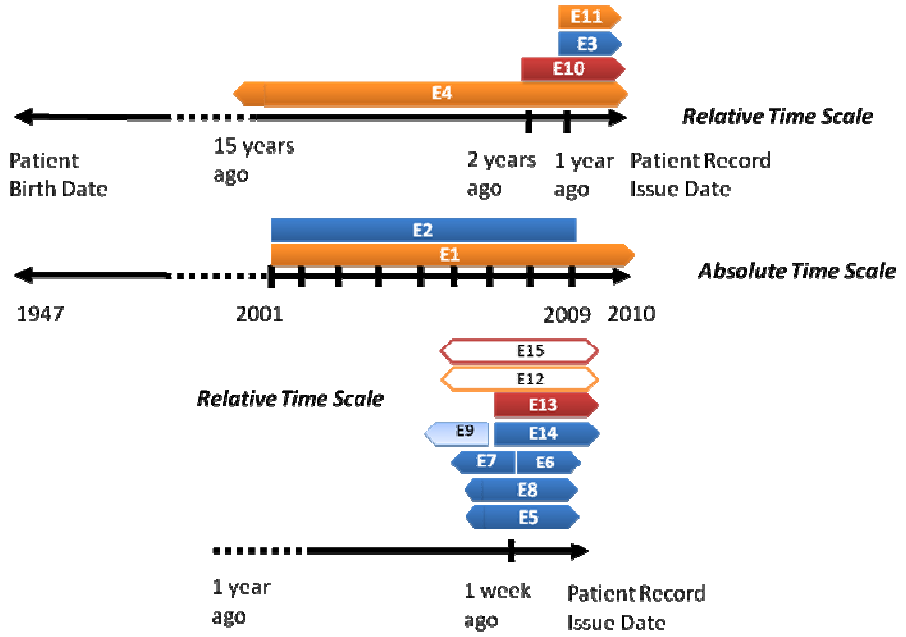

Fig. 2. Relative and Absolute time scales for primitive events from PR ID 03211

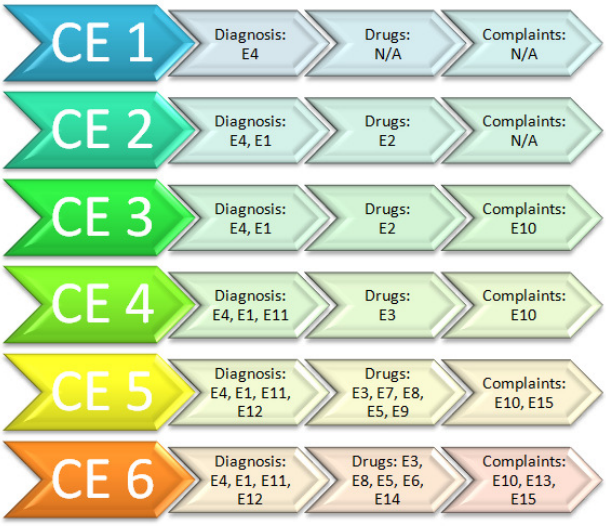

Fig. 3. Compound Events generated for PR ID 03211

3.3 Workbench Prototype

The workbench prototype is implemented in C#.Net as multidocument container. It provides functionalities for single and multiple PRs automatic analyses.

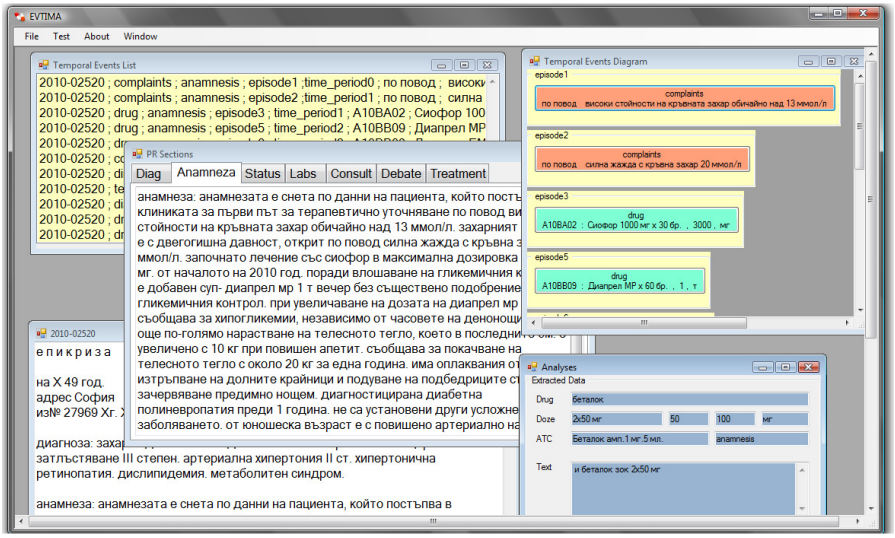

Fig. 4. Workbench for Temporal Events Information Extraction

For single PR processing the user can start manually each of the separate modules of the system and to monitor the analyses result in separate windows (see Fig. 4). One window contains the original PR text; a tabbed window displays the PR split into sections; another window shows step by step the extracted drug and diagnoses information. The list of extracted primitive events is presented in CSV format and the generated compound events and primitive events relations are presented in automatically generated graphical representation where different types of markers (for time, complaints, drugs, diagnoses) are shown by different colors and the compound events are shown in Group Boxes (see Fig. 4, the Temporal Events Diagram on the right side).

The multiple PRs processing mode generates CVS and XML files containing information about all PRs in a selected folder.

4 Evaluation and Results

In NLP the performance accuracy of text extraction procedures is usually measured by the *Precision P* (percentage of correctly extracted entities as a subset of all extracted entities), *Recall R* (percentage correctly extracted entities as a subset of all entities available in the corpus) and their harmonic mean f-score: $F1 = 2.P.R / (P+R)$.

The experiments were made with a training corpus containing 1,300 PRs and the evaluation results are obtained using a test corpus, containing 6,200 PRs. The evaluation results (Table 2) show high percentage of success in events and time information recognition in the PRs texts. These results are comparable with best systems accuracy like [5, 11] which reports results for relatively small corpora.

Table 2. Extraction sensitivity according to the IE performance measures

		Precision	Recall	F-Score
EVENT	**Drugs**	97.28%	99.59%	98.42%
	Diagnoses	97.30%	74.68%	84.50%
	Complaints	97.98%	96.82%	97.40%
TIME	**Dates**	98.86%	98.21%	98.53%
	Duration	99.14%	98.26%	98.70%
	Frequency	92.25%	95.51%	93.85%

For the Test set (6,200 PRs) were set 104,426 temporal markers in total and for the Training set (1,300 PRs) were set 24,924 temporal markers. The obtained results show that these markers are distributed approximately in the same percentage for both sets – about 38% of events present diagnosis descriptions, 47% are drug events and the remaining 15% are complaints. Table 3 presents summary statistics for the corpus. The average amount of primitive events per PRs is 20.69. The training sets contains

information about 530 different drug names and 371 different diagnoses; in the test set data were found about 666 different drugs and 565 different diagnoses. The results are not surprising, because we are processing data for patients from a specialized hospital for treatment of endocrine disorders and most of them have similar diagnoses and treatment.

	Sentences	**Tokens**	**EVENTS**	**TIME**
Training	27,155	250,773	14,137	10,796
Test	108,424	1,010,996	58,472	45,954
Total	**135,579**	**1,261,769**	**72,609**	**56,750**

Table 3. Summary statistics for the corpus

In the test set the IE prototype has identified 1,349 temporal markers for complete date information (day/month/year), 2,698 absolute time markers with incomplete date information (year and/or month only), 2,362 markers for relative time periods and only 2,351 concerning the admission date.

Most incorrect event recognitions are due to: misspelling errors, unrecognized drug events for allergies, incorrect detection of negation scope, drug events occurrence in other context, unrecognized abbreviations, incorrect transliteration of Latin terminology and descriptions of specific pathological states which are hard to classify according to ICD-10 even for humans.

5 Conclusion and Further Work

The paper presents software modules which support the automatic extraction of temporal events information from PR texts.

The IE modules are strictly oriented to Bulgarian language and the structure of Bulgarian discharge letters. The plans for their further development and application are connected primarily to Bulgarian local context. Future enhancements are planned for the extension of the drug name and dosage recognition rules, to cope with certain specific exceptions. The preliminary correction of spell errors and other kinds of typos will also increase the IE accuracy. Regarding the diagnoses recognition task we plan improvement of the rules for more precise code assignments. Further work includes improvement of compound events generation and events ordering algorithms.

6 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.

7 References

1. Adlassnig K.-P., C. Combi, A.K Das, E.T. Keravnou, and G. Pozzi. Temporal Representation and Reasoning in Medicine: Research Directions and Challenges, *AI in Medicine*, 38(2):101-113, 2006.
2. Allen, James F.. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11): 832– 843, 1983.
3. Angelova G. and S. Boytcheva, Towards Temporal Segmentation of Patient History in Discharge Letters. In *Proc. of Biomedical NLP Workshop*, in conjunction with the Int. Conference RANLP-2011, Hissar, Bulgaria, 49-54.
4. Anscombe, G.E.M. Before and after. *The Philosophical Review*, 73:3–24, 1964.
5. Boland MR, SW Tu, S. Carini, I. Sim, and C. Weng, *EliXR-TIME: A Temporal Knowledge Representation for Clinical Research Eligibility Criteria*, Proc. of AMIA 2012 Clinical Research Informatics Summit, San Francisco, CA, 71-80, 2012.
6. Boytcheva, S. Shallow Medication Extraction from Hospital Patient Records. In: Koutkias, V., J. Niès, S. Jensen, N. Maglaveras, and R. Beuscart (Eds.), *Studies in Health Technology and Informatics series*, Vol. 166, IOS Press, 119-128, 2011.
7. Boytcheva S., Automatic Matching of ICD-10 codes to Diagnoses in Discharge Letters. In *Proc. of Biomedical NLP Workshop*, in conjunction with the Int. Conference RANLP-2011, 15 September 2011, Hissar, Bulgaria, 19-26, 2011.
8. Chakravarthy S., H. Lee, and H. Kim, Support for Temporal Events in Sentinel. Design, Implementation and Preprocessing, RL-TR-97-214 Final Technical Report, *Air Force Research Laboratory, Rome Research Site*, New York, February 1998.
9. Combi, C., E. Keravnou-Papailiou, and Y. Shahar, *Temporal Information Systems in Medicine,* Springer, 1st Edition, 397 pages, 2010.
10. Freksa, C.. Temporal reasoning based on semi-intervals. *J. of AI*, 54(1):199–227, 1992.
11. Galescu L. and N. Blaylock. A Corpus of Clinical Narratives Annotated with Temporal Information. *Proc. 2nd ACM SIGHIT Int. Health Inf. Symp. (IHI 2012)*. Miami, Jan. 2012.
12. Pustejovsky, J., J. Castaño, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer, and G. Katz. TimeML: Robust Specification of Event and Temporal Expressions in Text. *In: Proc. of the Fifth Int. Workshop on Computational Semantics (IWCS-5)*. Tilburg, 2003.
13. Pustejovsky, J., P. Hanks, R. Saurí, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, and M. Lazo. The TIMEBANK Corpus. *In: Proc. of Corpus Linguistics 2003*. Lancaster, 647–656, 2003.
14. Rink, B., S. Harabagiu, and K. Roberts. Automatic extraction of relations between medical concepts in clinical texts. *J. of AMIA*, 18:594-600, 2011.
15. 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*, 568-572, 2009.
16. Tcharaktchiev D., S. Boytcheva, and G. Angelova. Empirical Approach to Event Sequencing in Automatic Analysis of Patient Records. *Poster presentation at the 23rd Int. Conf. MIE 2011*, Norway August 2011.
17. Weng C., X. Wu, Z. Luo, M. Boland, D. Theodoratos, and S.B. Johnson, EliXR: An Approach to Eligibility Criteria Extraction and Representation. *JAMIA*, Suppl. 1:i116-i124, 2011.
18. Zadeh L.A.. A theory of approximate reasoning. In J.E. Hayes, D. Michie, and L.I Mikulich, editors, *Machine intelligence*, Elsevier, Amsterdam, pages 149-194, 1979.
19. Zhou L. and G. Hripcsak. Temporal reasoning with medical data - a review with emphasis on medical natural language processing. *J. Biom. Informatics*, 40(2), 183-202, 2007.