Automatic Analysis of Patient History Episodes in Bulgarian Hospital Discharge Letters

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

Galia Angelova, Ivelina Nikolova
Institute of Information and
Communication Technologies (IICT),
Bulgarian Academy of Sciences (BAS)
{galia,iva}@lml.bas.bg

Abstract

This demo presents Information Extraction from discharge letters in Bulgarian language. The *Patient history* section is automatically split into episodes (clauses between two temporal markers); then drugs, diagnoses and conditions are recognised within the episodes with accuracy higher than 90%. The temporal markers, which refer to *absolute* or *relative* moments of time, are identified with precision 87% and recall 68%. The direction of time for the episode starting point: *backwards* or *forward* (with respect to certain moment orienting the episode) is recognised with precision 74.4%.

1 Introduction

Temporal information processing is a challenge in medical informatics (Zhou and Hripcsak, 2007) and (Hripcsak et al., 2005). There is no agreement about the features of the temporal models which might be extracted automatically from free texts. Some sophisticated approaches aim at the adaptation of TimeML-based tags to clinically-important entities (Savova et al., 2009) while others identify dates and prepositional phrases containing temporal expressions (Angelova and Boytcheva, 2011). Most NLP prototypes for automatic temporal analysis of clinical narratives deal with discharge letters.

This demo presents a prototype for automatic splitting of the *Patient history* into episodes and extraction of important patient-related events that occur within these episodes. We process Electronic Health Records (EHRs) of diabetic patients. In Bulgaria, due to centralised regulations on medical documentation (which date back to the 60's of the last century), hospital discharge letters have a predefined structure (Agreement, 2005). Using the section headers, our Information Extraction (IE) system automatically identifies the *Patient history* (Anamnesis). It contains a summary, written by the medical expert who hospitalises the patient, and documents the main phases in diabetes development, the main interventions and their effects. The splitting algorithm is based on the assumption that the *Patient history* texts can be represented as a structured sequence of adjacent clauses which are positioned between two temporal markers and report about some important events happening in the designated period. The temporal markers are usually accompanied by words signaling the direction of time (backward or forward). Thus we assume that the episodes have the following structure: *(i)* reference point, *(ii)* direction, *(iii)* temporal expression, *(iv)* diagnoses, *(v)* symptoms, syndromes, conditions, or complains; *(vi)* drugs; *(vii)* treatment outcome. The demo will show how our IE system automatically fills in the seven slots enumerated above. Among all symptoms and conditions, which are complex phrases and paraphrases, the extraction of features related to polyuria and polydipsia, weight change and blood sugar value descriptions will be demonstrated. Our present corpus contains 1,375 EHRs.

2 Recognition of Temporal Markers

Temporal information is very important in clinical narratives: there are 8,248 markers and 8,249 words/phrases signaling the direction backwards or forward in the corpus (while the drug name occurrences are 7,108 and the diagnoses are 7,565).

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 77–81,
Avignon, France, April 23 - 27 2012.

In the hospital information system, there are two explicitly fixed dates: the patient birth date and the hospitalisation date. Both of them are used as antecedents of temporal anaphora:

- the hospitalisation date is a reference point for 37.2% of all temporal expressions (e.g. '*since 5 years*', '*(since) last March*', '*3 years ago*', '*two weeks ago*', '*diabetes duration 22 years*', '*during the last 3 days*' etc.). For 8.46% of them, the expression allows for calculation of a particular date when the corresponding event has occurred;
- the age (calculated using the birth date) is a reference point for 2.1% of all temporal expressions (e.g. '*diabetes diagnosed in the age of 22 years*').

Some 28.96% of the temporal markers refer to an explicitly specified year which we consider as an *absolute* reference. Another 15.1% of the markers contain reference to day, month and year, and in this way 44.06% of the temporal expressions explicitly refer to dates. Adding to these 44.06% the above-listed referential citations of the hospitalization date and the birth date, we see that 83.36% of the temporal markers refer to explicitly specified moments of time and can be seen as *absolute* references. We note that diabetes is a chronicle disease and references like '*diabetes diagnosed 30 years ago*' are sufficiently precise to be counted as explicit temporal pointers.

The anaphoric expressions refer to events described in the *Patient history* section: these expressions are 2.63% of the temporal markers (e.g. '*20 days after the operation*', '*3 years after diagnosing the diabetes*', '*about 1 year after that*', '*with the same duration*' etc.). We call these expressions *relative temporal markers* and note that much of our temporal knowledge is relative and cannot be described by a date (Allen, 1983).

The remaining 14% of the temporal markers are undetermined, like '*many years ago*', '*before the puberty*', '*in young age*', '*long-duration diabetes*'. About one third of these markers refer to periods e.g. '*for a period of 3 years*', '*with duration of 10-15 years*' and need to be interpreted inside the episode where they occur.

Identifying a temporal expression in some sentence in the *Patient history*, we consider it as a signal for a new episode. Thus it is very important to recognise automatically the time anchors of the events described in the episode: whether they happen *at* the moment, designated by the marker, *after* or *before* it. The temporal markers are accompanied by words signaling time direction *backwards* or *forward* as follows:

- the preposition '*since*' (от) unambiguously designates the episode startpoint and the time interval when the events happen. It occurs in 46.78% of the temporal markers;
- the preposition '*in*' (през) designates the episode startpoint with probability 92.14%. It points to a moment of time and often marks the beginning of a new period. But the events happening after '*in*' might refer backwards to past moments, like e.g. '*diabetes diagnosed in 2004, (as the patient) lost 20 kg in 6 months with reduced appetite*'. So there could be past events embedded in the '*in*'-started episodes which should be considered as separate episodes (but are really difficult for automatic identification);
- the preposition '*after*' (след) unambiguously identifies a relative time moment oriented to the immediately preceding event e.g. '*after that*' with synonym '*later*' e.g. '*one year later*'. Another kind of reference is explicit event specification e.g. '*after the Maninil has been stopped*';
- the preposition '*before*' or '*ago*' (преди) is included in 11.2% of all temporal markers in our corpus. In 97.4% of its occurrences it is associated to a number of years/months/days and refers to the hospitalisation date, e.g. '*3 years ago*', '*two weeks ago*'. In 87.6% of its occurrences it denotes starting points in the past after which some events happen. However, there are cases when '*ago*' marks an endpoint, e.g. '*Since 1995 the hypertony 150/100 was treated by Captopril 25mg, later by Enpril 10mg but two years ago the therapy has been stopped because of hypotony*';
- the preposition '*during, throughout*' (в продължение на) occurs relatively rarely, only in 1.02% of all markers. It is usually associated with explicit time period.

3 Recognition of Diagnoses and Drugs

We have developed high-quality extractors of diagnoses, drugs and dosages from EHRs in Bulgarian language. These two extracting components are integrated in our IE system which processes *Patient history* episodes.

Phrases designating diagnoses are juxtaposed to ICD-10 codes (ICD, 10). Major difficulties in matching ICD-10 diseases to text units are due to *(i)* numerous Latin terms written in Latin or Cyrillic alphabets; *(ii)* a large variety of abbreviations; *(iii)* descriptions which are hard to associate to ICD-10 codes, and *(iv)* various types of ambiguity e.g. text fragments that might be juxtaposed to many ICD-10 labels.

The drug extractor finds in the EHR texts 1,850 brand names of drugs and their daily dosages. Drug extraction is based on algorithms using regular expressions to describe linguistic patterns. The variety of textual expressions as well as the absent or partial dosage descriptions impede the extraction performance. Drug names are juxtaposed to ATC codes (ATC, 11).

4 IE of symptoms and conditions

Our aim is to identify diabetes symptoms and conditions in the free text of *Patient history*. The main challenge is to recognise automatically phrases and paraphrases for which no "canonical forms" exist in any dictionary. Symptom extraction is done over a corpus of 1,375 discharge letters. We analyse certain dominant factors when diagnosing with diabetes - values of blood sugar, body weight change and polyuria, polydipsia descriptions. Some examples follow:

(i) *Because of polyuria-polydipsia syndrome, blood sugar was - 19 mmol/l.*
(ii) *... on the background of obesity - 117 kg...*

The challenge in the task is not only to identify sentences or phrases referring to such expressions but to determine correctly the borders of the description, recognise the values, the direction of change - increased or decreased value and to check whether the expression is negated or not.

The extraction of symptoms is a hybrid method which includes document classification and rule-based pattern recognition. It is done by a 6-steps algorithm as follows: *(i)* manual selection of symptom descriptions from a training corpus; *(ii)* compiling a list of keyterms per each symptom; *(iii)* compiling probability vocabularies for left- and right-border tokens per each symptom description according to the frequencies of the left- and right-most tokens in the list of symptom descriptions; *(iv)* compiling a list of features per each symptom (these are all tokens available in the keyterms list without the stop words); *(v)* performing document classification for selecting the documents containing the symptom of interest based on the feature selection in the previous step and *(vi)* selection of symptom descriptions by applying consecutively rules employing the keyterms vocabulary and the left- and right-border tokens vocabularies. For overcoming the inflexion of Bulgarian language we use stemming.

The last step could be actually segmented into five subtasks such as: focusing on the expressions which contain the terms; determining the scope of the expressions; deciding on the condition worsening - increased, decreased values; identifying the values - interval values, simple values, measurement units etc. The final subtask is to determine whether the expression is negated or not.

5 Evaluation results

The evaluation of all linguistic modules is performed in close cooperation with medical experts who assess the methodological feasibility of the approach and its practical usefulness.

The temporal markers, which refer to *absolute* or *relative* moments of time, are identified with precision 87% and recall 68%. The direction of time for the episode events: backwards or forward (with respect to certain moment orienting the episode) is recognised with precision 74.4%.

ICD-10 codes are associated to phrases with precision 84.5%. Actually this component has been developed in a previous project where it was run on 6,200 EHRs and has extracted 26,826 phrases from the EHR section *Diagnoses*; correct ICD-10 codes were assigned to 22,667 phrases. In this way the ICD-10 extractor uses a dictionary of 22,667 phrases which designate 478 ICD-10 disease names occurring in diabetic EHRs (Boytcheva, 2011a).

Drug names are juxtaposed to ATC codes with f-score 98.42%; the drug dosage is recognised with f-score 93.85% (Boytcheva, 2011b). This result is comparable to the accuracy of the best

systems e.g. MedEx which extracts medication events with 93.2% f-score for drug names, 94.5% for dosage, 93.9% for route and 96% for frequency (Xu et al., 2010). We also identify the drugs taken by the patient at the moment of hospitalisation. This is evaluated on 355 drug names occurring in the EHRs of diabetic patients. The extraction is done with f-score 94.17% for drugs in *Patient history* (over-generation 6%) (Boytcheva et al., 2011).

In the separate phases of symptom description extraction the f-score goes up to 96%. The complete blood sugar descriptions are identified with 89% f-score; complete weight change descriptions - with 75% and complete polyuria and polydipsia descriptions with 90%. These figures are comparable to the success of extracting conditions, reported in (Harkema et al., 2009).

6 Demonstration

The demo presents: *(i)* the extractors of diagnoses, drugs and conditions within episodes and *(ii)* their integration within a framework for temporal segmentation of the *Patient history* into episodes with identification of temporal markers and time direction. Thus the prototype automatically recognises the time period, when some events of interest have occurred.

Example 1. (April 2004) Diabetes diagnosed last August with blood sugar values 14mmol/l. Since then put on a diet but without following it too strictly. Since December follows the diet but the blood sugar decreases to 12mmol/l. This makes it necessary to prescribe Metfodiab in the morning and at noon 1/2t. since 15.I. Since then the body weight has been reduced with about 6 kg. Complains of fornication in the lower limbs.

This history is broken down into the episodes, imposed by the time markers (table 1). Please note that we suggest no order for the episodes. This should be done by a temporal reasoner.

However, it is hard to cope with expressions like the ones in Examples 2-5, where more than one temporal marker occurs in the same sentence with possibly diverse orientation. This requires semantic analysis of the events happening within the sentences. *Example 2*: Since 1,5 years with growing swelling of the feet which became permanent and massive since the summer of 2003. *Example 3*: Diabetes type 2 with duration 2 years, diagnosed due to gradual body weight reduction during the last 5-6 years. *Example 4*: Secondary amenorrhoea after a childbirth 12 months ago, after the birth with ceased menstruation and without lactation. *Example 5*: Now hospitalised 3 years after a radioiodine therapy of a nodular goiter which has been treated before that by thyreostatic medication for about a year.

Ep	reference	August 2003
	direction	forward
	expression	last August
	condition	blood sugar 14mmol/l
Ep	reference	August 2003
	direction	forward
	expression	Since then
Ep	reference	December 2003
	direction	forward
	expression	Since December
	condition	blood sugar 12mmol/l
Ep	reference	15.I
	direction	forward
	expression	since 15.I
	treatment	Metfodiab A10BA02
		1/2t. morning and noon
Ep	reference	15.I
	direction	forward
	expression	Since then
	condition	body weight reduced 6 kg.

Table 1: A patient history broken down into episodes.

In conclusion, this demo presents one step in the temporal analysis of clinical narratives: decomposition into fragments that could be considered as happening in the same period of time. The system integrates various components which extract important patient-related entities. The relative success is partly due to the very specific text genre. Further effort is needed for ordering the episodes in timelines, which is in our research agenda for the future. These results will be integrated into a research prototype extracting conceptual structures from EHRs.

Acknowledgments

This work is supported by grant DO/02-292 EVTIMA funded by the Bulgarian National Science Fund in 2009-2012. The anonymised EHRs are delivered by the University Specialised Hospital of Endocrinology, Medical University - Sofia.

References

Allen, J. *Maintaining Knowledge about Temporal Intervals*. Comm. ACM, 26(11), 1983, pp. 832-843.

Angelova G. and S. Boytcheva. *Towards Temporal Segmentation of Patient History in Discharge Letters*. In Proceedings of the Second Workshop on Biomedical Natural Language Processing, associated to RANLP-2011. September 2011, pp. 11-18.

Boytcheva, S. *Automatic Matching of ICD-10 Codes to Diagnoses in Discharge Letters*. In Proceedings of the Second Workshop on Biomedical Natural Language Processing, associated to RANLP-2011. September 2011, pp. 19-26.

Boytcheva, S. *Shallow Medication Extraction from Hospital Patient Records*. In Patient Safety Informatics - Adverse Drug Events, Human Factors and IT Tools for Patient Medication Safety, IOS Press, Studies in Health Technology and Informatics series, Volume 166. May 2011, pp. 119-128.

Boytcheva, S., D. Tcharaktchiev and G. Angelova. *Contextualization in automatic extraction of drugs from Hospital Patient Records*. In A. Moen at al. (Eds) User Centred Networked Health Case, Proceedings of MIE-2011, IOS Press, Studies in Health Technology and Informatics series, Volume 169. August 2011, pp. 527-531.

Harkema, H., J. N. Dowling, T. Thornblade, and W. W. Chapman. 2009. *ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports*. J. Biomedical Informatics, 42(5), 2009, pp. 839-851.

Hripcsak G., L. Zhou, S. Parsons, A. K. Das, and S. B. Johnson. *Modeling electronic dis-charge summaries as a simple temporal con-straint satisfaction problem*. JAMIA (J. of Amer. MI Assoc.) 2005, 12(1), pp. 55-63.

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 Sympo-sium 2009, pp. 568-572.

Xu.H., S. P Stenner, S. Doan, K. Johnson, L. Waitman, and J. Denny. *MedEx: a medication information extraction system for clinical narratives*. JAMIA 17 (2010), pp. 19-24.

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.

Agreement fixing the sections of Bulgarian hospital discharge letters. Bulgarian Parliament, Official State Gazette 106 (2005), Article 190(3).

ICD v.10: International Classification of Diseases http://www.nchi.government.bg/download.html.

ATC (Anatomical Therapeutic Chemical Classification System), http://who.int/classifications/atcddd/en.