User Centred Networked Health Care
A. Moen et al. (Eds.)
IOS Press, 2011

© 2011 European Federation for Medical Informatics. All rights reserved.
doi:10.3233/978-1-60750-806-9-527

Contextualization in Automatic Extraction of Drugs from Hospital Patient Records

Svetla BOYTCHEVA [a, c], Dimitar TCHARAKTCHIEV [b], Galia ANGELOVA [a]
[a] *Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Sofia, Bulgaria*
[b] *University Specialized Hospital for Active Treatment of Endocrinology (USHATE), Medical University, Sofia, Bulgaria, dimitardt@gmail.com*
[c] *University of Library Studies and IT, Sofia, Bulgaria*

Abstract. Information Extraction (IE) from medical texts aims at the automatic recognition of entities and relations of interests. IE is based on shallow analysis and considers only sentences containing important words. Thus IE of drugs from discharge letters can identify as *'current'* some past or future medication events. This article presents heuristic observations enabling to filter drugs that are taken by the patients during the hospitalization. These heuristics are based on the default PR structure and linguistic expressions signaling temporal and conditional markers. They are integrated in a system for drug extraction from hospital Patient Records (PRs) in Bulgarian language. Present evaluation results are summarized as well.

Keywords. Natural Language Processing (NLP), Automatic IE from Patient Records, Structuring and Contextualization of medication events

1. Introduction

NLP is viewed as a promising technology that can help to acquire structured information by (partial) understanding of free-text medical documents [1]. Usually only extraction of entities and relations of interest is implemented since full text understanding is hard to achieve. In general the IE success is limited; the overview [2] suggests that the limitations of semantic analysis (the so-called '60-percent barrier') are perhaps due to the shallow processing which tackles only 'what the text wears on its sleeve'. However, the extraction correctness is much higher for well-defined tasks and in well-defined domains. Usually 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 relevant entities available in the corpus) and their harmonic mean called *f-score*. A successful systems is 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 [3]. The systems presented at the NLP shared task 'Medication Extraction Challenge' in 2009 also achieve best scores about 90% [4]. Our extractor identifies 1537 drug names in 6200 Bulgarian PRs with f-score 98,42% and dosage with f-score 93,85% [5]. However, recognition of the mere drug name occurrences delivers no information regarding the *'present treatment'* since numerous past or future medication events might be discussed in a single discharge letter. This article summarizes recent results related to the following research challenge: given an

IE component which recognizes medication events with high accuracy, how to filter drugs taken by the patient at the moment when the discharge letter is composed.

2. Project Background

Joining the project PISP [6] via a FP7 ICT Call for extension of running projects, our aim is to extract drugs from hospital PRs in order to fill in a PSIP-compliant repository and to validate PSIP rules for Adverse Drug Events. The drugs, prescribed through the Hospital Pharmacy of USHATE, are sent to the PSIP repository via the Hospital Information System. But USHATE is a specialized hospital which treats endocrine diseases of patients coming from allover the country; due to this fact drugs for the accompanying diseases are often brought in by the patient and taken without records in the USHATE Computerized Physician Order Entry (CPOE). In these cases the medication is documented in the PR texts so IE from discharge letters is the only means to generate a full picture of patient treatment during the hospitalization period.

The Semantic Mining results achieved in PSIP mark the state-of-the-art for French language. Merlin et al. [7] present a detailed evaluation where the extracted drugs are compared to the suggestions by human experts or the already encoded EHR content. The extraction of ATC codes from French text is performed with f-score 88% when compared to the manual extraction and with f-score 49% compared to the CPOE content. The miners exploit no PR structuring and name searching is done in the whole PR text. But for Bulgarian we can split the input PR text into sections because some default headers exist, and try to contextualize the extracted facts.

3. Material and Methods

The input texts in our experiment are free-text paragraphs of discharge letters. In Bulgaria the discharge letter structure is mandatory for all hospitals (it is published in the Official State Gazette, as Article 190(3) of the legal Agreement between the National Health Insurance Fund and the Bulgarian Medical and Dental Associations [8]). The PR text should contain the following sections: *(i)* personal details; *(ii)* diagnoses; *(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 extraction of medication events relevant for the respective hospitalization but in reality it is not strictly kept. Table 1 shows some statistics about availability of the above-listed sections in a training corpus of 1300 USHATE PRs. Although the structure is mandatory, many PRs are structured differently due to the following reasons: section merging, changing the section headers, skipping (empty) sections and replacing the default section sequence.

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

Diag-noses	Anam-nesis	Past di-seases	Allergies risk factors	Family medical history	Patient status	Lab tests	Exami-ners comments	Debate	Treat-ment
100	100	88,52	43,56	52,22	100	100	59,95	100	26,70

Thus inventing an algorithm for automatic recognition of *'current treatment'* is a non-trivial task which can be tackled only by heuristics collected on a representative corpus.

Drug names as tokens participate in all PR sections, including in *(ii) Diagnoses* (e.g. 'Amiodaron-induced hypothyroidism'). After removal of repeating drug names in each PR section, the recognised drugs in 1300 PRs are 10493 in total. Some 70% of them (7332) are unique in the respective PR and the remaining 30% (3161) repeat in several sections. The extractor has found 19 tokens in the *Diagnoses* section. Figure 1 summarizes the frequency of occurrences of drug names in the PR sections.

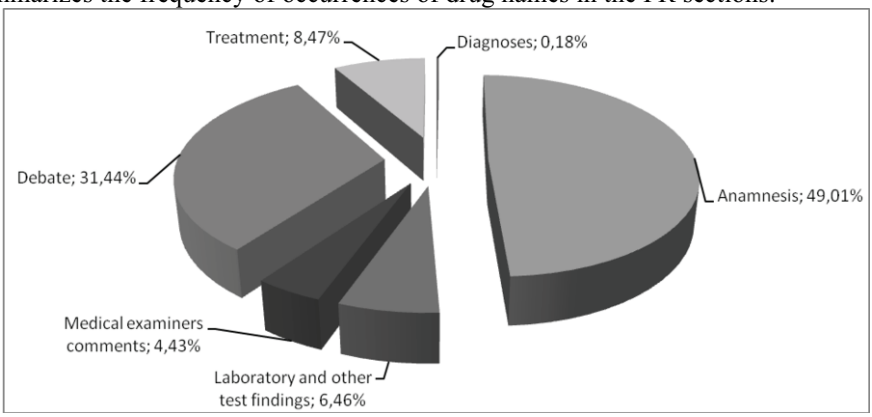

Figure 1. Potential descriptions of medication events in the various PR sections

4. Heuristics Supporting the Selection of *'Current Medication Events'*

Human experts have studied section by section the drugs extracted from 1300 anonymized PRs. The finings can be summarized as follows:

- Section '*Anamnesis*' discusses past medication events using phrases in past or present tense, hence the grammatical tense itself is not a filter for the event time. Among 5143 drug occurrences, only for 462 (9%) the local context contains explicit statements in that the drug is taken during the hospitalization;
- Section '*Lab data and other tests*' contains some 678 drug occurrences, especially in phrases like '*blood sugar has values X,Y while the patient is taking Z'*. These events are related to the period of hospitalization when the lab tests are made;
- Section '*Medical Examiners comments*' contains 465 drug name occurrences. About 90% of them are expressed by simple noun phrases with drug names and dosages, which can be interpreted as medication events related to the period of hospitalization. The remaining 10% contain longer conditional expressions like e.g. '*if X is ineffective, replace it by Y'* or '*if blood pressure increases include Z'*. These phrases and the related drugs cannot be automatically interpreted as medication events happening in USHATE;
- In section '*Debate*' some 3299 drug name occurrences are met. Some 542 (16,5%) are found in local contexts signaling therapy changes: start, stopping, increase, decrease, replacement of drugs. By default these events happen in USHATE;
- In section '*Treatment*', which is available as a separate paragraph in about 26% of the PRs in our training corpus, 889 drug occurrences are encountered. Some 236 of them (26,5%) concern future events as the local contexts contain typical

expressions signaling recommendations for further treatment. The remaining 73,5% of the events take place in USHATE.

Our IE system wrongly interpreted 74 drug name occurrences as medication events: 1,36% participate in description of allergies, 0,08% in descriptions of sensibility, and 0,01% in phrases signaling drug intolerance.

In principle an IE system, based on shallow analysis in the local context, might suggest each drug name occurrence as an actual event. After the statistical observations in a corpus of 1300 PRs, we consider as 'current' the following medication events:

- drugs in '*Anamnesis*', only if they are listed under the headers 'medication at the moment of hospitalization' or 'accompanying treatment'. Some phrases like 'started treatment with' can be also interpreted as hints for 'current medication' but only if they are not followed by phrases including 'replaced by' which signal past events;
- drugs in '*Medical Examiners comments*' which are recognized by the system with high confidence as present events (and not future prescriptions), except cases with 'stop' and 'replace' phrases;
- drugs in '*Debate*' which are recognized by the system with 100% confidence as present events (and not future prescriptions);
- drugs in '*Treatment*' which are not recognized as future events.

5. Evaluation Results, Impact and Conclusion

Our aim in the PSIP project is to extract information about drugs, taken by the patient, which are not prescribed via the Hospital Pharmacy. Usually these are drugs for accompanying and chronic diseases. Our extractor, integrating the heuristics defined at the end of section 3, has found 355 such drugs in the experimental test corpus of 6200 PRs (in addition to the 1182 drugs that are in use in USHATE during the period relevant for our experiment). These 'external' drugs might be listed in the *Anamnesis*, in a special paragraph with a subtitle which human experts recognize easily, and also seen in the *Debate* or *Treatment*. The automatic recognition of drugs for the accompanying diseases is somewhat simpler since they are often documented as simple enumeration (because the accompanying diseases attract less attention in a specialized hospital). The extraction algorithm tackles some negative phrases as single expressions according to a study of negative forms in Bulgarian medical texts [9].

The simple rules, summarized above, help to reduce the over-generation of medication events concerning these 355 'external' drugs. We have performed evaluation of our heuristic strategy on 1300 PRs; they contain in total 1648 names of drugs which are not in use by the USHATE Pharmacy. Table 2 present the percentage of 'external' drugs mentioned in different PRs sections in comparison to all drug events in that sections. We see that medication events for 'external' drugs are described mainly in the *Anamnesis*, *Medical examiners comments*, *Debate* and *Treatment* sections.

Table 2. Occurrences of drug names, which are not in use by the USHATE Pharmacy, in the PR sections

In the *Anamnesis* under header "Accompanying treatment"	As prescription by *external medical examiner*	In the *Debate*	In the *Treatment*
17,42%	16,34%	12,28%	21,48%

Table 3 proves the feasibility of our approach to mine the local context by searching typical phrasal expressions which are learnt from a representative training corpus. We have noted about 6% over-generation for two categories of events: *(i)* in the

Anamnesis, when a past event is considered as a present treatment, and *(ii)* in the *Debate* and *Treatment*, when a recommended medication event is interpreted as a present one (but often these expressions are ambiguous even for human readers).

Table 3. Accuracy of automatic extraction of medication events, related to 355 drugs

Precision	Recall	f-score
97,92%	90,69%	94,17%

Comparing our results to other experiments reported in the literature, we think that the relatively established PR structure is one of the main factors for the high accuracy.

Despite the limitations of the NLP technologies, IE is applied to discharge letters in many languages other than English, French and German: there are prototypes in Greek [10], Polish, Hungarian etc. The long-term objective is to enable the automatic or semi-automatic filling of big specialized scientific databases by extracting data from patient-related texts. However there is an inevitable percentage of extraction errors, which might be due to unrecognized entities (false negative) or over-generated entities (false positive). We believe that the NLP results should be embedded into large-scale experiments where the IE-induced noise will be statistically insignificant. Such data driven activities are related to the secondary use of EHR data.

Acknowledgments: The research tasks leading to these results have received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 216130 PSIP (Patient Safety through Intelligent Procedures in Medication).

References

[1] Demner-Fushman, D., W. Chapman and C. McDonald. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5), October 2009, pp. 760-772.

[2] Hobbs, J. and E. Riloff. Information Extraction, In: Indurkhya, N. and F. J. Damerau (Eds.) *Handbook of Natural Language Processing, 2nd Ed.*, Chapman & Hall/CRC Press, Taylor & Francis Group, 2010.

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

[4] *Third i2b2 Shared-Task and Workshop "Challenges in Natural Language Processing for Clinical Data: Medication Extraction Challenge"*, https://www.i2b2.org/NLP/Medication/, last visited April 2011.

[5] Boytcheva, S. Shallow Medication Extraction from Hospital Patient Record, To appear in the *Proc. 2nd Int. PSIP Workshop on Patient Safely through Intelligent Procedures in medication,* Paris, May 2011.

[6] PSIP project: *Patient Safely through Intelligent Procedures in medication*, http://www.psip-project.eu, European Community's 7FP, Information and Communication Technologies Programme.

[7] Merlin B., E. Chazard, S. Pereira, E. Serrot, S. Sakji, R. Beuscart, and S. Darmoni. Can F-MTI semantic-mined drug codes be used for Adverse Drug Events detection when no CPOE is available? In *Studies in Health Technology and Informatics*, *Proc. 13th World Congress on Medical Informatics*, Cape Town, South Africa, Volume 160, Number pt 1, 2010, pp. 1025-1029.

[8] National Framework Contract between the National Health Insurance Fund, the Bulgarian Medical Association and the Bulgarian Dental Association, *Official State Gazette* №106/30.12.2005, updates №68/22.08.2006 and №101/15.12.2006, Sofia, Bulgaria, http://dv.parliament.bg/.

[9] Boytcheva, S., A. Strupchanska, E. Paskaleva, and D. Tcharaktchiev, Some Aspects of Negation Processing in Electronic Health Records. In *Proc. of International Workshop Language and Speech Infrastructure for Information Access in the Balkan Countries*, 2005, Borovets, Bulgaria, pp. 1-8.

[10] Karanikolas, N. and C. Skourlas. Automatic Diagnosis Classification of patient discharge letters. In *Health Data in the Information Society, Proceedings of MIE2002*, *Stud. Health Technology and Informatics*, vol. 90, IOS Press, 2002, pp. 444-449.