LANGUAGE TECHNOLOGIES IN HEALTHCARE

GALIA ANGELOVA

Linguistic Modelling Department
Institute of Information and Communication Technologies
Bulgarian Academy of Sciences, Sofia, Bulgaria
galia@lml.bas.bg

Extended Abstract

Information Extraction (IE) is the dominating text analysis approach that is currently applied in the biomedical domain: due to the complexity of the narratives, shallow analysis is performed in order to extract automatically important entities, skipping the remaining text fragments. The IE systems operate further on the extracted text units relying on partial text understanding. Performance evaluation is done in terms of *precision* and *recall,* two widely-used indicators for the extraction accuracy and success. Here we present recent achievements in automatic IE from hospital discharge letters and outpatient records in Bulgarian language. Currently we deal mostly with records of diabetic patients, having in mind that diabetes is a chronical disease of major social importance. Several types of entities are essential for the IE tasks we tackle: *(i)* patient's principal diagnosis and diagnoses of the accompanying diseases; *(ii)* names of drugs, admitted by the patient, in particular those drugs that are discussed in free text descriptions with their dosage, frequency and route of admission; *(iii)* values of clinical tests and lab data that are documented in the patient records as free texts; *(iv)* patient status descriptions; *(v)* opinions of specialists concerning the patient status and diagnoses; *(vi)* family history and risk factors (e.g. smoking status, hypertony, etc.). Extracting these entities and events, the IE components need to cope with the negation. Timing of events is essential as well, so building timelines and temporal models is another important challenge for the medical IE applications.

Vocabulary. Starting with IE for Bulgarian patient records, we had to develop a relevant set of Bulgarian language resources: *medical terminology* including names of diseases (we used the Bulgarian version of ICD, the International Classification of Diseases), as well as *key phrases* that are used as typical names of entities and stable collocations. These phrases have to be learnt from the patient records' texts since they are not included in medical dictionaries. A language-independent extractor of phrasal units was developed (Boytcheva, 2012) that examines the frequency distribution of *N*-grams and suggests domain-specific collocations for final expert inspection. In this way meaningful collocations were learnt, like *'visible age'*, an important diabetic patient attribute, with its values *'corresponding to the real/passport/calendar one'*, *'about the real/calendar one'* and so on. Another important *attribute-value* pair is the *skin-characteristic* pair, as well as the *thyroid-status* pair, etc. This domain-specific vocabulary was acquired from the text of 6200 hospital discharge letters.

Grammar rules. In addition to the extraction of phrasal units, we needed to develop grammar rules for shallow analysis of text fragments that contain potentially interesting entities. In general, these rules are regular expressions that help the system to group alpha-numeric literals into meaningful text units. The extractor of lab data and clinical test values analyses the paragraphs where these values are enumerated without predetermined order and without standardised names of the indicators (Tcharaktchiev et al., 2011). The extractor recognises at first the indicator (i.e. the *name of the tested characteristic*), as well as the *value* related to the corresponding indicator. The *measure* and *interval limits* are desirable features, and the *time*, *condition* and *explanation* of further details are optional features. The extractor copes with *(i)* the variety of name writings (abbreviations, omitted words in the name, joined words in the name, typos), *(ii)* various symbols used as separators, *(iii)* the varying format of the numeric values, *(iv)* arbitrary replacements of Cyrillic and Latin letters which look identical and *(v)* ambiguity in the lab data recognition and the scoping of phrases related to certain indicators that have specific values. As an illustration, we show a rule for packing tokens into a structural group:

<(> <n> <v> <s> <v> <)> => <N> which means the following:

Find a sequence of tokens which:
starts with '('
followed by a phrase signalling referential values <n>,
followed by a number <v>,
followed by a separator <s>,
followed by a number <v>,
followed by a ')'.

If all tokens occur in the given order than this expression defines the group <N>.

This simple rule is used to group the literals *'(norm – 8,7-42)'* in the text fragment *'testosterone -3.2 (norm – 8,7-42)'*. The rule has 18 variants reflecting the various separators and delimiters learnt from a training set of 1000 discharge letters; it is used for the shallow analysis of lab data in 6200 hospital discharge letters of diabetic patients.

Similarly, using rule-based shallow analysis, values of blood sugar are extracted (Nikolova, 2012), as well as names of drugs, dosage, frequency and admission route (Boytcheva, 2011) and temporal expressions in Bulgarian discharge letters (Boytcheva et al., 2012). Negation in the Bulgarian clinical texts is also treated by shallow IE analysis (Boytcheva et al., 2005).

Corpora of medical records. Our first corpus of clinical narratives contains 6200 anonymised discharge letters of patients diagnosed with diabetes and other endocrinal diseases, provided by the University Specialised Hospital for Active Treatment of Endocrinology "Acad. I. Penchev" (USHATE), Medical University Sofia, Bulgaria. These letters are texts with length of 2-3 pages. Due to centralised national regulations, they consist of predefined sections like *Diagnoses*, *Anamnesis (Case History)*, *Patient Status*, *Lab data & clinical tests* and *Debate* which are available in 100% of the records in the corpus. Discharge letters are written in telegraphic style, discussing mostly positive findings, and can be treated more successfully by shallow analysis instead of full sentence parsing. Latin terms are commonly used, written in Latin alphabet (3% of all words in the corpus) or transliterated to Cyrillic alphabet (34% of all words). Spelling errors are common, too.

The current corpus with patient records contains more than 37.9 million pseudonymised reimbursement requests (outpatient records) submitted to the National Health Insurance

Fund (NHIF) in 2013 for more than 5 million patients, including 436000 diabetic ones. These records are semi-structured files with predefined XML-format, produced by the General Practitioners (GPs) and the Specialists from Ambulatory Care for every contact with the patient. They contain sufficient text explanations to summarise the case and to motivate the requested reimbursement, so IE is again the best analysis tool. Most important patient indicators like *Age*, *Gender*, *Location*, *Diagnoses* are easily seen since they are stored with explicit tags. The *Case history* is presented quite briefly in the *Anamnesis* section as free text with description of previous treatments, including drugs taken by the patient beyond the ones that are to be reimbursed by NHIF. The values of *Clinical tests and lab data* are enumerated in arbitrary order as free text in another section. The *Prescribed treatment* is described under a special tag. Only the drugs prescribed by the GPs and reimbursed by the NHIF are coded, the other medication is described as free text. Latin terms are relatively rarely used and spelling errors are rare too, compared to the discharge letters written as hospital documentation. An outpatient record might include about 160 tags.

Applications. Our initial tasks in research projects were oriented towards extraction of isolated entities and their attributes, such as status of patient skin, neck, thyroid gland, limbs and patient age (Boytcheva et al., 2010). These IE prototypes achieved accuracy of 83-92%.

Automatic recognition of temporal expressions helped to identify the drugs admitted by the patient at the moment of hospitalisation, i.e. at hospital stay Day 0 (Boytcheva at al., 2011). The extractor recognises about 350 drugs with accuracy of 90.17%, which are not prescribed via the Hospital Pharmacy, but are taken by the patients during the period of hospitalisation. This adds value to the data recorded in the Hospital Information System, where Day 0 is typically not reflected, and helps to search for Adverse Drug Events.

In (Boytcheva and Angelova, 2012) we present a prototype that constructs timelines of events that are described in the *Anamnesis* of hospital discharge letters; there are two timelines organised for the absolute and relative temporal markers. All clauses between two temporal markers are called “an episode” where drugs, diagnoses and patient conditions are recognised with accuracy higher than 90%.

The drug extractor, initially tested on discharge letter texts (Boytcheva, 2011), is elaborated to tackle drugs in the outpatient records. For diabetic patients, currently the extractor handles 2239 drugs names included in the NHIF nomenclatures. Recent extraction experiments deal with large-scale analysis of the outpatient records of 33641 diabetic patients. The drug extractor finds in the *Anamnesis* section drug names, daily dosages, frequency and route of admission with precision 95.2% and recall 93.7%, and replaces the drug names by their ATC[1] codes.

The extractor of values of lab tests and clinical examinations was elaborated too and now it copes successfully with the outpatient record texts. Major difficulties encountered in the entity recognition are due to the large variety of expressions

[1] Anatomical Therapeutic Chemical (ATC) Classification System for the classification of drugs, see http://www.who.int/classifications/atcddd/en/

describing the laboratory tests and clinical examinations. It tackles more than 40 types of clinical tests and works with precision that exceeds 98%.

Our recent efforts are focused on extraction of entities from the outpatient record Repository, in order to facilitate the construction of the Bulgarian diabetic register (Nikolova et al., 2014). Integrating language technologies and a business intelligence tool, we discover potential diabetic patients who are not formally diagnosed with diabetes. Diabetes can be suggested, for instance, by the values of the patients' clinical tests, or because the patients admit drugs that treat higher levels of blood sugar, or because some medical specialist has doubts that the patient might have diabetic complications (e.g. diabetic retinopathy). The Repository of outpatient records is pseudonymised, i.e. we can track the multiple visits of the same patient to his/her GP, hospital, etc. In addition, we can identify the important risk factors and the family history, in case they are described in the outpatient record text. Finally, the records of potential diabetic patients are classified according to the hypothesis "*having diabetes*" with precision 91.5% and the findings are delivered to the medical authorities for further checks. Discovering several hundreds of potential diabetic patients shows the importance of automatic text analysis in large medical Repositories.

Acknowledgements

The contribution summarises the results jointly obtained with Dr. Svetla Boytcheva, Dr. Dimitar Tcharaktchiev MD, Ivelina Nikolova, and Dr. Zhivko Angelov. The research work presented here is partially supported by the projects *EVTIMA* DO 02-292 (Effective search of conceptual information with applications in medical informatics), funded by the Bulgarian National Science Fund in 2009-2012; *PSIP* (Patient Safety through Intelligent Procedures in Medication) grant agreement 216130, funded by FP7 ICT in 2010-2011; and *AComIn* (Advanced Computing for Innovation) grant agreement 316087, funded by FP7 Capacity in 2012-2016.

References

Boytcheva, S. (2011). Shallow Medication Extraction from Hospital Patient Records. In: Koutkias, V. et al. (Eds.), *Patient Safety Informatics - Adverse Drug Events, Human Factors and IT Tools for Patient Medication Safety*, IOS Press, *Studies in Health Technology and Informatics* series 166, 119-128.

Boytcheva, S. (2012). Structured Information Extraction from Medical Texts in Bulgarian. *Cybernetics and Information Technologies 12:4*, 52-65, available at http://www.cit.iit.bas.bg/CIT_2012/v12-4/4Boicheva4-2012-Gotovos.pdf

Boytcheva, S., Angelova, G. (2012). A workbench for temporal event information extraction from patient records. *Proceedings of the 15th Int. Conference on Artificial Intelligence: Methodology, Systems, and Applications AIMSA 2012*, Springer, *Lecture Notes in Artificial Intelligence* 7557, 48-58.

Boytcheva, S., Angelova, G., Nikolova, I. (2012). Automatic Analysis of Patient History Episodes in Bulgarian Hospital Discharge Letters. *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the*

Association for Computational Linguistics, Avignon, France, 77-81, available at http://www.aclweb.org/anthology-new/E/E12/E12-2016.pdf

Boytcheva S., Nikolova, I., Paskaleva, E., Angelova, G., Tcharaktchiev, D., Dimitrova, N. (2010). Obtaining Status Descriptions via Automatic Analysis of Hospital Patient Records. In: V. Fomichov (Ed.), *Special Issue on Semantic Technologies, Informatica (Slovenia)*, Issue 4, December 2010, 269-278.

Boytcheva, S., Strupchanska, A., Paskaleva, E., Tcharaktchiev, D. (2005). Some Aspects of Negation Processing in Electronic Health Records. *Proceedings of the Int. Workshop Language and Speech Infrastructure for Information Access in the Balkan Countries*, associated to RANLP-2005, Borovets, Bulgaria, 1-8.

Boytcheva, S., Tcharaktchiev, D., Angelova, G. (2011). 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, the 23th Int. Conf. of the European Federation for Medical Informatics, Norway, IOS Press, *Studies in Health Technology and Informatics* series 169, 527-531.

Nikolova, I. (2012). Unified Extraction of Health Condition Descriptions, *Proceedings of the NAACL HLT 2012 Student Research Workshop*, Montreal, Canada, 23-28, available at http://aclweb.org/anthology-new/N/N12/N12-2005.pdf

Nikolova, I., Tcharaktchiev, D., Boytcheva, S., Angelov, Z., Angelova, G. (2014). Applying Language Technologies on Healthcare Patient Records for Better Treatment of Bulgarian Diabetic Patients. In: G. Agre et al. (Eds.): *Proceedings of AIMSA 2014*, Springer, *Lecture Notes in Artificial Intelligence* 8722, 92–103.

Tcharaktchiev, D., Angelova, G., Boytcheva, S., Angelov, Z., Zacharieva, S. (2011). Completion of Structured Patient Descriptions by Semantic Mining. In: Koutkias, V. et al. (Eds.), *Patient Safety Informatics - Adverse Drug Events, Human Factors and IT Tools for Patient Medication Safety*, IOS Press, *Studies in Health Technology and Informatics* series 166, 260–269.