Extracting Medical Information Using Linked Data

Jakub Kozák*, Martin Nečaský, and Jaroslav Pokorný

Faculty of Mathematics and Physics, Charles University in Prague,
Malostranské náměstí 25, 118 00 Prague, Czech Republic
{kozak,necasky,pokorny}@ksi.mff.cuni.cz
http://www.ksi.mff.cuni.cz

Abstract. Medical documentation is usually represented as a free text. That makes it almost unusable for all other analytical purposes. It would be a great advantage to have at least some information structured in a machine readable form. In this contribution we introduce a system which performs information extraction from partially structured texts such as discharge summaries. It is a rule-based system using ontologies even for rule formulation. We take advantage of Linked Data principles, represent all knowledge as ontologies and make output data available in RDF. We hope this is a right direction for medical information publication and enrichment with other data e.g., from Linked Open Data cloud.

Keywords: information extraction, discharge summaries, Linked Data

1 Introduction

Medical documentation offers a rich source of information which has a great potential for further analysis. It could lead to improvement of the quality of healthcare. Unfortunately, the documentation is usually represented as a free text which means the data has to be extracted first.

Free text representation of medical documentation prevails also in healthcare in the Czech Republic. In this contribution we will discuss the issue of information extraction (IE) from partially structured texts such as discharge summaries written in Czech language, which has its own specifics and special tools for natural language processing (NLP) has to be used. By partially structured text we mean text composed of sections with headings which can be recognized with regular expressions. Discharge summaries usually contain a finite set of sections e.g., anamnesis, examination and procedures, medication etc.

Information extraction can be performed using basically two methods. It is either based on rules or on machine learning (ML) or their combination. We have decided to use a rule-based approach which was also mostly used and among

* This research was partially supported by The Grant Agency of Charles University, grant GAUK no. 572212. Authors also thank Internal Clinic of 2nd Faculty of Medicine, Charles University in Prague for collaboration and input data.

best performing in The Third i2b2 Workshop on Natural Language Processing Challenges for Clinical Records which was focused on medication information extraction [1]. Recently, a well performing rule-based system for IE from medical records has also been developed for Polish language which is close to Czech language [2].

Discharge summaries contain various data about patients. Therefore a vast domain model would be needed for the extraction of all information. There exist a number of different medical dictionaries, thesauri and ontologies. We can point out at least some of them. SNOMED CT[1] or UMLS[2] are well-known controlled vocabularies. Most of the vocabularies are available in English but not in Czech. It makes them almost unusable unless translated.

Previous information led us to a decision to build our own domain model. We have decided to adopt principles of Linked Data which are intended to become a standard for publishing data on the Web [3]. This approach is quite beneficial when thinking about complicated queries with the use of external knowledge bases e.g., from Linked Open Data[3] (LOD) cloud. Enrichment of data using LOD cloud has been proven to be successful at Mayo Clinic [4].

We will present a system which is able to extract information from partially structured texts. The system uses domain ontologies and rules described in RDF[4] (Resource Description Framework). External knowledge can be included e.g., from Linked Data cloud. Preliminary application to discharge summaries will be presented.

2 Information Extraction System

We have developed a system that allows IE from partially structured texts using the advantages of ontologies and Linked Data. Linked Data principles allow enrichment of data in internal system by linking them to other data and therefore run more complex queries. The system itself is domain independent because the knowledge is only contained in configurable ontologies in the underlying data store.

The process of extraction can be summarized into a pipeline:

- natural language processing
- section matching using templates
- entity and more complex structures recognition
- RDF output

Czech language is quite complicated in comparison with English. Therefore at least basic NLP procedures have to be conducted. Our system performs tokenization first and then lemmatization using the tool MORCE[5] (Czech Morphology - software for morphological disambiguation (tagging) of Czech text).

[1] http://www.ihtsdo.org/snomed-ct/
[2] http://www.nlm.nih.gov/research/umls/
[3] http://linkeddata.org
[4] http://www.w3.org/RDF/
[5] http://ufal.mff.cuni.cz/morce

For definition of input text structure, i.e. section recognition, a special template, which should be common for one type of texts, is used. Templates are implemented using XML and have an XML schema. Each section is represented with one entity or property in ontology which describes the type of documents e.g. discharge summaries. The following example shows definition of pharmacological anamnesis which is preceded by the heading defined with the regular expression "FA:" and is represented with the range of property "hasPharmacologicalAnamnesis":

```
<def:section def:scope="PARAGRAPH" def:outputElementType=
  "...#hasPharmacologicalAnamnesis">
  <def:header>
    <def:expression>FA:</def:expression>
  </def:header>
</def:section>
```

After matching the section, entities are being extracted based on the rules. We recognize three types of rules regular expression, taxonomy or complex structure. The first type seeks for entities described with a regular expression. The second one looks up a literal property from ontology. Finally, complex structure combines the previous two approaches.

If we want to extract basic medication information from the piece of text "Anopyrin 100 mg 1-0-0" contained in the section describing pharmacological anamnesis, we need to combine a taxonomy lookup for the name of the medication, regular expression for the dosage and another complex structure for the strength (number plus unit). Complex structures are more or less trees and allow using constructs such as optional, after or close-to, which refer to the position of previously found part of information. Example of informal model of complex structure rule for medication extraction is given in Figure 1.

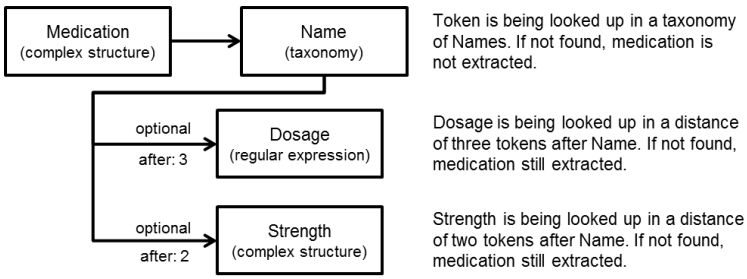

Fig. 1. Complex structure model for medication extraction

The system is able to extract quite complex structures using rules described above. The rules are defined also in RDF and form an ontology of rules, which is pointing to other ontologies. Schema in Figure 2 shows how the document is linked to ontologies and extraction rules.

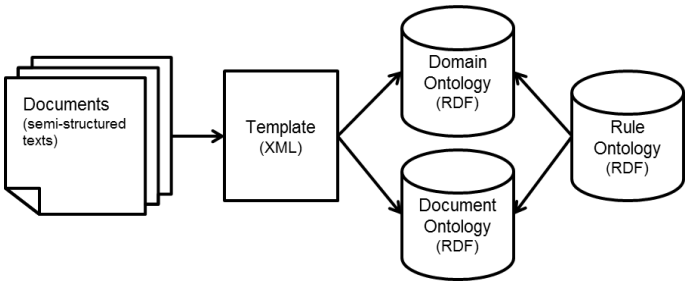

Fig. 2. Schema showing connection between documents and ontologies

Extracted information is made available as RDF and can be easily inserted into a triple store. The presented system uses Virtuoso[6] for ontology storage.

3 Conclusion

The system was originally developed as a universal tool for IE from partially structured texts which can be used in various domains (job descriptions, CVs, public contract offers, discharge summaries etc.). The only necessary thing to do is to connect domain ontology and to define templates and extraction rules.

We are now trying to use the system on discharge summaries from an internal clinic, which usually contain these sections: reason for acceptance, anamnesis, examinations and procedures, therapy, conclusion and recommendation. Each of these sections can have many subsections annotated with some kind of headings which we are trying to recognize using our IE tool.

Application to medication extraction and laboratory results (also free text) extraction shows promising result. The research is still in progress. It involves a lot of activities e.g. domain knowledge representation using ontologies.

We believe that systems based on ontologies and using Linked Data principles, which allow data enrichment, have a bright future in medicine and we plan to aim our further efforts this way.

References

1. Uzuner Ö., Solti I., Cadag E.: Extracting medication information from clinical text. J Am Med Inform Assoc 17, 514-518 (2010)
2. Mykowiecka A., Marciniak M., Kup A.: Rule-based information extraction from patients' clinical data. J. of Biomedical Informatics 42, 5, 923-936 (2009)
3. Bizer C., Heath T., Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems (IJSWIS) 5, 3, 1-22 (2009)
4. Pathak J, Kiefer RC, Chute CG.: Applying Linked Data principles to represent patient's electronic health records at Mayo Clinic: A case report. 2nd ACM SIGHIT International Health Informatics Symposium (IHI), 455-464 (2012)

[6] http://virtuoso.openlinksw.com