Applying Language Technologies on Healthcare Patient Records for Better Treatment of Bulgarian Diabetic Patients

Ivelina Nikolova, Dimitar Tcharaktchiev,
Svetla Boytcheva, Zhivko Angelov
and Galia Angelova

The problem

Building the Bulgarian Diabetic Register

- *(i)* keep the established practice of patient registration without burdening the medical experts with additional paper work;
- *(ii)* reuse the existing standard records in compliance with all legal requirements for safety and data protection;
- *(iii)* save time and resources by avoiding multiple patient registrations and disturbance of the diagnostic and treatment process.

Outline

- NLP in the biomedical domain
- Integration platform
- NLP modules
  - Medication extraction
  - Diabetic positive statements extraction
- Experiments and results

Background

Example: “The anamnesis is taken from the patient and medical documentation. He enters the clinic again on the occasion of decompensated diabetes mellitus. Complains from burning pain in the lower limbs…”

- The most important findings about the patients are kept as free texts in various documents and languages.
- These text descriptions are usually oriented to human readers.
- Thus Information Extraction (IE) becomes the dominating natural language processing (NLP) approach to biomedical texts.

BITool

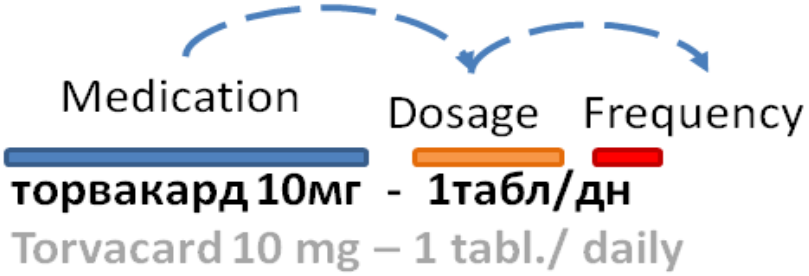
- Integration platform of the analyses performed on the medical data.
- Business Intelligence tool can deliver various types of findings to decision makers in order to improve the public health policy and the management of Bulgarian healthcare system.
- The data of the Health Insurance Fund contains a lot of information that is structured using codes of medical classifications and nomenclatures.
- In this study we are interested in the analysis of free texts and capturing some essential entities described there.
- By means of NLP techniques integrated with the *BITool* we discover the potential diabetic patients which were not formally diagnosed with diabetes.

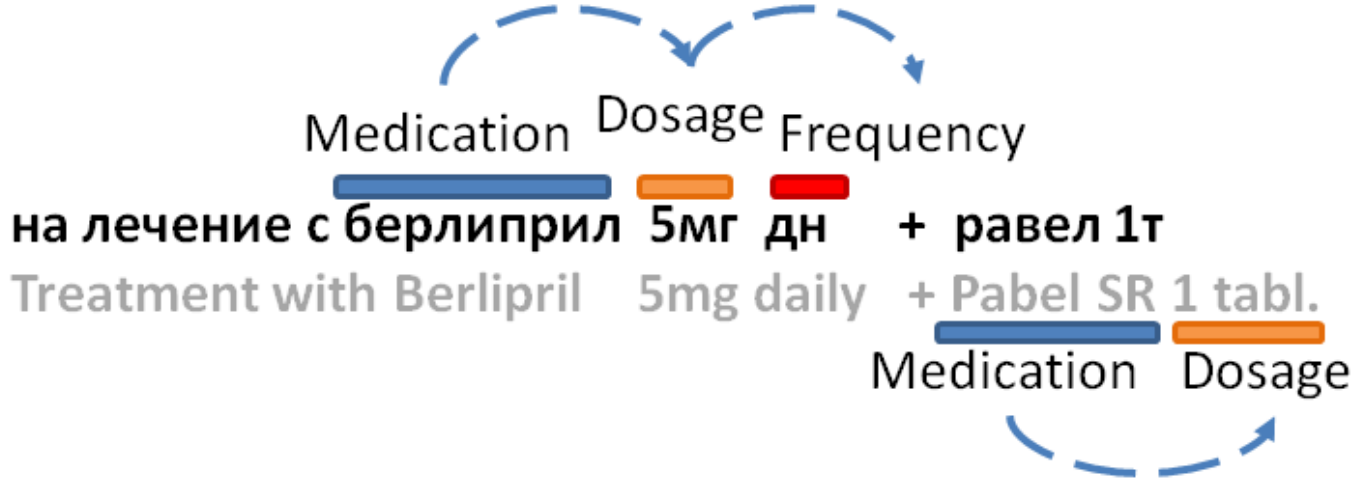
Knowledge Discovery

(*i*) the **medical treatment** – if the patient has diabetes he/she would also take appropriate drugs

(*ii*) **statements in the anamnesis** about the patient having diabetes or its complications.

Regular expressions of linguistic patterns for analysis of Dosage





Structuring drug treatment information

- Automatic procedure analyses the free texts in the *Prescribed treatment* section:
  - drug names;
  - dosages;
  - modes of admission;
  - frequency and treatment duration
- Assigns the corresponding ATC and NHIF codes to each medication event.
- Using regular expressions to describe linguistic patterns.
- More than 80 different patterns for matching text units deal with the ATC and NHIF code, medication name, dosage and frequency.

Drug extractor performance

- The extractor handles 2,239 drugs names included in the NHIF nomenclatures
- Manual evaluation on 33,641 diabetic patients for 2013
  - precision - 95.2%
  - sensitivity - 93.7%
  - The labelled data is split to 20 equal subsets and randomly selected records are evaluated by an expert (about 40% of each subset). The average of the subset evaluation is the final score of the module.

Drug extraction - error analysis

The major reasons for incorrect recognition of drug events are:

- misspelling of drug names;
- drug names occurring in the contexts of other descriptions;
- undetected descriptions of drug allergies, sensibility, intolerance and side effects;
- drug treatment described by (exclusive) *OR*;
- negations and temporally interconnected events of various kinds:

  - undetected descriptions of cancelled medication events;

  - of changes or replacements in therapy;

  - of insufficient treatment effect and change of therapy.

Drug extraction - error analysis

- About 30% of the medication events in the test corpus are described without any dosage.
- Lack of explicit descriptions occurs mostly for treatment of accompanying diseases.
- After applying the recognition algorithm and default daily dosage, the number of records lacking dosage reduces to 15.7% in the final result.

Discovering potential diabetic patients

- Medical experts propose criteria for happening of the event "*having diabetes*": e.g. high blood sugar or high glycated hemoglobin in the text of the section *Lab test results* or
- Admission of drugs used for diabetes treatment mentioned in the *Anamnesis, or*
- Statements in the *Anamnesis* describing diabetes or its symptoms.

Input data

- Chunks extracted from a concordancer built for the string *диабет* (*diabetes*).
- 67,904 distinct chunks extracted from the records of 156,310 patients who are not formally diagnosed with diabetes.
- Chunks contain the word *diabetes* and a 6-token window of its left and right context.

Our goal is: classify the chunks according to the hipotesis “has diabetes”.

Example chunks

Sample chunks which demonstrate the variety of positive and negative examples:

(i) NEG Фамилност- обременен/а-**диабет**ици по майчина линия/.
*Family - heredity - **diabet**ic on maternal line.*

(ii) NEG Необходимо е изключване на стероиден **диабет**; насочва се към ТЕЛК...
*It is necessary to exclude steroid **diabet**es; re-directing to TEMC...*

(iii) POS Покачва кръвно налягане. Има **диабет**. Оплаква се от сърцебиене...
*Raises the blood pressure. Has **diabet**es. Complains of palpitation...*

Rule-based rough filtering

- Most of the input chunks are negative examples.
- About 10% of the input records match several patterns of negative examples

  “*no evidence about diabetes*”,

  “*no diabetes in the family*” etc.
- With a set of 41 expressions in the filter, the number of chunks was reduced to 26,000 (about 1/3 of the initial corpus size).

Supervised classification of positive/negative examples

- Two random subsets were annotated:
  - one of 282 documents – development set used for feature selection
    - 74 positive and ;
    - 208 negative examples;
  - one of 1,000 documents – used for testing
    - 187 positive and;
    - 813 negative examples.

Classification overview

- Several algorithms on the same dataset: NaiveBayes, J48, SVM and JRip, all with boolean features - JRip and J48 performed best.

- Classification with nominal features with MaxEnt algorithm - outperformed all the rest in means of precision.

Experiment 1

- 93 features which correspond to stems of terms occurring in the text of **positive examples** (excluding numbers).
- J48 recognised only 63.1 % of the positive examples
- JRip recognised 91.2%.
- The rules inferred by JRip are only 2 but obviously they fit well the data.

Rule 1*:* (*family*) = *false*) and (*sugar* = *true*) and (*noninsulindependent*= *true*)

=> class=pos (30.0/2.0)

Rule 2*:* (*family* = false) and (*sugar* = true) and (*treat* = true)

=> class=pos (6.0/0.0)

Experiment 2

- 112 textual features which correspond to the stems of terms occurring in **positive and negative examples**.
- The terms are pre-filtered manually by an expert.
- JRip and J48 performed worse than in the first experiment and scored precision under 70%.

Results from Experiments 1 and 2

	Exp. 1: 93 pos features; stems only; 10-fold cross-validation						Exp. 2: 112 pos/neg features; stems only; 10-fold cross-validation					
	JRip			J48			JRip			J48		
Class	P	R	F	P	R	F	P	R	F	P	R	F
positive	91.2	16.6	28.1	66.7	21.4	32.4	61.1	29.4	39.7	65.8	52.4	58.3
negative	83.9	99.6	91.1	84.4	97.5	90.5	85.5	95.7	90.3	89.5	93.7	91.6
w eighted avg	85.2	84.1	84.1	81.1	83.3	79.6	80.9	83.3	80.8	85.1	86	85.4

Experiment 3

- Automatic feature-selection
- The initial feature set contained 10,576 attributes
- Applied on it the chi-squared attribute evaluator implemented in Weka
- 151 features were selected
- JRip - **65.3%** precision when adding bigrams and **73.1%** when adding trigrams
- J48 - **86.1%** precision and **87.2** when adding trigrams
- In the tree built by J48 one could clearly see the importance of bigrams and trigrams features
- Out of 17 tree leaves, only 4 are unigrams; the rest are bigrams and trigrams.

Results from Experiment 3

	Exp. 3a: 151 automatically selected stem features + bigrams; 10-fold cross-validation						Exp. 3b: 151 automatically selected stem features + bigrams + trigrams; 10-fold cross-validation					
	JRip			J48			JRip			J48		
Class	P	R	F	P	R	F	P	R	F	P	R	F
positive	65.3	41.2	50.5	**86.1**	36.4	51.1	73.1	40.6	52.2	**87.2**	36.4	51.3
negative	87.5	95	91.1	87.1	98.6	92.5	87.6	96.6	91.9	87.1	98.8	92.6
weighted avg	83.4	84.9	83.5	86.9	87	84.8	84.9	86.1	84.5	87.1	87.1	84.9

Experiment 4

- Trained a model with **MaxEnt** using
  - all textual features plus bigrams and
  - all textual features plus bigrams and trigrams as nominal values
- Similar results in both experiments.
- Outperformed J48 and JRip in means of precision.
- The best precision on positive examples was reached when including bigrams and trigrams 91.5%
- MaxEnt with nominal features has the advantage that the features are not pre-set

Results from Experiment 4

	MaxEnt					
	Exp. 4a: all stems + bigrams; 10-fold cross validation			Exp. 4b: all stems+bigrams + trigrams; 10-fold cross validation		
Class	**P**	**R**	**F**	**P**	**R**	**F**
positive	91.3	22.6	36.2	**91.5**	20	32.8
negative	85.6	84.2	85	85.6	84.2	84.9
weighted avg	88.45	53.4	60.6	88.55	52.1	58.9

Results from Experiments 1 and 2

	Exp. 1: 93 pos features; stems only; 10-fold cross-validation						Exp. 2: 112 pos/neg features; stems only; 10-fold cross-validation					
	JRip			J48			JRip			J48		
Class	P	R	F	P	R	F	P	R	F	P	R	F
positive	91.2	16.6	28.1	66.7	21.4	32.4	61.1	29.4	39.7	65.8	52.4	58.3
negative	83.9	99.6	91.1	84.4	97.5	90.5	85.5	95.7	90.3	89.5	93.7	91.6
w eighted avg	85.2	84.1	84.1	81.1	83.3	79.6	80.9	83.3	80.8	85.1	86	85.4

Results from Experiment 3

	Exp. 3a: 151 automatically selected stem features + bigrams; 10-fold cross-validation						Exp. 3b: 151 automatically selected stem features + bigrams + trigrams; 10-fold cross-validation					
	JRip			J48			JRip			J48		
Class	P	R	F	P	R	F	P	R	F	P	R	F
positive	65.3	41.2	50.5	**86.1**	36.4	51.1	73.1	40.6	52.2	**87.2**	36.4	51.3
negative	87.5	95	91.1	87.1	98.6	92.5	87.6	96.6	91.9	87.1	98.8	92.6
weighted avg	83.4	84.9	83.5	86.9	87	84.8	84.9	86.1	84.5	87.1	87.1	84.9

Results from Experiment 4

	MaxEnt					
	Exp. 4a: all stems + bigrams; 10-fold cross validation			Exp. 4b: all stems+bigrams + trigrams; 10-fold cross validation		
Class	**P**	**R**	**F**	**P**	**R**	**F**
positive	91.3	22.6	36.2	**91.5**	20	32.8
negative	85.6	84.2	85	85.6	84.2	84.9
weighted avg	88.45	53.4	60.6	88.55	52.1	58.9

Error analysis

- The positive examples in our database are so rare that the data quality has major impact on the training – misspellings, mixed Cyrillic and Latin names, numbers.

- Having a larger corpus and having more golden data will help for learning better the patterns of positive examples.

- Nevertheless the current results show that such a hybrid method combining rule-based and machine learning approach can be used to prove the hypothesis *having diabetes* with high precision.

Summary

- The IE modules are exploited quite carefully, for extraction of a limited number of entities and events only.
- They are tested in various scenarios and gradually improve their performance using hybrid rule-based and machine learning approaches.
- Automatic extraction from the records' free text essential entities related to the drug treatment such as **drug names**, **dosages**, **modes of admission**, **frequency** and **treatment duration** with precision **95.2%**;
- Classification the records according to the hypothesis "*having diabetes*" with precision **91.5%**
- Deliver these findings to decision makers in order to improve the public health policy and the management of Bulgarian healthcare system.
- Large-scale analysis of medical texts can be viewed as a reliable technology if the input is well-structured into zones and the extraction task has clear and well-defined target entities.

Acknowledgements

This research work is partially supported by the :

FP7 grant **AComIn** No. 316087, funded by the European Commission in the FP7 Capacity Programme in 2012–2016.

And also:

- Medical University – Sofia

- The Bulgarian Ministry of Health

- The Bulgarian National Health Insurance Fund.