Unified Extraction of Health Condition Descriptions

Ivelina Nikolova
Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
2, Acad. G. Bonchev Str, 1113 Sofia
iva@lml.bas.bg

Abstract

This paper discusses a method for identifying diabetes symptoms and conditions in free text electronic health records in Bulgarian. The main challenge is to automatically recognise phrases and paraphrases for which no ”canonical forms” exist in any dictionary. The focus is on extracting blood sugar level and body weight change which are some of the dominant factors when diagnosing diabetes. A combined machine-learning and rule-based approach is applied. The experiment is performed on 2031 sentences of diabetes case history. The F-measure varies between 60 and 96% in the separate processing phases.

1 Introduction

Electronic Health Records (EHRs) are a rich source of information regarding patient's health condition and treatment over time but they often exist as free text only. Currently great efforts are put into structuring such data and making them available for further automatic processing, the so-called *secondary use of EHRs*. Following this line of work in this paper we present a pilot study for extracting condition descriptions from EHRs in Bulgarian with the help of NLP techniques thus making a step toward the structuring of the free text. The specificity of the EHRs as a combination of biomedical terminology in an underresourced language and a source of valuable health-care data makes them attractive for various medical and language research tasks. We present an algorithm which comprises machine learning (ML) techniques and rule-based analysis to automatically identify phrases and paraphrases, for which no ”canonical forms” exist in any dictionary, with minimal effort. We analyse anonymous EHRs of patients diagnosed with diabetes.

We focus on extracting the levels of blood sugar and body weight change (examples are given in table 1) which are some of the dominant factors when diagnosing diabetes but we believe this approach can extend to recognise also other symptoms or medication expressions which have similar record structure. We extract information which is on one hand very important for the professionals and on the other hand not directly observable in a collection of unstructured documents because of its composite meaning. In Bulgarian EHRs laboratory data is sometimes present inline in the text only and means for extracting such information from the plain text message are often needed.

The paper is structured as follows: section 2 presents related studies, section 4 describes the method, and section 3 the experiments. The results are given in section 5 and the conclusion in section 6.

2 Related Work

There are several successful systems for identifying patient characteristics and health conditions, mostly in English documents. The one presented by Savova et al. (2008) solves the task of identifying the smoking status of patients by accurately classifying individual sentences from the patient records. They achieve F-measure 85.57. One of the limitations is the lack of negation detection. Similarly to their approach our source documents are decomposed into sentences which are to be classified. The symptom

Proceedings of the NAACL HLT 2012 Student Research Workshop, pages 23–28,
Montréal, Canada, June 3-8, 2012. ©2012 Association for Computational Linguistics

descriptions are short and always written within a single sentence, therefore it is important to filter out the irrelevant sentences. We employ ML techniques and rule-based analysis and in addition deal with negation detection.

Harkema et al. (2009) presents an algorithm called ConText, which determines whether clinical conditions mentioned in clinical reports are negated, hypothetical, historical, or experienced by someone other than the patient. The system is entirely rule-based and infers the status of a condition from simple lexical clues occurring in the context of the condition. This algorithm proves successful in processing different clinical report types with F-measure for negation (75-95%), historical (22-84%), hypothetical (86-96%) and experiencer (100%) depending on the report types. Our work rests on a similar idea – we prepare a set of vocabularies which are learned from data and are used for determining the scope of the expressions of interest but we focus on extracting health conditions, their status, values and negation.

Negation is one of the most important features to be recognized in medical texts. There is a work for Bulgarian by Boytcheva (2005) which specifically tackles the negation by the presence of triggering expression as we do too.

Many systems implement isolated condition identification and rarely complete semantic model of all conditions, e.g. MedLEE (Friedman, 1994), MEDSYNDIKATE (Hahn, 2002) etc. identify the status condition and also modifying information like anatomic location, negation, change over time. In Boytcheva et al. (2010) the authors extract from Bulgarian EHRs the status of the patient skin, limbs, and neck with thyroid gland with high accuracy.

3 Experimental Data

Source Data This work is done on free text EHRs of diabetic patients submitted by the Endocrinology Hospital of the Medical University in Sofia. The health conditions are written in the case history which describes the diabetes development, complications, their corresponding treatment, etc. Symptom descriptions are written within a single sentence (sometimes other symptoms are described in the same sentence too) as shown in table 1.

Ex. 1. При изследване кръвната захар е била - 14 ммол/л. (*After examination the blood sugar was - 14 mmol/l.*)

Ex. 2. Постыпва по повод на полиурично-полидипсичен синдром, редукция на теглото и кетоацидоза. (*Enters hospital because of polyuria-polydipsia syndrome, weight reduction and ketoacidosis.*)

Table 1: Examples of symptom descriptions.

Our training corpus is a subset of anamnesis sentences regarding only symptom descriptions. It is annotated with symptom type on sentence level and with symptom description on token level. These are excerpts from from 100 epicrises. All sentences are marked with class ”bs” (blood sugar), ”bwc” (body weight change) or another symptom. The sentences that describe more symptoms have more than one label. These data was used for learning the rules and the vocabularies. The experimental/test dataset consists of 2031 anamnesis sentences annotated with symptoms. The documents are manually sentence split and automatically tokenized. To overcome the inflexion and gain a wider coverage of the rules we also use stemmed forms (Nakov, 2010).

Vocabularies The algorithm relies on a set of specific vocabularies manually built from the annotated training set. We build a *Focal Term Vocabulary* which contains words and phrases signalling the presence of the health condition description (e.g. ”glycemic control”, ”hypoglycemia” etc.). It is used for defining the condition in *phase 2*. All single words which appear in this vocabulary except for the stop words form the so called *Key Term Vocabulary* used in the *phase 1* classification task.

There are two vocabularies containing border terms: one with rightmost context border expressions (*Right Border Vocabulary*); and one with left border expressions (*Left Border Vocabulary*). These are conjunctions and phrases separating the blood sugar level description from another observation preceding it. Both vocabularies are sorted in descending order by the probability of occurrence associated with each expression as border term.

A *Vocabulary of Negation Expressions* is also compiled as well as a *Vocabulary of Condition Statuses* (e.g. ”good”, ”bad”, ”increased” etc.).

4 Methodology

We aim at identifying health conditions in the EHR case history. The problem can be broken down into the following subtasks: **Phase 1**: *identify the relevant sentences*; **Phase 2**: *identify the condition and its status*; **Phase 3**: *identify the values related to the symptom of interest* - mmol/l, kg etc.; **Phase 4**: *identify negation*; **Phase 5**: *identify the scope of the description - match left and right border terms*.

Two experiments for accomplishing *phase 1* have been carried out: a rule-based one and ML-based one. In the ML setting we train a binary classification algorithm. We experiment with 3 feature sets: *(i)* all tokens in the corpus; *(ii)* the tokens from *Key Term Vocabulary* and *(iii)* restricted subset of the *Key Term Vocabulary*. In all cases each sentence is considered a document and each document feature vector contains the following boolean values for each feature: *the feature value is true iff the document contains the corresponding token from the feature set, otherwise it is false*. In this setting we use the experimental corpus which we split in folds for training and testing and the vocabularies which are learned from the training corpus.

In the rule-based experiment, we construct lightweight regular expressions that match symptom descriptions in the training set. We model them in context window of up to 5-7 tokens to the left and right of the focal terms depending on the kind of expression. When composing the rules like in figure 1 we introduce new links between tokens which are not subsequent in the training dataset, if the newly created token sequences would be meaningful focal expressions. The black edges are obtained from the training corpus and the dashed grey one is manually added. This approach would not harm any identification procedure because it can match only an existing sequence in the target text therefore we can only benefit from such augmented rules. Moreover these rules are crafted for stemmed text which partially overcomes the morphological agreement problem (Bulgarian is a highly inflective language) thus they have wider coverage on the possible signalling words (see table 2). The sentences matching these rules are passed to *phase 2*.

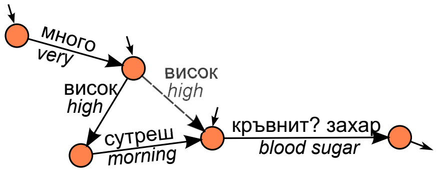

Figure 1: Adding new edges between tokens.

кръвн[аи](т)? захар (*the blood sugar*)
((незадоволителен) OR (добър) OR (лош) OR (отлич)) (гликемич контрол) (*not satisfactory OR good OR bad OR excellent glycemic control*)

Table 2: Phase 1 rules after stemming.

At *phase 2* the condition status is recognised. The blood sugar level is most often cited as *low*, *high* or *normal* and could be also *bad* or *good*, body weight can be *increased* or *decreased*. The context words which signal the status of the condition appear on the left side of the focal terms, such as: с високи стойности на кр. захар (*with high values of the blood sugar*); лош гликемичен контрол (*bad glycemic control*).

Phase 3 analysis is related to the dynamic extension of the right context of the analysed expression in order to cover all necessary attributes. At this phase we aim at identifying the value of the blood sugar test if there is such. The values of this test are given in various ways – as an interval of values; as a maximal value reached during some period or a concrete value. At this step we apply rules matching vocabulary words signalling the type of value representation e.g. между (*between*); до (*up to*); над (*above*); около (*around*).

When the algorithm recognises a word signalling interval value representation such as между (*between*), it takes action to expand the right context to the next two numbers and measuring unit after the second one, but with no more than 7 tokens. If the numbers fall out of this 7-token window they are ignored and the value identification algorithm fails. We determined the 7-token window experimentally by analysing the training set of EHRs where often verbs expressing temporality are connecting/separating the focal terms from the ones describing lab test values (as shown in table 3).

кръвнозахарна стойност до 10-11 ммол/л (*bloodsugar level up to 10-11 mmol/l*)
стойностите на кръвната захар са били (между 4 и 6,5 ммол/л) (*level of the blood sugar has been between 4*) (*and 6,5 mmol/l*)

Table 3: Recognition of lab test values.

In *phase 4* we recognise negation. We observe only limited occurrences of negations in the text. This is due to the fact that in Bulgarian practice mostly medical conditions with pathological changes are described. The expressions signalling negation appear on the left context of the phrases marked at *phase 1* and they modify the expressions identified at *phase 2*. Some examples are: не съобщава за... (*does not inform about...*); не [много] високи стойности на ... (*not [very] high values of...*).

Phase 5 identifies the symptom description scope. It is determined by the context words which signal the beginning of the expression, its ending and the already identified attributes. The expression of interest either starts at the beginning of the sentence or follows another description and conjunctions. The end of the expression is either coinciding with the end of the sentence, or is signalled by a value of the blood sugar test, or a description of another symptom (see table 4). The border identification rules are applied on the right and on the left of the already identified attributes starting from the rule having highest probability and continue in descending order until a match is found. If no match is found in 7-token context window the right border is considered the right most token of the current expression and the left border of the expression is either the first token of the focal term or negation of the expression or status of the condition.

5 Evaluation

5.1 Phase 1 - Rules vs ML

The evaluation of our approach is performed from several different perspectives. We compare text classification versus rule-based approach at *phase 1*. In the ML setting each input document (meaning each sentence) has a boolean feature vector representing the presence of each token of the feature set in that sentence. The concrete attribute position x_i is $false$ if the sentence does not contain the corresponding feature token and is $true$ if it contains it.

Beginning of expressions of interest
при кръвна захар... (*with blood sugar...*)
на фона на лош гликемичен контрол... (*on the background of bad glycemic control...*)
с високи стойности на кръвната захар... (*with high values of the blood sugar...*)
Ending of expressions of interest
...кръвна захар - 14 ммол/л. (*...blood sugar - 14 mmol/l.*)
...лош гликемичен контрол и кетоацидоза. (*...bad clycemic control and ketoacidosis.*)

Table 4: Beginning and ending of expressions.

The applied classification algorithm is a standard J48 decision tree which we consider appropriate, given the fact we supply binary data (Visa et al., 2007). We used Weka Data Mining Software (Hall, 2007) for performing the tests. The results with best settings are shown in table 5.

To achieve these results we did several experiments on the test set, using the features selected from the training set. The initial test was done with a feature set comprising all tokens in the text collection except for stop words. The achieved F-measure was about 82 in 10-fold cross-validation, to 89% in isolated experiments and up to 92% on balanced datasets. The precision was as high as 92% and the recall varying from 73 to 85% in the different symptoms. In the second round the feature set contained only tokens from the *Key Term Vocabulary*. This boosted up the classification performance to 90% F-measure for blood sugar and body weight change. When we restricted the feature space once again leaving only the most significant symptom words in the feature space the performance was about 89% F-measure. In all cases the precision varied about 92-94%, and up to 98% when classifying blood sugar level with the full keyword set, which is encouraging. At that time the recall was about 75% in blood sugar identification and this could be explained with the highly imbalanced dataset. Only about 20% of the sentences were blood sugar related and 6% body weight change related. These results

Method %	Precision	Recall	F-measure
J48 bs 22 feat.	94.80	80.00	86.80
J48 bwc 16 feat.	94.30	85.30	89.60
Rule-based bs	96.40	90.00	93.09
Rule-based bwc	98.50	92.00	95.14

Table 5: Level 1 evaluation. ML vs Rule-based best performance.

Phase	Precision	Recall	F-measure
Blood sugar (bc)			
Ph.1 Focus	96.4	90.0	93.09
Ph.2 Status	91	45.5	60.6
Ph.3 Values	88.9	77.8	83
Ph.4 Neg.	96.3	94.2	95.2
Ph.5 Scope	97	96	96.5
Body weight change (bwc)			
Ph.1 Focus	96.6	90.6	93.5
Ph.2 Status	86.2	78.1	82
Ph.3 Values	87.5	70	77.8
Ph.4 Neg.	NA	NA	NA
Ph.5 Scope	82.7	75	78.7

Table 6: Rule performance by level

show that even without introducing domain knowledge and using the full feature space the positive output predictions are reliable. SVM classification was also tested but it was outperformed by J48.

Table 5 shows that the precision of the rule-based approach is higher than the one obtained by automatic classification. However during the error analysis we noticed that in the rule-based setting some true positives were wrongly classified as such because they matched non symptom related expressions in sentences where the symptoms occur and respectively are annotated as positive. In means of precision both approaches differ only in about 2 points which invokes the assumption that they are comparable to each other and could be used as alternatives for experiments on a larger scale even without incorporating domain knowledge, especially in such a task where the accuracy of the extraction is more important than the coverage.

5.2 Phase by Phase Evaluation

Results from the separate phases of rule-based analysis are shown in table 6.

At *phase 2* the tokens available in the training set are completely recognised; there is a group of tokens which are not available in the training set, but during the *phase 1* processing fall into the scope of the expression of interest. These ones are included to the condition description without having assigned any status class. Tokens may not be identified for two reasons – they are not available in the training set or they are misplaced (e.g. the target adjective is following the focal expression instead of preceding it, as it happens in all training set examples). 45% of the attributes expressing blood sugar status are recognised and 78.1% espressing body weight. Although the recall for blood sugar seems to be low at this phase, the result is actually good because during the error analysis we found out that 60% of the tokens which were not identified were equivalent.

At *phase 3* the main problem for value recognition were the alphanumerical expressions of lab test values which occur comparatively rare and have a wide range of spelling variants (and errors). Thus few extraction errors have high influence on the precision. This problem can be easily overcome by pregenerating a list of alphanumeric expressions and their variations. The negation at *phase 4* was recognised with high accuracy.

At *phase 5* all scope problems for blood sugar related expressions are resolved successfully except for one. The interval describing the value of the blood sugar was written as ”от 12 ммол/л до 14 ммол/л” *(from 12 mmol/l to 14 mmol/l)* instead of ”*from 12 to 14 mmol/l*” like all such examples in the training set. This lead to wrongly recognised right border and only partial recognition of the blood sugar level value. However this issue could be easily overcome by extending the recognition rules with additional ”cosmetic” clauses for processing of alphanumeric values as suggested above. It would be helpful for recognition of any symptom to add new lexical alternations and paraphrases in addtion to the stemmed forms in the regex. Our approach is completely driven by the training set analysis because our goal is to see how far do we get on that base.

The extension of the rules as shown on figure 1 helped identifying blood sugar descriptions twice. We believe that such extensions in feature will have higher impact on a larger scale experiment.

6 Conclusion and Future Work

We proposed a unified approach to the recognition of medical descriptions with composite meaning, which are represented by a wide range of paraphrases in the free EHR texts in Bulgarian. The results show a relatively high precision in identifying health condition descriptions in the EHR texts. This was achieved with the use of shallow rules and minor additional effort to extend the rules coverage - stemming of the source documents and adding new meaningful links to the rules where possible. The sentence identification task has nearly the same accuracy in terms of precision when performed with a binary J48 classifier and with the rule-based *phase 1* analysis even without incorporating key terms in the classification. These results give an insight into the possibilities of a further usage of automatic classification for such tasks, due to its flexibility.

As a follow up to this study we will try to generalise this algorithm to a more abstract level so that it can be transferable for the identification of other health conditions, medication etc. We will also put effort in the automatic extraction of symptom identification rules by analysing the classification predictions and the corresponding document feature vectors.

Acknowledgments

The research work presented in this paper is partly supported by grant DO 02-292/December 2008 ”Effective search of conceptual information with applications in medical informatics”, funded by the Bulgarian National Science Fund in 2009-2012 (EVTIMA).

References

Boytcheva, S., A. Strupchanska, E. Paskaleva, D. Tcharaktchiev, 2005. *Some Aspects of Negation Processing in Electronic Health Records.* In Proc. Int. Workshop LSI in the Balkan Countries, 2005, Borovets, Bulgaria, pp. 1-8.

Boytcheva S., I. Nikolova, E. Paskaleva, G. Angelova, D. Tcharaktchiev and N. Dimitrova, 2010. *Obtaining Status Descriptions via Automatic Analysis of Hospital Patient Records.* Informatica, 34(3):269-278.

Chapman, W., W. Bridewell, P. Hanbury, G. Cooper and B. Buchanan, 2001. *A simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries.* J Biomed Inf, 34(5):301-310.

Chapman D. Chu, J.N. Dowling and W.W. Chapman, 2006. *Evaluating the Effectiveness of Four Contextual Features in Classifying Annotated Clinical Conditions in Emergency Department Reports.* AMIA Annu Symp Proc, pp. 141-145.

Elkin, P.L., S.H. Brown, B.A. Bauer, C.S. Husser, W. Carruth and L.R. Bergstrom, et al., 2005. *A Controlled Trial of Automated Classification of Negation From Clinical Notes.* BMC Med Inform Decis Mak, 2005, 5(1), p. 13.

Friedman, C., P.O. Alderson, J.H. Austin, J.J. Cimino and S.B. Johnson, 1994. *A General Natural-Language Text Processor for Clinical Radiology.* JAMIA, 1994 Mar-Apr, 1(2), pp. 161-74.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, 2009. *The WEKA Data Mining Software: An Update*; SIGKDD Explorations, 11(1).

Harkema, H., J. Dowling, T. Thornblade, W. Chapman, 2009. *ConText: An Algorithm for Determining Negation, Experiencer, and Temporal Status from Clinical Reports.* J Biomed Inf, 2009, 42(5), pp. 839-51.

Hahn, U., M. Romacker and S. Schulz, 2002. *MEDSYNDIKATE - a Natural Language System for the Extraction of Medical Information from Findings Reports.* Int J Med Inf, 2002, 67(1-3), pp. 63-74.

Huang, Y. and H.J. Lowe, *A Novel Hybrid Approach to Automated Negation Detection in Clinical Radiology Reports.* JAMIA, 2007, 14(3), pp. 304-11.

Mutalik, P., A. Deshpande, P. Nadkarni, *Use of General-purpose Negation Detection to Augment Concept Indexing of Medical Documents: a Quantitative Study using the UMLS.* JAMIA, 2001, 8(6), pp. 598-609.

Nakov, P., 2003. *BulStem: Design and Evaluation of Inflectional Stemmer for Bulgarian.* In Proc. of Workshop on Balkan Language Resources and Tools (1st Balkan Conference in Informatics), Thessaloniki, Greece, November, 2003.

Savova G., P. Ogren, P. Duffy, J. Buntrock, C. Chute, 2008. *Mayo Clinic NLP System for Patient Smoking Status Identification* JAMIA, 2008, 15(1), pp. 25-28.

Visa, S., A. Ralescu, M. Ionescu, 2007. Investigating Learning Methods for Binary Data In Proc. Fuzzy Information Processing Society, 2007. NAFIPS '07. Annual Meeting of the North American June 2007, pp. 441-445

Project EVTIMA – ”Effective search of conceptual information with applications in medical informatics”, `http://www.lml.bas.bg/evtima`