Extraction of Named Entities from Semi-Structured Texts for Medical Domain

Natalia Zhukova[1], Maksim Berezov[1], Sergey Lebedev[1] and Ekaterina Zavadskaya[2]

[1] ITMO University, Kronverksky Pr. 49, 197101 St. Petersburg, Russia
[2] National Research University Higher School of Economics, Myasnitskaya ulitsa 20, 101000, Moscow, Russia

Abstract. There are many unstructured text data stored in health information systems (HIS). To use this data for automatic processing it is necessary to be able to extract specific medical entities such as prescribed drugs, diagnosis, body conditions and so on. The article is concerned with building discriminatory models for solving the problem of named-entity extraction (NEE) for medical texts in Russian. Such models as Markov random fields and support vector machines (SVM) are considered. These methods showed better results in comparison with other NEE methods for English language corpuses. The application of these methods to text in Russian and moreover to medical text that burdened with specific medicine terminology is still a problem. To solve this problem the processes of feature extraction and models building are described in the context of said texts. Methods are evaluated on a corpus received from Federal Almazov North-West Medical Research Center. As a result, the most accurate method in according to F1-measure is chosen.

Keywords: Named-entity extraction, SVM, Markov Random Fields, Russian medical texts

1 Introduction

Natural Language Processing is a field of computer science, computational linguistics concerned with the interactions between computers and human languages, and concerned with programming computers to fruitfully process large natural language corpora [1]. NLP methods are especially acute for medical texts processing. There are a lot of texts have been stored in health information systems. But this data is weakly structured and cannot be used for automated analysis without preliminary processing. From the other hand, these texts contain very important expertise that can be used for data analyses or health care quality evaluation. That is why it is very important to be able to correctly extract useful information, which in the future will facilitate the work of medical or insurance personnel.

The task of the NLP for semi-structured medical texts differs from the processing of conventional texts because of the specific characteristics – absence of verbs, lack of emotional coloring, lack of homonymy, presence of standard patterns. Such properties make effective processing of available texts with the help of dictionaries and rule-based approaches. However, the perspective development direction is the use of ML methods. ML, in combination with available methods, is necessary for selecting those entities that are not in the dictionary, and can also be used to increase the prediction accuracy of entities, which are already in the dictionary.

Identifying special entities is a kind of information extraction (IE). For NLP it is one of the basic tasks. Identifying named entities in text is called Named Entity Recognition (NER). More accurate, NER is the task that seeks to locate and classify entities in text into pre-defined categories. This task is extremely significant for the processing of semi-structed medical texts. The most important difference is the obligatory presence of the context dependency. Therefore, to effectively solve this problem, we can use only those methods that take into account the context (CRF) or partially consider the context (SVM). In contrast to texts of general orientation, where we can achieve great accuracy using almost any ML method.

The objective of this research is context-dependent NER from medical texts in Russian gotten from one of the Russian medical center. Contextual dependency is provided by the type of records (diaries, complaints, etc.) and the available dictionaries obtained earlier. And also by the specific features of each named entity. For example, for drugs – this is the use of the active substance to find an analogue in the dictionary. The following requirements for NER-module can be singled out: replenishment of the dictionary (based on new entities) and integration with other classifiers and modules in the existing NLP system.

The article is organized as follows. First, the existing methods for NER are presented and the scientific researches in this area are analyzed. Specification of texts and technology – is in the third section. The description of the methods used, the construction of a feature space and the adjustment of models are presented in the fourth section. After that, the results of the experiment will be presented and conclusions will be drawn about the applicability of these methods for solving NER problem.

2 Existing NER approaches

There are three main approaches for solving this problem: rule-based, ML based and mixed approach. We will consider in detail only ML methods, which are divided into generative and discriminative models (see Fig.1).

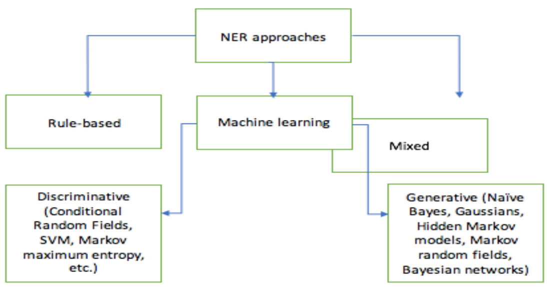

Fig. 1. NER approaches

Generative models randomly generate observable data values, typically given some hidden parameters. It specifies a joint probability distribution over observation and label sequences. Discriminative models directly estimate posterior probabilities, they don't try to model underlying probability distributions. This is the key difference between them.

In [2-3], the authors conclude that discriminative models are more effective for NER than generative models.

As we can see in [4], authors reach good results using SVM-classifier for NER in English medical texts. They consider it suitable for this problem.

In [5], author comes to the conclusion that, in general, CRFs outperformed SVMs for clinical texts. But both are effective.

In [6], author achieves results more than 90% for all measures. This gives us reason to assume that these methods are highly effective for this task.

Thus, we can distinguish two most suitable discriminative models for NER for clinical texts – CRF, SVM.

However, these methods do not find proper reflection in research [7] in Russian. Authors prefer to use other approaches and the accuracy of their results is not given.

Research that does not relate to medical data processing deserves special attention.

In [9] authors proposed two baselines (knowledge-based and statistical) for Russian language NER. They obtained results of 62.17% and 75.05% F1-measure. They find these results very promising, given that neither of our baselines employs morphological or syntactical analysis.

In article [10] authors extracted names of organizations, media, locations, and geopolitical entities using CRF. They come to the conclusion that this approach is suitable and gives high accuracy.

In [11] authors also used CRF. They have explored the task of recognizing opinion expressions in social media associated with diseases and drugs. Authors demonstrated the superiority of CRF as compared to a dictionary-based method and recurrent neural networks.

This research [12] is related to fact extraction system. This paper was distinctive and extremely useful for our research. We used the same four-level markup and the same methodology for our experiment.

One can draw a conclusion that CRF, SVM showed better results in comparison with other NER methods for English language corpuses. The application of these methods to text in Russian and moreover to medical text that burdened with specific medicine terminology is still a problem. So, the main goal of this article is to apply the most efficient discriminative models for increasing recognition accuracy for medical texts in Russian. The obtained results of methods evaluation can be useful for creators of modern health information systems.

3 Technology of medical texts processing

3.1 Data representation

We use a corpus received from Federal Almazov North-West Medical Research Center. The initial data includes prescribed by the doctor medicines, in what dosage, how often the patient should take them and the active substance of the medication. All received texts are semi-structured. The main difficulty in the processing of texts is their preprocessing. The training sample should be clearly and correctly labeled. Since the volume of the training sample is large, this task becomes time-consuming. The format of corpus markup will be discussed below in the corresponding section. For testing the statistical model, manual BIO-marking of about 700 sentences containing about 2000 named-entities was performed.

In this article, the objective was to recognize the following named-entities:

Prescribed drugs: DRUG

Dosage: DOSAGE

Active substance: SUBSTANCE

Frequency of use: FREQUENCY

In the NER task, the machine learning algorithm is used to define labels for classes of named entities for each word of the text. The label in this case can be the type of the named entity or the label "unnamed entity". However, this approach to the selection of labels does not allow us to establish the boundaries of entities in the case when two named entities of the same type are near

In the BIO representation, the region information is represented as the prefixes “B”,” I”,” O”. Prefix “B” (Beginning) means that the current word is at the beginning of a named entity, “I” (Inside) means that the current word is in a named entity, “O” (Outside) means that word does not belong to named entity.

Example (in origin language): [B-SUBSTANCE Метопролол] [O ( ] [B-DRUG: эгилок-ретард] [O)] [B-DOSAGE 50] [I-DOSAGE мг] [B-FREQUENCY 1] [I-FREQUENCY раз] [I-FREQUENCY утром]

3.2 Evaluation

As evaluation metrics, we use: precision, recall, F1-score

$$P = \frac{TP}{TP+FP}$$

Where FP – number of false positive samples, TP – number of true positive samples.

$$R = \frac{TP}{TP+FN}$$

Where FN – number of false negative samples, TP – number of true positive samples.

$$F = \frac{2RP}{R+P}$$

3.3 Medical texts processing algorithm

In the article the processing of medical texts can be represented in the form of an algorithm.

1) Manual or partially automated marking of training medical data.
2) Construction of the classification model.
3) Construction of the feature space.
4) Classifier training.
5) Classification of the test set.
6) Evaluation of the results

4 Machine learning based methods

In this section, a detailed description of each method used I given. We show the implementation of this model for solving the problem and explain which features we used, in other words we discuss the 2nd-5th steps of the algorithm. For the fourth and fifth steps, we give the mathematical apparatus of the methods used.

4.1 CRF

Definition and constructing of the classification model. Conditional Random Field - statistical method of classification. A distinctive feature of this method is the ability to take into account the context of an object. CRF is a discriminative non-directional probabilistic graphic model.

Formally, the Markov random field consists of the following components:

- Unoriented graph or factor graph G = (V, E) , where $\forall$ vertex is random variable X and each edge is a relation between the random variables u and v.

- A set of potential functions $\{\varphi_k\}$, one for each clique (full subgraph G of the undirected graph) in the graph. The function φ_k puts each possible state of the clique elements into a certain non-negative real-valued number.

Vertexes that are not contiguous must correspond to conditionally independent random variables. The group of adjacent vertices forms a clique, the set of states of vertices is the argument of the corresponding potential function.

Potential function should be chosen equal to

$$\varphi\,(\mathrm{Yt}-1\,,\mathrm{Yt}\,,\mathrm{Xt}\,) \;=\; \exp\{\textstyle\sum_{n=1}^{N} \Theta\mathrm{n}\, f(\,\mathrm{Yt}-1\,,\mathrm{Yt}\,,\mathrm{Xt}\,)\}$$

Where f (Yt-1, Yt , Xt) is feature function, and $\boldsymbol{\Theta}$n are the corresponding parameters of the model which should be evaluated in the learning process. Then the probability of a chain of hidden variables (Y) under the condition of a chain of observables variables (X) is:

$$P(y|x) = \frac{\prod_{t=1}^{T} \varphi\,(Yt-1\,,Yt\,,Xt\,)}{\sum_{y\prime} \prod_{t=1}^{T} \varphi\,((Yt-1)\prime\,,(Yt)\prime\,,Xt\,)}$$

According the above, to solve NER problem we use linear CRF model with the size of the maximum clique equals 3 (see Fig.2).

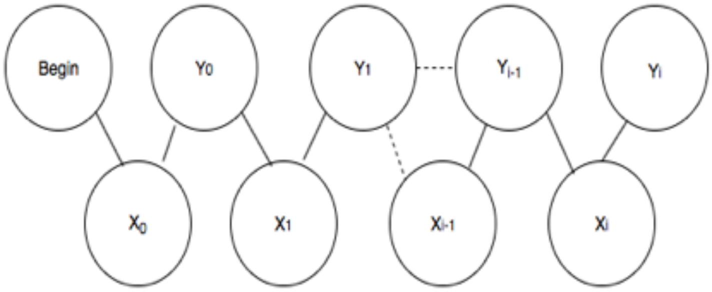

Fig. 2. Linear CRF Model

Feature space. In the NER problem, features are understood as the characteristics of words used by machine learning algorithms during the learning process and entity recognition.

The following set of characteristics was proposed:

Current word Xi

2 previous labels Yi-1 , Yi-2

Type of current word Xi (all letters are capitalized, the first letter is capitalized, all symbols are digits, word contains digit, word is alphanumerical, word contains dash, word contains Roman numericals, words contain one capital letter in in the middle, digits-coma-digits, digits-dot-digits, one digit, word is the first word of sentence, word contains special symbols, word contains brackets in front of, word contains brackets behind)

Part of speech

Suffixes of length 2

Prefixes of length 2

Suffixes of length 3

Prefixes of length 3

Suffixes of length 4

Prefixes of length 4

Neighborhood words N = (Xi-3, Xi-2, Xi-1, Xi+1, Xi+2, Xx+3)

Conjunction Yi-1 with N

Type of word and POS for each word in N

The size of the admissible set of X signs is large enough, it was decided to use Random Forest algorithm for selecting the features.

4.2 Support vector machines

Definition and constructing of the classification model. Support vector machines (SVM) is binary classifiers, which outputs are +1 or -1 given a sample vector x. The decision bases on separating hyperplane h(x).

$$h(x) = \begin{cases} +1, \; if \; wx + b > 0 \;\; b \in R, w \in \text{Rn} \\ \qquad\qquad -1, otherwise \end{cases}$$

The class for an input is determined by side of the space separated by the hyper-plane, if h(x) = 0 it means that input sample lies on separating line. The key idea is to find optimal hyperplane with the maximum margin (distance between nearest data sample and plane). Speaking formally, we should minimize:

$$\frac{||w||^2}{2} \rightarrow \min$$

Also,

$$\text{hi}(\text{w} * \text{xi} - \text{b}) \geq 1, 1 \leq \text{i} \leq \text{n}$$

The solution of this problem is known and can be written in this form:

$$\text{f}(\text{x}) = \text{w} * \text{x} + \text{b} = \sum_{i \in SVs} \text{yi ai x} * \text{xi} + \text{b}$$

Multi-class classifier. As we can see, classical SVM is suitable for binary classification, but it is not our case. We should use method for constructing multi-class classifier. The most progressive way is to use pairwise method proposed by Krebel and extend the BIO representation to enable the training with the entire corpus. The idea of this method is to combine a lot of binary classifiers. We construct N(N-1)/2 binary SVMs, each of them votes (makes decision is sample belongs to i or j class). After that we should choose class with maximum number of votes. But we can encounter with unbalanced class distribution. Efficient way to solve it is to split the class "Outside" into several sub- classes according to part-of-speech (POS) information of the word. This approach was applied in the study [8]

Feature extracting. Input vector x – is feature representation of current word Xi and its context in some area. To correctly convey the context, we are using 4 relatives' positions of this word (2 in front of and 2 behind). For example, we denote word's position as i index. If it's negative it means that we are talking about previous word (which was i words before current word). Similarly, for positive indices (see fig.3).

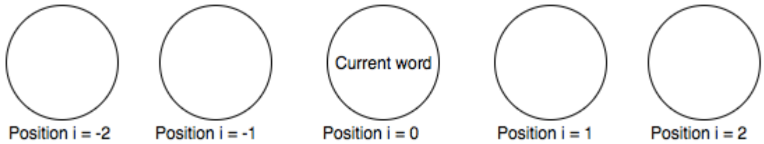

Fig. 3. Word's position for SVM

$$\text{Part of speech feature i, j} = \begin{cases} 1, & \textit{if word in position i is assigned the j} - \textit{th position in POS list} \\ & 0, \textit{otherwise} \end{cases}$$

$$\text{Prefix feature i, j} = \begin{cases} 1, & \textit{if word in position i is assigned the j} - \textit{th position in prefix list} \\ & 0, \textit{otherwise} \end{cases}$$

$$\text{Suffix feature i, j} = \begin{cases} 1, & \textit{if word in position i is assigned the j} - \textit{th position in suffix list} \\ & 0, \textit{otherwise} \end{cases}$$

$$\text{Previous class feature i, j} = \begin{cases} 1, & \textit{if word in position i } (i < 0) \textit{ is assigned the j} - \textit{th class} \\ & 0, \textit{otherwise} \end{cases}$$

It's very useful feature, we can use classes' of already predicted words as features. Note that the selection of the first three features requires the compilation of lists of possible characteristic values (in contradistinction to CRF).

Also, we use 14 type of word features, which we described in section about CRF feature extracting.

So, after feature extracting the dimension of input vector equals 5*(19+|POS|+|Prefix|+|Suffix|). We can reduce this dimension by using Random Forest.

5 Results

The medical corps was provided partly by Federal Almazov North-West Medical Research Center. and partly taken from open sources. The medical corps was manually marked with a BIO markup. For testing the statistical model, manual BIO-marking of about 700 sentences containing about 2000 named-entities was performed. It should be noted that we singled out in a separate class every word that did not belong to any of the entities

The table below shows the results of a computational experiment on a training set for CRF.

Table 1. Results for CRF

	Precision	Recall	F1-score
Drug	67,23	74,20	70,54
Dosage	63,33	60,11	61,67
Substance	70,1	69,33	69,71
Frequency	61,1	62,54	61,81

The table below shows the results of a computational experiment on a training set for SVM.

Table 2. Results for SVM

	Precision	Recall	F1-score
Drug	65,01	66,80	65,89
Dosage	57,15	55,66	56,39
Substance	67,10	67,50	67,29
Frequency	58,83	59,82	59,01

As we can see, quite good results were achieved in recognizing such entities as Drug and Substance. This is due to the fact that only one tag was used for their marking. CRF is completely superior to its competitor.

5.1 Cross-validation

Cross-validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice.

5-fold cross-validation was used in the project. It means that the original sample is randomly partitioned into 5 equal sized subsamples. Of the 5 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 4 subsamples are used as training data. The cross-validation process is then repeated 5 times, with each of the 5 subsamples used exactly once as the validation data. The 5 results from the folds are averaged to produce a single estimation.

6 Conclusion and future work

In this paper, two ML methods for solving NER problems have been described: CRF and SVM. The article describes the general model of CRF, describes the details of using and tuning the parameters of the linear model of CRF. Also, the concept of using SVM for multiclass classification taking into account the context was presented. As can be seen from the results of the experiment, CRF showed the best performance for almost all named entities due to deeper consideration of the context. But we can conclude that both approaches are applicable to the solution of our problem. Entities with the highest recognition accuracy are drugs and substances. This may be due to the fact that for markup we use, as a rule, only one tag, which essentially simplifies the task.

The results obtained tell us about the possibility of integration these methods into our NLP system as new named-entity recognizer. Of course, it's possible to improve the results by analyzing typical classifier's errors. This will be our further work in this area, also we are trying to test the consistency of the idea of combining classifiers in a cascade to increase the accuracy.

References

1. *Wikipedia: The Free Encyclopedia*. Wikimedia Foundation, Inc. 22 July 2004. Web. 10 Aug. (2004). https://en.wikipedia.org/wiki/Natural_language_processing
2. T. Jebarra, Machine Learning: Discriminative and Generative, Kluwer, (2004)
3. Huang Z., Xu W., Yu K. Bidirectional LSTM-CRF models for sequence tagging //arXiv preprint arXiv:1508.01991, (2015)
4. Kazama, Jun'ichi, et al. "Tuning support vector machines for biomedical named entity recognition." Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3. Association for Computational Linguistics, (2002)
5. Li, Dingcheng, Karin Kipper-Schuler, and Guergana Savova. "Conditional random fields and support vector machines for disorder named entity recognition in clinical texts." Proceedings of the workshop on current trends in biomedical natural language processing. Association for Computational Linguistics (2008)
6. Settles B. Biomedical named entity recognition using conditional random fields and rich feature sets //Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, pp. 104-107, Association for Computational Linguistics (2004)
7. Богатырев М. Ю., Вакурин В. С. Концептуальное моделирование в исследовании биомедицинских данных //Математическая биология и биоинформатика, pp.340-349, (2013) – Т. 8. – №. 1
8. Lee, Ki-Joong, Young-Sook Hwang, and Hae-Chang Rim. "Two-phase biomedical NE recognition based on SVMs." Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13. Association for Computational Linguistics (2003)
9. Gareev R. et al. Introducing baselines for Russian named entity recognition //International Conference on Intelligent Text Processing and Computational Linguistics, pp. 329-342, Springer Berlin Heidelberg (2013)
10. Mozharova V. A., Loukachevitch N. V. Combining Knowledge and CRF-based Approach to Named Entity Recognition in Russian //International Conference on Analysis of Images, Social Networks and Texts, pp. 185-195, Springer, Cham (2016)
11. Miftahutdinov Z. Sh. et al. Identifying Disease-Related Expressions in Reviews Using Conditional Random Fields //Computational Linguistics and Intellectual Technologies, pp. 155-166, Proceedings of the Annual International Conference «Dialogue» (2017) - Т. 1
12. Starostin A. S. et al. FactRuEval 2016: Evaluation of Named Entity Recognition and Fact Extraction Systems for Russian //Computational Linguistics and Intellectual Technologies, pp. 702-720, Proceedings of the Annual International Conference «Dialogue»(2016)