pHealth 2021
B. Blobel and M. Giacomini (Eds.)

doi:10.3233/SHTI210579

An Unsupervised Approach to Structuring and Analyzing Repetitive Semantic Structures in Free Text of Electronic Medical Records

Varvara KOSHMAN[a,1], Anastasia FUNKNER[a], Sergey KOVALCHUK[a,b]
[a] *ITMO University, Saint Petersburg, Russia*
[b] *Federal Almazov North-west Medical Research Centre, Saint Petersburg, Russia*

Abstract. Electronic Medical Records (EMR) contain a lot of valuable data about patients, which is however unstructured. There is a lack of labeled medical text data in Russian and there are no tools for automatic annotation. We present an unsupervised approach to medical data annotation. Morphological and syntactical analyses of initial sentences produce syntactic trees, from which similar subtrees are then grouped by Word2Vec and labeled using dictionaries and Wikidata categories. This method can be used to automatically label EMRs in Russian and proposed methodology can be applied to other languages, which lack resources for automatic labeling and domain vocabularies.

Keywords. syntactical parsing, natural language processing, electronic health records, node2vec, automatic text labeling, graph algorithms

1. Introduction

It has been shown that the use of textual content of EMRs for training language models significantly improves model performance [1]. However, it is currently hardly feasible to include it when working with Russian, because of very few labeled datasets available. One reason for it is that manual labeling requires time and great effort by domain experts. Also, the idea of automatic annotation faces a difficulty that there are no ready-to-use medical terminologies in Russian. A specific syntactic structure with free word order missing subject naming and omitting conjunctions makes automatic annotation difficult. An attempt to extract deterministic characteristics from EMRs in Russian was proposed by A. Funkner [2]. However, results contained many incorrect and unnecessary constructions and it was concluded that a syntactic parsing should be used for discovering sentence structure. Morphological parsing was applied to improve labeling EMR data in Arabic in a task of assigning English labels to textual data in Arabic [3]. Arabic and Russian have in common that there are no resources for language processing and no structured medical databases like PubMed for English. They use Wikipedia to link entities in Arabic with their corresponding English entities. The application of a Wiki-based approach to Russian was studied by Sysoev A., who used Russian Wikidata graph

[1] Corresponding Author: Varvara Koshman ITMO University, 49 Kronverksky Pr., St. Petersburg, 197101, Russia; E-mail:207955@niuitmo.ru

for training word embeddings to improve performance of entity linking [4]. J. . Raiman suggested using Wikidata graph's parental relations as categories for entity linking [5]. However, aforementioned methods have not been utilized yet together for annotation of EMRs. The aim of this paper is a design and development of a method for automatic detection of repetitive semantic constructions in unstructured EMR text data.

2. Methods

The idea behind grouping similar semantic constructions is putting together similar symptom terms, word abbreviations and drug names, including those not present in the knowledge base. With this done, some of the words and phrases get relevant labels as members of a labeled group. To this end, the structure of a sentence is analyzed with morphological and syntactic parsers; cosine distance between word embeddings is used as a similarity metric between words in parsed trees. Figure 1 shows the detailed method schema.

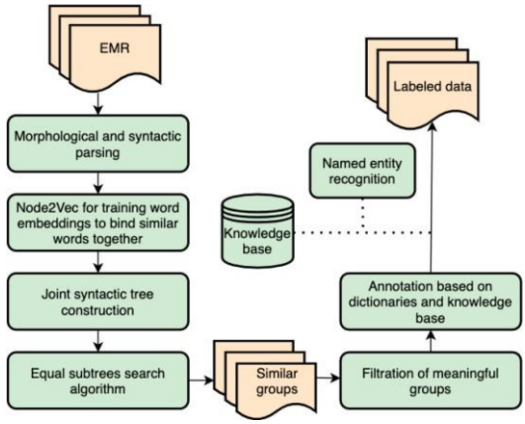

Figure 1. Method schema of sequential modules with EMR as input and labeled groups as output.

2.1. Morphological and Syntactic Parsing

As a morphological and a syntactic parsers BERT-based models implemented by DeepPavlov [6] were used. Models used in this research were trained on UD Russian SynTagRus corpus (version 2.3) and were not additionally trained on medical data.
First paragraph.

2.2. Node2Vec on Syntactic Trees

We use Node2Vec to train a CBOW model. Node2Vec [7] is commonly used with graphical structures [8,9], its strategy of random sampling helps to preserve hierarchical relations between nodes in word embeddings. The graph was created by connecting the roots of all syntactic trees with a virtual node with syntactic relations as weights. Stop words were removed for training and all words were converted to their normal forms. Node2Vec was executed with non-normalized probabilities p=2, q=3, with five random walks per root and five words in one walk at most. When q is higher than p, the algorithm's behavior is similar to local search. Such behavior is more beneficial when

dealing with syntactic relations. The resulting vector space contains embeddings trained on medical data and 50k embeddings pre-trained on Russian fiction dataset.

2.3. Algorithm for Search of Similar Subtrees in a Tree

Our algorithm for grouping similar subtrees in a tree is inspired by equal subtree search [10], but extended to the version with a dependency on multiple node's heights, not a single one, i.e. it allows a subtree (a phrase) occur in different parts of a tree (a sentence) instead of a fixed position. This is especially useful for languages with free word order like Russian. The main idea is shown in Figure 2.

```
1: G ← joint syntactic tree
2: H ← dictionary of nodes' heights of G
3: groups ← {}
4: extendTree() ▷ create new nodes in G for synonymous words
   to incorporate similarity between them
5: for h from 0 to max (H) do
6:   representations ← computeRepresentations()
     ▷ compute string representations of subtrees for H(h) nodes
7:   combinations ← generateCombinations(representations)
     ▷ generate possible subtree combinations C_n^k, n – number of
       children, k = 1..n
8:   groups ← groups ∪ groupSubtrees(combinations)
9: stringGroups ← DFS(G, groups)
   ▷ traverse G to find initial word sequences
```

Figure 2. Main idea of similar subtree search module

2.4. Annotation Process

Some group names were decided manually based on the specific context of EMR: 'Event', 'Time stamp', 'Drug', 'Sign and symptom', 'Disease' and 'Physician specialization'. The belonging to these groups is decided by simple rules. A phrase gets an 'Event' label if it contains a verb in passive voice indicating that some action was applied, 'Time stamp' – if there is temporal information contained. For assigning 'Drug', 'Sign and symptom', 'Disease' dictionaries of labels were prepared. Our dictionary of drugs contains a parsed set of names listed in Vidal.ru reference book (6360 names), data for dictionaries of diseases (4657 names), sign and symptoms (355 names) and physician specializations (41 names) was crawled from Russian medical web sites.

Although dictionaries cover an extensive amount of information in each group, they do not contain synonyms and most common abbreviations for domain terms. This is crucial, as they are very often used in EMRs. Besides, 'disease' and 'Sign and symptom' are too general labels and make it impossible to distinguish groups of related terms. Also, medical procedures, medical organizations, body organs and etc. need to be included. As a solution a medical database was created based on the data gathered and structured from different medical databases by Wikidata. Using SPARQL queries and public MediaWiki API medical Wikidata entities were fetched. An entity is considered medical if it has a link to a medical database as a property. 32 databases that were considered most relevant for EMRs were chosen from the initial list [11]. The resulting number of medical entities was 18.9k entities and 17.1k synonyms for them. Parental relations between entities in the Wikidata graph can be used to define categories for each entity. Among parental relations we picked 'instance of', 'subclass of' and 'part of' as the most defining [5] and used entities they point to for labeling. Figure 2 represents database schema. For example,

according to Wikidata entity for 'electrocardiogram' has medical property 'P486' (MeSH descriptor ID), has synonyms 'EKG', 'ECG' and categories 'medical testtype' (instance of), 'medical test' (subclass of), 'electrophysiology' (part of). This waya mention of 'EKG' gets a 'medical test type' label as the closest parental relation.

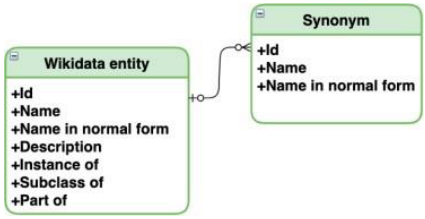

Figure 3. Medical database schema

Though the prepared database has a medical domain, there are a few hundreds of names that point to multiple entities. As these ambiguous cases are quite rare, it was decided to define a rule that prefers those entities that are closer to the context of a corpus being annotated. Concretely, a skip-gram model was trained with Node2Vec on a database graph and a forest of initial syntactic trees (Node2Vec parameters: p=1, q=2, number of walks per root=3, walk length=5) and a decision between possible entities was made in favor of the one with the highest cosine similarity score.

3. Results

3.1. Data

Experiments were conducted on a corpus of 5k sentences with time expressions in Russian of anonymized EMRs of patients with acute coronary syndrome, who were under observation in Almazov National Medical Research Centre in 2010-2015.

3.2. Extracted Groups

Our algorithm extracted nearly 11k groups in total. Figure 4 shows bar charts with frequency statistics of size of groups: commonly groups are small and in most consist of up to 10 repeated phrases. The maximum repeat length was limited to five words. Table 1 represents several examples of repeats in groups and Table 2 – synonymous words found with Node2Vec.

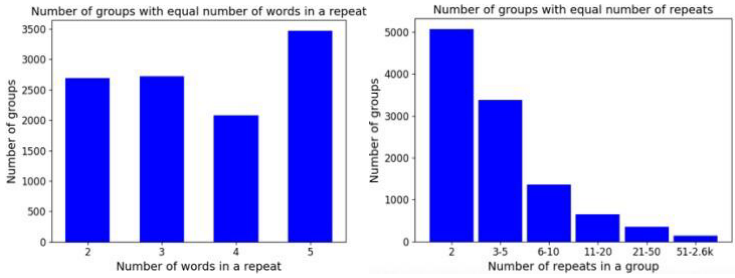

Figure 4. Bar charts with descriptive statistics of the annotated groups

Table 1. Examples of repeats in groups with labels

Group Russian	Group English	Label Russian	Label English
в больницув поликлинику в МСЧв ЦРБ	in hospital in policlinic in medical unit in hospital	медицинская организация, больница, общественный институт	medical organization, hospital, public institution
ухудшение состояния	deterioration	характеристика заболевания	disease characteristic
перелом бедра	hip fracture	анатомическая структура, кость, болезнь	anatomical structure, bone, disease

Table 2. Examples of words similar by cosine distance defined by Node2Vec

Word Russian	Word English	Synonyms Russian	Synonyms English
жена	wife	дочь, сестра, мать	daughter, sister, mother
мрт	mri	томография, флюорография, экг	tomography, fluorography, ecg

3.3. Annotation

Using dictionaries 6.8k out of 11k got annotated. Labelling with Wikidata increased the annotated number of groups to 8.6k and number of groups with more than one label grew from 1.2k to 4k. Figure 5 shows 30 most common labels.

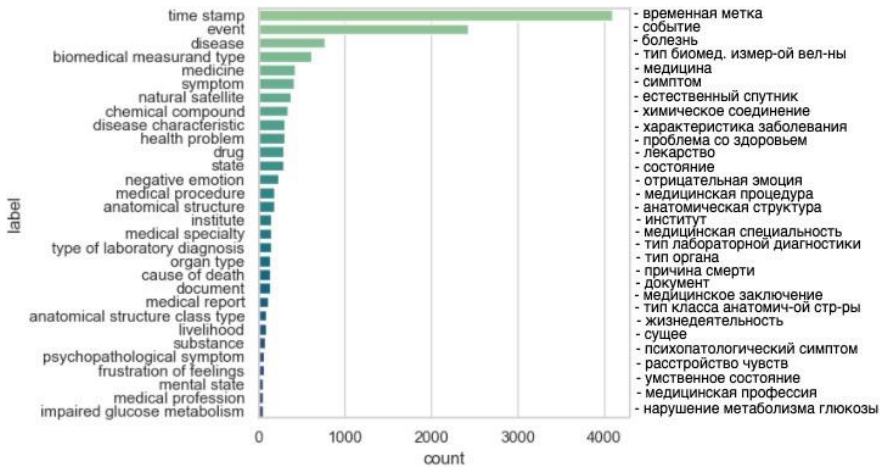

Figure 5. 30 most common labels sorted by their frequency in a result set

4. Discussion

The method we developed succeeded in joining semantically close phrases: some common abbreviations (for example, for medical organizations and lab tests), word reductions (for example, ‘department’ and ‘dep' in Russian) and minor typos. Diseases, organs, body parts, geographical places were grouped by the system. To the best of our knowledge, this is a first attempt of grouping medical free text by semantic similarity before automatic annotation intending to cover more words.

5. Conclusions

The Key contributions of this work are a design of a new methodology for automatic annotation of EMRs, a proposed method for finding similar subtrees in a tree, asuccessful application of a classic Node2Vec algorithm to syntactic trees and a creationof a medical Wikidata-based database for labeling in Russian. The whole pipeline can beadapted to other languages by changing the language-specific preprocessing module, bychanging a language code a corresponding database can be created. For Russian a graphicinterface was implemented for annotating new datasets with statistics representation.

Current limitation of a method is a limit in number of words in a sentence equal to 25. Our algorithm gets slower on longer sentences, as it gets more string combinations to compute and compare on big trees. In the nearest future it is planned to avoid this limitation and provide a fast performance for all sentences. The tool can generally increase the number of labelled datasets available, which can be used later by researches in machine learning problems related to the medical domain. This, in turn, can broad the scope of problems and save time for both domain experts preventing them from huge manual work and for researches who get their data labeled quickly.

Acknowledgement

This research is financially supported by The Russian Scientific Foundation, Agreement #17-15-01177.

References

[1] J. Liu, Z. Zhang, N. Razavian, Deep EHR: Chronic Disease Prediction Using Medical Notes, *Proceedings of the 3rd Machine Learning for Healthcare Conference* (2018).
[2] A. A. Funkner, S. V. Kovalchuk, Time Expressions Identification Without Human-Labeled Corpus for Clinical Text Mining in Russian, *Computational Science - ICCS 2020* (2020), 591-602.
[3] A. Bouziane, D. Bouchiha, N. Doumi, Annotating Arabic Texts with Linked Data, *The 4th International Symposium on Informatics and its Applications (ISIA)* **4** (2020), 1-5. Available from: http://dx.doi.org/10.1109/ISIA51297.2020.9416543
[4] A. Sysoev, I. A. Andrianov, Named Entity Recognition in Russian: the Power of Wiki-Based Approach,
[5] *Proceedings of the International Conference "Dialogue 2016"* (2016).
[6] J. Raiman, O. Raiman, DeepType: Multilingual Entity Linking by Neural Type System Evolution, *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)* (2018).
[7] Y. Kuratov, M. Arkhipov, Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language, *Proceedings of the International Conference "Dialogue 2019"* (2019).
[8] A. Grover, J. Leskovec, node2vec: Scalable Feature Learning for Networks, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), 855–864.
[9] Available from: https://doi.org/10.1145/2939672.2939754
[10] M. Kim, S. Baek, M. Song, Relation extraction for biological pathway construction using node2vec, *BMC Bioinformatics* **19** (2018). Available from: https://doi.org/10.1186/s12859-018-2200-8
[11] F. Shen, S. Liu, Y. Wang, L. Wang, A. Wen, A. H. Limper, H. Liu, Constructing Node Embeddings for Human Phenotype Ontology to Assist Phenotypic Similarity Measurement, *IEEE International Conference on Healthcare Informatics Workshop (ICHI-W)* (2018), 29-33.
[12] M. Christou, M. Crochemore, T. Flouri, C. Iliopoulos, J. Janoušek, B. Melichar, S. Pissis, Computing all subtree repeats in ordered ranked trees, *String processing and information retrieval. Lecture Notes in Computer Science* **7024** (2011), 338–343.
[13] H. Turki, T. Shafee, M. A. H. Taieb, M. B. Aouicha, D. Vrandečić, D. Das, H. Hamdi, Wikidata: A large-scale collaborative ontological medical database, *Journal of Biomedical Informatics* **99** (2019). Available from: https://doi.org/10.1016/j.jbi.2019.103292