Health Informatics Vision: From Data via Information to Knowledge
J. Mantas et al. (Eds.)
IOS Press, 2019

doi:10.3233/SHTI190051

An Information Extraction Algorithm for Detecting Adverse Events in Neurosurgery Using Documents Written in a Natural Rich-in-Morphology Language

Gleb DANILOV[a,1], Michael SHIFRIN[a], Uliya STRUNINA[a], Tatyana PRONKINA[a] and Alexander POTAPOV[a]
[a] *National Medical Research Center for Neurosurgery named after N.N. Burdenko, Moscow, Russian Federation*

Abstract. Rich-in-morphology language, such as Russian, present a challenge for extraction of professional medical information. In this paper, we report on our solution to identify adverse events (complications) in neurosurgery based on natural language processing and professional medical judgment. The algorithm we proposed is easily implemented and feasible in a broad spectrum of clinical studies.

Keywords. Electronic Health Records, Neurosurgery, Natural Language Processing, Adverse Events

1. Introduction

Electronic Health Records (EHR) contain a lot of unstructured data significantly challenging for certain information extraction. The processing of medical records may potentially be enhanced using computer automation. However, this task is more difficult when processing rich-in-morphology language, such as Russian. The identification and classification of postoperative complications in neurosurgery is a topical non-resolved issue that may take advantage of text mining[1]–[3]. In this paper, we report on our solution to identify adverse events (complications) in neurosurgery based on natural language processing and professional medical judgment.

2. Methods

N.N. Burdenko National Medical Research Center for Neurosurgery is one of the largest neurosurgical facilities in the world with the electronic database containing data in structured and unstructured formats [4]. In our research the text data from EHRs of N.N. Burdenko Neurosurgery Center related to 77,865 patients underwent 104,489 operations and discharged between 2000 and 2017 were processed [5].

[1] Corresponding Author, Gleb Danilov, National Medical Research Center for Neurosurgery named after N.N. Burdenko, Moscow, Russian Federation; E-mail: glebda@yandex.ru.

Below we propose an algorithm for explicit identification of in-hospital adverse events in medical records which implies seven essential steps:

1) Source documents selection to build the general corpus of medical texts
2) Text preprocessing and word tokenization
3) Stratification of tokens by concepts (entities)
4) Selecting the concepts (entities) potentially related to adverse events
5) Mapping the tokens of adverse events concepts to n-grams containing them
6) Labeling n-grams containing the tokens related to adverse events concepts
7) Mapping the labeled n-grams into documents containing them to conclude on adverse events identification

The idea of this method is to preselect a lexicon which could be potentially utilized in reporting of a specific adverse event in EHRs and explore short word sequences (n-grams) containing the lexicon items. This straightforward technique also enables the identification of all spectrum of adverse events that were specified by doctors.

All the data were initially extracted from the relational database of EHRs. The data processing and analysis was done with R programming environment (version 3.5.0) in RStudio IDE for MacOS (version 1.1.453) using *tidyr, dplyr, tidytext, tm, readr, stringr, stringdist, SnowballC, shiny, ggplot2* packages. We briefly describe each step of the proposed algorithm and our experience with it below.

2.1. Shaping the corpus from source documents

Each table in the relational database is queried and screened by the data manager to select the text data fields. This process is augmented with the R script iteratively.

2.2. Text preprocessing and word tokenization

The next step is done to preprocess the text before constructing a comprehensive dictionary with the following:

- Transformation to lower case
- Removing all the characters and symbols from the texts except for letters and single spaces
- Tokenization with a space separator
- Removing “stop-words” and meaningless words (single letters, artifacts, etc.)
- Excluding words found less than N times (seldom) in all texts
- This step could be tuned depending on the task specifications not to lose valuable information in a final lexicon, simultaneously not inflating it.

2.3. Stratification of tokens by concepts (entities)

In this step the tokens are normalized, so that different forms of a word are mapped to the “initial” form based on a common root. The step consists of several following items:

- Lemmatization
- Clustering (matching) the lemmas beginning with the same letters and/or sharing the same root using Damerau-Levenshtein distance
- Naming each cluster by its most frequent lemma and labeling the corresponding initial tokens with the cluster name

- Checking if the lemmas and the corresponding class labels have the same root using stemming
- Manually checking and correcting the labels if they have different root with initial tokens
- Manually checking and modifying the unique labels to name certain concepts (entities) properly

At the end of this step we have a dictionary with each word from the raw text mapped to a unique class by the common root (**Table 1**). The lemmatization appears to be more feasible than stemming to create interpretable normalized forms in morphologically complex Russian language. We found "mystem" lemmatizer by Yandex company (Russian Federation) most effective for this purpose. The Porter stemmer was applied for stemming.

Table 1. The example of morphological forms and typos for the word "status" in Russian. The English translation is given in brackets.

	Tokens in Russian (translation into English)	Counts	Class label
1	состояние (status)	596,807	status
2	состоянии (in status)	54,199	status
3	состояния (of status)	9,930	status
4	состояний (statuses)	2,655	status
5	состояние (statuc)	2,305	status
6	состоянию (to status)	1,310	status
7	состяние (staus)	1,249	status
8	состоянеи (statsu)	1,020	status
9	состяоние (stauts)	973	status
10	состоянием (by status)	874	status
…	…	…	…

2.4. Selecting the concepts (entities) potentially related to adverse events

This step is accomplished by browsing through the list of unique classes (now we call them "concepts") to identify all or searching for specific related to adverse events. This procedure is done and agreed by several experts (doctors) independently and is supported by a user-friendly Shiny application.

2.5. Mapping the tokens of a concept to n-grams containing them

When all the single-word concepts related to the adverse event(s) are selected and agreed, the tokens belonging to these concepts are found in short word sequences (n-grams). So, the text is tokenized into n-grams with n typically set to 3,4 or 5.

2.6. Labeling n-grams containing the tokens from a concept

N-grams, selected at the previous step and indicating the presence of complications, are labeled accordingly. This step should be worked out and agreed by several doctors independently facilitated with a special Shiny application.

2.7. Mapping the labeled n-grams into documents containing them

That is the final automatic procedure to judge whether medical cases are presented with adverse events if the previously selected n-grams are reflected in related documents.

3. Preliminary results of algorithm implementation

The initial corpus of ~13 million texts retrieved from EHR was split into 167,955 unique tokens occurred more than five times. The algorithm (with edit distance set to 3) produced a list of 29,743 unique concepts matching the tokens. The validation of the classification was suggested for 95,364 (57%) tokens with the roots not precisely equal to the roots of the corresponding concepts proposed by the algorithm. The typos in tokens less than seven characters in length, as well as the application of the fixed edit distance for matching words independent of their lengths, were the main reasons. Based on preliminary results we expect the necessity of classification result correction with a special Shiny application in ~20-30% of all tokens. The corrected list of concepts will be mapped into SNOMED CT (research license) and processed by doctors. The evaluation and fine-tuning of the algorithm are ongoing.

4. Discussion (on further improvement)

This algorithm can be modified in step 2.3, e.g., substituting lemmas sharing a common root by most frequent lemma and then grouping the substituted lemmas with edit distance. The edit distance may be applied dependent on word lengths. The concepts appeared in similar contexts may be arranged in larger classes using word embeddings (e.g., word2vec) [6]. The labeled texts will be used for machine learning to identify complications in neurosurgery as a part of our project in artificial intelligence [3].

5. Conclusions

The algorithm we proposed to extract entities from unstructured medical records explicitly, could be easily implemented and is feasible in a broad spectrum of clinical studies. *This project was supported by the Russian Foundation for Basic Research (grant 18-29-22085).*

References

[1] F.A. Landriel Ibanez et.al., A new classification of complications in neurosurgery, *World Neurosurgery* **75** (2011), 709–711.

[2] M. Simmons, A. Singhal and Z. Lu, Text Mining for Precision Medicine: Bringing Structure to EHRs and Biomedical Literature to Understand Genes and Health, *Adv. Exp. Med. Biol.* **939** (2016), 139–166.

[3] G. Danilov, K. Kotik, M. Shifrin, U. Strunina and T. Pronkina, Prediction of postoperative hospital stay with deep learning based on 101 654 operative reports in neurosurgery, *Stud. Health Technol. Inform.* **258** (2019), 125–129.

[4] A. Potapov, L. Likhterman and G. Danilov, Great Hospitals of the Russian Federation: National Medical Research Center for Neurosurgery Named After N. N. Burdenko: History and Contemporaneity, *World Neurosurg*. **120** (2018), 100–111.

[5] M.A. Shifrin, E.E. Kalinina and E.D. Kalinin, A sustainability view on the EPR system of N.N. Burdenko Neurosurgical Institute, *Stud. Health Technol. Inform.* **129** (2007), 1214–1216.

[6] M. Topaz et al., Mining fall-related information in clinical notes: Comparison of rule-based and novel word embedding-based machine learning approaches, *J. Biomed. Inform.* **90** (2019), 103.