E. Tutubalina[1,2]

TEXT MINING IN BIOMEDICAL RESEARCH

[1] Chemoinformatics and Molecular Modeling Laboratory,
Kazan (Volga Region) Federal University, Kremlyovskaya Str., 18,
Kazan, Russia;
elvtutubalina@kpfu.ru

Modern biomedical studies increasingly use nonstandard sources of information to obtain new data related to medical conditions, efficiency of drugs, their adverse effects, interactions between different genes/proteins, chemicals, drugs, and so on. Since academic publications and health records are primarily written in text, natural language processing (NLP) becomes increasingly important in biomedical research. Moreover, another source of information can be provided by the drug users themselves, in the form of free-text web reviews, social media posts, and other user-generated texts. These sources have been successfully used, for instance, to monitor adverse drug reactions (ADRs), making it possible to detect rare and underestimated ADRs through the users complaining about their health on social networks or specialized forums [1].

In the first part of my talk, I will describe basic components of a NLP system for various kinds of biomedical applications including preprocessing and feature extraction for supervised methods. In particular, I will focus on recent advances in distributed word representations (also called word embeddings or continuous space representation of words) that have also been applied to numerous NLP problems [2, 3]. Distributed word representations are models that map each word to a Euclidean space, attempting to capture semantic relationships between the words as geometric relationships in the Euclidean space. Word embedding models represent each word using a single real-valued vector. Word representations trained on a large collection of raw texts further employed by a machine learning method (as features) and deep neural networks (as input layer).

In the second part of the talk, I will bring attention to a number of task and community challenges where NLP-based methods help find relevant information in literature and social media data [4]. In particular, I will describe two essential areas of research: named entity recognition (NER) and text classification. The goal of NER is to extract important biological entities such as genes, proteins, drugs and clinical entities such as diseases, medical problems, tests, treatments. All entities are further mapped to the medical concepts to allow an integration between patient experiences and biomedical databases. The goal of text classification is to automatically determine whether a document contains relevant information or a target of interest. For example, this field includes models for determining smoking status, emotional stability, mood, for classifying drug reactions and health-related issues, for monitoring health discussions and so on. In the task, I will especially focus on extraction of disease-related entities from user reviews.

1. Martınez P. et al. *Computers in Industry*, 2016, **78**: 43–56.
2. Mikolov T. et al. *CoRR*, 2013, abs/1301.3781.
3. Mikolov T. et al. CoRR, 2013, abs/1310.4546.
4. Huang C.C. et al. *Briefings in bioinformatics*, 2016, **17**, **1:** 132-144.

The work is supported by the Russian Science Foundation grant no. 15-11-10019 and performed according to the Russian Government Program of Competitive Growth of Kazan Federal University