Journal of Biomedical Informatics 84 (2018) 93–102

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Medical concept normalization in social media posts with recurrent neural networks

Elena Tutubalina[a,d,*], Zulfat Miftahutdinov[a], Sergey Nikolenko[a,b,c], Valentin Malykh[e,b]

[a] Kazan Federal University, 18 Kremlyovskaya street, Kazan 420008, Russian Federation
[b] St. Petersburg Department of the Steklov Mathematical Institute, 27 Fontanka, St. Petersburg 191023, Russian Federation
[c] Neuromation OU, Tallinn 10111, Estonia
[d] Insilico Medicine, Baltimore, MD 21218, United States
[e] Neural Systems and Deep Learning Laboratory, Moscow Institute of Physics and Technology, 9 bld. 7 Instituski per., Dolgoprudny 141700, Russian Federation

ARTICLE INFO

Keywords:
Natural language processing
Information extraction
Medical concept normalization
Recurrent neural networks
User reviews
Social media

ABSTRACT

Text mining of scientific libraries and social media has already proven itself as a reliable tool for drug repurposing and hypothesis generation. The task of mapping a disease mention to a concept in a controlled vocabulary, typically to the standard thesaurus in the Unified Medical Language System (UMLS), is known as medical concept normalization. This task is challenging due to the differences in the use of medical terminology between health care professionals and social media texts coming from the lay public. To bridge this gap, we use sequence learning with recurrent neural networks and semantic representation of one- or multi-word expressions: we develop end-to-end architectures directly tailored to the task, including bidirectional Long Short-Term Memory, Gated Recurrent Units with an attention mechanism, and additional semantic similarity features based on UMLS. Our evaluation against a standard benchmark shows that recurrent neural networks improve results over an effective baseline for classification based on convolutional neural networks. A qualitative examination of mentions discovered in a dataset of user reviews collected from popular online health information platforms as well as a quantitative evaluation both show improvements in the semantic representation of health-related expressions in social media.

1. Introduction

Recent years have seen many new applications of natural language processing (NLP) to biomedical information. Much of this work has been focused on the central task of information extraction, in particular *named entity recognition* (NER) from scientific literature and electronic health records. However, comparatively little work has been carried out to automatically process social media comments of individuals undergoing medical treatment.

Social media nowadays is a virtually inexhaustible source of people's opinions on a wide variety of topics. In this work, we are focusing on patients' opinions on drug effects, i.e., patient reports. In effect, social media provide huge datasets of people's opinions complete with demographic information and often much more detailed data regarding a specific user. We expect that continuous advancement and improvement in the accuracy of text mining approaches applied to patient reports in social media will have a significant impact in several areas including pharmacovigilance (especially for new drugs), drug repurposing, and understanding drug effects in the context of other factors such as concurrent use of other drugs, diet, and lifestyle.

In this work, we study the problem of discovering disease-related medical concepts from patients' comments on social media. In the context of this problem, we translate a text written in "social media language" (e.g., "I can't fall asleep all night" or "head spinning a little") to "formal medical language" (e.g., "insomnia" and "dizziness", respectively). This goes beyond simple matching of natural language expressions and vocabulary elements: string matching approaches are not able to link social media language to medical concepts since the words may not overlap at all. We call the task of mapping everyday life language to medical terminology *medical concept normalization*. The main benefit of solving this task is bridging the gap between the language of the lay public and medical professionals.

This task is difficult given that patients use social media to discuss different concepts of illness (ranging from well-defined conditions such as "major depressive disorder" to informal phrases describing specific symptoms such as "woke up too early" or "mucus building up in my

* Corresponding author at: Kazan Federal University, 18 Kremlyovskaya street, Kazan 420008, Russian Federation.
E-mail addresses: ElVTutubalina@kpfu.ru (E. Tutubalina), zulfatmi@gmail.com (Z. Miftahutdinov), sergey@logic.pdmi.ras.ru (S. Nikolenko), valentin.malykh@phystech.edu (V. Malykh).

https://doi.org/10.1016/j.jbi.2018.06.006
Received 30 October 2017; Received in revised form 24 April 2018; Accepted 10 June 2018
Available online 12 June 2018
1532-0464/

lungs") and a wide diversity of drug reactions (e.g., "excessive sweating at night", "slept like a baby", or "clearing up an infection"). Moreover, social network data usually contains a lot of noise, such as misspelled words, incorrect grammar, hashtags, abbreviations, and different variations of the same word.

Formally, this task is related to several NLP challenges, including paraphrase detection, word sense disambiguation, and entity linking where an entity mention is mapped to a unique concept in an ontology after solving the disambiguation problem [1,2]. To address the challenges described above, recent studies treat the task of linking a one- or multi-word expression to a knowledge base as a supervised sequence labeling problem. Miftahutdinov and Tutubalina [3] proposed an encoder-decoder model based on bidirectional recurrent neural networks (RNNs) to translate a sequence of words from a death certificate into a sequence of medical codes. Two research groups [4,5] presented two systems with similar performances that utilize RNNs for normalization of tweets' phrases at the AMIA 2017 Social Media Mining for Health Applications workshop, while Limsopatham and Collier [6] experimented with convolutional neural networks (CNNs) on social media data. These works demonstrate the first attempts to use deep learning methods for medical concept normalization.

2. Background

Automatic extraction of health-related information from social media is a strong trend in related research nowadays. This task provides a challenging and rich context to explore computational models of natural language, motivating new research in computer science and computational linguistics. For an excellent overview of the work on social analytics for healthcare done up to 2015, see [7], which demonstrates how social media data can be used to mine health-related knowledge.

There exist many applications where a system needs to mediate between natural language expressions and elements of a vocabulary in an ontology. Huang and Lu [8] survey the work done in the organization of biomedical NLP (BioNLP) challenge evaluations up to 2014.

In this section, we give an overview of major findings in previous research on terminology association. In biology, a common task is to identify gene and protein names in text and link them to standard sources such as *Entrez Gene*. Biomedical researchers have addressed the needs to automatically detect diseases as well as corresponding acronyms and abbreviations in the scientific literature (e.g., *BioCreative V* lab). Recent open challenge evaluations have also focused on named entity recognition (NER) of disease names in clinical notes (e.g., *ShARe/CLEF eHealth*, *SemEval 2014*). Ontologies of medical concepts such as the Unified Medical Language System (UMLS) [9], the Systematized Nomenclature Of Medicine Clinical Terms (SNOMED-CT) [10], the Medical Dictionary for Regulatory Activities (MedDRA) [11], and the International Classification of Diseases (ICD-9, ICD-10) are widely used for this task. We note that there is no agreed definition of a disease in general, and diseases may be classified differently by different clinicians [12]. UMLS is undoubtedly the largest lexico-semantic resource for medicine which represented over more than 150 lexicons with terms from 25 languages. In particular, ICD and SNOMED-CT are subsets of UMLS. Every concept is represented by its Concept Unique Identifier (CUI). UMLS has integrated resources used worldwide in clinical care, public health, and epidemiology.

Automatic approaches to BioNLP tasks roughly fall into two categories: (i) linguistic approaches based on dictionaries, association measures, morphological and syntactic properties of texts [13–17]; (ii) machine learning approaches [18–21,6,3]. Recent studies have employed deep learning models such convolutional neural networks [6,5] or recurrent neural network architectures [3,5,4].

In the rest of the section, we describe methods that were trained on publicly available data of different health-related sources and related to the shared tasks such as Social Media Mining shared tasks, *CLEF eHealth* tasks, and *SemEval* tasks.

2.1. Bio- and medical natural language processing

A lot of work in bio- and medical NLP has been focused on language evaluation, information retrieval, and extraction from electronic medical records and biomedical academic literature. In their book, Cohen and Demner-Fushman [22] gave an overview of major challenges and the work done in biomedical NLP up to 2014.

The most popular knowledge-based system for mapping texts to UMLS concept identifiers (CUI) is MetaMap [13]. MetaMap was developed by the National Library of Medicine (NLM) in 2001 and has become a de facto baseline method for many recent studies. This system is based on a linguistic approach using lexical lookup and variants by associating a score with phrases in a sentence. General limitations of the linguistic method include low recall of information extraction from social media and unavailability for under-resourced natural languages.

The *ShARe/CLEF eHealth 2013* lab addressed the problem of identification and normalization of disorders from clinical reports in Task 1 [23]. The corpus consists of discharge summaries and electrocardiogram, echocardiogram, and radiology reports received from US intensive care. Each disorder mention is mapped to a UMLS code or a SNOMED-CT code. The best results were achieved with a DNorm system by NCBI team [18]. Leaman et al. introduced a DNorm system for assigning disease mentions from PubMed abstracts a unique identifier from a MEDIC vocabulary, which combines terminology from Medical Subject Headings (MeSH) and Online Mendelian Inheritance in Man (OMIM) [19]. DNorm consists of a text processing pipeline, including the named entity recognizer to locate diseases in the text, and a normalization method. The normalization method is based on a pairwise learning-to-rank technique using the tokens from all mentions as features. DNorm outperformed MetaMap as the baseline.

The *SemEval 2014* lab addressed the problem of analysis of clinical data in Task 7 [1]. This task was a follow-up to the ShARe/CLEF eHealth 2013 task 1 using a larger test set and a larger set of unlabeled MIMIC notes to inform models and generalize lexical features. The best results were obtained by UTH_CCB and UWM teams [17,16]. In [16], the UWM team present a pattern-based system that consists of several steps. First, the system looks for exact matches with disorder mentions in the training data and in the UMLS. Second, for every mention without exact matches, suitable variations were generated based on Levenshtein distances between the variations. Edit distance patterns were computed between all synonyms of the disorder concepts is UMLS as well as between their mentions in the training data. In [17], the UTH_CCB team used the cosine similarity scores between disorder entity and all UMLS terms to rank candidate terms.

The *CLEF Health* 2016 and 2017 labs addressed the problem of mapping death certificates to ICD codes. Death certificates are standardized documents filled by physicians to report the death of a patient [24]. Most submitted methods utilized dictionary-based semantic similarity and, to some extent, string matching. Mulligen et al. [14] obtained the best results in task 2 by combining a Solr tagger with ICD-10 terminologies. The terminologies were derived from the task training set and a manually curated ICD-10 dictionary. They achieved an F-measure of 84.8%. Mottin et al. [15] used pattern matching approach and achieved the F-measure of 55.4%. Dermouche et al. [20] applied two machine learning methods: (i) a supervised extension of Latent Dirichlet Allocation (LDA), i.e., Labeled-LDA and (ii) Support Vector Machine (SVM) based on bag-of-words features. For Labeled-LDA, they used ICD-10 codes from the training set as classes. The Labeled-LDA and SVM classifier achieved F-measures of 73.53% and 75.19%, respectively. This study did not focus on designing effective features to obtain better classification performance. Zweigenbaum and Lavergne [25] utilized a hybrid method combining simple dictionary projection and mono-label supervised classification. They used Linear SVM trained on the full training corpus and the 2012 dictionary provided for CLEF participants. This hybrid method obtained an F-measure of 85.86%. The participants of task 2 did not use word embeddings or deep neural