Journal of Biomedical Informatics xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

# Counting trees in Random Forests: Predicting symptom severity in psychiatric intake reports

Elyne Scheurwegs [a,b,c,*], Madhumita Sushil [a,c], Stéphan Tulkens [a], Walter Daelemans [a], Kim Luyckx [c]

[a] University of Antwerp, Computational Linguistics and Psycholinguistics (CLiPS) Research Center, Lange Winkelstraat 40-42, B-2000 Antwerp, Belgium
[b] University of Antwerp, Advanced Database Research and Modelling Research Group (ADReM), Middelheimlaan 1, B-2020 Antwerp, Belgium
[c] Antwerp University Hospital, ICT Department, Wilrijkstraat 10, B-2650 Edegem, Belgium

ARTICLE INFO

*Article history:*
Received 1 February 2017
Revised 31 May 2017
Accepted 5 June 2017
Available online xxxx

*Keywords:*
Symptom severity identification
Natural language processing
Concept detection
Psychiatry
Positive valence

ABSTRACT

The CEGS N-GRID 2016 Shared Task (Filannino et al., 2017) in Clinical Natural Language Processing introduces the assignment of a severity score to a psychiatric symptom, based on a psychiatric intake report. We present a method that employs the inherent interview-like structure of the report to extract relevant information from the report and generate a representation. The representation consists of a restricted set of psychiatric concepts (and the context they occur in), identified using medical concepts defined in UMLS that are directly related to the psychiatric diagnoses present in the Diagnostic and Statistical Manual of Mental Disorders, 4th Edition (DSM-IV) ontology. Random Forests provides a generalization of the extracted, case-specific features in our representation. The best variant presented here scored an inverse mean absolute error (MAE) of 80.64%. A concise concept-based representation, paired with identification of concept certainty and scope (family, patient), shows a robust performance on the task.



## 1. Introduction

A large portion of information in hospitals is stored in the form of unstructured clinical documents. Examples are doctor's notes, daily progress reports, and intake reports. To assist decision making, relevant information in such reports needs to be extracted, for which machine learning methods are typically used. This extracted information can be used as input for several applications, such as the identification of diagnostic criteria for heart failure [1,2] or the prediction of diagnosis and procedure codes [3]. Since the effectiveness of data-driven algorithms not only depends on data size, but also on the quality of data representation, robust information extraction techniques are required [4,5]. These techniques need to extract qualitative concepts from clinical texts which tend to have misspellings and ungrammatical sentences.

### 1.1. Background

One of the basic methods of representing information in a text is as a 'bag of words', considering individual words as the primary source of meaning. Medical information, however, is often present in the form of a medical multi-word expression (mMWE) such as 'alcohol abuse'. Techniques to extract these mMWEs often use knowledge resources such as the UMLS metathesaurus to map medical concepts onto texts fragments [6].

Medical concept detection has been the target of several shared tasks. The first track of the 2010 i2b2/VA shared task [7] challenged the participants to identify medical entities and label them as a 'problem', 'treatment', or 'test'. The best-performing system used a wide range of features for this purpose, including n-grams and document-level features [8]. Their results demonstrated the beneficial impact of using semantic and syntactic features based on concepts derived from UMLS and concept mapping modules in cTAKES [9]. In the ShARe/CLEF 2013 eHealth shared task [5], the goal was to recognize mMWEs in clinical notes and normalize them to UMLS Concept Unique Identifiers (CUIs). The best-ranking system applied supervised learning techniques, representing candidates as a tf-idf weighted bag of words [10]. The detection of concepts can also be supported using distributional semantics. Concepts can then be assigned to words that have a similar semantic context to certain concepts, but are not present themselves in a lexicon [11].

Aiming for a complete and accurate representation of all concepts occurring in a text is a different focus than applying concept detection methods to a specific task (such as scoring the severity for a specific symptom). In that case, an unfiltered list of detected concepts may not be the optimal representation for a clinical text.

* Corresponding author at: University of Antwerp, Computational Linguistics and Psycholinguistics (CLiPS) Research Center, Lange Winkelstraat 40-42, B-2000 Antwerp, Belgium.
*E-mail address:* elyne.scheurwegs@uantwerpen.be (E. Scheurwegs).

http://dx.doi.org/10.1016/j.jbi.2017.06.007
1532-0464/

Different methods can be suggested to zoom in on the task-specific relevant information. Often, such mechanisms improve algorithm performance, particularly in situations where training data is limited [12,13]. This motivates us to look for relevant medical factors, yielding a more concise document representation.

The second track of the 2014 i2b2/UTHealth shared task [2] hypothesized that, to identify medical risk factors in longitudinal medical records, and in particular to assess the severity of a risk factor, identifying indicators of a disease would be more informative than identifying the actual diagnosis. If hypertension, for instance, can be managed through diet and exercise, it would be considered less severe than when it requires anticoagulants. The best performing team did not directly rely on syntactic or semantic information, but rather focused on improving the annotated training data by annotating the negation markers present in the text but absent in the annotated indicators [14]. For instance, the concept 'diabetes' was extended to 'no diabetes' if it was used in a negated context. This demonstrates that having detailed concept annotations can be useful to identify risk factors in a patient's medical record.

The detection of negation, social context, and family history is crucial for several tasks. Garcelon et al. used unstructured information within the electronic health record (EHR) to identify rare diseases [15]. Negation detection is crucial, since mentions of rare diseases are often negated, and negated symptoms, combined with other factors, can be an indicator of a rare disease. Intake reports that have an interview style also benefit from such context detection, especially because a large proportion of the questions are answered negatively, and/or often refer to patient's surroundings and family.

An (already sparse) concept-based representation of clinical text becomes more sparse with the addition of context to the feature representation. To adequately deal with an approach that uses sparse features, we need to either generalize these features, or use a classifier that is able to deal with sparsity without overfitting. Random Forests [16] is one algorithm that deals well with sparsity. The individual trees are designed to overfit on features (making very specific decisions that only account for part of the dataset), while the voting strategy later mitigates these effects (by generalizing over the decisions of multiple trees).

### 1.2. CEGS N-GRID 2016 shared task

The Research Domain criteria (RDoc), a framework established by the National Institute of Mental Health (NIMH), describe human behavior, from normal to abnormal, by means of functional constructs. The RDoC domain in focus for the task at hand is positive valence, which refers to "Systems primarily responsible for responses to positive motivational situations or contexts, such as reward seeking, consummatory behavior, and reward/habit learning" [17]. The severity of positive valence, as a symptom, describes the ability of a person to control the processes influencing these systems (e.g., addictions, obsessive compulsive disorders).

The CEGS N-GRID 2016 Shared Task focuses on classifying symptom severity in the positive valence domain based on psychiatric intake reports [18]. In this paper, we target the task of symptom severity prediction as a multi-class classification task where the different severity scores are the labels we assign to an intake report. We have developed techniques to present the intake report to a classifier in a simple, interpretable manner. We attempt to minimize the manual collection of expert knowledge for this task specifically, to allow for scalability. The approach we present strongly builds on the questionnaire structure of the documents and on the ability for Random Forests to generalize over case-specific features generated using UMLS as a source collection of ontologies for medical concept detection.

## 2. Materials and methods

### 2.1. Dataset

The dataset available from the shared task consists of a training set of 600 psychiatric intake reports and a test set of 216 reports. The training set itself has 325 reports annotated by two or more annotators, 108 reports annotated by one annotator, and 167 unannotated reports. The annotation consists of one score per report, indicating the severity of symptoms in the positive valence domain exhibited by that person. This score corresponds to four distinct categories: *absent*, *mild*, *moderate*, and *severe*, presented as discrete classes 0, 1, 2, and 3.

A psychiatric intake report documents a first interview conducted with the patient. This interview contains case-based semi-standardized questions, which can contain a short or elaborate answer. It always contains a section in which the patient explains in his own words his or her motivation for coming to the hospital. A question is often preceded by the category on which the question applies, such as 'OCD' or 'Bipolar disorder'. In Fig. 1, an excerpt of an intake report is shown.

### 2.2. Medical information extraction

Raw text needs to be processed to generate feature vectors before feeding it to a machine learning algorithm. Different levels of preprocessing are performed: text normalization, identification of crucial information like medical concepts, identification of the scope of a certain medical concept, and identification of linguistic markers - such as negation cues and family cues - that influence the factuality or context of the medical concept.

In our pipeline,[1] linebreaks are restored (e.g., 'heart burnPsychiatric') using capitalization cues. The raw texts are then processed using the cTAKES Natural Language Processing (NLP) pipeline [9] to identify sentence boundaries, token boundaries, part-of-speech tags, and syntactic chunks such as noun phrases. The resulting text is then processed further (outside of cTAKES) to extract information in the form of medical concepts with a custom concept detection algorithm.

### 2.3. Medical Multi-Word Expression (mMWE) detection

A weighted partial matching algorithm is used to detect medical Multi-Word Expressions (mMWE) in the psychiatric intake report. The individual words in each noun phrase are matched to definitions of concepts extracted from the UMLS metathesaurus [6]. These matches result in a weighted sum of the scores for all word matches, where each word is weighted according to its Inverse Document Frequency (IDF), for which each concept definition was considered a document [19]. Using IDF limits the contribution of frequent words while mapping concept definitions. A concept definition is mapped onto a noun phrase when a score of at least 80% is achieved. Multiple concepts can be assigned to a single noun phrase. For example, when mapping the noun phrase 'obsessive compulsive spectrum deviation' to the concept 'obsessive compulsive disorder', both 'obsessive' and 'compulsive' have an IDF of 2. The word 'spectrum' does not occur in the definition and as such does not contribute to the total score. 'Disorder' is a more common word, with an IDF of 1, but it is not mapped either, since 'deviation' is present in the noun phrase instead. Using the IDF scoring method, we can map the noun phrase to the concept with a certainty of 80%, since the sum of IDF of the matched words is 4, out of 5 as the total sum of IDF for all words in the definition. This

[1] Code can be found here: https://github.com/Elyne/rdocChallenge.

Please cite this article in press as: E. Scheurwegs et al., Counting trees in Random Forests: Predicting symptom severity in psychiatric intake reports, J Biomed Inform (2017), http://dx.doi.org/10.1016/j.jbi.2017.06.007