Journal of Biomedical Informatics 69 (2017) 118–127

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

# Assigning clinical codes with data-driven concept representation on Dutch clinical free text

Elyne Scheurwegs [a,b,*], Kim Luyckx [c], Léon Luyten [d], Bart Goethals [a], Walter Daelemans [b]

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

ARTICLE INFO

*Article history:*
Received 26 February 2016
Revised 6 March 2017
Accepted 7 April 2017
Available online 8 April 2017

*Keywords:*
Clinical coding
Data mining
Text mining
Unsupervised learning
International classification of diseases
Electronic health records
Distributional semantics
Word2vec

ABSTRACT

Clinical codes are used for public reporting purposes, are fundamental to determining public financing for hospitals, and form the basis for reimbursement claims to insurance providers. They are assigned to a patient stay to reflect the diagnosis and performed procedures during that stay. This paper aims to enrich algorithms for automated clinical coding by taking a data-driven approach and by using unsupervised and semi-supervised techniques for the extraction of multi-word expressions that convey a generalisable medical meaning (referred to as *concepts*). Several methods for extracting concepts from text are compared, two of which are constructed from a large unannotated corpus of clinical free text. A distributional semantic model (i.c. the word2vec skip-gram model) is used to generalize over concepts and retrieve relations between them. These methods are validated on three sets of patient stay data, in the disease areas of urology, cardiology, and gastroenterology. The datasets are in Dutch, which introduces a limitation on available concept definitions from expert-based ontologies (e.g. UMLS). The results show that when expert-based knowledge in ontologies is unavailable, concepts derived from raw clinical texts are a reliable alternative. Both concepts derived from raw clinical texts perform and concepts derived from expert-created dictionaries outperform a bag-of-words approach in clinical code assignment. Adding features based on tokens that appear in a semantically similar context has a positive influence for predicting diagnostic codes. Furthermore, the experiments indicate that a distributional semantics model can find relations between semantically related concepts in texts but also introduces erroneous and redundant relations, which can undermine clinical coding performance.



## 1. Introduction

Medical knowledge is electronically stored in a high number of complex data sources, such as electronic health records (EHRs), electronic archives, ontologies, and scientific publications [1–3]. In a modern hospital setting, clinical codes are determined based on information found in the electronic health record. These clinical codes are assigned primarily for the purpose of reporting and reimbursement from health care providers or governments. Their widespread adoption in clinical environments allows for the usage as an important and complementary factor in research applications (e.g. identifying acute venous thromboembolisms) [4]. While clinical codes are often assigned manually by a team of specialized coders, techniques that can (semi-)automatically predict these codes can lower the burden of this codification process.

Most data stored in hospitals is not annotated due to the large effort that is required from physicians to accurately annotate this data. This limits the usability of this data for supervised machine learning techniques. We investigate unsupervised and semi-supervised techniques to create appropriate text representations for use in a prediction pipeline for clinical codes. The objective of this paper is to improve automated prediction of clinical codes by (I) introducing methods that are independent of expert-created ontologies to extract these concepts from the source documents of this patient stay and (II) using a distributional semantics model to generalize and represent concepts associated with a patient stay.

* Corresponding author at: University of Antwerp, Advanced Database Research and Modelling Research Group (ADReM), Middelheimlaan 1, B-2020 Antwerp, Belgium.
*E-mail address:* elyne.scheurwegs@uantwerpen.be (E. Scheurwegs).

http://dx.doi.org/10.1016/j.jbi.2017.04.007
1532-0464/

### 1.1. Background

Adequate feature engineering is arguably one of the most important steps for any sort of machine learning task, including trying to learn from unstructured clinical documents. The feature engineering task here can be defined as the conversion of unstructured data into a structured representation suitable to make predictions. A simple strategy is to use a lexical representation by using a library of relevant tokens (words occurring in a medical dictionary), or just using the text itself as the library (e.g. bag-of-words in which all unique tokens occurring in the document are counted and directly used as a representation). Other strategies include using a syntactic representation or a semantic representation. A lexical representation can be enhanced (or filtered) with semantic and/or syntactic properties as metadata (e.g. PoS-tags).

In this work, a series of documents, associated with a patient stay, is represented by a series of extracted concepts. These concepts are in essence multi-word expressions that convey a generalisable medical meaning.

#### 1.1.1. Medical information extraction

The approach chosen to extract information from clinical free texts is largely determined by the intended purpose. This purpose can be a specific case (e.g. identifying heart failure diagnostic criteria [5]), or a generic task (e.g. extracting medication terms from clinical narratives) [6]. In the ShARe/CLEF 2013 eHealth shared task [7], entities were recognized in clinical notes and subsequently normalized to UMLS identifiers (i.e. CUI codes) [8]. The best-ranking system found entities with supervised machine learning techniques, for which candidate CUIs were represented as Bag-of-Words, weighted with their TF-IDF score [9,10].

Pathak et al. mapped structured and unstructured data onto the UMLS identifier structure for the purpose of high-throughput phenotyping [11]. This approach allows for the integration of multiple types of data sources, but is substantially dependent on the existence of predefined expert knowledge in ontologies and vocabularies. This is particularly problematic for languages with a smaller number of medical lexicons, such as Dutch.

#### 1.1.2. Distributional semantics in medical corpora

A distributional semantic model (DSM) acquires a semantic representation for tokens by looking at the surrounding tokens in a large corpus [12]. Tokens are assumed to be semantically related if they are often surrounded by similar context. Antonyms and frequently co-occurring tokens are thus also marked as semantically related (e.g. 'white' and 'black','dear and 'colleague in headings). Jonnalagadda et al. extracted medical information from clinical narratives with a Random Indexing (RI) DSM [13,14]. They retrieved semantically related tokens in an unannotated corpus of Medline abstracts with the RI model, after which they supplemented the basic features (i.e. dictionary- and pattern-matched features and Part-of-Speech tags) in a machine learning algorithm with the related tokens. Including semantically related tokens increased their achieved F-measure to 91.3% for inexact matches (an increase of 2%). Henriksson et al. similarly applied RI to enhance a medical lexicon with synonyms and abbreviations [15].

Moen et al. applied two distributional semantic models (Random Indexing and a word2vec model [16]) to retrieve care episodes that are similar to the care episode under review [17]. A care episode consisted of a free text summary. Their most successful variant modified the network creation of the word2vec skip-gram model by introducing feedback that takes the ICD-10 code assigned to the training samples into account [18]. While this method significantly improved results, it also required an ICD-code to be linked to each document used to train the word2vec method. This is often not the case with archived documents. The second best variant was the unmodified word2vec skip-gram model.

In this study, we chose to use the word2vec skip-gram model [16,19]. Word2vec is an implementation of two vector representation algorithms (CBOW and skip-gram) for tokens. These algorithms both encompass a neural network, consisting of one input layer, one hidden layer, and one output layer. The vocabulary items are mapped to each input node, and a hidden layer within the model is shaped with n nodes (with n representing the number of dimensions requested), with each node representing one dimension of the desired vector. The models are then trained by presenting them with each vocabulary item and the context in which it occurs. This process is repeated until the network converges to a predetermined output error. A trained model provides a multi-dimensional space in which each word and/or token is represented by an individual dense vector with a relatively low number of dimensions.

#### 1.1.3. Automated clinical coding

Current automated clinical coding approaches are often used in controlled environments, with strongly normalized data and a limited scope in document type (e.g. radiology reports) and disease area (e.g. oncology) [20]. The most successful approaches are (partially) handcrafted, which renders them harder to port to different languages or medical specialties [21]. Perotte et al. predicted 5030 unique ICD-9-CM codes on discharge files from the MIMIC-II dataset by exploiting the ICD-9-CM hierarchy, with a resulting F-measure of 39% [22,23]. Scheurwegs et al. integrated structured and unstructured data sources to assign clinical codes to patient stays for multiple specialties [24]. They confirmed the large difference in achieved F-measure between specialties and presented a technique that is portable over medical specialties. The texts in the discharge files of the latter approaches was represented with a Bag of Words (BoW) approach, rendering the approach more portable over different languages.

This paper aims to show the feasibility of using unsupervised methods for representing unstructured data in automated clinical coding approaches. Unsupervised methods are used to both detect concepts in texts and represent those concepts in a dense vector space. The proposed methods are mainly dependent on unannotated resources (raw text) instead of on annotated resources (such as ontologies, hand-crafted rules for information extraction, and annotated training data) and are thus easily deployable on languages with limited coverage in ontologies. These methods are evaluated on a medical dataset in Dutch.

## 2. Materials and methods

### 2.1. Dataset

Two types of datasets are derived from the clinical data warehouse at the Antwerp University Hospital. The first dataset consists of an unannotated corpus of 2,374,723 automatically de-identified texts (with an average of 152 words per text). This data in this corpus is essentially raw text and covers multiple medical specialties. The second dataset consists of a randomized subset of anonymized patient stays with associated documents (radiology reports, requests, surgery reports, notes, letters, and attestations) and ICD-9-CM codes [18]. This dataset is divided into three specialties (i.e. cardiology, gastroenterology, and urology, with respectively 10,000, 7440, and 3440 patient stays). In Table 1, we show the total number of texts, patient stays and the properties of both diagnostic and procedural codes in each dataset.

The task is defined as predicting all clinical codes (i.e. procedural codes, primary as well as secondary diagnosis codes) associated