Clustering of Medical Free-Text Records Based on Word Embeddings

Adam Gabriel Dobrakowski  
University of Warsaw  
ad359226@students.mimuw.edu.pl

Agnieszka Mykowiecka  
Institute of Computer Science  
Polish Academy of Sciences  
agn@ipipan.waw.pl

Małgorzata Marciniak  
Institute of Computer Science  
Polish Academy of Sciences  
mm@ipipan.waw.pl

Wojciech Jaworski  
University of Warsaw  
wjaworski@mimuw.edu.pl

Przemysław Biecek  
University of Warsaw  
P.Biecek@mimuw.edu.pl

Abstract

Is it true that patients with similar conditions get similar diagnoses? In this paper we show NLP methods and a unique corpus of documents to validate this claim. We (1) introduce a method for representation of medical visits based on free-text descriptions recorded by doctors, (2) introduce a new method for clustering of patients’ visits and (3) present an application of the proposed method on a corpus of 100,000 visits. With the proposed method we obtained stable and separated segments of visits which were positively validated against final medical diagnoses. We show how the presented algorithm may be used to aid doctors during their practice.

1 Introduction

Processing of free-text clinical records play an important role in computer-supported medicine (Apostolova et al., 2009; Ganesan and Subotin, 2014). A detailed description of symptoms, examination and an interview is often stored in an unstructured way as free-text, hard to process but rich in important information. Some attempts of processing medical notes exist for English, while for other languages the problem is still challenging (Orosz et al., 2013).

A clustering task is easier for structured data such as age, sex, place, history of diseases, ICD-10 code etc. (an example of patients clustering based only on their history of diseases is Ruffini et al., 2017), but it is far from being solved for unstructured free-texts.

The clustering of visits has many applications. If we are able to group visits into clusters based on interview with a patient and medical examination then (1) we can follow recommendations that were applied to patients with similar visits in the past to create a list of possible diagnoses, (2) reveal that the current diagnosis is unusual, (3) identify subsets of visits with the same diagnosis but different symptoms.

A desired goal in the clustering is to divide patients into groups with similar properties. In the case of clustering hospitalized patients one of the most well-known examples are Diagnosis Related Groups (Fetter et al., 1980) which aim to divide patients into groups with similar costs of treatment. Grouping visits of patients in health centers is a different issue. Here most of the information is unstructured and included in the visit’s description written by a doctor: the description of the interview with the patient and the description of a medical examination of the patient.

In general, there are two main approaches to generate vector representations of a text. The first takes into account the occurrence and the frequency of words in the considered text. The simplest example is one-hot encoding or weighting by Term Frequency-Inverse Document Frequency (TF-IDF, Salton and Buckley, 1988). The main disadvantage of these methods is that they do not take into account the semantic similarity between words. Thus very similar texts that do not have common words but do have some synonyms can be represented by totally different vectors. This is
an especially serious problem in creating short text representations, like medical descriptions, where two random visits very often do not have any common word.

We achieved better results from the second approach which generates text representation based on words/concepts embeddings.

Medical concepts to be extracted from texts very often are taken from Unified Medical Language System (UMLS, Bodenreider, 2004), which is a commonly accepted base of biomedical terminology. Representations of medical concepts are computed based on various medical texts, like medical journals, books, etc. (Minarro-Giménez et al., 2014; De Vine et al., 2014; Newman-Griffis et al., 2017; Choi et al., 2016c; Chiu et al., 2016) or based directly on data from Electronic Health Records (Choi et al., 2016a,b,c).

An interesting algorithm for patient clustering is given in (Choi et al., 2016a). A subset of medical concepts (e.g. diagnosis, medication, procedure codes) and computed embeddings is aggregated for all visits of a patient. This way we get patient embedding that summarizes the medical history of a patient.

In this work we present a different approach. Our data contains medical records for different medical domains. This allows us to create a more comprehensive description of a patient. A second difference is that the aim is a clustering of visits, not patients. This way a single patient may belong to several clusters.

Our clustering is also based on a dictionary of medical concepts. We create our own dictionary as for Polish does not exist any classification of medical concepts like UMLS. It will be presented in details in the section Methodology.

3 corpus of free-text clinical records

The clustering method is developed and validated on a new dataset of free-text clinical records of about 100,000 visits.

The data set consists of about 100,000 patients’ visits from different primary health care centers and specialist clinics in Poland. The first part of the data contains descriptions of visits, which have a free-text form. They are written by doctors representing a wide range of medical professions, e.g. general practitioners, dermatologists, cardiologists or psychiatrists. Each description is divided into three parts: interview, examination, and recommendations.

3.1 Extraction of medical concepts

As there are no generally available terminological resources for Polish medical texts, the first step of data processing was aimed at automatic identification of the most frequently used words and phrases. The doctors’ notes are usually rather short and concise, so we assumed that all frequently appearing phrases are domain related and important for text understanding. The notes are built mostly from noun phrases so it was decided to extract simple noun phrases which consist of a noun optionally modified by a sequence of adjectives (in Polish they can occur both before and after a noun) or by another noun in the genitive. We only extracted sequences that can be interpreted as phrases in Polish, i.e. nouns and adjectives have to agree in case, number and gender.

Phrase extraction and ordering was performed by TermoPL (Marciniak et al., 2016). The program processes text which is tokenized, lemmatized and tagged with POS and morphological features values. It allows for defining a grammar describing extracted text fragments and order them according to the modified version of the C-value coefficient (Frantzi et al., 2000). We preprocessed texts using Concraft tagger (Waszczuk, 2012) and we used the standard grammar for describing noun phrases included in TermoPL.

In order to get the most common phrases, we processed 220,000 visits by TermoPL. The first 4800 phrases (with C-value equal at least 20) from the obtained list were manually annotated with semantic labels. The list of labels covered most general concepts like anatomy, feature, disease, test. It contained 137 labels. Some labels were assigned to multi-word expressions (MWEs), in some cases all or some of their elements were also labeled separately, e.g. left hand is labeled
as anatomy while hand is also labeled as anatomy and left as lateralization. The additional source of information was the list of 9993 names of medicines and dietary supplements.

The above list of terms together with their semantic labels was then converted to the format of lexical resources of Categorial Syntactic-Semantic Parser „ENIAM” (Jaworski and Kozakoszczak, 2016; Jaworski et al., 2018). The parser recognized lexemes and MWEs in visits according to the provided list of terms, then the longest sequence of recognized terms was selected, and semantic representation was created. Semantic representation of a visit has a form of a set of pairs composed of recognized terms and their labels (not recognized tokens were omitted). The average coverage of semantic representation was 82.06% of tokens and 75.38% of symbols in section Interview and 87.43% of tokens and 79.28% of symbols in section Examination.

Texts of visits are heterogeneous as they consist of: very frequent domain phrases; domain important words which are too infrequent to be at the top of the term list prepared by TermoPL; some general words which do not carry relevant information; numerical information; and words which are misspelled. In the clustering task we neglect the original text with inflected word forms and the experiment described in this paper is solely performed on the set of semantic labels attached to each interview and examination.

3.2 Embeddings for medical concepts

Three most common classic non-contextual approaches to obtain word embeddings are skip-gram, Continuous Bag of Words (two algorithms from Mikolov et al., 2013) and GloVe (Global Vectors, Pennington et al., 2014, where higher accuracy than in previous algorithms was proved). Some authors use pretrained embeddings (especially when their data set is too small to train their own embeddings) or try to modify these embeddings and adjust to their set. But the biggest drawback of these approaches is that the corpus for training embeddings can be not related to the specific task where embeddings are utilized. A lot of medical concepts are not contained in well-known embeddings bases. Furthermore, the similarity of words may vary in different contexts.

In the experiments, we reduce the description of visits to extracted concepts. Furthermore we choose only unique concepts and abandon their original order in the description. During creating the term cooccurrence matrix the whole visit’s description is treated as the neighbourhood of the concept.

Then, we compute embeddings of concepts (by GloVe) for interview descriptions and for examination descriptions separately. We compute two separate embeddings, because we want to catch the similarity between terms in their specific context, i.e. words similar in the interview may not be similar in the examination description (for example we computed that the nearest words to cough in interview descriptions was runny nose, sore throat, fever, dry cough but in examination description it was rash, sunny, laryngeal, dry cough).

3.3 Visit embeddings

The simplest way to generate text embeddings based on term embeddings is to use some kind of aggregation of term embeddings such as an average. This approach was tested for example by Banea et al. (2014) and Choi et al. (2016b). De Boom et al. (2016) computed a weighted mean of term embeddings by the construction of a loss function and training weights by the gradient descent method.

Final embeddings for visits are obtained by concatenation of average embeddings calculated separately for the interview and for the medical examination, see Figure 1.

3.4 Visits clustering

Based on Euclidean distance between vector representations of visits we applied and compared two clustering algorithms: k-means and hierarchical clustering with Ward’s method for merging clusters (Ward Jr, 1963). The similarity of these clusterings was measured by the adjusted Rand index (Rand, 1971). For the final results we chose the hierarchical clustering algorithm due to greater stability.
| Domain            | # clusters | # visits | clusters’ size            | K-means - hclust |
|-------------------|------------|---------|---------------------------|------------------|
| Cardiology        | 6          | 1201    | 428, 193, 134, 303, 27, 116 | 0.87             |
| Family medicine   | 6          | 11230   | 3108, 2353, 601, 4518, 255, 395 | 0.69             |
| Gynecology        | 4          | 3456    | 1311, 1318, 384, 443       | 0.8              |
| Internal medicine | 5          | 6419    | 1915, 1173, 1930, 1146, 255 | 0.76             |
| Psychiatry        | 5          | 1012    | 441, 184, 179, 133, 75     | 0.81             |

Table 1: The statistics of clusters for selected domains by hierarchical clustering algorithm. The last column shows adjusted Rand index between k-means and hierarchical clustering.

| Type of relationship | # Pairs | Term Pair 1          | Term Pair 2          |
|----------------------|---------|----------------------|----------------------|
| Body part – Pain     | 22      | eye                  | foot                 |
| Specialty – Adjective| 7       | dermatologist        | neurologist          |
| Body part – Right side| 34     | hand                 | knee                 |
| Body part – Left side| 32      | thumb                | heel                 |
| Spec. – Consultation | 11      | surgeon              | gynecologist         |
| Specialty - Body part| 9       | cardiologist         | oculist              |
| Man - Woman          | 9       | patient (male)       | brother              |
|                      |         |                      | eye                  |

Table 2: The categories of questions in term analogy task, total number of pairs in every category and two examples.

(a) Body part – Right side  
(b) Body part – Pain  
(c) Specialty – Adjective  
(d) Specialty – Body part

Figure 2: Visualization of analogies between terms. Pictures show term embeddings projected into 2d-plane using PCA. Each panel shows a different type of analogy.
We would like embeddings to be able to catch relationships between terms. A question in the term analogy task is computing a vector: $\text{vector}(\text{left foot}) - \text{vector}(\text{foot}) + \text{vector}(\text{hand})$ and checking if the correct $\text{vector}(\text{left hand})$ is in the neighborhood (in the metric of cosine of the angle between the vectors) of this resulting vector.

For clustering, we selected only visits where the description of recommendation and at least one of interview and examination were not empty (that is, some concepts were recognized in the text). It significantly reduced the number of considered visits.

Table 1 gives basic statistics of obtained clusters. The last column contains the adjusted Rand index. It can be interpreted as a measure of the stability of the clustering. The higher similarity of the two algorithms, the higher stability of clustering.

For determining the optimal number of clusters, for each specialty we consider the number of clusters between 2 and 15. We choose the number of clusters so that adding another cluster does not give a relevant improvement of a sum of differences between elements and clusters’ centers (according to so called Elbow method).

4 Results

The evaluation of results is based on empirical studies described in separate subsections.

4.1 Analogies in medical concepts

To better understand the structure of concept embeddings and to determine the optimal dimension of embedded vectors we use word analogy task introduced by Mikolov et al. (2013) and examined in details in a medical context by Newman-Griffis et al. (2017). In the former work the authors defined five types of semantic and nine types of syntactic relationship.

We propose our own relationships between concepts based on the fact that a lot of concepts contain the same words.
Figure 3: Clusters of visits for selected domains. Each dot correspond to a single visit. Colors correspond to segments. Visualization created with t-SNE.

| Cluster | The 5 most common recommendations in a given cluster |
|---------|-----------------------------------------------------|
| 1       | holter ekg (7.5%), continuation of treatment (7%), ekg examination (5.6%), everyday walk (5.1%), abidance of recommended diet (5.1%) |
| 2       | echo of the heart (8.8%), visit (3.1%), b12 (1%), ekg examination (0.5%), everyday walk (0.5%) |
| 3       | meal (18.7%), sugar-free diet (17.2%), physical activity (14.2%), programmable effort (14.2%), regular measurement of blood pressure (12.7%) |
| 4       | control of blood pressure (12.9%), echo of the heart (11.2%), holter ekg (9.2%), repeat control (7.6%), urgent consultation (6.6%) |
| 5       | helpline (100%), cardiomedical (100%), body weight reduction (66.7%), performing the examination (48.1%), isotonic effort (44.4%) |
| 6       | helpline (90.5%), cardiomedical (90.5%), body weight reduction (61.2%), performing the examination (39.7%), isotonic effort (36.2%) |

Table 4: The most common recommendations for each segment derived for cardiology. In brackets we present a percentage of visits in this cluster which contain a specified term. We skipped terms common in many clusters, like control, treat, medicament, morphology, next visit and therapy.
Figure 4: Correspondence analysis between clusters and doctors’ IDs for psychiatry clustering (panel a) and between clusters and ICD-10 codes for family medicine clustering (panel b). Clusters 2 and 3 in panel a are perfectly fitted to a single doctor.

Figure 5: Map of ICD-10 codes in the space of embeddings for all visits. More popular codes have larger labels.
For computing visit embeddings we chose embeddings of dimensionality 20, since this resulted in the best accuracy of the most restrictive analogy task and it allowed us to perform more efficient computations than higher dimensional representations.

Figure 2 illustrates term embeddings from four categories of analogies. Embeddings are projected using PCA method.

4.2 Visits clustering
Clustering was performed separately for each specialty of doctors. Figure 3 illustrates two-dimensional projections of visit embeddings coloured by clusters. The projections were created by t-SNE algorithm (Maaten and Hinton, 2008). For some domains clusters are very clear and separated (Figure 3a). This corresponds with the high stability of clustering measured by Rand index.

In order to validate the proposed methodology we evaluate how clear are derived segments when it comes to medical diagnoses (ICD-10 codes). No information about recommendations or diagnosis is used in the phase of clustering to prevent data leakage.

Figure 4 (b) shows correspondence analysis between clusters and ICD-10 codes for family medicine clustering. There appeared two large groups of codes: first related to diseases of the respiratory system (J) and the second related to other diseases, mainly endocrine, nutritional and metabolic diseases (E) and diseases of the circulatory system (I). The first group corresponds to Cluster 1 and the second to Cluster 4. Clusters 3, 5 and 6 (the smallest clusters in this clustering) covered Z76 ICD-10 code (encounter for issue of repeat prescription).

We also examined the distribution of doctors’ IDs in the obtained clusters. It turned out that some clusters covered almost exactly descriptions written by one doctor. This situation took place in the specialties where clusters are separated with large margins (e.g. psychiatry, pediatrics, cardiology). Figure 4 (a) shows correspondence analysis between doctors’ IDs and clusters for psychiatry clustering.

4.3 Recommendations in clusters
According to the main goal of our clustering described in Introduction, we would like to obtain similar recommendations inside every cluster. Hence we examined the frequency of occurrence of the recommendation terms in particular clusters.

We examined terms of recommendations related to one of five categories: procedure to carry out by patient, examination, treatment, diet and medicament. Table 4 shows an example of an analysis of the most common recommendations in clusters in cardiology clustering. In order to find only characteristic terms for clusters we filtered the terms which belong to one of 15 the most common terms in at least three clusters.

4.4 ICD-10 codes embeddings
Embeddings for visits can be used to generate vector representations of ICD-10 codes. For every ICD-10 code we computed an average of embeddings of all visits assigned by the doctor to this code. Figure 5 shows t-SNE visualisation of these embeddings. We can see clear groups of codes from the same categories of diseases.

5 Conclusions and applications
We proposed a new method for clustering of visits in health centers based on descriptions written by doctors. We validated this new method on a new large corpus of Polish medical records. For this corpus we identified medical concepts and created their embeddings with GloVe algorithm. The quality of the embeddings was measured by the specific analogy task designed specifically for this corpus. It turns out that analogies work well, what ensures that concept embeddings store some useful information.

Clustering was performed on visits embedding created based on word embedding. Visual and numerical examination of derived clusters showed an interesting structure among visits. As we have shown obtained segments are linked with medical diagnosis even if the information about recommendations nor diagnosis were not used for the clustering. This additionally convinces that the identified structure is related to some subgroups of medical conditions.

Obtained clustering can be used to assign new visits to already derived clusters. Based on descriptions of an interview or a description of patient examination we can identify similar visits and show corresponding recommendations.
References

E. Apostolova, D. S. Channin, D. Demner-Fushman, J. Furst, S. Lytinen, and D. Raicu. 2009. Automatic segmentation of clinical texts. In 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pages 5905–5908.

Carmen Banea, Di Chen, Rada Mihalcea, Claire Cardie, and Janyce Wiebe. 2014. Simcompass: Using deep learning word embeddings to assess cross-level similarity. In Proceedings of the 8th international workshop on semantic evaluation (semeval 2014), pages 560–565.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research, 32(suppl_1):D267–D270.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampol Pyysalo. 2016. How to train good word embeddings for biomedical nlp. In Proceedings of the 15th workshop on biomedical natural processing, pages 166–174.

Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. 2016a. Multi-layer representation learning for medical concepts. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1495–1504. ACM.

Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016b. Medical concept representation learning from electronic health records and its application on heart failure prediction. arXiv preprint arXiv:1602.03686.

Youngduck Choi, Chill Yi-I Chiu, and David Sontag. 2016c. Learning low-dimensional representations of medical concepts. AMIA Summits on Translational Science Proceedings, 2016:41.

Cedric De Boom, Steven Van Canneyt, Thomas De meester, and Bart Dhoedt. 2016. Representation learning for very short texts using weighted word embedding aggregation. Pattern Recognition Letters, 80:150–156.

Lance De Vine, Guido Zuccon, Bevan Koopman, Lurianne Sitbon, and Peter Bruza. 2014. Medical semantic similarity with a neural language model. In Proceedings of the 23rd ACM international conference on conference on information and knowledge management, pages 1819–1822. ACM.

R. Fetter, A. Novák, and G. Prószéký. 1980. Case mix definition by diagnosis related groups. volume 18, pages 1—53.

Katerina Frantzì, Sophia Ananiadou, and Hideki Mima. 2000. Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method. Int. Journal on Digital Libraries, 3:115–130.

K. Ganesan and M. Subotin. 2014. A general supervised approach to segmentation of clinical texts. In 2014 IEEE International Conference on Big Data (Big Data), pages 33–40.

Wojciech Jaworski and Jakub Kozakoszcza. 2016. Eniam: Categorial syntactic-semantic parser for polish. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, pages 243–247.

Wojciech Jaworski, Daniel Oklesiński, Jan Lupa, Szymon Rutkowski, Jakub Kozakoszcza, Joanna Przetacka, Helena Teleżyńska, Bartłomiej Antonowicz, Adam Markiewicz, Jakub Kowalewski, Michał Pieńkosz, and Agnieszka Morusiewicz. 2018. Categorial parser. CLARIN-PL digital repository.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. Journal of machine learning research, 9(Nov):2579–2605.

Malgorzata Marciniak, Agnieszka Mykowiecka, and Piotr Rychlik. 2016. TermoPL — a flexible tool for terminology extraction. In Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016, pages 2278–2284, Portoroz, Slovenia. ELRA, European Language Resources Association (ELRA).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

José Antonio Minarro-Giménez, Oscar Marin-Alonso, and Matthias Samwald. 2014. Exploring the application of deep learning techniques on medical text corpora. Studies in health technology and informatics, 205:584–588.

Denis Newman-Griffiths, Albert M Lai, and Eric Fosler-Lussier. 2017. Insights into analogy completion from the biomedical domain. arXiv preprint arXiv:1706.02241.

G. Orosz, Y. Shin, J. Freeman, and R. Averill. 2013. Hybrid text segmentation for hungarian clinical records. In Advances in Artificial Intelligence and Its Applications: MICAi 2013. Springer.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543.

William M Rand. 1971. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association, 66(336):846–850.

Matteo Ruffini, Ricard Gavalda, and Esther Limón. 2017. Clustering patients with tensor decomposition. arXiv preprint arXiv:1708.08994.
Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Jakub Waszczuk. 2012. Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 2789–2804.