Interpretable Segmentation of Medical Free-Text Records Based on Word Embeddings

Adam Gabriel Dobrakowski[1(✉)], Agnieszka Mykowiecka[2],
Małgorzata Marciniak[2], Wojciech Jaworski[1], and Przemysław Biecek[1]

[1] University of Warsaw, Banacha 2, Warsaw, Poland
ad359226@students.mimuw.edu.pl,
{W.Jaworski,P.Biecek}@mimuw.edu.pl
[2] Institute of Computer Science Polish Academy of Sciences,
Jana Kazimierza 5, Warsaw, Poland
{agn,mm}@ipipan.waw.pl

Abstract. Is it true that patients with similar conditions get similar diagnoses? In this paper we present a natural language processing (NLP) method that can be used 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 segmentation of patients' visits, (3) present an application of the proposed method on a corpus of 100,000 medical visits and (4) show tools for interpretation and exploration of derived knowledge representation. With the proposed method we obtained stable and separated segments of visits which were positively validated against medical diagnoses. We show how the presented algorithm may be used to aid doctors in their practice.

1 Introduction

Processing of free-text clinical records plays an important role in computer-supported medicine [1,13]. 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. Although there exist some attempts to process medical notes for English and some other languages, in general, the problem is still challenging [23]. The most straightforward approach to the processing of clinical notes could be their clustering with respect to different features like diagnosis or type of treatment. The process can either concentrate on patients or on their particular visits.

Grouping of visits can fulfil many potential goals. If we are able to group visits into clusters based on interview with a patient and medical examination then we can: follow recommendations that were suggested to patients with similar history to create a list of possible diagnoses; reveal that the current diagnosis is unusual; identify subsets of visits with the same diagnosis but different symptoms. A desired goal in the patients' segmentation is to divide them into groups with

similar properties. In the case of segmentation hospitalized patients one of the most well-known examples are Diagnosis Related Groups [11] 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.

Segmentation (clustering) is a well studied task for structured data such as age, sex, place, history of diseases, ICD-10 code etc. (an example of patients segmentation based only on their history of diseases is introduced in [26]), but it is far from being solved for unstructured free-texts which requires undertaking many decision on how the text and its meaning is to be represented. Medical concepts to be extracted from texts very often are taken from Unified Medical Language System (UMLS, see [4]), 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. [5,8,10,21,22] or based directly on data from Electronic Health Records [6–8]. Other approach for patient segmentation is given in [6]. A subset of medical concepts (e.g. diagnosis, medication, procedures) and embeddings is aggregated for all visits of a patient. This way we get patient embedding that summaries patient medical history.

In this work we present a different approach. Our data include medical records for the medical history, description of the examination and recommendations for the treatment. Complementary sources allow us to create a more comprehensive visit description. The second difference is grouping visits, not patients. In this way a single patient can belong to several clusters. Our segmentation is based on a dictionary of medical concepts created from data, as for Polish does not exist any classification of medical concepts like UMLS or SNOMED. Obtained segments are supplemented with several approaches to visual exploration that facilitate interpretation of segments. Some examples of visual exploration of supervised models for structured medical data are presented in [3,14,17]. In this article we deal with a problem of explainable machine learning for unsupervised models.

2   Corpus of Free-Text Clinical Records

The clustering method is developed and validated on a dataset of free-text clinical records of about 100,000 visits. The data set consists of descriptions of patients' visits from different primary health care centers and specialist clinics in Poland. They have a free-text form and 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   Methodology

In this section we describe our algorithm for visits clustering. The process is performed in the following four steps: (1) Medical concepts are extracted from free-text descriptions of an interview and examination. (2) A new representation of identified concepts is derived with concepts embedding. (3) Concept embeddings are transformed into visit embeddings. (4) Clustering is performed on visit embeddings.

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 which consist of a noun optionally modified by a sequence of adjectives or by another noun in the genitive. We only extracted sequences that can be interpreted as phrases in Polish.

To get the most common phrases, we processed 220,000 visits' descriptions. First, we preprocessed texts using Concraft tagger [28] which assigns lemmas, POS and morphological features values. It also guesses descriptions (apart from lemmas) for words which are not present in its vocabulary. Phrase extraction and ordering was performed by TermoPL [19]. The program allows for defining a grammar describing extracted text fragments and order them according to a version of the C-value coefficient [12], but we used the built-in grammar of noun phrases. The first 4800 phrases (all with C-value equal at least 20) from the obtained list were manually annotated with semantic labels. The list of 137 labels covered most general concepts like *anatomy, feature, disease, test*. Many labels were assigned to multi-word expressions (MWEs). In some cases phrases 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 list of terms together with their semantic labels was then converted to the format of lexical resources of Categorial Syntactic-Semantic Parser "ENIAM" [15,16]. The parser recognized lexemes and MWEs in texts 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

(a) Body part – Right side

(b) Body part – Pain

(c) Specialty – Adjective

(d) Specialty – Body part
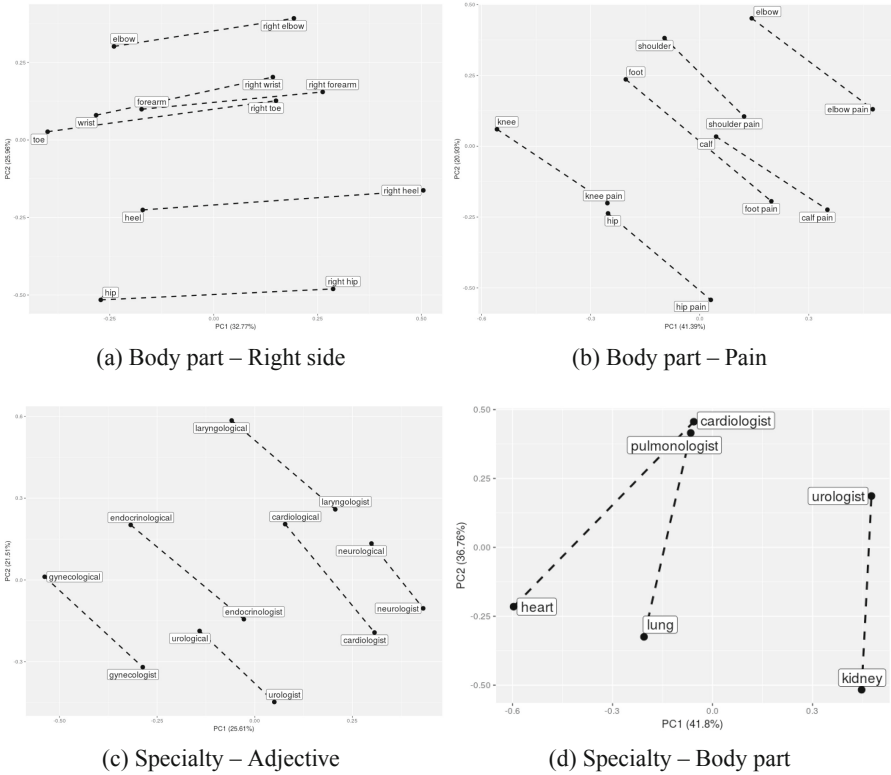
Fig. 1. Visualization of analogies between terms. The pictures show term embeddings projected into 2d-plane using PCA. Each panel shows a different type of analogy.

the clustering task we neglect the original text with inflected word forms and the experiments are solely performed on the set of semantic labels attached to each interview and examination.

3.2   Embeddings for Medical Concepts

Operating on relatively large amount of very specific texts, we decided not to use any general model for Polish. In the experiments, we reduce the description of visits to extracted concepts and train on them our own domain embeddings. During creating the term co-occurrence matrix the whole visit's description is treated as the neighbourhood of the concept. Furthermore we choose only unique concepts and abandon their original order in the description (we follow this way due to simplicity).

We compute embeddings of concepts by GloVe [24] for interview descriptions and for examination descriptions separately. Computing two separate embeddings we aim at catching the similarity between terms in their specific context. For example, the nearest words to *cough* in the interview descriptions is *runny nose*, *sore throat*, *fever* but in the examination description it is *rash*, *sunny*, *laryngeal*.

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 in [2] and [7]. In [9] the authors computed a weighted mean of term embeddings by the construction of a loss function and training weights by the gradient descent method. Thus, in our method we firstly compute embeddings of the descriptions (for interview and examination separately) as a simple average of concepts' embeddings. Then, the final embeddings for visits are obtained by concatenation of two descriptions' embeddings.

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 [27]. The similarity of these clusterings was measured by the adjusted Rand index [25]. For the final results we chose the hierarchical clustering algorithm due to greater stability.

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

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 2. The categories of questions in term analogy task with example pairs.

Type of relationship	# Pairs	Term pair 1		Term pair 2	
Body part – Pain	22	Eye	Eye pain	Foot	Foot pain
Specialty – Adjective	7	Dermatologist	Dermatological	Neurologist	Neurological
Body part – Right side	34	Hand	Right hand	Knee	Right knee
Body part – Left side	32	Thumb	Left thumb	Heel	Left heel
Spec. – Consultation	11	Surgeon	Surgical consult.	Gynecologist	g. consult.
Specialty – Body part	9	Cardiologist	Heart	Oculist	Eye
Man – Woman	9	Patient (male)	Patient (female)	Brother	Sister

Table 3. Mean accuracy of correct answers on term analogy tasks. Rows show different embeddings sizes, columns correspond to size of neighborhoods.

Dim./Context	1	3	5
10	0.1293	0.2189	0.2827
15	0.1701	0.3081	0.4123
20	**0.1702**	0.3749	0.4662
25	0.1667	0.4120	0.5220
30	0.1674	0.4675	0.5755
40	0.1460	**0.5017**	0.6070
50	0.1518	0.4966	**0.6190**
100	0.0435	0.4231	0.5483
200	0.0261	0.3058	0.4410

Table 4. The most common recommendations for each segment derived for gynecologyy. In brackets we present a percentage of visits in this cluster which contain a specified term. We skipped terms common in many clusters, like: *treatment, ultrasound treatment, control, morphology, hospital, lifestyle, zus (Social Insurance Institution)*.

Cluster	Size	Most frequent recommendations
1	1311	Recommendation (16.6%), general urine test (5.6%), diet (4.1%), vitamin (4%), dental prophylaxis (3.4%)
2	1318	Therapy (4.1%), to treat (4%), cytology (3.8%), breast ultrasound (3%), medicine (2.2%)
3	384	Acidum (31.5%), the nearest hospital (14.3%), proper diet (14.3%), health behavior (14.3%), obstetric control (10.2%)
4	443	To treat (2%), therapy (2%), vitamin (1.8%), diet (1.6%), medicine (1.6%)

For clustering, we selected visits where the description of recommendation and at least one of interview and examination were not empty (some concepts were recognized). 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

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 in [20] and examined in details in a medical context in [22]. In the former work the authors defined five types of semantic and nine types of syntactic relationship.

We propose our own relationships between concepts, more related to the medical language. We exploit the fact that in the corpus we have a lot of multiword concepts and very often the same words are included in different terms. We would like the embeddings to be able to catch relationships between terms. A question in the term analogy task is computing a vector: $vector(left\ foot) - vector(foot) + vector(hand)$ and checking if the correct $vector(left\ hand)$ is in the neighborhood (in the metric of cosine of the angle between the vectors) of this resulting vector.

We defined seven types of such semantic questions and computed answers' accuracy in a similar way as in [20]: we created manually the list of similar term pairs and then we formed the list of questions by taking all two-element subsets of the pairs list. Table 2 shows the created categories of questions.

We created one additional task, according to the observation that sometimes two different terms are related to the same object. This can be caused for example by the different order of words in the terms, e.g. *left wrist* and *wrist left* (in Polish both options are acceptable). We checked if the embeddings of such words are similar.

We computed term embeddings for terms occurring at least 5 times in the descriptions of the selected visits. The number of chosen terms in interview descriptions was equal to 3816 and in examination descriptions – 3559. Among these there were 2556 common terms for interview and examination. Embeddings of the size 10 to 200 were evaluated. For every embedding of interview terms there was measured accuracy of every of eight tasks. Table 3 shows the mean of eight task results. The second column presents the results of the most restrictive rule: a question is assumed to be correctly answered only if the closest term of the vector computed by operations on related terms is the same as the desired answer. The total number of terms in our data set (about 900,000 for interviews) was many times lower than sets examined in [20]. Furthermore, words in medical descriptions can have a different context that we expect. Taking this into account, the accuracy of about 0.17 is very high and better than we expected. We then checked the closest 3 and 5 words to the computed vector and assumed a correct answer if in this neighbourhood there was the correct vector. In the biggest neighbourhood the majority of embeddings returned accuracy higher than 0.5.

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 1 illustrates PCA projection of term embeddings from four categories of analogies.
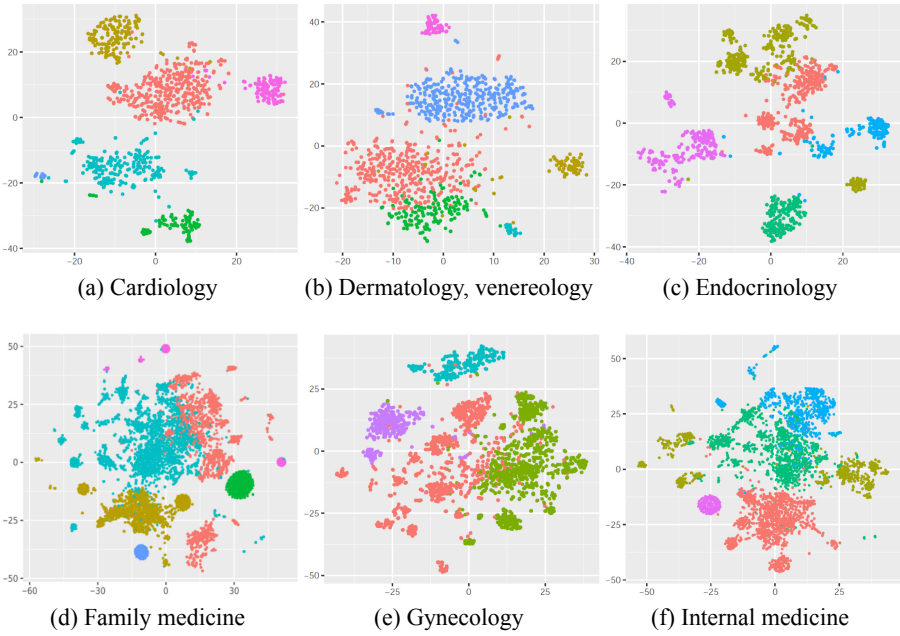
Fig. 2. Clusters of visits for selected domains. Each dot corresponds to a single visit. Colors correspond to segments. Visualization created with t-SNE. (Color figure online)
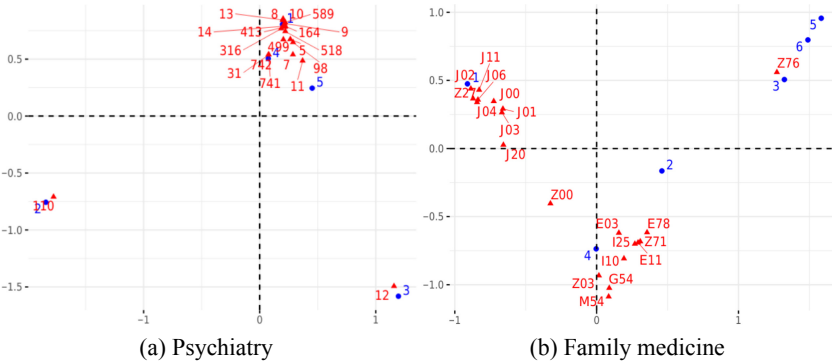
Fig. 3. Correspondence analysis between clusters and doctors' IDs for psychiatry clustering (panel a) and between clusters and ICD-10 codes for family medicine clustering (panel b). Clusters 2 and 3 in panel a are perfectly fitted to a single doctor.

4.2   Visits Clustering

Clustering was performed separately for each specialty of doctors. Figure 2 illustrates two-dimensional t-SNE projections of visit embeddings coloured by clusters [18]. For some domains clusters are very clear and separated (Fig. 2a). 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). No information about recommendations nor diagnosis is used in the phase of clustering to prevent data leakage.

Figure 3(b) shows correspondence analysis between clusters and ICD-10 codes for family medicine clustering. There appeared two large groups of codes: the 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 happen in the specialties where clusters are separated with large margins (e.g. psychiatry, pediatrics, cardiology). Figure 3(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 gynecology 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.

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 or 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

1. Apostolova, E., Channin, D.S., Demner-Fushman, D., Furst, J., Lytinen, S., Raicu, D.: Automatic segmentation of clinical texts. In: Proceedings of EMBC, pp. 5905–5908 (2009)
2. Banea, C., Chen, D., Mihalcea, R., Cardie, C., Wiebe, J.: Simcompass: using deep learning word embeddings to assess cross-level similarity. In: Proceedings of SemEval, pp. 560–565 (2014)
3. Biecek, P.: DALEX: explainers for complex predictive models in R. J. Mach. Learn. Res. **19**(84), 1–5 (2018)
4. Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. **32**(suppl-1), D267–D270 (2004)
5. Chiu, B., Crichton, G., Korhonen, A., Pyysalo, S.: How to train good word embeddings for biomedical NLP. In: Proceedings of BioNLP, pp. 166–174 (2016)
6. Choi, E., et al.: Multi-layer representation learning for medical concepts. In: SIGKDD Proceedings, pp. 1495–1504. ACM (2016)
7. Choi, E., Schuetz, A., Stewart, W.F., Sun, J.: Medical concept representation learning from electronic health records and its application on heart failure prediction. arXiv preprint arXiv:1602.03686 (2016)
8. Choi, Y., Chiu, C.Y.I., Sontag, D.: Learning low-dimensional representations of medical concepts. AMIA Summits Transl. Sci. **2016**, 41 (2016)
9. De Boom, C., Van Canneyt, S., Demeester, T., Dhoedt, B.: Representation learning for very short texts using weighted word embedding aggregation. Pattern Recogn. Lett. **80**, 150–156 (2016)
10. De Vine, L., Zuccon, G., Koopman, B., Sitbon, L., Bruza, P.: Medical semantic similarity with a neural language model. In: Proceedings of CIKM, pp. 1819–1822. ACM (2014)
11. Fetter, R.B., Shin, Y., Freeman, J.L., Averill, R.F., Thompson, J.D.: Case mix definition by diagnosis-related groups. Med. Care **18**(2), i-53 (1980)
12. Frantzi, K., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms: the C-value/NC-value method. Int. J. Digit. Libr. **3**, 115–130 (2000)
13. Ganesan, K., Subotin, M.: A general supervised approach to segmentation of clinical texts. In: IEEE International Conference on Big Data, pp. 33–40 (2014)
14. Gordon, L., Grantcharov, T., Rudzicz, F.: Explainable artificial intelligence for safe intraoperative decision support. JAMA Surg. **154**(11), 1064–1065 (2019)

15. Jaworski, W., Kozakoszczak, J.: ENIAM: categorial syntactic-semantic parser for Polish. In: Proceedings of COLING, pp. 243–247 (2016)
16. Jaworski, W., et al.: Categorial parser. CLARIN-PL digital repository (2018)
17. Kobylińska, K., Mikołajczyk, T., Adamek, M., Orłowski, T., Biecek, P.: Explainable machine learning for modeling of early postoperative mortality in lung cancer. In: Marcos, M., et al. (eds.) KR4HC/TEAAM -2019. LNCS (LNAI), vol. 11979, pp. 161–174. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-37446-4_13
18. Maaten, L.V.D., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(Nov), 2579–2605 (2008)
19. Marciniak, M., Mykowiecka, A., Rychlik, P.: TermoPL – a flexible tool for terminology extraction. In: Proceedings of LREC, pp. 2278–2284. ELRA, Portorož, Slovenia (2016)
20. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
21. Minarro-Giménez, J.A., Marin-Alonso, O., Samwald, M.: Exploring the application of deep learning techniques on medical text corpora. Stud. Health Technol. Inform. **205**, 584–588 (2014)
22. Newman-Griffis, D., Lai, A.M., Fosler-Lussier, E.: Insights into analogy completion from the biomedical domain. arXiv preprint arXiv:1706.02241 (2017)
23. Orosz, G., Novák, A., Prószéky, G.: Hybrid text segmentation for Hungarian clinical records. In: Castro, F., Gelbukh, A., González, M. (eds.) MICAI 2013. LNCS (LNAI), vol. 8265, pp. 306–317. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-45114-0_25
24. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: Proceedings of EMNLP, pp. 1532–1543 (2014)
25. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**(336), 846–850 (1971)
26. Ruffini, M., Gavaldà, R., Limón, E.: Clustering patients with tensor decomposition. arXiv preprint arXiv:1708.08994 (2017)
27. Ward Jr., J.H.: Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. **58**(301), 236–244 (1963)
28. Waszczuk, J.: Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In: Proceedings of COLING, pp. 2789–2804 (2012)