Learning Portuguese Clinical Word Embeddings: a multi-specialty and multi-institutional corpus of clinical narratives supporting a downstream biomedical task

Lucas Emanuel Silva e Oliveira[a], Yohan Bonescki Gumiel[a], Arnon Bruno Ventrilho dos Santos[a], Lilian Mie Mukai Cintho[a], Deborah Ribeiro Carvalho[a], Sadid A. Hasan[b], Claudia Maria Cabral Moro[a]

[a] Health Technology Program, Pontifical Catholic University of Paraná, Curitiba, PR, Brazil,
[b] AI Lab, Philips Research North America, Cambridge, MA, USA

Abstract

In this paper, we train a set of Portuguese clinical word embedding models of different granularities from multi-specialty and multi-institutional clinical narrative datasets. Then, we assess their impact on a downstream biomedical NLP task of Urinary Tract Infection disease identification. Additionally, we intrinsically evaluate our main model using an adapted version of Bio-SimLex for the Portuguese language. Our empirical results show that the larger and coarse-grained model achieved a slightly better outcome if compared with the small and fine-grained model in the proposed task. Moreover, we obtained satisfactory results with Bio-SimLex intrinsic evaluation.

Keywords:

Natural Language Processing; Clinical Word Embeddings, Clinical narratives.

Introduction

The Electronic Health Record (EHR) was mainly designed to digitally store patient's data and improve healthcare operational efficiency. Moreover, researchers found in it a rich source to support several clinical informatics applications such as medical concept extraction, disorder reasoning, and patient history summarization [1].

Natural Language Processing (NLP) and Machine Learning (ML) techniques are widely used in order to extract, identify and summarize EHR data, despite the dependency in laborious manual annotation and hand-crafted features [2,3]. Recently, many studies applied Deep Learning (DL) approaches to process EHR data [4–6], achieving better performance than traditional NLP/ML methods requiring less time-consuming feature engineering.

An important component of DL for NLP methods is the use of Word Embeddings (WE) to represent each word as a vector in a low dimensional space [7] and employ this vector as an input feature. Besides, the resulting word vectors can be used to address many other NLP-related problems, like sentiment analysis [8] and paraphrase detection [9].

Several studies used WE to solve health-related tasks like drug name recognition [10], semantic similarity [11], biomedical named entity recognition or bio-NER [12], and patient outcome prediction [6].

To the best of our knowledge, only a few studies applied WE to the Portuguese language in the biomedical domain (e.g., [13,14]). Three main studies made a WE repository available for both European and Brazilian Portuguese (pt-br) languages using a multi-genre corpus with data from Wikipedia, GoogleNews etc. [15–17].

Despite the success of WE in the clinical NLP domain, it is difficult to find large representative corpora to address relevant tasks, especially based on EHR data. Wang et al. [18] provided a comprehensive set of WE training experiments from distinct resources, namely clinical notes, biomedical articles, Wikipedia and news. They found that the WE trained from EHR has the best results in the clinical information extraction task; the semantic similarity captured by the EHR embeddings is closer to human experts' judgments on all datasets, and together with PubMed WE, the EHR embedding model can find more relevant similar medical terms.

Roberts [19] evaluated the trade-offs between small (and representative) corpora against large (but unrepresentative) corpora for training a WE model for clinical NLP tasks. In fact, it is not easy to decide between a huge general-purpose corpus and a small highly representative corpus. For instance, one can decide for a medium-sized clinical notes corpus instead of a corpus of a varying set of documents, or a large scientific corpus, or even a combination of both. They found that merging multiple corpora is the best option when generating embeddings.

Thus, there exists a gap in building Portuguese clinical WE models for research, that is, we could not find a clinical WE model available for NLP tasks in Portuguese. However, to provide a consistent WE model a set of experiments are required in order to prove the usefulness of the model. It is possible to evaluate the model *extrinsically* by applying it to a downstream task, or *intrinsically,* measuring the innate quality of word representations through syntactic and semantic analogies (e.g., [15,20,21]).

Chiu et al. [22] developed comprehensive resources targeting the intrinsic evaluation of word representations in biomedicine (Bio-SimLex and Bio-SimVerb). The Bio-SimLex is a list of words (nouns) disposed in pairs with their respective similarity score (defined by a group of expert annotators), which can be used to compare the similarity scores between a WE model and the ones defined in Bio-SimLex. Besides, some other studies affirm that intrinsic and extrinsic evaluation scores do not always correlate [23–25]; the authors claim that their evaluation resources can serve as a predictor of performance on downstream tasks. This is especially true for the Bio-SimLex set and bio-NER task, which presented a high correlation.

In this paper, we address the research gap in pt-br clinical WE model generation and investigate an important research question: can a clinical WE model trained with multi-specialty and

multi-institutional clinical narratives achieve good results in downstream biomedical NLP tasks? We trained a preliminary multi-institutional and multi-specialty clinical WE model and assess its performance by (i) checking if such a big coarse-grained model applied to a Deep Learning algorithm performs equally well to a small fine-grained model for predicting a specific disease, and (ii) analyzing the results on Bio-SimLex evaluation set.

Table 1– Example narratives from each dataset and their granularity. Note that, narratives from ANN-UTI contains an ICD-10 code at the beginning of the note, which is the result of expert annotation i.e. labeling with the corresponding diagnosis

Dataset	Sample narrative	Granularity
GROUP-ALL	# RETORNO PARA REAVALIAÇÃO DE GLAUCOMA # GPAA EM USO DE DUO TRAVATAN E BRIMONIDINA - EM USO IRRGEULAR! # DMRI SECA AO - PACIENTE ESTÁ USANDO TRAVATAN A NOITE E LACRIFILM 3X/DIA AV CC 20/60 - 20/40 PIO 21/16 AO ESCAVAÇÃO DE 0,9, DRUSAS EM POLO POSTERIOR ORIENTO NECESSIDADE DE USO DOS COLIRIOS - CIENTE DO PROGNOSTICO PRESCREVO NOVAMENTE DUOTRAVATAN (E FORNEÇO MAIS UMA AMOSTRA GRATIS) + BRIMONIDINA 12/12 + LACRIFIM AO SOLICITO NOVO CAMPO VISUAL. RETORNO EM 1 MÊS PARA REAVALIAR PIO	**Coarse-grained**: Various medical specialties and institutions
GROUP-UTI	PACIENTE 62A, COM QUADRO DE INCONTINÊNCIA URINÁRIA DE ESFORÇO ASSOCIADA A URGÊNCIA HÁ APROXIMADAMENTE 01 ANO, PCTE RELATA TER INFECÇÃO DO TRATO URINÁRIO NÃO TRATADA ENCAMINHO A SEU MÉDICO SOLICITO EPU	**Medium-grained**: Narratives that contain ICD-10 N39 group of diseases, multi-institutional.
ANN-UTI	[N39.0] Paciente submetida a correção cirúrgica de incontinência urinária com Safyre transobturatório. Procedimento cirúrgico evoluiu sem intercorrências. Recebe condições e orientações quanto aos cuidados pós-operatórios. Agendar consulta em 2 semanas.	**Fine-grained:** Narratives of single specialty and institution

Methods

In this section, we describe our pt-br clinical WE model training process including the dataset, preprocessing steps, and parameter space; the deep learning algorithm used to identify Urinary Tract Infection disease (UTI); the Bio-SimLex pt-br adaptation process; and the experimental setup.

Data

To generate the WE model, we used a collection of de-identified clinical narratives obtained from a group of three hospitals in south of Brazil from January 2013 until December 2017. These narratives comprehend 745,731 documents of different types (nursing notes, discharge summaries and ambulatory records) and various medical specialties (Cardiology, Nephrology, Endocrinology, etc.). We will call this entire col-lection, GROUP-ALL. We classify this collection as coarse-grained because of the generalized nature of its data, which does not contain specific data of only one type of document, specialty or institution.

As we intend to run an algorithm to detect UTI disease and compare the results between a fine- and coarse-grained WE model (see next sections for details), we used a subset of data from GROUP-ALL. All narratives that contained a disease name corresponding to the ICD-10 code N.39, denoting UTI (and the corresponding subclass codes) were filtered, forming the new dataset called GROUP-UTI.

Besides, another dataset from a Hospital in southeast of Brazil was obtained to train the UTI disease detection algorithm (see the following sections). This dataset, named ANN-UTI, consisted of narratives annotated with corresponding ICD-10 codes related to UTI diseases. We present a few sample narratives and dataset statistics in Table 1 and Table 2.

Preprocessing

We preprocessed the narratives of GROUP-ALL dataset using the following steps sequentially: (1) sentence parsing, (2) sentence tokenization, (3) lowercasing, (4) accentuation removal, (5) numeric characters removal, and (6) stopwords removal (using NLTK stopwords[1])

Embedding parameters

To train a preliminary WE model, we followed recommendations and guidelines provided by prior studies that deeply analyzed the impact of hyperparameters on word vector quality. Unlike Beam et al. [26] who strictly reproduced the parameters of Levy et al. [27], we opt to select values from clinical WE studies as the default algorithmic configuration in case of inconsistencies among the guidelines.

CBOW vs. Skip-gram: Several studies affirm that, in general, Skip-gram model performs better than CBOW [26,28,29]. Hence, we used it to train our model.

Negative sampling: Levy et al. [27] recommended using multiple negative samples [30] when using Skip-gram. Boag and Kané [31] used 8 negative samples in their work for training WE based on a Clinical Metathesaurus. We also used the same number of negative samples in our training.

Minimum count: Chiu et al. [28] showed the limited effect of this parameter on overall scores. Therefore, we used the default value of 5. This reduced the GROUP-ALL vocabulary size to 56,195 unique tokens, GROUP-UTI to 5,125 and ANN-UTI to 3,203.

Sub-sampling: Chiu et al. [28] described that this parameter does not have a significant impact on extrinsic evaluation, so we set the default value of 1e-3.

[1] http://www.nltk.org/howto/portuguese_en.html

Context-size: Similar to Boag and Kané [31], we set the context/window size to 8.

Vector dimension: Following Boag and Kané [31], we used 300 as the dimension size of the vectors since Chiu et al. [28] and Fanaeepour et al. [32] did not find much improvement using 200 as the dimension size. Furthermore, the size of 300 also corresponds with the configuration requirements of the GloVe model (that is the UTI detection algorithm baseline).

DeepCoder: an Algorithm to Identify Urinary Tract Infection

UTI is defined as "an infection anywhere in the urinary tract (urethra, bladder, ureters, or kidneys)" [33] or "the clinical syndromes of acute, uncomplicated, urinary infection" [34]. ICD-10 reserves a specific class for such problems: "N39 - Other disorders of urinary system" and 7 subclasses (N39.0, N39.1, N39.2, N39.3, N39.4, N39.8 and N39.9) to provide more detailed information for clinical evaluation.

Table 2 – Dataset sizes by number of tokens and sentences

Dataset	#Sentences	#Tokens	#Unique tokens
GROUP-ALL	2,412,055	32,023,244	287,495
GROUP-UTI	26,719	319,203	17,518
ANN-	2,030	205,318	11,494

DeepCoder was developed using as input a GloVe trained WE model based on ANN-UTI texts following the preprocessing steps and with hyperparameters configured as described in previous sections. The algorithm is composed of a neural network architecture formed with an embedding layer as input with 500 dimensions (e), four convolutional layers with kernel = 128 (k), a Rectified Linear Unit (ReLU) and window sizes of 5, 8, 10, 12. Between each convolutional layer, there exists a 1-max pooling layer. Finally, a global max pooling layer, followed by dropout of 0.2 (d) that feeds a 128-sized Dense (i) layer and softmax activation with 8 possible outputs (7 for the ICD-10 N39 code and its subclasses, and 1 for a generic non-UTI code).

Experimental setup

We based our experiments on an extrinsic evaluation using the DeepCoder algorithm, and in order to predict the model performance in other biomedical NLP downstream tasks, we performed an intrinsic evaluation using the Bio-SimLex set.

Extrinsic evaluation

The Word2Vec models trained with GROUP-ALL, GROUP-UTI and ANN-UTI datasets, in addition to the original DeepCoder GloVe model (trained with ANN-UTI), were utilized in DeepCoder in order to perform the extrinsic evaluation and compare relative performance. All results were calculated using a 10-fold cross-validation. Figure 1 presents the extrinsic experimental setup overview.

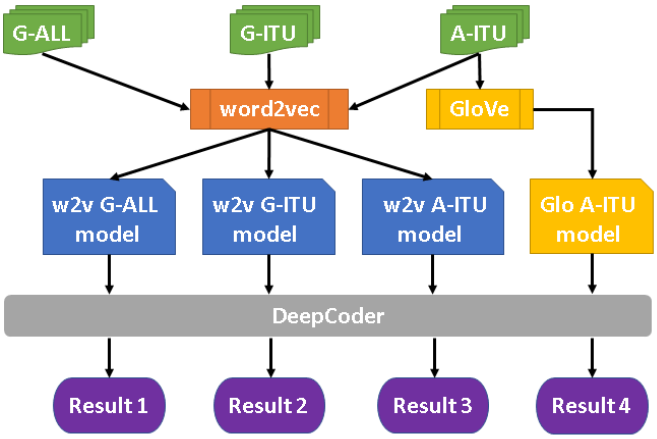

Figure 1 – Extrinsic experimental setup: each dataset was used to obtain different WE models using Word2Vec and GloVe algorithms. We used the four pre-trained WE models to create input embeddings for DeepCoder

Intrinsic evaluation and Bio-SimLex adaptation

To perform the intrinsic evaluation, we used the Bio-SimLex evaluation set, which contains 988 word pairs in English language, associated with a score concerning the semantic similarity and relatedness of the word pairs. Then to evaluate a WE model, it is necessary to calculate the Spearman's correlation coefficient between the similarity ratings found in the model versus the ratings defined by experts and available in Bio-SimLex.

Two researchers (one with Medical, and the other with Health Informatics background) translated and adapted the terms to pt-br, observing the following instructions.

- Check if the English term exists in the Unified Medical Language System (UMLS)[2] and has a pt-br translation. If not, discuss the best possibility consensually between the two translators. If yes, then pick the preferable option. If multiple options, then prioritize terms:
  - Labeled as ISPREF='Y'
  - With a higher number of occurrences in GROUP-ALL dataset
- Label the word pair as General, EHR or Biomedical – where the first stands for words from the general domain, the second represents words seen in EHR texts, and the third category contains words that are not present in EHR, but used in biomedical context (e.g., biomedical articles).

In Table 3 we present some translation examples.

Table 3 – Bio-SimLex translation examples

Term1	Term2	Score	Type
therapy/terapia	treatment/tratamento	9.32	EHR
oxide/óxido	sucrose/sacarose	0.00	BioM
lake/lago	river/rio	2.65	General

It is important to highlight some issues found during the translation process, like: (i) the difficulty to translate ambiguous terms that have multiple meanings or translations, depending on the context (e.g., abstract); (ii) word pairs without well-known synonyms in pt-br, like hindbrain and rhombencephalon, which we found only one possible word to translate both

[2] https://www.nlm.nih.gov/research/umls/

(*rombencéfalo*). In this case, we removed the word pair to avoid equal words in the evaluation set.

Due to the differences between the general domain, biomedical literature, and EHR texts, we performed three separate intrinsic evaluation runs. One using all the translated Bio-SimLex data; another with word pairs categorized as Biomedical and EHR, excluding General category; and the last one using only the word pairs labeled as EHR. By doing that, we are trying to give a more fair evaluation of our model, which was trained exclusively with EHR text data.

Results

We summarize the results from our extrinsic evaluation in Table 4. It is possible to verify that the scores are very similar across all models. The GROUP-ALL model is the largest and less representative compared to the gold standard, and ANN-UTI is the smallest and more representative (because it contains texts from the same dataset as the gold standard). This suggests that the GROUP-ALL model takes advantage of its size to yield

Table 4 – Extrinsic experiments results. F1-score mean (F1 with cross-validation with 10 folds) and standard deviation (SD)

WE model	F1	SD
Word2Vec GROUP-ALL	**0.95**	**0.04**
Word2Vec GROUP-UTI	0.91	0.14
Word2Vec ANN-UTI	**0.95**	0.05
GloVe ANN-UTI	**0.95**	0.07

the good performance for the UTI identification task, and the ANN-UTI models (Word2Vec and GloVe), as the more representative ones, achieve good results as expected. The GROUP-UTI embeddings yielded an inferior result, most likely due to less representative texts of less than medium-size.

The results of the intrinsic evaluation of the GROUP-ALL model, using three subsets of the Bio-SimLex are shown in Table 5. The Spearman's correlation coefficient increases when we use more specific subsets (Biomedical and EHR). Using only the word pairs labeled as EHR our model achieved a correlation score of 0.6419, which is similar to the results obtained by Chiu et al. [22] with their Skip-gram and PubMed-w2v models, varying 0.07 and 0.05 respectively.

Table 5 – Intrinsic experiments results of Spearman's correlation coefficient (RHO) divided by the categories of Bio-SimLex word pairs. RHO

Categories	RHO
General+Biomedical+EHR	0.4558
Biomedical+EHR	0.5679
EHR	**0.6419**

Discussion

The results from our empirical extrinsic evaluation confirm previous findings [19] suggesting that we have indeed a trade-off between corpus size and similarity when it comes to WE. The results imply that the answer to our research question is: yes, a large coarse-grained WE model can yield good results for a downstream biomedical NLP task.

We can highlight a few limitations of this study and consequently propose some future work. For example, the models were extrinsically evaluated for one task only; and to overcome this limitation we opted to use Bio-SimLex to emulate the results to another variety of bio-NLP tasks. But due to the difficulties found in the translation process, we think that the evaluation set lost some of its reliability, then would be better to build a specific evaluation resource for EHR pt-br data, although the obtained results are similar to Chiu et al. [22].

We would also explore hyperparameter tuning and other WE algorithms such as fastText [35], and wang2vec [36] in the future.

In this paper, we explored WE models trained with words only, although it is possible to use several approaches that focus on enhancing WE with clinical knowledge by concatenating extra information to the vector space [11,21,26,31,37,38]. It is also worth noting that despite the existence of various approaches to generate clinical embeddings, there is limited consensus among researchers on what is the state-of-the-art for each bio-NLP task.

Besides some authors (e.g., [39]) discuss the reliability on ICD-10 coding, our work relied on a simple annotation process containing only one disease and its specializations, which did not lead us to uncertainties and the complex ICD environment that build-up in some cases.

We also plan to enlarge our corpus by adding biomedical publications, Wikipedia and other open source datasets in Portuguese.

Conclusions

In this paper, we built WE models with different granularities, and extrinsically evaluated them using a disease prediction algorithm (DeepCoder) to assess the performance variation due to different word embeddings. We used an adapted version of Bio-SimLex set to intrinsically evaluate a large and coarse-grained model, in order to predict the model's performance in other biomedical downstream tasks.

We concluded that it is possible to achieve similar results using a large coarse-grained WE model and a small fine-grained alternative to facilitate a bio-NLP task; however, robustness of our models could be ensured by applying them to a wide range of clinical prediction tasks, as the Bio-SimLex adaptation to pt-br has some limitations and reliability issues.

Acknowledgments

Philips Research North America and Pontifical Catholic University of Paraná supported this work. "*Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001*" partially financed this study also.

References

[1] B. Shickel, P.J. Tighe, A. Bihorac, and P. Rashidi, Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis, *IEEE J. Biomed. Heal. Informatics*. **22** (2017) 1589–1604.

[2] F. Liu, J. Chen, A. Jagannatha, and H. Yu, Learning for Biomedical Information Extraction: Methodological Review of Recent Advances, *Comput. Res. Repos.* **1606.07993** (2016)..

[3] C. Friedman, T.C. Rindflesch, and M. Corn, Natural language processing : State of the art and prospects for significant progress , a workshop sponsored by the

National Library of Medicine q, *J. Biomed. Inform.* **46** (2013) 765–773.

[4] R. Miotto, L. Li, B.A. Kidd, and J.T. Dudley, Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records, *Sci. Rep.* **6** (2016) 1–10.

[5] A.N. Jagannatha, and H. Yu, Structured prediction models for RNN based sequence labeling in clinical text., *Proc. Conf. Empir. Methods Nat. Lang. Process. Conf. Empir. Methods Nat. Lang. Process.* **2016** (2016) 856–865.

[6] Y. Feng, X. Min, N. Chen, H. Chen, X. Xie, H. Wang, and T. Chen, Patient outcome prediction via convolutional neural networks based on multi-granularity medical concept embedding, in: 2017 IEEE Int. Conf. Bioinforma. Biomed., IEEE, 2017: pp. 770–777.

[7] Y. Goldberg, A Primer on Neural Network Models for Natural Language Processing, *J. Artif. Intell. Res.* **57** (2016) 345–420.

[8] A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, and C. Potts, Learning Word Vectors for Sentiment Analysis, *Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol.* (2011) 142–150.

[9] R. Socher, E. Huang, and J. Pennington, Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection., *Adv. Neural Inf. Process. Syst.* (2011) 801–809.

[10] I. Segura-bedmar, and P. Mart, Exploring Word Embedding for Drug Name Recognition, *Sixth Int. Work. Heal. Text Min. Inf. Anal.* (2015) 64–72.

[11] L. De Vine, G. Zuccon, B. Koopman, L. Sitbon, and P. Bruza, Medical semantic similarity with a neural language model, *Proc. 23rd ACM Int. Conf. Inf. Knowl. Manag. - CIKM '14*. (2014) 1819–1822.

[12] Y. Wu, J. Xu, M. Jiang, Y. Zhang, and H. Xu, A Study of Neural Word Embeddings for Named Entity Recognition in Clinical Text., *AMIA Annu. Symp. Proc.* **2015** (2015) 1326–33..

[13] F. Duarte, B. Martins, C.S. Pinto, and M.J. Silva, A Deep Learning Method for ICD-10 Coding of Free-Text Death Certificates, in: Prog. Artif. Intell. EPIA 2017. Lect. Notes Comput. Sci., 2017: pp. 137–149.

[14] M. V. Treviso, C.D. Shulby, and S.M. Aluisio, Evaluating Word Embeddings for Sentence Boundary Detection in Speech Transcripts, *Proc. 11th Brazilian Symp. Inf. Hum. Lang. Technol.* (2017) 151–160.

[15] J. Rodrigues, A. Branco, S. Neale, and J. Silva, LX-DSemVectors: Distributional Semantics Models for Portuguese, in: Comput. Process. Port. Lang. PROPOR 2016. Lect. Notes Comput. Sci., 2016: pp. 259–270.

[16] N.S. Hartmann, E. Fonseca, C.D. Shulby, M. V Treviso, J.S. Rodrigues, and S.M. Aluísio, Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks, *Proc. 11th Brazilian Symp. Inf. Hum. Lang. Technol.* (2017).

[17] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, Learning Word Vectors for 157 Languages, *Comput. Res. Repos.* **1802.06893** (2018).

[18] Y. Wang, S. Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen, P. Kingsbury, and H. Liu, A comparison of word embeddings for the biomedical natural language processing, *J. Biomed. Inform.* **87** (2018) 12–20.

[19] K. Roberts, Assessing the Corpus Size vs. Similarity Trade-off for Word Embeddings in Clinical NLP, *Proc. Clin. Nat. Lang. Process. Work.* (2016) 54–63.

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, Efficient Estimation of Word Representations in Vector Space, *ArXiv Prepr. ArXiv1301.3781*. (2013).

[21] Y. Choi, C.Y.-I. Chiu, and D. Sontag, Learning Low-Dimensional Representations of Medical Concepts., *AMIA Jt. Summits Transl. Sci. Proceedings. AMIA Jt. Summits Transl. Sci.* **2016** (2016) 41–50.

[22] B. Chiu, S. Pyysalo, I. Vulić, and A. Korhonen, Bio-SimVerb and Bio-SimLex: wide-coverage evaluation sets of word similarity in biomedicine, *BMC Bioinformatics*. **19** (2018) 33.

[23] B. Chiu, A. Korhonen, and S. Pyysalo, Intrinsic Evaluation of Word Vectors Fails to Predict Extrinsic Performance, in: Proc. 1st Work. Eval. Vector-Sp. Represent. NLP, ACL, 2016: pp. 1–6.

[24] A. Gladkova, and A. Drozd, Intrinsic Evaluations of Word Embeddings: What Can We Do Better?, in: Proc. 1st Work. Eval. Vector-Sp. Represent. NLP, ACL, 2016: pp. 36–42.

[25] M. Faruqui, Y. Tsvetkov, P. Rastogi, and C. Dyer, Problems With Evaluation of Word Embeddings Using Word Similarity Tasks, in: Proc. 1st Work. Eval. Vector-Sp. Represent. NLP, ACL, 2016: pp. 30–35.

[26] A.L. Beam, B. Kompa, I. Fried, N. Palmer, X. Shi, T. Cai, and I.S. Kohane, Clinical Concept Embeddings Learned from Massive Sources of Medical Data, *Comput. Res. Repos.* **1804.01486** (2018).

[27] O. Levy, Y. Goldberg, and I. Dagan, Improving Distributional Similarity with Lessons Learned from Word Embeddings, *Trans. Assoc. Comput. Linguist.* **3** (2015) 211–225.

[28] B. Chiu, G. Crichton, A. Korhonen, and S. Pyysalo, How to Train good Word Embeddings for Biomedical NLP, *Proc. 15th Work. Biomed. Nat. Lang. Process.* (2016) 166–174..

[29] T.H. Muneeb, S.K. Sahu, and A. Anand, Evaluating distributed word representations for capturing semantics of biomedical concepts, *Proc. 2015 Work. Biomed. Nat. Lang. Process. (BioNLP 2015)*. (2015) 158–163.

[30] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, Distributed Representations of Words and Phrases and their Compositionality, in: NIPS'13 Proc. 26th Int. Conf. Neural Inf. Process. Syst., 2013: pp. 3111–3119.

[31] W. Boag, and H. Kané, AWE-CM Vectors: Augmenting Word Embeddings with a Clinical Metathesaurus, *NIPS 2017 Work. Mach. Learn. Heal.* (2017).

[32] M. Fanaeepour, A. Makarucha, and J.H. Lau, Evaluating Word Embedding Hyper-Parameters for Similarity and Analogy Tasks, *Comput. Res. Repos.* **1804.04211** (2018).

[33] B. Foxman, Urinary tract infection syndromes. Occurrence, recurrence, bacteriology, risk factors, and disease burden, *Infect. Dis. Clin. North Am.* **28** (2014) 1–13.

[34] L.E. Nicolle, Urinary Tract Infection, *Crit. Care Clin.* **29** (2013) 699–715.

[35] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, Enriching Word Vectors with Subword Information, *Comput. Res. Repos.* **1607.04606** (2016).

[36] W. Ling, C. Dyer, A.W. Black, and I. Trancoso, Two/Too Simple Adaptations of Word2Vec for Syntax Problems, in: Proc. 2015 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang.

Technol., ACL, 2015: pp. 1299–1304.
[37] Z. Yu, T. Cohen, E. V Bernstam, T.R. Johnson, and B.C. Wallace, Retrofitting Word Vectors of MeSH Terms to Improve Semantic Similarity Measures, *Proc. Seventh Int. Work. Heal. Text Min. Inf. Anal.* (2016) 43–51.
[38] T. Bai, A.K. Chanda, B.L. Egleston, and S. Vucetic, Joint learning of representations of medical concepts and words from EHR data, in: 2017 IEEE Int. Conf. Bioinforma. Biomed., IEEE, 2017: pp. 764–769.
[39] J. Stausberg, N. Lehmann, D. Kaczmarek, and M. Stein, Reliability of diagnoses coding with ICD-10, Int. J. Med. Inf. **77**(1), 50–57 (2008).

Address for correspondence

Lucas Emanuel Silva e Oliveira.
tel: +55 41 992872702.
email: lucas.oliveira@pucpr.br