BioVec-Ita: Biomedical Word Embeddings for the Italian Language

(Discussion Paper)

Marcello Bavaro[1,*], Tommaso Dolci[1] and Davide Piantella[1]

[1] *Politecnico di Milano – Department of Electronics, Information and Bioengineering*

Abstract

In the healthcare field, the information created by digital technologies that collect clinical care notes, health service reports and patients' records, is generating terabytes of data, a great part of which is in textual format. These datasets may become an incredibly valuable asset only if the knowledge they carry is extracted using the appropriate artificial intelligence techniques, and specifically natural language processing (NLP) ones. Unfortunately, most existing tools support NLP for the English language, while local administrations and hospitals typically work in their native language, and therefore it becomes very important to have NLP tools to process biomedical data written also in these languages. Word embeddings are a popular and powerful NLP technique to extract semantics from textual data that could be very useful to solve the problem, but unfortunately for the Italian language there are no such tools specialized in the biomedical field. In this paper we propose BioVec-Ita, a new word embedding model for Italian, specialized in the biomedical field and designed using Word2vec, a flexible model for semantic representation that can be easily integrated with other pipelines. We also evaluate the performance of our word embeddings model in capturing the semantic similarities of biomedical terms, using three very popular test datasets translated into Italian.

Keywords

Word embeddings, Word2vec, natural language processing, healthcare

1. Introduction

In the field of biomedicine, the digitization of clinical care processes and health services is producing more and more medical data, much of it in textual format: reports, nursing notes, discharge letters and emergency room reports are just a few of the digital documents that are generated every day in hospitals. Moreover, humans are being digitized through new medical devices, apps, and monitoring technologies, which track, analyze and store a massive amount of data. It has been estimated that by 2025 the annual growth rate of data for healthcare will reach 36%, more than the general growth rate estimated at 27% [1].

Biomedical pieces of information are a huge asset for medical researchers and, since much of them are in unstructured form, it is necessary to leverage artificial intelligence (AI) and

SEBD 2023: 31st Symposium on Advanced Database System, July 02–05, 2023, Galzignano Terme, Padua, Italy

[*] Corresponding author.

marcello.bavaro@mail.polimi.it (M. Bavaro); tommaso.dolci@polimi.it (T. Dolci); davide.piantella@polimi.it (D. Piantella)

0000-0002-1403-7766 (T. Dolci); 0000-0003-1542-0326 (D. Piantella)

natural language processing (NLP) techniques to extract and create knowledge from them. This knowledge can then be used to improve patients' care and management, reduce costs, speed up procedures, and more. Among the most popular NLP models for biomedicine are pre-trained language representations such as BioBERT [2]. This model is able to perform text mining tasks like named entity recognition, relation extraction, and question answering. Moreover, language models can be used to classify documents based on diseases described in the text [3], or to extract information from Electronic Health Records to predict future diseases [4].

Despite a widespread use of the English language, local administrations and hospitals still work in their native language. For this reason, biomedical researchers need tools for data analysis in other languages, such as Italian. For example, word embeddings are a powerful semantic representations used for many different NLP tasks, but there exists few word embeddings for the Italian language [5, 6], and none specifically trained for the biomedical field, except for some preliminary work [7]. Word embeddings for the biomedical field are commonly used as feature input to machine learning and deep learning models, enabling techniques for the contextualization of textual data to be used for many tasks, such as readmission prediction [8].

In this work, we introduce a biomedical word embeddings model for the Italian language, named BioVec-Ita. Our model is based on Word2vec [9] and takes inspiration from previous literature on English biomedical word embeddings [10]. A key step of creating powerful language models is to search for well-documented biomedical data in text format to be used for the training phase. To train our model, we use a set of corpora offered by OPUSnlp [11], an online repository of textual data, and additional data extracted from Wikipedia dumps of thousands of pages in Italian belonging to biomedical topics. Moreover, we include material from the International Classification of Diseases (ICD9) translated into Italian.[1] Finally, we test the performance of BioVec-Ita on three popular test datasets for biomedical word embeddings evaluation, manually translated into Italian.[2]

The rest of the paper is organized as follows: Section 2 explores the state of the art, Section 3 introduces the BioVec-Ita embeddings and its training phase, Section 4 describes results of the experimental evaluation, Section 5 concludes the the paper and outlines future work.

2. Related Work

Over the years, there have been several studies on the creation of specialized word embeddings for the biomedical field and on their usage in a variety of downstream medical tasks. For instance, Xiao et al. [8] combined word embeddings and recurrent neural networks to predict future readmissions of patients after their discharge, while Liu et al. [12] integrated word embeddings in a drug name recognition system. Katikapalli et al. [13] studied the use of different pre-trained language models, including the famous ELMo, BERT, and sentenceBERT, in capturing the semantic of biomedical terms. Similar results on intrinsic performance of biomedical word embeddings were obtained by Chiu et al. [10] using a much simpler Word2vec model trained on a big text corpus of titles and abstracts of papers from PubMed[3], totaling 2.7 billion tokens.

[1] The Italian version of ICD9 is available at https://www.salute.gov.it/portale/temi/manuale-icd9cm/

[2] BioVec-Ita embeddings and test datasets are available at https://github.com/MarcelloBavaroff/BioVec-Ita

[3] https://pubmed.ncbi.nlm.nih.gov/

Although there are some works about NLP for biomedical applications regarding the Italian language (e.g., extraction of medical concepts [14] or medical notes understanding [15]) the limited amount of freely available textual data and resources hinder the research in this area. In fact, there is a small number of works even on general-purpose Italian word embeddings: Berardi et al. [6] first addressed the issue in 2015, and more recently Di Gennaro et al. [5]. The only exception is represented by the recent preliminary work by Bondarenko et al. [7], where the researchers tried to improve a fine-tuned version of BERT on Italian medical texts by combining contrastive learning and knowledge graph embeddings. However, their work leaves room for improvements considering the performance score obtained by equivalent English biomedical word embeddings.

3. BioVec-Ita

Biomedical word embeddings are a popular and powerful tool for language understanding, but there is a lack of research regarding the Italian language. For this reason, we take inspiration from the much richer English literature on biomedical word embeddings, particularly regarding Word2vec models for biomedicine. Word2vec [9] is a flexible and powerful model to retrieve accurate semantic representations of words, and it is easy to train and retrain even with a big amount of data. It is part of the family of static vector models: contextualized vector models provide instead representations based also on the context of the sentence. However, despite contextualized models being generally considered more powerful for language understanding tasks, biomedical terms are rarely synonymical, and for many medical tasks static embeddings obtain comparably high results in intrinsic performance tests [16].

Our work starts from the experience of [10], whose authors provide an in-depth overview of parameters setup for biomedical Word2vec model training in English. In this section, we first illustrate how we identified and pre-processed the training data used for the creation of BioVec-Ita. Then, after describing the training parameters adopted, we proceed with a preliminary evaluation on the intrinsic quality of the word embedding model produced.

3.1. Training Data

The main problem when it comes to biomedical word embeddings is that there are no resources in Italian like those offered in English by PubMed or MIMIC III [17], containing huge amounts of textual and medical data. Therefore, our training data mainly comes from OPUSnlp, an open-source collection of text corpora in different languages [11]. Specifically, we selected three corpora on biomedical topics, namely *ELRC-wikipedia_health*, *EMAv3 from European Medicines Agency*, and *Tilde MODEL-multilingual open data for EU languages*. Additionally, we include the Italian version of ICD9.

Finally, following the methodology described in [18], we integrate a number of Wikipedia pages on medical concepts in Italian. In particular, we download the Italian Wikipedia pages belonging to the following categories: “Salute”, “Medicina”, “Procedure mediche”, “Diagnostica medica”, “Specialità mediche”, “Farmacologia”, “Farmaci”, “Chirurgia”, and “Infermieristica”. Moreover, we add all the respective subcategories at depth two: for instance, considering the category “Salute”, its subcategories include “Alimentazione” at depth one, which in turn includes

Table 1
Overview of BioVec-Ita training parameters.

Parameter	Setting
Architecture	Skip-gram
Negative sampling	10
Vector dimensionality	200
Learning rate	0,05
SubSampling	1e-4
Window	30
Min_count	4-5
Training epochs	10-100

the subcategory “Alimentazione animale” at depth two. Duplicated pages are properly removed. In total, 14,395 pages are considered.

The overall resulting training data is then split into sentences, with each sentence tokenized to separate its words (also called tokens). In addition, all special characters (e.g., emojis and other non-character symbols) are removed and the order of the sentences in the datasets is randomized. The overall training dataset contains about 41M tokens.

3.2. Parameters and Setup

Two important parameters to adjust are the number of training epochs and the minimum number of occurrences (`min_count`) required for a word to appear in BioVec-Ita. If min_count is too large, important words may be left out of the model, if it is too small, the size of the vocabulary increases too much without accurate representations. Regarding the number of epochs, we carry out several tests by varying the number of epochs between 10 and 100. Additionally, we test the training process with both 4 and 5 as `min_count` values. Table 1 shows an overview of all the parameters involved in the training of BioVec-Ita.

We train BioVec-Ita on a machine equipped with an NVIDIA GPU RTX 360 (12Gb VRAM) and an Intel I5-12400F, using the *Gensim* library in version 4.2.0 in an environment with Python 3.9.15. The resulting word embeddings contain 158,251 words with `min_count` equal to 4, and 134,697 words with `min_count` equal to 5.

4. Experiments and Results

In this section, we illustrate the tests used to assess the performance of BioVec-Ita. We start from a description of the test datasets and the evaluation metrics, followed by the presentation and discussion of the experimental results.

4.1. Test Datasets

We evaluate the results BioVec-Ita on the translated version of the following three datasets: *MayoSRS* [19], *UMNSRS-similarity* and *UMNSRS-relatedness* [20].

MayoSRS and UMNSRS-similarity in Italian are provided in [7] and we make only few changes to certain terms to improve the quality of the data. For instance, we decide not to translate American drug names into the equivalent Italian active ingredient, but into the equivalent most popular drug name in our country. For example, Coumadin (original drug) is not translated into Warfarina (active ingredient), but instead into Warfarin (Italian equivalent drug). Concerning UMNSRS-relatedness, we manually translated the medical terms from English to Italian. All three datasets are composed of pairs of medical terms, associated with a score indicating the similarity between the two. Medical terms may consist of a single word (e.g., “fever”), or multi-word expressions (e.g., “portal hypertension”).

4.2. Metrics

To evaluate the quality of word embeddings in representing the semantics of biomedical words, we calculate the cosine similarity between the word pairs in each tuple of the test datasets, comparing it to their gold similarity score. Considering two embeddings $\vec{a}$ and $\vec{b}$, their cosine similarity is defined as:

$$sim_{cos}(a, b) = cos(\phi) = \frac{a \cdot b}{||a|| \cdot ||b||} ,$$

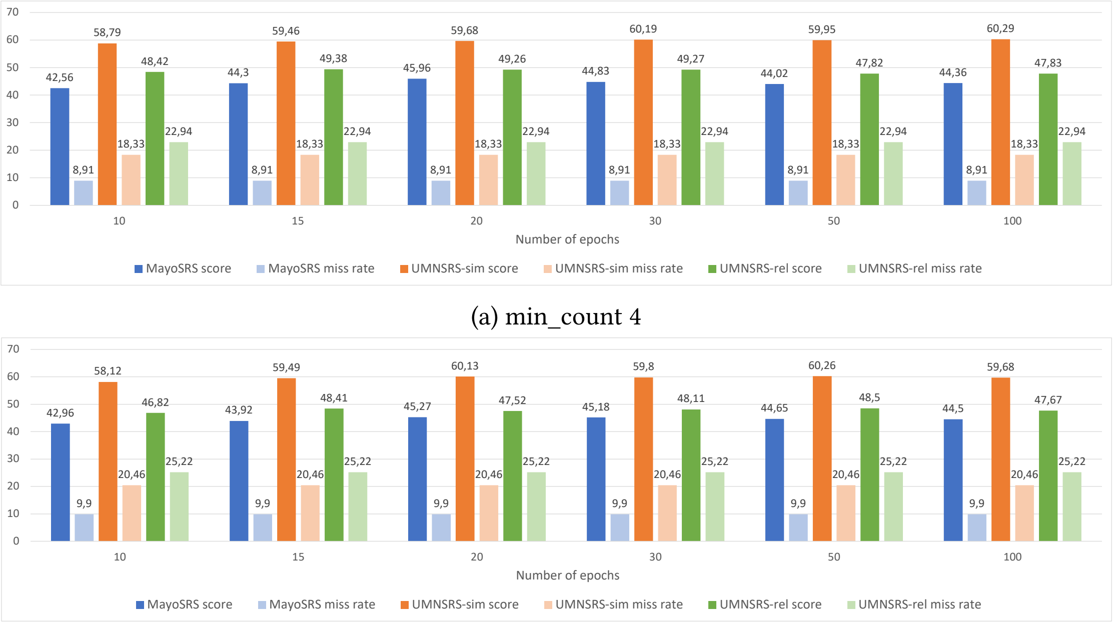
where $||\vec{a}||$ is the magnitude of $\vec{a}$. Once the similarity is calculated for all the tuples, the Spearman function is used to assess whether the distribution of the obtained scores follows the distribution of the gold scores indicated in the dataset. The final score is between 1 and 100. Since BioVec-Ita represents only single words, for multi-word expressions we take the average vector of all the word vectors that compose the expression. While averaging, semantically meaningless words such as prepositions are excluded. Tests are carried out with the help of BioNLP-2016[4], a toolkit for word embeddings evaluation.

4.3. Performance Results

Figures 1a and 1b show the results of testing BioVec-Ita on the three test datasets. Different training parameters are compared: number of epochs equal to 10, 15, 20, 30, 50 and 100, and `min_count` equal to 5 and 4. In addition to the final score, for each dataset, the corresponding miss rate is given, i.e., for how many tuples the similarity cannot be calculated because one or more words are not present in BioVec-Ita. Results are fairly homogeneous among each other, and no parameter configuration overcomes the others on all three test datasets. Only training for 10 epochs seems to be insufficient to achieve high enough results; on the opposite side, 100 epochs appears also similarly counterproductive. As expected, the miss rate is higher with `min_count` equal to 5, in all tests. Considering the average scores returned by the Spearman function and the miss rate, the best model is the one trained for 20 epochs with `min_count` equal to 4, achieving 45.96 on MayoSRS, 49.26 on UMNSRS-relatedness and 59.68 on UMNSRS-similarity, with an average miss rate of 16.73.

Figure 2 makes a comparison between our best model of BioVec-Ita (20 epochs, `min_count` = 4) and the word embeddings proposed by Bondarenko et al. [7]. Their model is based on sub-tokens, meaning that it produces a vector representation of any word, thus their miss rate is

[4] https://github.com/cambridgeltl/BioNLP-2016



(a) min_count 4

(b) min_count 5

Figure 1: Performance results of BioVec-Ita with different number of epochs and `min_count`.

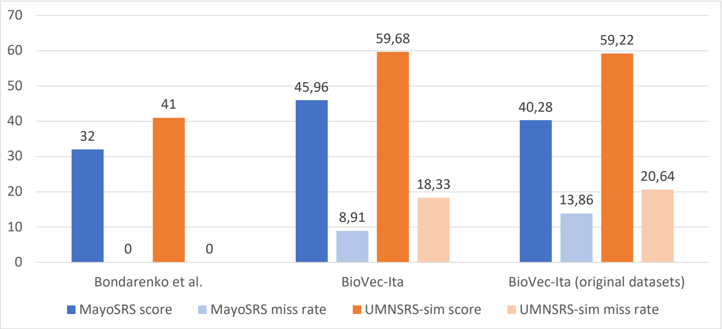


Figure 2: Comparison between BioVec-Ita and the model proposed in [7].

0. To ensure a fair comparison, we also calculate BioVec-Ita scores using the original versions of MayoSRS and UMNSRS-similarity without any of our changes. In both cases, our model obtain significantly higher results. Finally, Figure 3 compares the results of our best BioVec-Ita model with the state-of-the-art English biomedical word embeddings obtained from replicating the methodology described in [10]. BioVec-Ita obtains fairly comparable results, despite a difference in the size of the training data of two orders of magnitude. Moreover, our model has a lower miss rate on MayoSRS compared to the English embeddings.

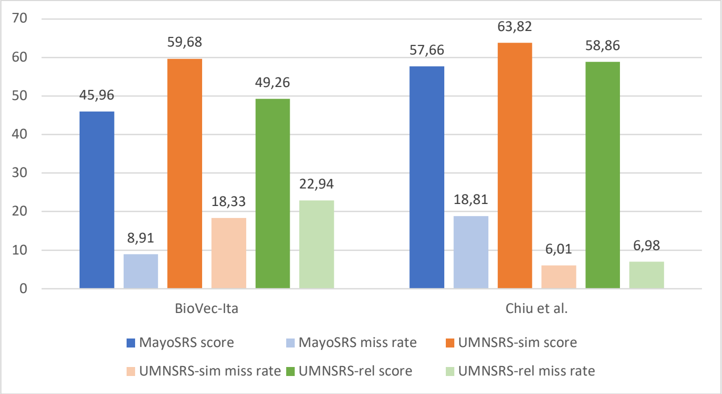


Figure 3: Comparison between BioVec-Ita and the English biomedical embeddings from [10].

5. Conclusions

The ever-growing amount of healthcare data demands advanced techniques to be properly exploited. Word embeddings represent a powerful and flexible NLP tool for managing all sorts of biomedical tasks involving language understanding, but there is a chronic lack of research on biomedical word embeddings for the Italian language. In this paper, we introduce BioVec-Ita, an Italian word embedding model trained specifically on biomedical data. BioVec-Ita is based on Word2vec, thus providing a flexible and simple, yet powerful model for a variety of NLP tasks. After describing the textual data used to train it, and the overall methodology adopted, we tested our model on three translated datasets to assess the quality of its semantic representations. Additionally, we compared BioVec-Ita both with previous Italian biomedical word embeddings from literature and with state-of-the-art biomedical embeddings in English, showing that our model achieves high-quality semantic representations comparable to those of its English counterpart, given the smaller size of the training data.

Future work includes retrieving more textual data in Italian regarding the biomedical field, for instance by including the digitized version of medical books. Moreover, we plan to train and test BioVec-Ita with different parameters. For example, increasing the dimensionality of the vectors to 300 or even higher, to try improving the semantic granularity of the representations. On the other hand, this would also increase the training time and the complexity of the model.

Acknowledgments

We are grateful to Letizia Tanca for helping us in the definition and realization of this work. We also thank Carmela Giannelli for supervising the translation of the test datasets into Italian.

References

[1] D. R.-J. G.-J. Rydning, J. Reinsel, J. Gantz, The digitization of the world from edge to core, Framingham: International Data Corporation 16 (2018).

[2] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.

[3] L. Yao, C. Mao, Y. Luo, Clinical text classification with rule-based features and knowledge-guided convolutional neural networks, BMC Medical Informatics and Decision Making 19 (2019) 31–39.

[4] J. Liu, Z. Zhang, N. Razavian, Deep ehr: Chronic disease prediction using medical notes, in: Machine Learning for Healthcare Conference, PMLR, 2018, pp. 440–464.

[5] G. Di Gennaro, A. Buonanno, A. Di Girolamo, A. Ospedale, F. A. Palmieri, G. Fedele, An analysis of word2vec for the italian language, Progresses in Artificial Intelligence and Neural Systems (2021) 137–146.

[6] G. Berardi, A. Esuli, D. Marcheggiani, Word embeddings go to italy: A comparison of models and training datasets., in: Proceedings of the 6th Italian Information Retrieval Workshop, 2015.

[7] D. A. Bondarenko, R. Ferrod, L. Di Caro, Combining contrastive learning and knowledge graph embeddings to develop medical word embeddings for the italian language, arXiv preprint arXiv:2211.05035 (2022).

[8] C. Xiao, T. Ma, A. B. Dieng, D. M. Blei, F. Wang, Readmission prediction via deep contextual embedding of clinical concepts, PloS one 13 (2018) e0195024.

[9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in Neural Information Processing Systems 26 (2013).

[10] B. Chiu, G. Crichton, A. Korhonen, S. Pyysalo, How to train good word embeddings for biomedical nlp, in: Proceedings of the 15th Workshop on Biomedical Natural Language Processing, 2016, pp. 166–174.

[11] J. Tiedemann, Parallel data, tools and interfaces in opus, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation, 2012, pp. 2214–2218.

[12] S. Liu, B. Tang, Q. Chen, X. Wang, Effects of semantic features on machine learning-based drug name recognition systems: Word embeddings vs. manually constructed dictionaries, Information 6 (2015) 848–865.

[13] K. S. Kalyan, S. Sangeetha, A hybrid approach to measure semantic relatedness in biomedical concepts, arXiv preprint arXiv:2101.10196 (2021).

[14] P. Agnello, S. M. Ansaldi, F. Azzalini, G. Gangemi, D. Piantella, E. Rabosio, L. Tanca, Extraction of medical concepts from italian natural language descriptions, in: 29th Italian Symposium on Advanced Database Systems, SEBD, volume 2994, 2021, pp. 275–282.

[15] R. Ferrod, E. Brunetti, L. Di Caro, C. Di Francescomarino, M. Dragoni, C. Ghidini, R. Marinello, E. Sulis, A support for understanding medical notes: Correcting spelling errors in italian clinical records., in: SMARTERCARE@ AI* IA, 2021, pp. 19–28.

[16] F. K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney, F. Rudzicz, A survey of word embeddings for clinical text, JBI 100 (2019) 100057.

[17] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R. G. Mark, Mimic-iii, a freely accessible critical care database, Scientific data 3 (2016) 1–9.

[18] G. Calarota, Domain-specific word embeddings for ICD-9-CM classification, Ph.D. thesis,

Alma Mater Studiorum, Università di Bologna, 2018. URL: http://amslaurea.unibo.it/16714/.

[19] T. Pedersen, S. V. Pakhomov, S. Patwardhan, C. G. Chute, Measures of semantic similarity and relatedness in the biomedical domain, JBI 40 (2007) 288–299.

[20] S. Pakhomov, B. McInnes, T. Adam, Y. Liu, T. Pedersen, G. B. Melton, Semantic similarity and relatedness between clinical terms: an experimental study, in: AMIA Annual Symposium Proceedings, 2010, p. 572.