Italian Word Embeddings for the Medical Domain

Franco Alberto Cardillo, Franca Debole
Institute for Computational Linguistics, Institute of Information Science and Technologies
Consiglio Nazionale delle Ricerche
Via G. Moruzzi 1, Pisa, Italy
francoalberto.cardillo@ilc.cnr.it, franca.debole@isti.cnr.it

Abstract

Neural word embeddings have proven valuable in the development of medical applications. However, for the Italian language, there are no publicly available corpora, embeddings, or evaluation resources tailored to this domain. In this paper, we introduce an Italian corpus for the medical domain, that includes texts from Wikipedia, medical journals, drug leaflets, and specialized websites. Using this corpus, we generate neural word embeddings from scratch. These embeddings are then evaluated using standard evaluation resources, that we translated into Italian exploiting the concept graph in the UMLS Metathesaurus. Despite the relatively small size of the corpus, our experimental results indicate that the new embeddings correlate well with human judgments regarding the similarity and the relatedness of medical concepts. Moreover, these medical-specific embeddings outperform a baseline model trained on the full Wikipedia corpus, which includes the medical pages we used. We believe that our embeddings and the newly introduced textual resources will foster further advancements in the field of Italian medical Natural Language Processing.

1. Introduction

Contemporary approaches to Natural Language Processing (NLP) heavily rely on word embeddings, which are vector representations of words that attempt to capture their semantic aspect according to the distributional hypothesis, suggesting that words used in similar contexts express and convey similar meanings (Harris, 1954). In their simplest form, neural word embeddings correspond to learnt parameters of an artificial neural network trained to predict either a target word given its context or the reverse. Pioneering models like word2vec (Mikolov et al., 2013a,b) and fastText (Bojanowski et al., 2017) provide a static, fixed embedding for every word encountered during training. For every word, the embedding vector is always the same, regardless of the context where the word occurs.

More recent models like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) produce dynamic, contextual word embeddings, but they usually require very large training corpora and higher computational resources. Neural word embeddings (hereafter named simply word embeddings) have proven to be highly effective in many downstream tasks across both general (Torregrossa et al., 2021) and specialized fields, like the medical domain (Khattak et al., 2019). However, as a unique lexicon characterizes specialized domains, the word embeddings should be computed from domain-specific training data. Medical word embeddings or large corpora of medical texts are available for numerous widely spoken language, such as English, Spanish and French (Khattak et al., 2019; Yijia et al., 2019).

For some languages, including Italian, there are neither specialized embeddings nor public text corpora.

Our contributions:

- We built a corpus of public medical texts, that can be used to train and evaluate word embeddings.
- We translated into Italian the standard English resources that are commonly used to evaluate the quality of medical word embeddings. The translation relies on a robust automatic procedure.
- We trained and evaluated word embeddings using two algorithms (word2vec, fastText) with the models CBOW and SG.

The rest of the paper is organized as follows. Section 2 describes the dataset, the pre-processing procedure, and the embedding models we experimented with. Section 3 describes the resources used to evaluate our medical word embeddings. Section 4 presents the results of the evaluation, where we compared the medical word embeddings to word embeddings trained on the Italian Wikipedia. Section 5 draws the conclusions and describes future work.

2. Material and Methods

The corpus contains official and formal documents, scientific articles and web pages. The data sources are described in Section 2.1. The cleaning and preprocessing procedures applied to the raw files are described in Section 2.2. As the size of the corpus is not very large, we limited our experiments to fixed word embeddings, as described in Section 2.3.

2.1. Material

We collected medical documents written in Italian from the four different sources **AIFA**,**HTML**, **OJ** and **WIKI**, described below.

LREC-COLING 2024, pages 9434–9440
20-25 May, 2024.

Corpus	Files	Type	Percentage
AIFA	23,246	pdf	42%
HTML	6,197	html	11%
OJ	5,354	pdf	10%
WIKI	20,042	html	37%
Total	54,839		100%

Table 1: Source, number of files, type of the original raw data, percentage of the dataset downloaded from the source.

Words / Corpus	Raw	Clean
AIFA	132,887,859	82,703,322
HTML	10,875,680	6,978,227
OJ	34,725,112	22,666,267
WIKI	12,923,312	7,609,021
Total	191,411,963	119,956,837

Table 2: Number of words for each source before pre-processing (raw) and after pre-processing (clean).

AIFA The Agenzia Italiana del Farmaco (Italian Medicines Agency, AIFA) is the public agency responsible for the regulatory activity of pharmaceuticals in Italy. We downloaded all the leaflets and the summaries of product characteristics (both in PDF format) of the drugs approved in Italy and available online in September 2022.

HTML We realize a customised web scraper to download and to extract data from two Italian web portals specialized in medical topics. **OJ** We downloaded full PDF issues of 20 online medical journals that are publicly available and accessible without any, paid or free, subscriptions.

WIKI We retrieved the URLs of articles accessible from the four top-level health-related categories *Biology*, *Medicine*, *Pharmacology* and *Pharmacy* using the Wikipedia dump[1]. The initial set of URLs was then refined by:

- removing duplicate items that could be accessed through multiple paths in the Wikipedia graph;
- filtering out loosely related pages, such as those describing movies about medicine, using a specially created stop-word list.

We used the Wikipedia API for Python[2] to extract the textual content from the HTML files.

Tables 1, 2, 3 show some statistics about the data.

2.2. Pre-processing

To maintain flexibility for a future expansion of the dataset, we chose to apply minimal pre-processing

[1]Dump date: 2022/09/01.

[2]https://pypi.org/project/wikipedia

Words / Corpus	Raw	Clean
AIFA	168,605	168,348
HTML	90,631	90,237
OJ	386,909	386,358
WIKI	244,053	243,776
Total	615,793	615,284

Table 3: Number of *unique* words for each source before pre-processing (raw) and after pre-processing (clean).

to the collected texts. This pre-processing[3] consists of the following steps:

- converting all text to lowercase;
- eliminating Italian stop-words[4];
- replacing URLs, e-mail addresses, and numbers with special tokens such as $\_N\_$ for numbers.

Due to the heterogeneity both in the data sources and in the types of data format, the output of the pre-processing contains some errors:

- The conversion from PDF to text results in many errors that are related to either incorrect detection of the page layout or incorrect tokenisation of words.
- The raw files contain many sentences in languages other than Italian. We have chosen not to perform language detection because none of the open source or free libraries available to us proved reliable in several tests we carried out.
- Some stop words are not recognised due to errors in converting the raw format to text.

These problems are also reflected in the data presented in Table 3, in fact the difference in the number of unique words between the two columns 'Raw' and 'Clean' is quite small (i.e. for **AIFA** Raw= 168.605 Clean=168.348). This is due to the fact that certain stop words and malformed tokens were not removed.

2.3. Methods

We experimented with the methods word2vec and fastText [5], both computing static word embeddings. Although more advanced and effective approaches have superseded these two methods, they are still valuable tools as they require smaller training sets and fewer computational resources.

The two methods train a shallow neural network in two different predictive models:

- in the Continuous Bag of Words (CBOW)

[3]Using the spacy (https://spacy.io) Python library.

[4]The set of Italian stop-words suggested by Spacy.

[5]We used the implementation in the Gensim Python Library (Rehurek and Sojka, 2011).

Hyperparameter	Values
alpha	0.025
dim	25 / 50 / 100 / 200
min count	3
neg	1 / 5 / 10 / 15
samp	0.001
window	5 / 10 / 15 / 30

Table 4: Hyperparameters and tested values.

model, the learner is trained to predict a target word given its context (surrounding words);
- in the Skip-Gram (SG) model, the learner is trained to predict the context given its centre word.

The main difference between word2vec and fastText lies in the atomic units they embed. Word2vec embeds whole words, while fastText works with subwords, i.e. n-grams of characters within words. Consequently, unlike word2vec, fastText can leverage the information at the subword level to generate embeddings of out-of-vocabulary words, assuming these words contain n-grams observed during training.

Word2vec and fastText share many hyperparameters that influence the quality of the final embeddings and that need to be evaluated (Rong, 2016). We follow (Chiu et al., 2016), but instead of varying a single hyperparameter while keeping the others fixed, we perform a full grid search over the hyperparameters listed in Table 4 for both word2vec and fastText in computing CBOW and SG embeddings.

3. Resources

The quality of the word embeddings is typically assessed using both extrinsic and intrinsic evaluation procedures (Wang et al., 2019). Extrinsic evaluations use word embeddings as input for downstream tasks such as Named Entity Recognition (NER) or document classification. The performance achieved on these tasks is taken as an indication of the quality of the embeddings: higher performance indicates a better quality of the embeddings. Intrinsic evaluations aim to assess the quality of embeddings in a way that is independent of any specific task, often by measuring semantic relationships between words. Although the performance on specific tasks is often the primary concern, it is important to note that extrinsic evaluations typically result in quality measures that are specific to the task at hand. As there are no readily available Italian datasets for downstream tasks in the medical domain, we have limited our evaluation to an intrinsic one, where the similarity between the the word embeddings is compared to the human perception of the association (similarity or relatedness) between the words.

We use a collection of reference standards created to test the degree of semantic relatedness and similarity between medical terms[6], composed by datasets such as the MayorSRS and MiniMayoSRS (Pedersen et al., 2007; Pakhomov et al., 2011) and the UMNSRS (Pakhomov et al., 2010). They contain pairs of English words whose degree of association (similarity or relatedness) was rated by human operators and converted to a numeric score. In details:

- MayoSRS: A set of 101 medical term pairs alongside their average rating of semantic relatedness assigned by a group of 13 medical coders. These coders were professionals without formal medical training, but with extensive experience in classifying clinical diagnoses.
- MiniMayoSRS: A set of 29 medical term pairs alongside two average scores of semantic relatedness assigned by three physicians and nine of the original 13 medical coders.
- UMNSRS_similarity: A set of 566 term pairs alongside their average score of semantic similarity assigned by eight physicians.
- UMNSRS_relatedness: A set of 588 term pairs alongside their average score of semantic similarity assigned by eight physicians.

These datasets have been used not only to evaluate word embeddings in English (Pakhomov et al., 2016) but also, after translation, in other languages such as French (Dynomant et al., 2018) and Spanish (Soares et al., 2019). They are considered standard benchmarks to use in the intrinsic evaluation of word embeddings within the medical domain (Chiu et al., 2016; Wang et al., 2018).

Each term in these datasets is paired with a Concept Unique Idenfier (CUI) in the Unified Medical Language System (UMLS) Metathesaurus [7]. Each term may consist of a single word or multiple words. The four resources described above contain a total number of 577 unique English terms (after lowercasing) associated with 586 unique CUIs; nine terms are each associated with two different CUIs. In order to evaluate intrinsically the Italian medical word embeddings using the mentioned datasets, we first need to translate the English terms into Italian. We implemented a fully automatic translation procedure based on the UMLS Metathesaurus Concepts Source Names and Codes (*MRCONSO*) (National Library of Medicine (US), 2009), which is a large multi-lingual vocabulary containing information about biomedical and health-related concepts (CUIs) and their different names in multiple

[6] Download link: https://doi.org/10.13020/D6CX04
[7] https://www.ncbi.nlm.nih.gov/books/NBK9684

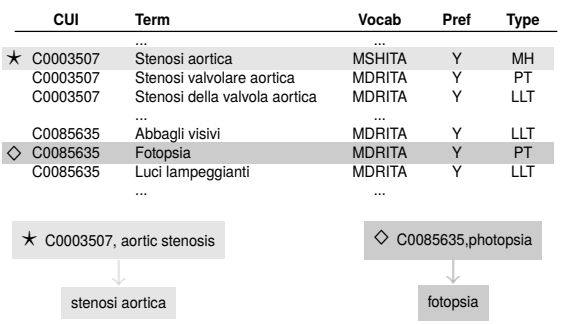

Figure 1: Examples of translations based on the CUI associated to a term. Columns: CUI, Term, Vocabulary, Preferred flag (Yes/No), Term Type.

national vocabularies, including five Italian vocabularies. Hereafter, we will refer to these five vocabularies as $\text{MRCONSO}_{\text{ITA}}$.

Here we give a general overview of the translation process, the details of which can be found in the published repository with the source code[8]. First, for a given pair (t_{en}, C), we select the set of Italian terms associated with the CUI C within $\text{MRCONSO}_{\text{ITA}}$, if any. Among these terms, we assign a preference to translations labelled as *preferred* in the vocabularies "MeSH[9] Italian" (MSHITA) or "MedDRA[10] Italian" (MDRITA), in that order. Among the preferred terms of the selected vocabulary, we finally choose the "main" form, i.e. a term defined of type *MH* (Main Heading) or *PT* (Preferred Term) in the two vocabularies, respectively (see two examples in Figure 1).

For the English terms associated to two different CUIs, when only one CUI is available in $\text{MRCONSO}_{\text{ITA}}$, we use the translation selected for the known CUI and use it for the missing CUI as well. For example, the term "weakness" is associated in MayoSRS and the two UMNSRS datasets, to respectively, CUIs C0004093 and C1883552. Since only the first CUI is found in $\text{MRCONSO}_{\text{ITA}}$, we translate both CUIs as "astenia", i.e. the Italian term associated to C0004093.

The previous steps leave 58 terms (10%) that cannot be translated into Italian using UMLS resources. Some of these terms correspond to American idiomatic expressions that are no longer part of the UMLS dictionaries (e.g. "banana bag" for "multivitamin"). For each of these CUIs we select the preferred English term in the current UMLS, translate it using three different translation services (Google Translate, DeepL, ChatGPT3.5) and select the translation with a majority vote (51 terms) or randomly when the three services disagree (7 terms). The full list of translations is available on the repository.

[8] https://github.com/med-nlp/italian-medical-word-embeddings

[9] Medical Subject Headings

[10] Medical Dictionary for Regulatory Activities

Dataset	ρ_p	ρ_s	n/m
MayoSRS	0.06	0.08	94/101
MiniMayoSRS coders	0.37	0.46*	29/29
MiniMayoSRS physicians	0.38*	0.40*	29/29
UMNSRS_rel	0.22*	0.23*	486/587
UMNSRS_sim	0.30*	0.29*	472/566

Table 5: Pearson (ρ_p) and Spearman (ρ_s) correlation coefficients with Wikipedia2Vec word embeddings (* indicates statistical significance). In the third column, *n* refers to the number of term pairs (t_1, t_2) used for computing the correlation and *m* is the total number of pairs in the dataset.

Dataset	ρ_p	ρ_s	n/m
MayoSRS	0.57*	0.58*	99/101
MiniMayoSRS coders	0.82*	0.84*	29/29
MiniMayoSRS physicians	0.78*	0.80*	29/29
UMNSRS_rel	0.49*	0.50*	544/587
UMNSRS_sim	0.60*	0.60*	472/566

Table 6: Best performance obtained by any model among those tested (* indicates statistical significance). Columns as in Table 5.

4. Evaluation

The evaluation of our embeddings has been conducted using the four datasets, described in Section 3. MiniMayoSRS contains two different similarity scores, that will be used separately in the evaluation.

The evaluation consists of calculating the cosine similarity between the vector embeddings for each pair of terms in the datasets. These similarities are then compared with the human-assigned scores using Pearson (ρ_p) and Spearman (ρ_s) correlation coefficients, which measure linear and rank correlation, respectively. Although the human-assigned scores are not normally distributed (Pakhomov et al., 2010), we included the Pearson correlation coefficient, as is done in similar works, e.g. (Soares et al., 2019), to facilitate a comparative evaluation of the results.

We use as baseline the word embeddings in the Wikipedia2Vec model [11] (Yamada et al., 2020), built using the text in Italian Wikipedia pages. Table 5 shows the results of the baseline evaluation.

In our experiments, all the fastText models exhibited a very low correlation with the evaluation corpus, which could likely be attributed to the limited size of the training corpus. For this reason, we will only focus on the word2vec models and present their results. We trained these models

[11] https://wikipedia2vec.github.io/wikipedia2vec/pretrained/, file: itwiki_20180420

Dataset	ρ_p	*d,e,w,n* (ρ_p)	ρ_s	*d,e,w,n* (ρ_s)
MayoSRS	0.49*	100,10,30,5	0.53*	100,2,10,1
MiniMayoSRS coders	0.82*	100,1,15,1	0.80*	25,1,10,1
MiniMayoSRS physicians	0.75*	100,30,15,5	0.74*	25,2,10,2
UMNSRS_rel	0.48*	100,5,15,5	0.50*	100,30,10,1
UMNSRS_sim	0.59*	100,25,10,5	0.60*	100,10,10,5

Table 7: Best performance obtained by any CBOW model among those tested. First and second column: ρ_p of the best model and corresponding hyperparameters. Third and fourth column: analogously for ρ_s (* indicates statistical significance). Hyperparameters: vector *d*imension, *e*poch, *w*indow size, and *n*egative samples.

Dataset	ρ_p	*d,e,w,n* (ρ_p)	ρ_s	*d,e,w,n* (ρ_s)
MayoSRS	0.57*	100,30,30,1	0.58*	100,30,30,1
MiniMayoSRS coders	0.78*	100,30,30,1	0.84*	100,30,30,1
MiniMayoSRS physicians	0.78*	100,30,30,1	0.80*	100,30,30,1
UMNSRS_rel	0.49*	50,25,5,5	0.50*	100,30,10,1
UMNSRS_sim	0.60*	100,30,10,1	0.59*	100,30,15,1

Table 8: Best performance obtained by any SG model among those tested (* indicates statistical significance). Columns as in Table 7.

for up to 200 epochs, saving the embeddings at epochs $\{1, 2, 5, 10, 15, 30, 50, 75, 100, 150, 200\}$. To optimize the training process, we implemented early stopping after 30 epochs if loss reached a plateau. We evaluated a total of 975 word embeddings.

Among the top-10 best-performing models across all datasets, SG occurs 40 times, whereas CBOW only 10 times. Analyzing the final epoch of the top-10 models, word2vec typically peaked at 25 or 30 epochs in 40 cases, while in 10 cases the best performance was reached in a single training epoch. Within this group, the most frequent value for the number of negative samples is equal to one.

Table 6 shows the two correlation coefficients of the best model for each dataset. Table 7 and Table 8 provide more details and show the best performance reached by, respectively, CBOW and SG embeddings. Our models demonstrate higher and statistically significant Pearson and Spearman correlations than the baseline model across all datasets. Furthermore, these correlations were determined using a larger set of term pairs than the baseline embeddings.

While we recognise the need for deeper and more extensive experimentations to draw concrete conclusions, the correlation values we obtained are consistent with previously published results obtained using larger or more complex corpora. Furthermore, the results underline the effectiveness of our new corpus in computing word2vec embeddings that capture the medical semantics of the word. We conclude with some examples that give additional highlights of the semantics stored in the embeddings. Similarity searches and word analogies give mixed results, some of the good ones being:

- most similar to *dottore* (doctor) and *bocca* (mouth) = *dentista* (dentist);
- most similar to *vitamina* (vitamine) and *ossa* (bones) = food supplement for the prevention of osteoporosis;
- most similar to *bambini* (children) and *medico* (physician) = *pediatra* (pediatrician);
- *antibiotico* (antibiotic) is to *batterio* (bacteria) as *antivirale* (antiviral) is to x, we get x = *virus*;
- *aerosol* (aerosol) is to *polmoni* (lungs) as x is to *occhi* (eyes), we get x = *collirio* (eye drops).

5. Conclusions and Future Work

We have introduced two novel resources tailored for NLP applications within the medical domain. The newly created text corpus, albeit smaller compared to similar datasets in other languages, enables the training of word2vec models that seem to capture the semantics of medical terms. Additionally, we have developed a reliable and automated procedure to translate into Italian widely-used resources for evaluating medical embeddings. Our preliminary results are encouraging, but a larger corpus and more comprehensive experiments are still required. A clear indication that we need more documents is the absence of the word *uomo* (*man*) in the vocabulary extracted from our corpus. We are currently collecting and structuring additional data both to improve the embeddings and to build resources for downstream tasks.

Data and code availability

We published the source code and the resources we are permitted to share on the repository at https://github.com/med-nlp/italian-medical-word-embeddings.

Acknowledgments

This work was partially supported by the H2020 DeepHealth Project (GA No. 825111), by TAILOR EU Horizon 2020 research and innovation programme under (GA No. 952215), by the H2020 STARWARS Project (GA No. 101086252). We wish to thank Centro Servizi CNR of the ICT-SAC Department of the National Research Council for the precious computing services and resources they made available. We wish to address a special thanks to Ing. Giorgio Bartoccioni (ICT-SAC) for his technical support.

6. Bibliographical References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 166–174, Berlin, Germany. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Emeric Dynomant, Romain Lelong, Badisse Dahamna, Clément Massonaud, Gaétan Kerdelhué, Julien Grosjean, Stéphane Canu, and Stefan Darmoni. 2018. Word embedding for french natural language in healthcare: a comparative study (preprint). *JMIR Medical Informatics*, 7.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Faiza Khan Khattak, Serena Jeblee, Chloé Pou-Prom, Mohamed Abdalla, Christopher Meaney, and Frank Rudzicz. 2019. A survey of word embeddings for clinical text. *Journal of Biomedical Informatics*, 100.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality.

National Library of Medicine (US). 2009. *UMLS® Reference Manual [Internet]*. Bethesda (MD). 2, Metathesaurus. [Updated 2021 Aug 20].

Serguei Pakhomov, Bridget McInnes, Terrence Adam, Ying Liu, Ted Pedersen, and Genevieve B. Melton. 2010. Semantic similarity and relatedness between clinical terms: An experimental study. In *AMIA Annual Symposium Proceedings*, pages 572–576. American Medical Informatics Association.

Serguei V.S. Pakhomov, Greg Finley, Reed McEwan, Yan Wang, and Genevieve B. Melton. 2016. Corpus domain effects on distributional semantic modeling of medical terms. *Bioinformatics*, 32(23):3635–3644.

Serguei V.S. Pakhomov, Ted Pedersen, Bridget McInnes, Genevieve B. Melton, Alexander Ruggieri, and Christopher G. Chute. 2011. Towards a framework for developing semantic relatedness reference standards. *Journal of Biomedical Informatics*, 44(2):251–265.

Ted Pedersen, Serguei V.S. Pakhomov, Siddharth Patwardhan, and Christopher G. Chute. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Xin Rong. 2016. Word2vec Parameter Learning Explained. (arXiv:1411.2738).

Felipe Soares, Marta Villegas, Aitor Gonzalez-Agirre, Martin Krallinger, and Jordi Armengol-Estapé. 2019. Medical word embeddings for Spanish: Development and evaluation. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 124–133, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

François Torregrossa, Robin Allesiardo, Vincent Claveau, Nihel Kooli, and Guillaume Gravier. 2021. A survey on training and evaluation of word embeddings. *International Journal of Data Science and Analytics*, 11(2):85–103.

Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C.-C. Jay Kuo. 2019. Evaluating word embedding models: Methods and experimental results. *APSIPA Transactions on Signal and Information Processing*, 8(1).

Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. 2018. A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*, 87:12–20.

Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2020. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 23–30. Association for Computational Linguistics.

Zhang Yijia, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong lu. 2019. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific Data*, 6.