applied sciences

MDPI

Article

Examining the Effect of the Ratio of Biomedical Domain to General Domain Data in Corpus in Biomedical Literature Mining

Ziheng Zhang, Feng Han, Hongjian Zhang, Tomohiro Aoki and Katsuhiko Ogasawara *

Graduate School of Health Science, Hokkaido University, Sapporo 060-0808, Japan; zhangzihengstudy@yahoo.co.jp (Z.Z.); hanfenghokudai@gmail.com (F.H.); zhanghongjianruanjian@yahoo.co.jp (H.Z.); tomohiro_aoki@hs.hokudai.ac.jp (T.A.)
* Correspondence: oga@hs.hokudai.ac.jp; Tel.: +81-011-706-3409

Abstract: Biomedical terms extracted using Word2vec, the most popular word embedding model in recent years, serve as the foundation for various natural language processing (NLP) applications, such as biomedical information retrieval, relation extraction, and recommendation systems. The objective of this study is to examine how changes in the ratio of the biomedical domain to general domain data in the corpus affect the extraction of similar biomedical terms using Word2vec. We downloaded abstracts of 214,892 articles from PubMed Central (PMC) and the 3.9 GB Billion Word (BW) benchmark corpus from the computer science community. The datasets were preprocessed and grouped into 11 corpora based on the ratio of BW to PMC, ranging from 0:10 to 10:0, and then Word2vec models were trained on these corpora. The cosine similarities between the biomedical terms obtained from the Word2vec models were then compared in each model. The results indicated that the models trained with both BW and PMC data outperformed the model trained only with medical data. The similarity between the biomedical terms extracted by the Word2vec model increased when the ratio of the biomedical domain to general domain data was 3:7 to 5:5. This study allows NLP researchers to apply Word2vec based on more information and increase the similarity of extracted biomedical terms to improve their effectiveness in NLP applications, such as biomedical information extraction.

Keywords: biomedical literature mining (BLM); natural language processing (NLP); Word2vec

Citation: Zhang, Z.; Han, F.; Zhang, H.; Aoki, T.; Ogasawara, K. Examining the Effect of the Ratio of Biomedical Domain to General Domain Data in Corpus in Biomedical Literature Mining. *Appl. Sci.* **2022**, *12*, 154. https://doi.org/10.3390/app12010154

Academic Editor: Evgeny Nikulchev

Received: 11 November 2021
Accepted: 21 December 2021
Published: 24 December 2021

1. Introduction

Owing to the rapid development of biomedical research, there is a large number of biomedical publications available online in an electronic format, and this number is increasing every year. For example, the number of articles in PubMed, a biomedical literature database, is increasing by approximately 1 million documents each year [1]. Biomedical reports, which contain valuable information on new discoveries and knowledge, have been continually added to the overwhelming amount of literature. Because medical literature publications contain a wealth of biomedical information, using publication data to solve a variety of biomedical problems, such as relationship extraction, has become a popular method in recent years [2,3]. As a result, there is a high demand for automatic knowledge extraction from biomedical literature highly demanding.

When processing large amounts of unlabeled unstructured data, such as treatises, word embeddings technology [4,5] is an ideal approach for obtaining semantic relationships between words. Word embeddings were introduced to represent a single word as a low-dimensional vector that captures the frequencies of co-occurring adjacent words, and the mathematical similarity between the word vectors captures the similarity in meaning between the words. Compared to ontology-based approaches, such as the Unified Medical Language System (UMLS) and WordNet [6–8], word embedding technology has the following advantages: (1) it saves time and resources because it does not require human

involvement; (2) it can analyze big data and produce results that humans are incapable of producing; (3) up-to-date results are available by feeding up-to-date corpora. The Word2vec [4,5] model proposed by Mikolov in 2013 is one of the most popular word embedding techniques because of its excellent term extraction performance [9]. The extracted terms form the basis of the application of biomedical natural language processing (NLP), such as medical information retrieval, relation extraction, and recommendation systems. More recently, deep masked language models pretrained with unsupervised learning have achieved state-of-the-art in a lot of NLP-related tasks. Bidirectional Encoder Representations from Transformers (BERTs) [10] were the first to apply the transformer architecture as a general framework for NLP. The pretrained model is available and can be fine-tuned for a different study. Some studies, like BioBERT [11], already pretrained such models based on BERTs using biomedical texts, fine-tuned them for downstream tasks, and performed better than most previous biomedical NLP models.

Previous studies in the biomedical field in recent years have investigated how to improve the performance of the Word2vec model in extracting similar biomedical terms. First, researchers studied the setting of model parameters [12–14]. Subsequently, studies were conducted from a data perspective. For example, a study by Pakhomov et al. [15] examined the corpus domain effects on measuring semantic relatedness and similarities between biomedical terms (i.e., clinical notes, PubMed Central (PMC) articles, and Wikipedia), and the results showed that the model trained on the PMC dataset was superior to that trained on clinical notes, and the performance of the model did not increase when the dataset reached a certain size (100 million words). Zhu [16] studied data size, recency, and section of publications, and their results for data size are consistent with those of Pakhomov [15]; they concluded that increasing the size of the dataset does not always enhance the performance. There is a certain point at which the performance reaches its peak, and as more data are added, more noise starts to affect the performance. These two studies demonstrate that large amounts of data cannot always guarantee effective word embedding in biomedical NLP. However, few studies have directly compared the effectiveness of word embedding in different resources (e.g., medical and general domains) [9]. In addition, an interesting phenomenon appeared in a previous study: the results of Habibi [17] showed that the combination of domain-specific and domain-unspecific data achieved the best performance (i.e., the results of Wiki + PubMed + PMC data were better than those of PubMed + PMC data). The reason for this was thought to be that the Wiki-PubMed-PMC dataset is larger than the PubMed-PMC dataset; however, as a limitation, no further investigation has been conducted in this regard. In fact, in the study by Habibi [17], the sizes of these two datasets exceeded the specific size reported in [16], where the model's performance reached its peak. In other words, the reason why the combined medical and non-medical data achieved the best performance is not the increase in data size, but other factors. In this study, we consider that the cause of the above results [17] is not an increase in data size, but rather the addition of a certain percentage of general domain data.

Therefore, the objective of this study is to examine how changes in the ratio of the biomedical domain to general domain data in the corpus affect the extraction of similar biomedical terms using Word2vec. The results of this study will help biomedical researchers apply Word2vec in a more informed manner and improve the method's performance in extracting similarity and relatedness information from biomedical publication data.

2. Materials and Methods

2.1. Workflow

The flow of the method of this study is illustrated in Figure 1.

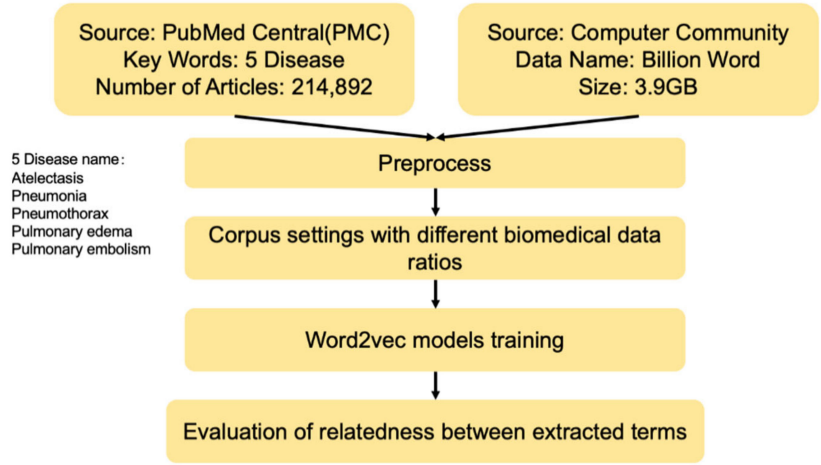

Figure 1. Flow of method of this study.

2.2. Data

2.2.1. Biomedical Data

PMC articles pertaining to five diseases (atelectasis, pneumonia, pneumothorax, pulmonary edema, and pulmonary embolism: these diseases are the subject of research by the research group) were accessed using the disease names as keywords, and 214,892 abstracts from the past five years were downloaded and used as biomedical domain data. Table 1 lists the specific number of abstracts for each disease.

Table 1. Number of abstracts for each disease.

Disease	Number of Abstracts
Atelectasis	8128
Pneumonia	128,086
Pneumothorax	13,307
Pulmonary edema	36,915
Pulmonary embolism	28,456
Total	214,892

2.2.2. General Data

Chelba [18] obtained the benchmark data Billion Word (BW) from the computer community, which contains approximately 1 billion words (3.9 GB) for statistical language modeling, as general domain data. For this BW standard data, text data obtained from the WMT11 [19] website were used, duplicates were deleted, and all words with less than three were discarded. Furthermore, it was obtained by performing processing such as randomizing the order of sentences.

2.3. Preprocessing

The downloaded initial treatise data are expected to contain several factors that can reduce the training performance; therefore, it is necessary to preprocess the initial data. The tools used were the NLTK library [20] on the Python 3.6 platform and regular expressions. The content that required to be deleted was that other than the body of the abstract, such as author information, URLs, references, symbols, and extra spaces. All numbers (integer or non-integer) were replaced by num. Then, the English alphabets were converted to lowercase and removed the extra spaces and line breaks were removed. In addition, stop words were set to exclude words that are useless or that adversely affect processing results (including words such as "introduction" and "background"). Finally, "Pulmonary edema"

was changed to "Pulmonaryedema" and "Pulmonary embolism" to "Pulmonaryembolism" because the Word2vec model can vectorize only a single word at a time. The BW data were preprocessed similarly.

2.4. Corpus Settings with Different Ratio of Biomedical Data

To examine the effect of the ratio of biomedical data in the corpus, first, using the stratified sampling method, one-tenth of the PMC data for each of the five diseases was randomly extracted, and combined into a single dataset. Thus, 10 datasets were generated (hereinafter referred to as PMC 1–10). Each PMC dataset contained 21,488 abstracts with a data size of approximately 31 MB. To prevent the influence of data size, 10 data sets with the same data size of approximately 31MB were extracted from the BW data (hereinafter referred to as BW 1–10).

Subsequently, a total of 10 datasets (BW datasets + PMC datasets = 10) were selected from the BW and PMC datasets based on the ratio ranging from 0:10 (BW:PMC) to 10:0, and 11 corpora were generated.

2.5. Word2vec

The 11 corpora obtained in the previous step were sorted and trained using the Word2vec model. Specifically, the BW data in the corpus were first input to the Word2vec model for training, the weight of the obtained model was used as the pre-training weight, and the PMC dataset remaining in the corpus was input to the model for retraining. For example, when the ratio was 8:2, BW 1 to 8 were input to Word2vec for training, and the obtained model weight was used as the pre-training weight, and PMC 1 to 2 were input to the model for retraining. With 11 corpora, 11 trained models were obtained.

The parameter settings of the model were set in accordance with Chiu [14], as the biomedical publication reported the optimal parameter settings for Word2vec after a number of detailed experiments. Specifically, the model used was Skip-Gram, the number of dimensions was 200, the window size was 30, and the optimization method was the Negative Sampling.

2.6. Evaluation of Model' Performance

The cosine similarity between word vectors ([0, 1] interval) was used to measure the similarity between two given biomedical terms. A value of 1 indicates that the terms are exactly the same, and 0 indicates that they are completely different. The five disease names were individually input into the model trained with only medical data (0:10), and the terms with the highest cosine similarity were output. Input words and output words were input separately into the other 10 models as medical term pairs, and the similarity was calculated and compared.

3. Results

3.1. Five Biomedical Term Pairs

In a model trained with medical data only (0:10), the pairs of five diseases and similar words with the highest individual similarity are shown in Table 2. Thus, five biomedical term pairs were obtained.

Table 2. Similar terms with the highest degree of similarity correspond to five diseases in the 0:10 ratio model.

Pair	Five Disease	Term with the Highest Degree of Similarity	Similarity
A	Atelectasis	Lobar	0.578
B	Pneumonia	Pneumonias	0.665
C	Pneumothorax	Pneumothoraxes	0.737
D	Pulmonary edema	Cardiogenic	0.608
E	Pulmonary embolism	Thromboembolism	0.782

3.2. Five Pairs Measure Similarity in Each Model

The five pairs were respectively input to the other ten models, and the obtained similarity of the pairs is shown in Figure 2. The vertical axis represents the cosine similarity of the pair, and the horizontal axis represents the data ratio (BW: PMC) of the model. The term was excluded because it could not be identified in the model with a ratio of 10:0.

Four pairs (pair A, pair B, pair C, and pair E) had higher similarity in the model trained with both domain data than in the model trained with biomedical data alone. On the contrary, the similarity of the three pairs (pair C, pair D, and pair E) in the model with a ratio of 9:1 was close to or higher than that in the 0:10 model.

Specifically, the similarity of pair A was 0.578 in the model trained only on biomedical data (ratio 0:10) and reached a peak of 0.580 in the model with a ratio of 5:5. As the proportion of general data increased above this point, the similarity began to decline significantly.

The similarity of pair B was 0.665 in the model trained only on biomedical data (ratio 0:10) and reached a peak of 0.687 in the model with a ratio of 4:6. As the ratio of general data increased above this point, the similarity began to decline significantly.

The similarity of pair C was 0.737 in the model trained only on biomedical data, reaching the similarity peak in this model, but in the case of the similarity in other models, the difference was small.

The similarity of pair D was 0.608 in the medically trained model (ratio 0:10) and reached a peak of 0.628 in the model with a ratio of 9: 1. In addition, the similarity between the 2:8 and 8:2 models was higher than that of the 0:10 model. The similarity in the other models was approximately the same.

The similarity of pair E was 0.782 in the medically trained model (ratio 0:10) and reached a peak of 0.793 in the 3:7 ratio model. Furthermore, it was revealed that the similarity in the model with a ratio of 2:8 was also higher than that in the model with a ratio of 0:10. The similarity in the other models was approximately the same.

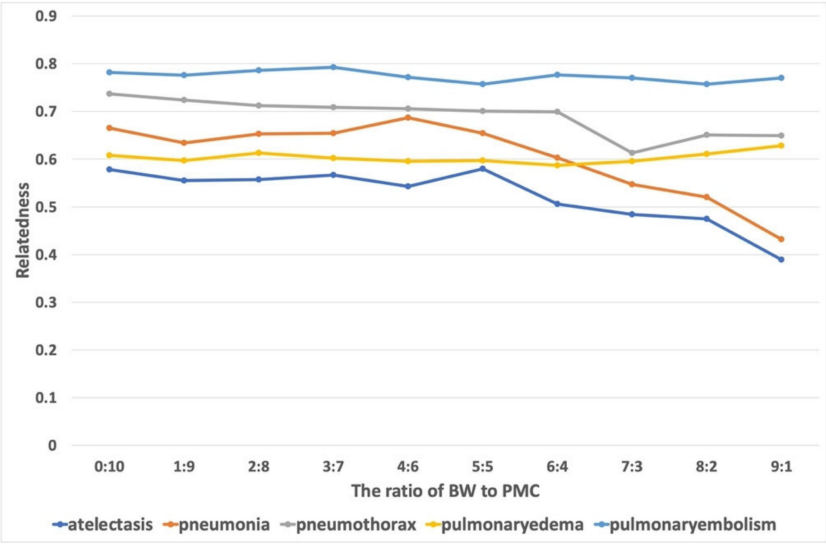

Figure 2. Similarity of five pairs in models.

4. Discussion

4.1. Principal Results

This study investigated how changes in the ratio of the biomedical domain to general domain data in the corpus affect the extraction of similar biomedical terms using Word2vec. This study is the first to compare the effectiveness of word embedding in terms of the impact of the ratio of biomedical domain data within the corpus. The Word2vec model was trained by setting a corpus with different ratios of PMC data to BW data, and the

performance of similar term extraction of the 11 models was measured. The results show that the performance of the model trained on biomedical domain and general domain data is better than that of the model trained only on biomedical domain data; thus, the effect of the ratio of biomedical data on the extraction of similar biomedical terms was revealed. In this study, the interval with the best ratio (BW: PMC) for the extraction of similar biomedical terms is approximately 3:7 to 5:5.

4.2. Comparison with Prior Work

Prior to this study, several researchers studied the effect of the corpus domain on the similarities between biomedical terms extracted by Word2vec.

Chen et al. [21] used health-related Wikipedia articles and general Wikipedia articles to train multiple word embeddings to evaluate performance in analogical and health-related relationship search tasks and compare word embedding performance. They evaluated whether the results depended on the domain of the text corpus. The results showed that word embeddings trained on health-related Wikipedia articles performed better on health-related relationship search tasks than those trained on general Wikipedia articles. This is consistent with the results of the present study. The results of this study show that models trained solely on biomedical data performed better in extracting biomedical terms than models trained solely on general domain data (five pairs cannot be recognized by the models with a ratio of 10:0).

Pakhomov [15] also investigated the corpus domain effects on semantic similarity and relatedness between biomedical terms (e.g., clinical notes, PMC articles, and Wikipedia). The 2010–2014 clinical notes obtained from the Fairview Health System were used. The PMC and Wikipedia datasets both contain all the data available as of September 2015. Modified versions of UMNSRS-Rel and UMNSRS-Sim were used as reference standards. The results showed that the model trained on the PMC dataset outperformed the one trained on the clinical notes (i.e., 0.62 vs. 0.60 for similarity and 0.58 vs. 0.57 for relatedness). This demonstrated the value of publication data for measuring semantic similarity and relatedness between biomedical terms. This is one of the reasons why the biomedical data used in this study were PMC data. This is significant, because access to clinical data is highly restricted, and difficult, to protect the confidentiality and security of patient health information. The PMC's open access biomedical articles have no such restrictions and are freely and easily accessible to all. In addition, PMC has many different research topics pertaining to new biomedical knowledge. Another important finding in the work of Pakhomov is that increasing the size of the corpus beyond a certain size (66 million tokens) would not enhance the performance of the model.

The same result was found in a study by Zhu [16], where there was a certain point at which the performance reached its peak and as more data were added, more noise started to affect the performance. The reason was thought to be that as more data were added, the number of vocabularies increased, and more noise was introduced.

Additionally, according to the results of Habibi [17], the combined medical and non-medical data achieved the best performance (i.e., the results of Wiki + PubMed + PMC data were better than those of PubMed + PMC data). The reason was thought to be that the Wiki-PubMed-PMC dataset was larger than the PubMed-PMC dataset; however, as a limitation, no further investigation has been conducted in this regard. In fact, in the study by Habibi [17], the sizes of these two datasets exceeded the specific size reported in [16], at which the model's performance reached its highest level. In other words, the reason why the combined biomedical and non-biomedical data reached the peak performance is not the increase in data size. According to the results of this study, the addition of a certain proportion of general domain data leads to the results reported by Habibi [17]. Some of these previous studies investigated (or did not investigate) the effect of the domain of the corpus on the semantic similarity between biomedical terms; however, the main difference from this study is that it is the first to compare the effectiveness of word embeddings from the perspective of the effect of the ratio of biomedical domain data.

4.3. Other Potential Factors That May Affect the Results

It was discussed whether other factors contributed to the results of this study. (1) Data size: The results of Habibi [17] were thought to be due to an increase in data size, and Pakhomov [15] showed an effect of the data size; however, in this study, the corpus sizes in 11 models were the same. Moreover, although the number of articles on the five diseases (Table 1) differed, this factor was excluded because the required relationship with the results could not be observed. (2) Amount of vocabulary: Because we considered that importing data from other domains will increase the number of vocabularies, we have stated the number of vocabularies in each model. The number of vocabularies in each model is shown in Figure 3. According to Figure 3, the number of vocabularies included in the model with a ratio of 7:3 was the largest; however, there was no necessary relationship between the interval (3:7 to 5:5) with the best ratio for extraction of similar medical words. Thus, this factor was also eliminated. Therefore, it is considered that the cause is the addition of a certain proportion of general domain data.

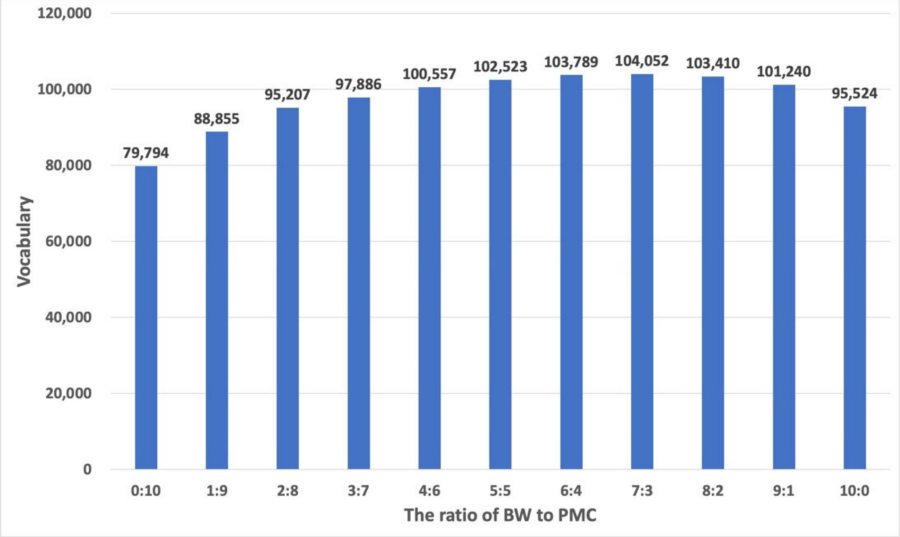

Figure 3. Vocabulary number statistics.

4.4. Limitation

This study has some limitations. (1) Difficulty in identifying the specific best ratio: It was proved that the performance of extraction of medical analogs can be improved by adding general data in a certain ratio; however, the specific ratio of the best is still unknown. The similarity was measured with the pairs of input words and output words; however, the gold standard provided by experts was not used in this study. Moreover, measurements with different gold standards produce different results and are difficult to measure. (2) Research target: Because only five diseases were selected, it is not possible to predict the results for expressions in other diseases. In addition, the size of the corpus is smaller, and using a larger corpus would increase the reliability of the results. (3) Data type: The data used in this study were PMC and BW; however, it cannot be predicted whether the results would be different if other corpora are used. In addition, the BW data statements were randomized, which could adversely affect the results. In the future, we will evaluate word embeddings trained on more corpus, both medical domain, and general domain, as we assume that different results may appear on different resources or different sizes of the training corpus. We also would like to further assess these results, such as using the gold standard provided by experts or using the downstream biomedical NLP applications like medial named entity recognition and clinical note summarization.

5. Conclusions

This study revealed the effect of the ratio of biomedical data on the extraction of biomedical terms in medical literature mining. When general domain data are added in a ratio between 3:7 to 5:5, the similarity of the biomedical terms extracted by the Word2vec model increases. The results of this study will help biomedical researchers apply Word2vec in a more informed manner and improve the method's performance in extracting similarity and relatedness information from biomedical publication data.

Author Contributions: Conceptualization, Z.Z. and F.H.; methodology, Z.Z.; software, Z.Z. and H.Z.; validation, Z.Z.; formal analysis, K.O.; investigation, Z.Z.; resources, Z.Z.; data curation, Z.Z.; writing—original draft preparation, Z.Z.; writing—review and editing, Z.Z. and T.A.; supervision, T.A. and K.O.; project administration, Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Statistical Reports on MEDLINE®/PubMed® Baseline Data. Available online: https://www.nlm.nih.gov/bsd/licensee/baselinestats.html (accessed on 30 April 2021).
2. Frijters, R.; van Vugt, M.; Smeets, R.; van Schaik, R.; de Vlieg, J.; Alkema, W. Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Comput. Biol.* **2010**, *6*, e1000943. [CrossRef] [PubMed]
3. Zhu, Y.; Song, M.; Yan, E. Identifying liver cancer and its relations with diseases, drugs, and genes: A literature-based approach. *PLoS ONE* **2016**, *11*, e0156091. [CrossRef] [PubMed]
4. Mikolov, T.; Chen, K.; Corrado, G.; Deal, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
5. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed representations of words and phrases and their compositionality. *arXiv* **2013**, arXiv:1310.4546.
6. Pesquita, C.; Faria, D.; Falcão, A.O.; Lord, P.; Couto, F.M. Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.* **2009**, *5*, e1000443. [CrossRef] [PubMed]
7. Sánchez, D.; Batet, M.; Isern, D.; Valls, A. Ontology-based semantic similarity: A new feature-based approach. *Expert Syst. Appl.* **2012**, *39*, 7718. [CrossRef]
8. Hadj Taieb, M.A.; Ben Aouicha, M.; Ben Hamadou, A. A new semantic relatedness measurement using WordNet features. *Knowl. Inf. Syst.* **2014**, *41*, 467. [CrossRef]
9. Wu, S.; Roberts, K.; Datta, S.; Du, J.; Ji, Z.; Si, Y.; Soni, S.; Wang, Q.; Wei, Q.; Xiang, Y.; et al. Deep learning in clinical natural language processing: A methodical review. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 457. [CrossRef] [PubMed]
10. Devlin, J.; Ming-Wei, C.; Kenton, L.; Kristina, T. BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 24 May 2019; Volume 1, pp. 4171–4186.
11. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. Biobert: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef] [PubMed]
12. Minarro-Giménez, J.A.; Marín-Alonso, O.; Samwald, M. Exploring the application of deep learning techniques on medical text corpora. *Stud. Health Technol. Inform.* **2014**, *205*, 584. [PubMed]
13. Muneeb, T.H.; Sahu, S.K.; Anand, A. Evaluating distributed word representations for capturing semantics of biomedical concepts. In Proceedings of the 2015 Workshop on Biomedical Natural Language Processing, Beijing, China, 30 July 2015; pp. 158–163.
14. Chiu, B.; Crichton, G.; Korhonen, A.; Pyysalo, S. How to train good word embeddings for biomedical NLP. In Proceedings of the 15th Workshop on Biomedical Natural Language Processing, Berlin, Germany, 12 August 2016; pp. 166–174.
15. Pakhomov, S.V.S.; Finley, G.; McEwan, R.; Wang, Y.; Melton, G.B. Corpus Domain effects on distributional semantic modeling of medical terms. *Bioinformatics* **2016**, *32*, 3635. [CrossRef] [PubMed]
16. Zhu, Y.; Yan, E.; Wang, F. Semantic relatedness and similarity of biomedical terms: Examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. *BMC Med. Inform. Decis. Mak.* **2017**, *17*, 95. [CrossRef] [PubMed]
17. Habibi, M.; Weber, L.; Neves, M.L.; Wiegandt, D.L.; Leser, U. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* **2017**, *33*, i37. [CrossRef] [PubMed]

18. Chelbaa, C.; Mikolov, T.; Schuster, M.; Ge, Q.; Brants, T.; Koehn, P.; Robinson, T. One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. In Proceedings of the 15th Annual Conference of the International Speech Communication Association, Interspeech 2014, Singapore, 14–18 September 2014; pp. 2635–2639.
19. EMNLP 2011 Sixth Workshop on Statistical Machine Translation. Available online: http://www.statmt.org/wmt11/ (accessed on 30 April 2021).
20. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*; O'Reilly Media: Sebastopol, CA, USA, 2009.
21. Chen, Z.; He, Z.; Liu, X.; Bian, J. Evaluating semantic relations in neural word embeddings with biomedical and general domain knowledge bases. *BMC Med. Inform. Decis. Mak.* **2018**, *18* (Suppl. S2), 65.