# Integrating word embedding neural networks with PubMed abstracts to extract keyword proximity of chronic diseases

Ahmad P. Tafti
*Department of Health Sciences Research*
*Mayo Clinic*
Rochester, USA
tafti.ahmad@mayo.edu

Yanshan Wang
*Department of Health Sciences Research*
*Mayo Clinic*
Rochester, USA
wang.yanshan@mayo.edu

Feichen Shen
*Department of Health Sciences Research*
*Mayo Clinic*
Rochester, USA
shen.feichen@mayo.edu

Elham Sagheb
*Department of Health Sciences Research*
*Mayo Clinic*
Rochester, USA
saghebhosseinpour.elham@mayo.edu

Paul Kingsbury
*Department of Health Sciences Research*
*Mayo Clinic*
Rochester, USA
kingsbury.paul1@mayo.edu

Hongfang Liu
*Department of Health Sciences Research*
*Mayo Clinic*
Rochester, USA
liu.hongfang@mayo.edu

***Abstract*—Chronic diseases are a leading cause of morbidity and mortality worldwide. They are common enough to affect large numbers of patients, and the chronic nature makes them costly to both patients and healthcare providers. Diagnosis of many chronic diseases is challenged by variability in their clinical manifestations. Although chronic diseases bear a set of structured terminology aiming to standardize nomenclature of the presentation and outcomes of the disease, in practice there is a wide spectrum of terminology associated with these diseases across different venues such as clinical notes, biomedical literature, and health-related social media. Among these sources, the scientific articles published in the biomedical literature usually follow principled approaches to terminology and are thus especially valuable for extracting diseases keywords. Given the fact that it is very costly and time-consuming to manually extract disease terminology from a large column of scientific articles, we aim to utilize artificial neural network strategies to automatically extract vocabularies associated with a set of chronic diseases. Our finding indicates the feasibility of developing word embedding neural nets for autonomous keyword extraction and abstraction of chronic diseases.**

***Index Terms*—Word embeddings, Word2vec, GloVe**

## I. Introduction

A vast amount of biomedical text data is distributed over a variety of data sources, including scientific articles, health-related social media, clinical narratives, and other natural language portions of EHRs, just to name a few. They cover valuable terms and phrases that could be used to accurately identify patients who suffer from different chronic diseases, such as diabetes, obesity, asthma, osteoporosis, and lupus. The cost and effort required to manually mine terminologies for chronic diseases is well-known, thus an automated system is needed to first extract the associated terms and then improve such vocabularies in a real time manner. Natural language processing (NLP) combined with machine learning algorithms have the power to extract these terms from unstructured text data, such as scientific articles and clinical notes. In recent years, word embeddings have been added to clinical text analytics as a powerful document representation model to address a variety of problems in named entity recognition [1], [2], text classification [3], [4], text summarization [5], [6], and bioNLP [7], [8]. These recent advances offer the opportunity to build efficient artificial neural networks on top of the large body of scientific articles to automatically detect, identify, and characterize terms relevant to chronic diseases. Automatic identification of chronic disease terms improves our understanding of such diseases and allows earlier and more precise diagnosis, prevention, and treatment. It motivates the main contribution of the current study which aims to combine context-based and count-based word embeddings, including Word2Vec [9], [10] and GloVe [11] together, and integrate them with large-scale analysis of PubMed abstracts to automatically calculate the semantic proximity of keywords for chronic diseases.

The organization of the paper is as follows. Section II explains the materials and methods. Experimental validation, performance analysis, and scientific visualization are presented in Section III. Section IV concludes the study.

## II. Materials and Methods

We first begin with a brief explanation of word embeddings, and then describe the proposed processing pipeline. Word embeddings refer to those computational text mining methods that map any words in a given corpus to vectors of real numbers. For example, applying word2vec word embedding on a given medical text corpus (e.g., clinical notes), represents the word “diabetes” as a vector of [0.2, -1.3, 1.5, 0.7]. This representation is quite important in artificial neural nets since

978-1-7281-0848-3/19/$31.00 ©2019 IEEE

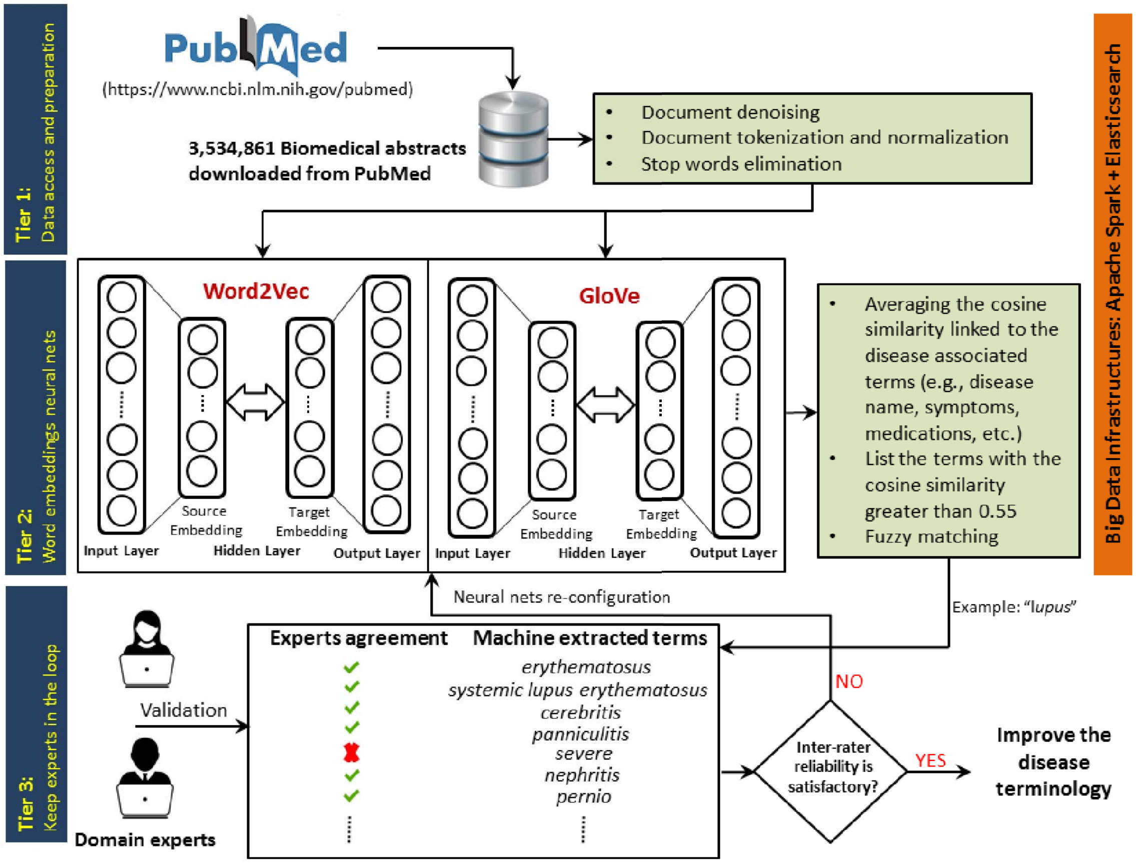

Tier 1: Data access and preparation
PubMed
(https://www.ncbi.nlm.nih.gov/pubmed)
3,534,861 Biomedical abstracts downloaded from PubMed
• Document denoising
• Document tokenization and normalization
• Stop words elimination
Tier 2: Word embeddings neural nets
Word2Vec
Source Embedding
Target Embedding
Input Layer
Hidden Layer
Output Layer
GloVe
Source Embedding
Target Embedding
Input Layer
Hidden Layer
Output Layer
• Averaging the cosine similarity linked to the disease associated terms (e.g., disease name, symptoms, medications, etc.)
• List the terms with the cosine similarity greater than 0.55
• Fuzzy matching
Big Data Infrastructures: Apache Spark + Elasticsearch
Neural nets re-configuration
Example: "lupus"
Tier 3: Keep experts in the loop
Validation
Domain experts
Experts agreement
Machine extracted terms
✓ erythematosus
✓ systemic lupus erythematosus
✓ cerebritis
✓ panniculitis
✗ severe
✓ nephritis
✓ pernio
NO
Inter-rater reliability is satisfactory?
YES
Improve the disease terminology

Fig. 1. The proposed software architectural model to extract keyword proximity of chronic diseases using word2vec and GloVe.

their architectures only process continuous numbers versus characters. There are mainly two methods in word embeddings: (1) *context-based*, which falls into supervised learning strategies, where given a corpus, an attempt will be made to build a predictive model to predict the target words, and (2) *count-based*, which mostly accounts for word frequency and it works as an unsupervised learning method [12].

### A. Word2vec and GloVe

Word2vec [9], [10] and GloVe [11], [12] neural network-based word embeddings that learn geometrical vectors of words within a document, and have shown promising results in a variety of applications, such as named entity recognition [1], [13], text classification [4], [16], and sentiment analysis [14], [15]. Instead of only capturing the word intensities across a corpus, they are also able to capture word order, higher-level syntax, and semantics. The general purposes of word embeddings, such as word2vec and GloVe are: (1) to create an input for machine learning methods, (2) to find nearest neighbors in the embedding space, and (3) to visualize the concepts and relations among words. GloVe is a count-based word embedding, while word2vec is a context-based method and has been widely used as a predictive model. Word2vec is itself grouped into two different learning methods, namely continuous bag-of-words (CBOW) and skip-gram models. The CBOW can predict a target word given a context, while conversely skip-gram predicts a target context given a word. Both models try to minimize a well-defined loss function (e.g., hierarchical softmax, full softmax, or noise contrastive estimation). For example, using word2vec skip-gram model, one loss function can be the full softmax, thus the very final output layer will apply softmax to estimate the probability of predicting the output word $W_{out}$ given $W_{in}$, as follows:

$$P(W_{out}|W_{in}) = \frac{\exp({v'_{W_{out}}}^{T} v_{W_{in}})}{\sum_{i=1}^{V} \exp({v'_{W_i}}^{T} v_{W_{in}})} \qquad (1)$$

where the embedding vector of each single word is defined by the matrix $W$, and the context vector is determined by the output matrix $W'$. Given an input word as $W_{in}$, we label the corresponding row of matrix $W$ as vector $v_{W_{in}}$, the embedding vector, and its corresponding column of $W'$ as $v'_{W_{in}}$, the context vector. In contrast, when the total size of the vocabulary is very large, a loss function such as hierarchical softmax [17] would be a better option.

In this work, we utilized skip-gram model since it suits large-scale data. GloVe works in a different way. Instead of extracting the embeddings from a neural net which is designed to predict neighbouring words, the embeddings are now optimized directly in a way that the dot product of two word vectors would be equal to the log of the frequency the two words will occur near each other. GloVe defines the co-occurrence probability as:

$$P_{co}(W_z|W_i) = \frac{C(W_i, W_z)}{C(W_i)} \tag{2}$$

where $C(W_i, W_z)$ counts the co-occurrence between two words $W_i$ and $W_z$. We employed $W_i$ and $W_z$, to differ than $P(W_{out}|W_{in})$, which is presented in equation 1. For example, if two terms as "diabetes" and "metformin" occur close to each other for 1000 times in a given a corpus, then $Vec(diabetes) \cdot Vec(metformin) = log(1000)$. This drives the vectors to encode the frequency distribution of which words lie near others. This paper does not have the space to go deeper into the details of developing word embeddings; interested readers are therefore referred to [9]–[12], [17] for further reading.

### *B. Proposed Software Architectural Model*

Figure 1 demonstrates the proposed processing pipeline to automatically extract keywords relevant to chronic diseases. This deploys across three different tiers: (1) Data access and preparation, (2) Word embeddings neural nets, and (3) Keep experts in the loop. Tier 1 is responsible to first get the data, and then perform preliminary text pre-processing steps, such as document denoising (e.g., deleting email addresses, digits, and characters like "[", "]", "%"), document tokenization and normalization (e.g., separating sentences into words, converting all of them into lowercase to ensure consistency), and finally stop word elimination to delete stop words (e.g., "a", "an", "the", "in") from the words list. In the next step, Tier 1 passes the words to Tier 2 to train both context-based and count-based word embeddings, including word2vec and GloVe. Once we trained the models, given a term, such as "lupus", the average of highest cosine distance value in the learning models generated by word2vec and GloVe is measured, and then the tier provides a list of terms with cosine similarity greater than 0.55, and does fuzzy matching to find matching phrases based on word-based matching queries from a database. For instance, given the term "lupus", the top terms with cosine similarity greater than 0.55 are "erythematosus", "systemic lupus erythematosus", "cerebritis", "panniculitis", and so on. In regard to fuzzy matching, it, for example, matches the term "SLE" to "systemic lupus erythematosus". We performed a set of experimental validations, and the result showed an average cosine similarity bigger than 0.51-0.55 tended to provide better inter-rater reliability among the experts and the proposed computational method.

Tiers 1 and 2 were developed on top of the big data infrastructures, including Apache Spark [18] and Elasticsearch [19]. In Tier 3, we keep two domain experts in the loop to validate the terms which have been automatically extracted by the proposed computational method. Tier 3 utilized Kappa measure [20] to calculate the inter-rater reliability among domain experts (human) and the method (machine), meaning that the machine and the human agreed that a pair of terms are related. If the inter-rater reliability is satisfactory, defined as kappa value higher than 0.80 meaning a strong level of agreement, then the list of extracted terms will be added to the disease terminology. If it was not satisfactory, we iteratively reconfigured the word embeddings internal parameters (e.g., epoch, window size, minimum word frequency, etc.) to get better inter-rater reliability among the experts and machine.

TABLE I
NUMBER OF ABSTRACTS IN EACH DISEASE CATEGORY.

| Chronic Disease | Number of Abstracts |
|---|---|
| Alzheimer | 69,982 |
| Arthritis | 149,383 |
| Asthma | 99,585 |
| Cancer | 2,711,246 |
| Diabetes | 275,093 |
| Lupus | 46,671 |
| Obesity | 145,288 |
| Osteoporosis | 37,613 |
| Total: 3,534,861 | |

### *C. Dataset*

We employed PubMed advanced search [21] to download 3,534,861 scientific abstracts published within eight different chronic diseases, including Alzheimer, Arthritis, Asthma, Cancer, Diabetes, Lupus, Obesity, and Osteoporosis. Number of words in the dataset was 395,904,432. The query used to download Alzheimer related scientific abstracts was as below:

```
(((alzheimer [MeSH Terms])
OR alzheimer [MeSH Major Topic]))
AND English[Language]
```

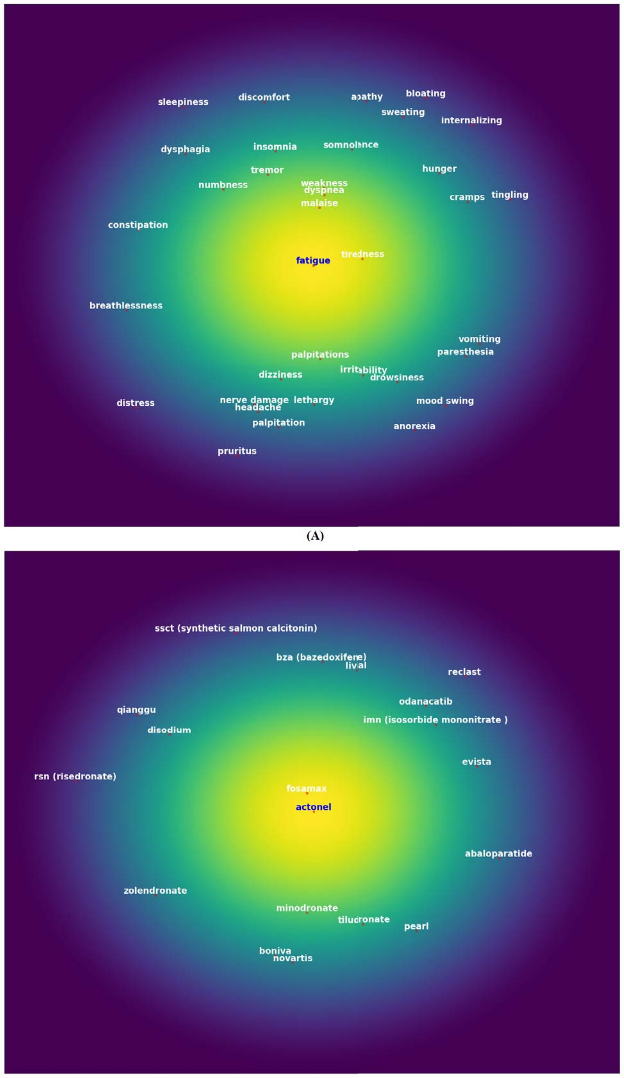
Similar queries with different disease name have been employed for other chronic diseases. The number of abstracts within each diseases category is shown in Table I.

## III. Experimental Validation and Scientific Visualization

From the computational perspective, a VM in a high performance cluster environment running a 64-bit CentOS operating system was used to run the models. All codes were implemented using Python 3.7. From the experimental validation side, an average cosine similarity bigger than 0.51-0.55 tends to provide better inter-rater reliability among domain experts and the proposed computational method. We tried different internal parameters within word2vec and GloVe (e.g., window size, embedding size, number of epochs). Utilizing the datasets illustrated in Table I, windows size of 5, epoch of 30, and iteration number of 15 to 20 help to get better inter-rater reliability. A set of scientific visualizations along with the achieved inter-rate reliabilities between the domain experts and the proposed method are shown in Figure 2.

## IV. Conclusion

A key area of high clinical interest is to automatically extract the keywords relevant to chronic diseases. In this study, we applied an integration of context-based and count-based word embeddings on a large-scale dataset downloaded from PubMed to cope with the problem. The results shows the feasibility



(A)

(B)

Fig. 2. (A) The scientific visualization results obtained by searching the term "fatigue", as one of the possible symptoms linked to diabetes disease. One can see almost all terms presented here are associated to diabetes symptoms. The inter-rater reliability between the domain experts and the terms automatically extracted by the proposed method was Kappa of 0.91 which indicates almost a perfect level of agreement [20]. (B) The scientific visualization results obtained by searching the term "actonel", as one of the possible medications uses to treat osteoporosis. One can see the terms presented here are all medications. These medications are most likely employed to treat and/or prevent osteoporosis. The inter-rater reliability between the domain experts and the terms automatically extracted by the proposed method was Kappa of 0.94 which indicates almost a perfect level of agreement.

of using these neural networks to this application area. Even though these word embedding neural nets are both relatively simple to implement and have brought excellent advancements in clinical NLP, there exists a set of limitations to this study. First, if the model has not seen a keyword term in the given corpus, it would not be able to interpret or generate a vector representation for such a term. Even though many words are similar morphologically, these models mostly represent each word as an independent vector. Finally, a cross-lingual utilization of such models is not imaginable.

## Acknowledgment

The authors declare that they have no competing interest.

## References

[1] Wu Y, Xu J, Jiang M, Zhang Y, Xu H. A study of neural word embeddings for named entity recognition in clinical text. In AMIA Annual Symposium Proceedings 2015 (Vol. 2015, p. 1326). American Medical Informatics Association.

[2] Zhao M, Masino AJ, Yang CC. A Framework for Developing and Evaluating Word Embeddings of Drug-named Entity. In Proceedings of the BioNLP 2018 workshop 2018 (pp. 156-160).

[3] Xu H, Dong M, Zhu D, Kotov A, Carcone AI, Naar-King S. Text classification with topic-based word embedding and convolutional neural networks. InProceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics 2016 Oct 2 (pp. 88-97). ACM.

[4] Tafti AP, Behravesh E, Assefi M, LaRose E, Badger J, Mayer J, Doan A, Page D, Peissig P. bigNN: an open-source big data toolkit focused on biomedical sentence classification. InBig Data (Big Data), 2017 IEEE International Conference on 2017 Dec 11 (pp. 3888-3896). IEEE.

[5] Chen Q, Sokolova M. Word2Vec and Doc2Vec in Unsupervised Sentiment Analysis of Clinical Discharge Summaries. arXiv preprint arXiv:1805.00352. 2018 May 1.

[6] Zhang L, Li J, Wang C. Automatic synonym extraction using Word2Vec and spectral clustering. InControl Conference (CCC), 2017 36th Chinese 2017 Jul 26 (pp. 5629-5632). IEEE.

[7] Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, Kingsbury P, Liu H. A comparison of word embeddings for the biomedical natural language processing. Journal of biomedical informatics. 2018 Nov 1;87:12-20.

[8] Moen SP, Ananiadou TS. Distributional semantics resources for biomedical text processing. In Proceedings of the 5th International Symposium on Languages in Biology and Medicine, Tokyo, Japan 2013 (pp. 39-43).

[9] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013 Jan 16.

[10] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. InAdvances in neural information processing systems 2013 (pp. 3111-3119).

[11] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. InProceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) 2014 (pp. 1532-1543).

[12] Baroni M, Dinu G, Kruszewski G. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. InProceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2014 (Vol. 1, pp. 238-247).

[13] Chalapathy R, Borzeshi EZ, Piccardi M. Bidirectional LSTM-CRF for clinical concept extraction. arXiv preprint arXiv:1611.08373. 2016 Nov 25.

[14] Poria S, Cambria E, Gelbukh A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. InProceedings of the 2015 conference on empirical methods in natural language processing 2015 (pp. 2539-2544).

[15] Xue B, Fu C, Shaobin Z. A study on sentiment computing and classification of sina weibo with word2vec. InBig Data (BigData Congress), 2014 IEEE International Congress on 2014 Jun 27 (pp. 358-363). IEEE.

[16] Iyyer M, Manjunatha V, Boyd-Graber J, Daum III H. Deep unordered composition rivals syntactic methods for text classification. InProceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) 2015 (Vol. 1, pp. 1681-1691).

[17] Morin F, Bengio Y. Hierarchical probabilistic neural network language model. InAistats 2005 Jan 6 (Vol. 5, pp. 246-252).

[18] Apache Spark. https://spark.apache.org.

[19] Elasticsearch. https://www.elastic.co.

[20] Eugenio BD, Glass M. The kappa statistic: A second look. Computational linguistics. 2004 Mar;30(1):95-101.

[21] PubMed Advanced Search. https://www.ncbi.nlm.nih.gov/pubmed/advanced.