Semantic Text Similarity on Medical Symptoms

Banafsheh Khazali

University of Calgary, Banafsheh.khazli@ucalgary.ca

Zahra Ghods

University of Calgary, zahra.ghods@ucalgary.ca

Text similarity measurement between words, sentences, paragraphs, and documents plays an important role in various tasks such as information retrieval, dialogue systems, word-sense disambiguation, automatic essay scoring, short answer grading, machine translation and document matching. In this proposal text similarity measurement methods are used to check the similarity between medical diseases.

Keywords: Language Processing, Text Similarity, Semantic Similarity, Lexical Similarity

1 INTRODUCTION

As the number of documents on the web increases, the need to retrieve the best relevant documents increases. Among the different techniques through which we can retrieve most relevant document from the large corpus, similarity between words, sentences, paragraphs, and documents is of importance in various tasks. Text similarity is defined as the commonness between the document text to the most relevant documents. Text similarity can be used in text and document categorization.

Text similarity is not only the similarity between two words (lexical similarity), but it also is about analyzing the shared semantic properties of two words (semantic similarity). Lexical similarity provides the similarity on the basis of character or statement matching, for e.g., “watery eyes that itches” and “watery or itchy eyes” are lexically similar to each other. Whereas Semantic similarity provide the similarity on the basis of meaning, for e.g., “Natural language Processing” and “NLP” both are semantic similar to each other.

The main purpose of this work is to identify the lexical and semantic similarity between medical symptoms and find the most similar diseases to each other based on the related factors and symptoms. This proposal is organized as follows: Section 2 presents a brief overview of the related work, and our proposed method is described in Section 3.

2 RELATED WORK

Text similarity is a crucial task in Natural Language Processing (NLP) and is used in a wide range of applications, including information retrieval, automatic question answering, machine translation, dialogue systems, and document matching. A number of techniques have been developed to measure text similarity, including semantic-based methods, such as Latent Semantic Analysis (LSA) [1], and syntax-based methods, such as Longest Common Subsequence (LCS) [2] and edit distance [3].

There are two ways to identify the similarity between two texts. Lexical text similarity, the task of determining the similarity between two texts based on their surface-level form such as word overlap and string similarity, has been widely studied in the field of Natural Language Processing (NLP). Some of the commonly used techniques for performing lexical text similarity include Jaccard Similarity [4], Cosine Similarity [5], and Edit Distance [6].

Unlike lexical text similarity, which focuses on surface-level form, semantic text similarity considers the underlying meaning of the words and the relationships between them. One popular approach to semantic text similarity is the use of word embeddings, such as word2vec [7] and GloVe [8]. Word embeddings are dense, continuous vector representations of words that capture the semantic relationships between them. By computing the similarity between the word embeddings of the words in two texts, it is possible to determine the overall semantic similarity between the texts.

Recently, deep learning models have shown promising results in semantic similarity tasks. Siamese networks [9], Convolutional Neural Networks (CNNs) [10], and Recurrent Neural Networks (RNNs) [11] are some of the most commonly used deep learning models for semantic similarity. Pre-trained transformer models, such as BERT [12], have achieved state-of-the-art performance on many text similarity benchmarks and have become the method of choice for many NLP applications.

There have been studies to further the understanding of diseases with studying the similarity between them. Studying disease similarities can be helpful to suggest treatment that can be appropriated from one disease to another [12]. Semantic similarity between diseases was assessed by computing the similarity between the sets of associated ontological terms [13] and similarity measures like KEGG (Kyoto Encyclopedia of Genes and Genomes) [13], co-occurrence of annotation (PMI), node-based measures like Resnik [14], and Lin [15] are used to calculate the similarity measures.

3 PROPOSED WORK

Failure of diagnose can cause a lot of problems. In a perfect world, when you are sick, hurt, or suffering from health complications, a doctor or other medical professional will be completely accurate in diagnosing sickness or suffer from health complications and lead to the right treatment plan. Unfortunately, this is not always the case. If any mistakes involve misdiagnosing a health condition, patients can suffer serious consequences such as a prolonged condition, further complications, ineffective treatment, or even death. The main goal for this project is to find the similarity of symptoms for different diseases and categorize them to help doctors and patients to understand the possibility of medical misdiagnosis.

3.1 Dataset

This website provides a resource for checking the factors causing some common disease in the human body; for instance, 'injury' as well as 'overuse' are resulting in pain or discomfort in the foot. Hence, the information provided in the site will be used as a source for collecting the data through web scraping methods. In this work, similar diseases are considered as a result of the text similarity between the documents containing their both related factors. The related factors documented in the second part specify each disease. Related factors include sections like the location of pain, other symptoms accompanying it, things that worsen the pain, etc. This variety helps us in creating more documents and a larger corpus as a result, from which we will generate the word vectors used for word embeddings. To be more specific, for a specific disease selected from the first section, we collect diverse set of factors that contribute to that pain as a document. Then these documents are combined and used as a corpus for that specific disease. The same approach will be used for another disease. The proposed semantic similarity will be applied to the group of these documents and corpuses of both diseases, resulting in the similarity between them. For instance, the combination of 'activity or overuse,' 'long period of rest', and 'joint deformity" are three factors of foot pain in adults and 'movement', 'prolonged sitting or standing', 'joint weakness' can be found in knee pain. Therefore, it can be concluded that the similarity score between 'foot pain' and 'knee pain' be very high.

3.2 proposed method

In the proposed method we first take a line of sentence and then transform it into a vector. The methods enhancing semantic similarity mostly focus on transformers. Transformers are used to encode sentences to get their embeddings and then compute the similarity score using the similarity metrics. These transformers have been fine-tuned on a variety of pre-trained models for different tasks [16]. We use the best models for semantic similarity like stsb-roberta-large [17], which uses ROBERTA-large as the base model and mean-pooling. Furthermore, the embeddings produced by the model encoder are great in maintaining the semantic information of the text. To calculate the similarity score, various metrics such as Cosine similarity [9], dot product [18], and Jaccard similarity [8] can be used. In a great scale, this project leads to better understanding and more accurate categorization for the factors give rise to distinct disorder and medical disease.

REFERENCES

[1] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25(2-3), 259-284. Sam Anzaroot and Andrew McCallum. 2013. UMass Citation Field Extraction Dataset. Retrieved May 27, 2019 from http://www.iesl.cs.umass.edu/data/data-umasscitationfield

[2] Bergroth L, Hakonen H, Raita T. A survey of longest common subsequence algorithms. InProceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000 2000 Sep 27 (pp. 39-48). IEEE. Chelsea Finn. 2018. Learning to Learn with Gradients. PhD Thesis, EECS Department, University of Berkeley.

[3] Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet physics doklady, 10(8), 707-710.

[4] Jaccard, P. (1901). Etude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles, 37, 547-579.

[5] Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.

[6] Bromley, J., Bottou, L., Hastings, J., LeCun, Y., & Howard, R. E. (1993). Signature verification using a siamese time delay neural network. In Advances in neural information processing systems (pp. 737-744).

[7] Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

[8] Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

[9] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

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

[11] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532-1543).

[12] A.J. Butte, I.S. Kohane, Creation and implications of a phenome–genome network, Nat Biotechnol, 24 (January) (2006), pp. 55-62

[13] Mathur, Sachin, and Deendayal Dinakarpandian. "Finding disease similarity based on implicit semantic similarity." Journal of biomedical informatics 45.2 (2012): 363-371.

[14] Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language J Artif Intell Res, 11 (1999), pp. 95-130

[15] Lin D. An information-theoretic definition of similarity. In: Proceedings of the fifteenth international conference on machine learning; 1998. p. 296–304.

[16] https://towardsdatascience.com/semantic-similarity-using-transformers-8f3cb5bf66d6

[17] https://huggingface.co/ sentence-transformers/stsb-roberta-large

[18] https://jamesmccaffrey.wordpress.com/2022/03/28/using-dot-product-as-a-measure-of