(temp) Using Word Embedding for Citation Recommendation in FinTech Scientific Articles

by Ting-Chun Chen,

Submitted to The University of Nottingham

September 2023

in partial fulfilment of the conditions for the award of the degree of

Master of Science in Data Science

I declare that this dissertation is all my own work, except as indicated in the text

Contents

Abstract

According to the guidelines the abstract should not exceed 300 words. However, it is unlikely that anyone will be counting when you submit. If you still wish to do so, then better use Emacs `count-words` command.

Acknowledgements

It is good to thank here your supervisors and any sponsoring bodies, as well as any family, friends, cats, dogs etc. that have been supportive during your time at University.

Chapter 1

Introduction

1.1 Citation in Scientific Articles

Citations are used to demonstrate research background, existing techniques and pieces of evidence for a statement, playing a critical role in scientific writing. Authors can properly acknowledge the source of information of a statement and avoid plagiarism with an accurate citation. It also serves as a verification that the idea of this statement is provided and supported by previous studies. These days, references and citations also include additional information for search engines and citation recommendation systems to recognise the subfields and similar research of this article.

There has been a noticeable increase in the number of scientific articles being published. 3.8 million scientific articles are being published in 2022, according to the Web of Science database. It is expected that there would be more scientific articles available on the internet. It can be harder for academics to absorb all the new perspectives with their exact sources.

When looking for reference papers, keywords are commonly used in most academic databases and search engines. The authors would then review the results of papers from search engines to evaluate their relevance to our research and reliability. It could cost a huge amount of time and effort to go through full papers, not to mention that the academic database and search engines might miss some relevant papers.

1.2 Natural Language Processing

Natural language processing (NLP) refers to the approach of giving computers the ability to understand the meaning and thoughts of words, sentences and articles just like human beings do, building a way of communication between computer and human language. However, human languages have several characteristics that make it tricky to process them. The variability and informality, such as different vocabularies, syntax and phrases in different cultures or individuals representing a similar meaning, bring difficulties to applying a general transformations pipeline for information extraction. A large amount of noise such as mis-spellings, irrelevant contexts and grammar errors confuse not only human readers but also machines and affect the performance of language processing algorithms, not to mention that people might have different understandings of the same sentence which yield even more inaccuracies. Some advanced circumstances such as sarcasm, irony and the speaker's intention, whose implied meanings are beyond plain text, are more difficult for algorithms to detect and extract straightforwardly. Problems in normal vector data such as missing values and lack of labelling occur in natural language as well.

NLP manages to overcome the challenges of processing human language mentioned above and enables machines to analyse, cluster or classify not only vector instances but also texts. Advanced applications of NLP, such as sentiment analysis, text clustering, text summarisation, human language generation and improving automatic translation, provide solutions and assistance to a wide range of tasks. It reduces our workloads and time-spending and helps improve the interaction between humans and computers by mimicking human language to express in an easy-understanding way for most o the people not only programmers.

1.3 Text Embedding and Similarity

Text embedding, also known as text vectorisation and text featurisation, aims to extract information from words and sentences and represent the semantics and meaning of words by numbers and vectors. The process of extracting information has been through significant development since the year 2000. Various techniques have been proposed, including naive

ways of bag-of-words model, TF-IDF encoding, LSA encoding, Word2Vec embeddings and GloVe, along with more advanced methods like BERT and GPT, which we are familiar with and commonly use today. After transforming every word into a vector, similarities between two words are to be calculated. The measurements of word similarity are used in word-embedding techniques to evaluate the similarity between prediction value and real value to update parameters in NN model training.

1.4 Semantic Search Engine

Techniques about search engines are a good starting point when it comes to matching a sentence to related articles. Search engines, especially full-text search engines, enable users to look for relative information in a huge full-text database. Among the tasks in a search process, the comparison between keyword and candidate results can be improved with text embedding techniques to retrieve semantic information from keyword and candidate documents, which is known as the semantic search engine.

Semantic search engines enable users to provide a sentence or description instead of just a keyword for querying, while text embedding methods are applied normally to both query texts and candidate documents to evaluate the similarity between their embedded vectors. Semantic search can better catch the contextual meaning within texts and understand the user's intent. Some advanced techniques such as WordNet can also be applied to the search engine to improve the accuracy and coverage of querying results and can serve partly as query expansion that expands a search query with its synonym words or semantic relative words.

1.5 Study Purpose

This project is about building a semantic search engine for scientific articles in the fintech field and comparing the performance of using different text embedding techniques. We want to discover and develop a text embedding method that can best describe sentences in FinTech research articles and yield the best result of finding related scientific articles by a citation sentence. This project aims to help academics find proper references for their

statement when writing literature reviews and introductions of scientific reports.

Citations are commonly used in writing scientific articles to acknowledge the source of a statement in our work. Though we usually file and organize papers well while doing literature reviews, it happens that some statements are based on months or years of experience or accumulated knowledge in a particular subject. It could be relatively hard work to find an appropriate source for this idea. Hence, we are proposing a method to find the best math articles for our writing. We want to look for articles that mainly describe similar thoughts to the statement we write down in a certain domain in scientific writing. Among all the processes in the natural language processing pipeline, the text embedding techniques would be mostly focused on in this research, which is to convert contents into meaningful numbers and vectors for algorithms to compute, transform and compare. We decide to apply this research in the fintech field and answer the question "What text embedding techniques can best represent a scientific sentence to derive a semantic search engine for the fintech academic articles?"

The detail of this project is recorded in this document. The main issues and goals of this project are stated in the Introduction chapter. Related works, commonly-used background techniques and open-source tools in NLP field are organised and described in the Literature Review chapter. Data collection criteria, preprocessing steps and experiment details are presented in the Material and Method chapter. The performances and analysis of modeling methods in each experiment are listed in the Results chapter. Last but not least, the Conclusion and Discussion chapter stated interesting or notable findings and simple summarisations of this project.

Chapter 2

Literature Review

2.1 Text Embedding

Text embedding methods can be roughly classified by the level of encoding unit (embed in a word, sentence, or article level), supervised or unsupervised, similarity measurement (such as Euclidean distance or cosine similarity) and other specific technique support such as term-based embedding in particular fields and graph-based embedding. Some advanced methods such as ELMo[1] and GPT[2] use autoregressive models and BERT[3] uses an autoencoder to do bidirectional context encoding. A more complex structure of neural networks and a bigger training corpus could derive a better and more robust model but a huge amount of time to train models as well. Hence, another improvement of embedding methods would be the trade-off between better performance and reasonable training time.

BOW & TF-IDF

Bag-Of-Word (BOW) embedding is one of the most straightforward methods to extract information from texts. BOW uses word counts as the representation of texts, not taking the order, grammar and semantic meaning of words into account. A fixed-length vector is used to record the appearance frequency of a word, encoding a word into a sparse vector.[4] BOW can successfully categorise texts[5], while text analysis from other aspects is still possible to be explored.

Term frequency-inverse document frequency (TF-IDF) model uses the word occurrence in sentences and documents to denote the importance of words. TF-IDF assumes a word to play an important role if it occurs frequently only in some sentences as the key idea of this document. An important word will have a higher term frequency and a lower document frequency, which makes the TF-IDF higher. On the opposite, a less important word equally occurs in most of the sentences, generally used in many situations and topics but implying fewer ideas.[6]

LSA Encoding

Sample text sample text sample text sample text sample text

Word2Vec Encoding

Sample text sample text sample text sample text sample text

Doc2Vec Encoding

Sample text sample text sample text sample text sample text

Universal Sentence Encoder

Sample text sample text sample text sample text sample text

GloVe

Sample text sample text sample text sample text sample text

Long-Short Term Memory

Sample text sample text sample text sample text sample text

BERT

Sample text sample text sample text sample text sample text

GPT

Sample text sample text sample text sample text sample text

ELMo

Sample text sample text sample text sample text sample text

Applications of Text Embedding Techniques

Text embedding techniques are improved for specific objectives, fields and styles and applied to a wide range of tasks, including various issues of tweets analysis[7], visualisation in the biomedical field[8] and sentiment analysis on movie reviews[9].

Khatua et al.[10] identified crisis-related tweets during the 2014 Ebola and 2016 Zika outbreaks with pre-trained Word2Vec and GloVe models. They found a better classification performance to have a small domain-specific corpus from tweets and scholarly abstracts from PubMed participated in the model. They also observed a higher accuracy from a higher dimension of word vector and skip-gram model than CBOW.

Lee et al.[11] utilized SentiWordNet 3.0 to analyse the effect of several negative emotions in hotel reviews. SentiWordNet 3.0[12] provided sentiment analysis as classification tasks and word embedding with a frequency-weighted bag-of-words model and the help of WordNet corpus. Onan[13] presented a sentiment analysis approach to product reviews from Twitter. This deep-learning-based method applied TF-IDF weighted GloVe to the CNN-LSTM architecture to do word embedding and outperformed conventional deep-learning methods.

2.2 Text Similarity

A pair of similar words should act similarly in most of the features we extracted, while a good similarity metric can aggregate the difference in every feature into a single value, which is comparable between each pair of words. Besides Euclidean distance and cosine similarity as semantic similarity measurements mentioned above, WordNet also provides path similarity to evaluate similarities of words with a lexical hierarchical structure.

Sentence similarity measurements are more complex and various compared to word similarity since sentences can be considered as combinations of words. The most straightforward approach is to aggregate words in the sentence as a representation to compare with other sentences, which can be seen as a baseline measurement of sentence similarity. The steps are to take averages of every word in each sentence, yield a single vector for each sentence and calculate the Euclidean distance or cosine similarity between two sentences with average vectors. This approach, however, doesn't consider the order of words, while it can have a huge effect on a sentence's meaning when the word order differs.

Several algorithms are proposed to map a sentence into a vector and calculate the similarity with the derived value directly from the embedding process. Much of them are extended from an existing word embedding method to apply to sentences or even documents, such as Word2Vec, Sent2Vec and Doc2Vec. Some Approaches consider different levels of embedding at the same time, such as word embedding and sentence embedding of BERT.

Lexical Relation and WordNet

Sample text sample text sample text sample text sample text

2.3 Text Embedding for Search Engine

With the increase of documents and web pages, traditional keywords-based search engines are thought to be powerless to correctly look for users' requirements. Text embedding techniques are applied to search engines to provide machine-readable web pages and se-

mantic annotations to the algorithm to yield a more accurate search result. Many websites are applying text embedding to their search engines, including Google Search, Microsoft's Bing Search, Amazon Search and many e-commerce websites and platforms. Not to confuse the search engine using text embedding and the search engine using semantic web search language, in this paper, we refer to the semantic search as the searching approach with text embedding or other semantic retrieval techniques.

Regarding the search engines for scientific articles, Eisenberg et al.[14] proposed a semantic search for Biogeochemical literature that undergoes paper research by comparing the concepts extracted from queries and academic literature. However, the concept extraction component in the workflow of this research is annotated by domain experts, which can be done automatically by NLP approaches mentioned in the future directions chapter. Fang et al.[15] performed a biomedical article-searching approach called Semantic Sequential Dependence Model (SSDM) that combined semantic information retrieval techniques and the traditional SDM model. Word embedding techniques of the Neural Network Language Model(NNLM)[16], Log-Bilinear Language Model(LBL)[17] and Word2vec are applied to the literature corpus to find synonyms by KNN classification algorithm and generate a domain-specific thesaurus. The thesaurus is then utilized to extend query keywords and the SDM played an important role in the combination strategy of the extended keywords for further comparison to documents.

Compare to the keyword embedding approaches, applying sentence embedding or other wider-level embedding techniques to search engines can better preserve information for query but be more challenging at the same time. Palangi et al.[18] managed to embed semantic information at a sentence level by using LSTM-RNN (Long Short-Term Memory - Recurrent Neural Network) framework to do web document retrieval, and compare the summarisation sentence vector and query sentence vector to yield the best searching result.

2.4 Searching-Based Citation Recommendation

Some academic databases come with citation recommendation tools, which provide relevant articles that share the same category with or are similar to our research. Citation recommendation tools would yield different results resulting from not only different algo-

rithms they used but also candidate papers' published journals, impact factors, times being referred to, and the availability if it's open to everyone or subscription-only.

Citation recommendation methods are always built with three main stages: (1) Generate candidate citations from the publication database (2) Create a recommendation list by ranking the candidate citations (3) Evaluate the accuracy of our recommendation system.[19] Text embedding techniques can be included generally in step 1: generation of candidate citation, that is, applying text embedding to filter out candidate citations from the publication database that better relate to keywords provided by users or meet users' needs.

Since the diversity of terms and patterns between scientific articles in different domains, researchers would tend to train a specific model with their domain data or use transfer learning to fine-tune the model parameters. Tshitoyan et al.[20] used modified Word2Vec embedding to successfully capture complex materials science concepts, without any additional chemistry knowledge insertion, and extract knowledge and relationship from scientific literature. Zhang et al.[21] used 15 text representation models, including 6 term-based methods, 2 word embedding methods, 3 sentence embedding methods, 2 document embedding methods and 2 BERT-based methods, to construct an article recommendation system in the biomedical field. They found BERT and BioSenVec, a Sent2Vec model trained on PubMed corpus, outperformed most of the online and offline citation recommendation systems and an improvement in BERT-based methods after fine-tuning to learn users' preferences.

Wang et al.[22] proposed a sentence-level citation recommendation system called SenCite that used CNN to recognise candidate citation sentences and FastText as the word embedding method to extract information from texts, without summarising an article into sentences. They evaluated the performance of SenCite with several evaluation metrics such as modified reciprocal rank, average precision and normalized discounted cumulative gain and human experts verification. The SenCite is shown to outperform most of the embedding methods and yield a comparable accuracy to BERT. The human experts also stated that SenCite provided the best top-1 citation recommendation.

Chapter 3

Material and Method

3.1 Dataset

Besides a less amount of fintech-specific journals, we assume that if a journal produces many fintech articles, most of the research in the journal could have highly relation to the fintech field even if the research themselves don't contain any fintech keyword in their main topics. To have the most articles from the least amount of journals, we decided to use the Web of Science[-citation-] search engine and database to find out the journals that produced the most fintech articles.

Reference papers were collected by the WOS search engine with rules listed as follows:

- Topic: "**Fintech**" or "**Financial technology**" or "**digital finance**" or "**electronic banking**" or "**cryptocurrency**" or "**Blockchain**"
- publication years: 2003-2023
- Document Types: **Article** or **Proceeding Paper** or **Review Article**
- Languages: English

We got 66,974 results on 23rd July 2023 from the Web of Science Core Collection and exported them to EndNote. The `Find Full Text` feature in EndNote is used to look for the full text of reference papers by their Accession numbers including Digital Object Identifier

(DOI), WOS ID and PubMed ID. This feature enables us to automatically batch-download loads of reference papers as PDF files if the journal is available in Open Access resources.

3.1.1 Second Level Heading

Sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text.

Third Level Headings Do Not Appear in the Table of Contents

Sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text [23]

Bibliography

[1] M. E. Peters, M. Neumann, M. Iyyer, *et al.*, "Deep contextualized word representations," *arXiv e-prints*, arXiv:1802.05365, arXiv:1802.05365, Feb. 2018. DOI: 10.48550/arXiv.1802.05365. arXiv: 1802.05365 [cs.CL].

[2] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[3] T. Niven and H.-Y. Kao, "Probing neural network comprehension of natural language arguments," *arXiv preprint arXiv:1907.07355*, 2019.

[4] A. Sethy and B. Ramabhadran, "Bag-of-word normalized n-gram models," in *INTERSPEECH 2008: 9TH ANNUAL CONFERENCE OF THE INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION 2008, VOLS 1-5*, 9th Annual Conference of the International-Speech-Communication-Association (INTERSPEECH 2008), Brisbane, AUSTRALIA, SEP 22-26, 2008, 2008, pp. 1594–1597, ISBN: 978-1-61567-378-0.

[5] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: A statistical framework," *INTERNATIONAL JOURNAL OF MACHINE LEARNING AND CYBERNETICS*, vol. 1, no. 1-4, pp. 43–52, Dec. 2010, ISSN: 1868-8071. DOI: 10.1007/s13042-010-0001-0.

[6] J.-R. Li, Y.-F. Mao, and K. Yang, "Improvement and application of tf * idf algorithm," in *INFORMATION COMPUTING AND APPLICATIONS*, B. Liu and C. Chai, Eds., ser. Lecture Notes in Computer Science, 2nd International Conference on Information Computing and Applications, Qinhuangdao, PEOPLES R CHINA, OCT 28-31, 2011, Natl Nat Sci Fdn China (NSFC); NE Univ Qinhuangdao; Yanshan Univ; Nanyang Technol Univ, vol. 7030, 2011, pp. 121+, ISBN: 978-3-642-25254-9.

[7] Z. Mottaghinia, M.-R. Feizi-Derakhshi, L. Farzinvash, and P. Salehpour, "A review of approaches for topic detection in twitter," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 33, no. 5, pp. 747–773, 2021.

[8] N. Oubenali, S. Messaoud, A. Filiot, A. Lamer, and P. Andrey, "Visualization of medical concepts represented using word embeddings: A scoping review," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, pp. 1–14, 2022.

[9] S. Sivakumar and R. Rajalakshmi, "Analysis of sentiment on movie reviews using word embedding self-attentive lstm," *International Journal of Ambient Computing and Intelligence (IJACI)*, vol. 12, no. 2, pp. 33–52, 2021.

[10] A. Khatua, A. Khatua, and E. Cambria, "A tale of two epidemics: Contextual word2vec for classifying twitter streams during outbreaks," *Information Processing & Management*, vol. 56, no. 1, pp. 247–257, 2019.

[11] M. Lee, M. Jeong, and J. Lee, "Roles of negative emotions in customers' perceived helpfulness of hotel reviews on a user-generated review website: A text mining approach," *International Journal of Contemporary Hospitality Management*, vol. 29, no. 2, pp. 762–783, 2017.

[12] S. Baccianella, A. Esuli, F. Sebastiani, *et al.*, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining.," in *Lrec*, vol. 10, 2010, pp. 2200–2204.

[13] A. Onan, "Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 23, e5909, 2021.

[14] J. D. Eisenberg, D. Banisakher, M. Presa, *et al.*, "Toward semantic search for the biogeochemical literature," in *2017 IEEE International Conference on Information Reuse and Integration (IRI)*, IEEE, 2017, pp. 517–525.

[15] F. Fang, B.-W. Zhang, and X.-C. Yin, "Semantic sequential query expansion for biomedical article search," *IEEE Access*, vol. 6, pp. 45 448–45 457, 2018.

[16] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," *Advances in neural information processing systems*, vol. 13, 2000.

[17] A. Mnih and G. Hinton, "Three new graphical models for statistical language modelling," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 641–648.

[18] H. Palangi, L. Deng, Y. Shen, *et al.*, "Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 694–707, 2016.

[19] S. Ma, C. Zhang, and X. Liu, "A review of citation recommendation: From textual content to enriched context," *Scientometrics*, vol. 122, pp. 1445–1472, 2020.

[20] V. Tshitoyan, J. Dagdelen, L. Weston, *et al.*, "Unsupervised word embeddings capture latent knowledge from materials science literature," *Nature*, vol. 571, no. 7763, pp. 95–98, 2019.

[21] L. Zhang, W. Lu, H. Chen, Y. Huang, and Q. Cheng, "A comparative evaluation of biomedical similar article recommendation," *Journal of Biomedical Informatics*, vol. 131, p. 104 106, 2022.

[22] H.-C. Wang, J.-W. Cheng, and C.-T. Yang, "Sentcite: A sentence-level citation recommender based on the salient similarity among multiple segments," *Scientometrics*, vol. 127, no. 5, pp. 2521–2546, 2022.

[23] L. Lamport, *LaTeX : A Document Preparation System*, Second. Addison-Wesley, 1994.

Appendix A

Supplementary Materials I

Sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text.

Appendix B

Supplementary Materials II

Sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text.