Construction and Application of a Semantic Retrieval Model for Ancient Agricultural Texts: Postprint

Authors: Liu Nanzhu, Cui Yunpeng, Wang Mo

Date: 2024-04-03T00:00:00+00:00

Abstract

[Purpose/Significance] This study constructs a semantic retrieval model that accepts modern vernacular Chinese as queries and automatically returns the most relevant ancient agricultural text passages, providing scholars with a more convenient method for retrieving and tracing ancient agricultural knowledge. [Method/Process] Using SikuBERT, pre-trained on the Siku Quanshu (Complete Library of the Four Treasuries), as the base model, and employing a contrastive learning approach, the model was further trained using a self-constructed ancient agricultural text dataset to develop a semantic retrieval model capable of supporting modern vernacular Chinese queries and returning ancient agricultural text passages most semantically similar to the queries. [Results/Conclusion] The Spearman coefficient of the ancient agricultural text semantic retrieval model reaches 86.51% on the test set, representing an improvement over the baseline model' s performance of 83.69%. The recall performance (recall@k) on the self-constructed ancient agricultural text retrieval test set also demonstrates improvement over the baseline model, indicating that the model achieves relatively good retrieval effectiveness on ancient agricultural texts. However, limited by the scale of the ancient agricultural text training corpus, considerable room for improvement in the model' s training effectiveness remains.

Full Text

Preamble

Authors: Liu Nanzhu[1],[2], Cui Yunpeng[1],[2]*, Wang Mo[1],[2]
Affiliations: 1. Institute of Agricultural Information, Chinese Academy of Agricultural Sciences, Beijing 100081; 2. Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing 100081

Abstract:
[**Purpose/Significance**] This study constructs a semantic retrieval model that enables users to query in vernacular Chinese and automatically returns the most relevant ancient agricultural text passages, providing scholars with more convenient methods for retrieving and tracing ancient agricultural knowledge. [**Method/Process**] Using SikuBERT—pre-trained on the Siku Quanshu corpus—as the base model, we employ contrastive learning methods to continue training the model on a self-constructed ancient agricultural text dataset, obtaining a semantic retrieval model that supports vernacular queries and returns semantically similar ancient agricultural passages. [**Results/Conclusions**] The Spearman coefficient of the ancient agricultural text semantic retrieval model reaches 86.51% on the test set, representing improvement over the baseline model' s 83.69%. Recall performance on our self-constructed ancient agricultural retrieval test set also shows enhancement over the baseline, demonstrating effective retrieval capability for ancient agricultural texts. However, limited by the scale of training corpora, there remains substantial room for improving the model' s recall@k performance.

Keywords: ancient agricultural texts; semantic retrieval; contrastive learning; model construction; deep learning
CLC Number: TP391
Document Code: A
Article ID: 1002-1248
Citation: Liu N, Cui Y, Wang M. Research on the construction and application of semantic retrieval models for ancient agricultural texts[J]. Journal of Agricultural Library and Information, 2023, 35(7): 52-62.

1 Introduction

China is one of the world' s great ancient civilizations with millennia of uninterrupted agricultural tradition and history. Ancient Chinese agricultural books serve as the primary carriers of traditional agricultural experience, knowledge, productivity, and historical essence—works on agricultural production knowledge written by Chinese scholars before modern Western agronomy influenced China. These texts preserve vast amounts of material awaiting exploration and represent precious historical cultural heritage. Traditional cultivation techniques remain largely adaptable for modern agriculture and continue to provide practical guidance for contemporary agricultural problems. However, ancient agricultural books are written in classical Chinese without punctuation, making them inaccessible to the general public. Their value nonetheless requires inheritance and promotion.

By leveraging machine learning and deep learning technologies to excavate knowledge from these texts, we can not only facilitate academic research but also promote the integration of outstanding cultural heritage with contempo-

rary social development and provide public knowledge services. Between 2017 and 2022, multiple national policy documents—including the "Opinions on Implementing the Excellent Traditional Chinese Culture Inheritance and Development Project," the "14th Five-Year Plan for National Economic and Social Development," and the "Outline for 2035 Long-Range Objectives" —have proposed advancing excellent traditional Chinese culture, strengthening protection and utilization of cultural relics and ancient books, promoting digitalization of ancient books, and excavating their contemporary value to bring cultural relics and written heritage to life.

To understand scholars' needs regarding ancient Chinese agricultural books, we conducted literature surveys of published papers on CNKI. Analysis reveals that agricultural history researchers and related students primarily use these texts, typically employing cataloging, collation, and compilation methods for analysis. Agricultural researchers generally consult ancient texts when modern agricultural problems arise, seeking historical precedents for current issues, related policies, and other information to develop new agricultural models. They require the ability to quickly query comprehensive ancient agricultural texts based on themes, keywords, or descriptions. Currently, specialized ancient agricultural book databases include the China Agricultural Civilization Network (providing search services for 16 agricultural texts), the Agricultural Heritage Information Platform of Nanjing Agricultural University (offering classified browsing and full-text search for over 200 agricultural classics), and the Chinese Academy of Agricultural Sciences Library' s digitization project for important agricultural texts. However, these systems still rely primarily on keyword or full-text retrieval, requiring complex search expressions and multiple rounds of searching for comprehensive results, with high barriers for scholars unfamiliar with ancient texts.

Semantic retrieval transcends literal matching to capture users' true intentions behind their queries, returning the most relevant results. Current implementation approaches include statistical feature-based, ontology-based, and vector-based semantic retrieval. This study employs vector semantic retrieval, where the system represents user queries and document collections as vectors reflecting their core features, indexes them in high-dimensional vector space, and calculates similarity using cosine similarity or Manhattan distance to identify semantically similar sentences and return relevant documents. This approach allows users to input natural language queries without specialized training or complex search syntax.

Our research focuses on constructing a semantic retrieval model for ancient agricultural texts based on the BERT framework. Using parallel corpora of 10 ancient agricultural texts and their translations—including *Lüshi Chunqiu*·*Shangnong*, *Guanzi*·*Diyuan*, *Kangcangzi*·*Nongdao*, and others—we build a model enabling semantic retrieval from vernacular to classical Chinese, automatically returning all ancient agricultural passages relevant to input queries. This provides scholars with convenient ancient agricultural knowledge

retrieval and traceability methods.

2.1 Semantic Retrieval

In industrial applications requiring retrieval of relevant documents from large databases, semantic retrieval typically employs a multi-stage ranking strategy to balance efficiency and effectiveness, consisting of recall and reranking stages. Both stages evaluate query-document relevance but use different models according to their purposes. The recall stage aims to retrieve all potentially relevant documents from vast collections for the reranking stage. With fewer documents to consider, the reranking stage employs more complex models to reorder the recalled documents, with one or more rerankers processing the list sequentially to generate final results. The recall stage prioritizes efficiency and high recall, while reranking emphasizes effectiveness.

Semantic retrieval has evolved through three stages: term-based retrieval, feature representation-based retrieval, and neural semantic retrieval. Initially, queries and documents were represented as discrete terms (e.g., BM25 + TF-IDF weights) managed through inverted indexing. While simple and powerful, this approach suffers from vocabulary mismatch and independence assumptions that may fail to capture document semantics. Feature representation methods used manually crafted features to build representation functions, such as query expansion, translation models, and document expansion techniques. After 2013, embedding technologies were applied to semantic retrieval, with dense word embeddings alleviating vocabulary mismatch. After 2016, deep learning enabled neural semantic retrieval methods that use neural networks to build representation and scoring functions, learning deep semantics and complex interactions end-to-end from data.

Currently, only vector semantic retrieval can support semantic similarity-based retrieval from vernacular to classical Chinese. Applications include: when writing and wanting to quote an ancient text but forgetting the exact wording while remembering its meaning; when reading ancient texts and wanting to check the evolution of a concept; or when needing to verify historical precedents for modern agricultural practices. Users can directly input modern descriptions into the semantic retrieval system to automatically return the most relevant ancient passages.

2.2 Related Pre-trained Models

With deep learning development, model parameters have increased dramatically, requiring larger datasets for full training and overfitting prevention. However, constructing large-scale annotated datasets presents enormous challenges. Pre-trained models can learn good language representations and better initialization

parameters from easily obtainable large-scale unannotated data, which can then be applied to other tasks. In ancient Chinese processing, where annotated resources are scarce, pre-trained models are essential for enhancing low-resource text processing.

Currently, few pre-trained language models exist for ancient Chinese. Available models include Beijing Institute of Technology' s GuwenBERT, SikuBERT and SikuRoBERTa, and Nanjing Agricultural University' s WANG and Bert-Ancient-Chinese—all BERT-based models trained on different corpora. GuwenBERT uses the Dazhige ancient Chinese corpus (15,694 books, 1.7B characters, simplified Chinese) for continued training on Chinese RoBERTa. SikuBERT and SikuRoBERTa, built on Chinese BERT and RoBERTa respectively, use the Siku Quanshu corpus (536,097,588 characters, traditional Chinese) through domain-adaptive training. Bert-Ancient-Chinese combines ancient corpora with Chinese BERT for continued training, supporting both traditional and simplified Chinese with a larger vocabulary (38,208 vs. SikuBERT' s 29,791).

For cross-lingual retrieval from vernacular to classical Chinese, only Nanjing Agricultural University' s BTfhBER and ZHANG' s XLsearch-cross-lang-search-zh-vs-classicical-cn exist. BTfhBER continues training Chinese BERT on 24 Histories parallel corpora, while XLsearch-cross-lang-search-zh-vs-classicical-cn trains on ~900,000 classical-vernacular sentence pairs. However, these models show limited effectiveness for retrieving within ancient agricultural texts specifically.

3.2 Construction of Ancient Agricultural Text Pre-trained Model

This study constructs an ancient agricultural text semantic retrieval model that uses vernacular queries to retrieve the k most semantically similar ancient agricultural passages from a database. The experiment comprises four stages: corpus preprocessing, model training, semantic retrieval task testing, and model evaluation.

We constructed a semantic retrieval dataset from 10 ancient agricultural texts and their translations, including *Qimin Yaoshu*, *Nongsang Jiyao*, *Nongshu*, *Nongzheng Quanshu*, *Lüshi Chunqiu · Shangnong* (four chapters), *Guanzi · Diyuan*, *Kangcangzi · Nongdao*, and others. Translations were sourced from Liang Le and Xu Hong' s vernacular translations (1995) and Donglu Wang' s annotated translations. After cleaning and alignment, we obtained 9,542 classical-vernacular parallel sentence pairs, divided into training, validation, and test sets at an 8:1:1 ratio.

3.2.1 Training Model Selection

Based on pre-experimental results, we selected PyTorch versions of SikuBERT and BERT-Base-Chinese as base models. BERT (Bidirectional Encoder Representations from Transformers), proposed by Google in 2018, uses Transformer as its bidirectional encoder. Through masked language modeling (MLM) and next sentence prediction (NSP) tasks, BERT learns contextual word representations and sentence relationships, supporting downstream tasks requiring deep semantic understanding. BERT-Base-Chinese is trained on Chinese Wikipedia, while SikuBERT is domain-adapted for ancient Chinese processing using the Siku Quanshu corpus.

3.2.2 Model Training Method

We employed the SimCSE supervised training framework. Native BERT is unsuitable for sentence embeddings as it lacks independent sentence representation calculation, instead using cross-encoding that concatenates sentence pairs—resulting in high computational cost and poor performance on similarity metrics like cosine similarity. SimCSE is a simple contrastive learning framework that significantly improves sentence embedding quality for semantic textual similarity tasks.

SimCSE pulls semantically close sentences together while pushing dissimilar ones apart. Besides using provided contradictory data as hard negatives, it treats other sentences in the same batch as negatives. During training, for N triplet sentence pairs in a batch, each sentence has one positive and N-1 negative examples. SimCSE also improves anisotropy through alignment (measuring expected distance between positive pairs) and uniformity (measuring evenness of embedding distribution). Empirical analysis shows that optimizing alignment and uniformity aligns with contrastive learning objectives and improves performance.

We used BERT-Base-Chinese and SikuBERT as base models with SimCSE' s supervised framework to train the ancient agricultural text semantic retrieval model on our dataset.

3.2.3 Model Evaluation Metrics

Spearman' s Rank Correlation Coefficient: This non-parametric measure evaluates dependence between two variables using monotonic functions. When data contain no repeated values, a Spearman coefficient of +1 indicates perfect monotonic correlation. It measures ranking rather than actual scores, making it suitable for evaluating sentence embeddings. We calculate the correlation between cosine similarity of sentence embeddings and gold labels to assess semantic similarity.

Recall@k: This measures the ratio of relevant results retrieved in the top-k results to all relevant results in the collection, indicating retrieval system compre-

hensiveness. The formula is: Recall@k = true_{positives}@k / all_{positive}, where true_{positives}@k represents positive examples in k predictions and all_{positive} represents all positives in the full collection.

3.2.4 Model Parameter Settings

Table 1 shows the hyperparameters used for training the ancient agricultural text semantic retrieval model. Due to SimCSE-SikuBERT' s superior performance, we adopted its parameters: maximum input sequence length of 128, training batch size of 64, learning rate of 5e-5, warmup steps of 10% of training data, and 3 training epochs.

4.1 Spearman Coefficient

Direct application of SikuBERT and BERT-Base-Chinese on the validation set yielded Spearman coefficients of 83.69% and 69.52%, respectively. After SimCSE training, SimCSE-BERT-Base-Chinese and SimCSE-SikuBERT achieved 86.14% and 86.51% on the validation set, respectively—substantial improvements demonstrating that the models learned relevant ancient agricultural knowledge.

4.2 Model Effect

As shown in Table 2 , trained SimCSE-BERT-Base-Chinese and SimCSE-SikuBERT demonstrate stronger recall on the test set expanded using the Bloom large language model compared to direct use of base models. SimCSE-SikuBERT shows the best overall performance.

4.3 Retrieval Effect

We conducted small-scale retrieval experiments using SikuBERT and the trained semantic retrieval model with a dual-encoder architecture. Queries and database passages were encoded into dense vectors using the same encoder, with cosine similarity calculating relevance. Given the small corpus size, we returned the top-2 scoring passages.

We collected 108 agricultural texts containing 14,133 ancient agricultural passages based on Wang Yuhu' s *Catalogue of Chinese Agricultural Books*. Table 3 presents retrieval examples (converted to simplified Chinese using OpenCC for readability). The trained semantic retrieval model retrieves higher-quality sentences, finding not only the exact classical matches to modern descriptions but also semantically similar passages, effectively enhancing retrieval effectiveness and enabling knowledge traceability.

However, limited by ancient agricultural corpus scale and quality, the model' s performance remains suboptimal. Training pairs often exhibit high lexical over-

lap, causing the model to seek surface similarity rather than semantic similarity. The absence of hard negatives in our dataset also constrains further improvement. Paragraph vector quality is heavily influenced by segmentation methods, and appropriate information organization is needed for better retrieval results.

Future work will explore suitable training paradigms for small samples, investigate transfer learning from cross-lingual pre-trained models, and incorporate existing ancient agricultural historical knowledge into model training to achieve better semantic retrieval performance.

References

[7] ZHOU X Y, ZHU L, SPENGLER R N, et al. Water management and wheat yields in ancient China: Carbon isotope discrimination of archaeological wheat grains[J]. The holocene, 2021, 31(2): 285-293.

[8] CHEN S C. Exploring the use of electronic resources by humanities scholars during the research process[J]. Electron libr, 2019, 37: 240-260.

[9] WANG S Y, CUI D A, LV Y N, et al. Cangpu oral liquid as a possible alternative to antibiotics for the control of undifferentiated calf diarrhea[J]. Frontiers in veterinary science, 2022, 9: 879857.

[10] XIA X Y, LIN Z C, SHAO K P, et al. Combination of white tea and peppermint demonstrated synergistic antibacterial and anti-inflammatory activities[J]. Journal of the science of food and agriculture, 2021, 101(6): 2500-2510.

[11] WANG N, LIU X, LI J G, et al. Antibacterial mechanism of the synergistic combination between streptomycin and alcohol extracts from the Chimonanthus salicifolius S. Y. Hu. leaves[J]. Journal of ethnopharmacology, 2020, 250: 112467.

[12] GE X H. Text, history date and knowledge - The paradigms of ancient agricultural books research in agricultural history of China[J]. Agricultural history of China, 2019, 38(2): 12-25.

[13] LI M J, CHEN M S, MENG B. Review on the collation of ancient Chinese scientific and technological documents in the past 70 years (1949-2019)[J]. Publishing journal, 2019, 27(5): 22-29.

[14] LIU S C, XIAO F, OU W W, et al. Cascade ranking for operational e-commerce search[J]. arXiv: 1706.02093, 2017.

[15] PEDERSEN J. Query understanding at being[R]. Invited Talk: SIGIR, 2010.

[16] FAN Y X, XIE X H, CAI Y Q, et al. Pre-training methods in information retrieval[M]. Beijing: Now Publishers, 2022.

[17] CHEN R C, GALLAGHER L, BLANCO R, et al. Efficient cost-aware cascade ranking in multi-stage retrieval[C]// Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2017: 445-454.

[18] LIANG D, XU P, SHAKERI S, et al. Embedding-based zero-shot retrieval through query generation[J]. arXiv preprint arXiv:2009.10270, 2020.

[19] FURNAS G W, LANDAUER T K, GOMEZ L M, et al. The vocabulary problem in human-system communication[J]. Communications of the ACM, 1987, 30(11): 964-971.

[20] ZHAO L, CALLAN J. Term necessity prediction[C]// Proceedings of the 19th ACM international conference on Information and knowledge management. New York: ACM, 2010: 259-268.

[21] LI H, XU J. Semantic matching in search[J]. Foundations and trends in information retrieval, 2014, 7(5): 343-469.

[22] LAVRENKO V, CROFT W B. Relevance based language models[C]// Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2001: 120-127.

[23] LESK M E. Word-word associations in document retrieval systems[J]. American documentation, 1969, 20(1): 27-38.

[24] QIU Y G, FREI H P. Concept based query expansion[C]// Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 1993: 160-169.

[25] XU J X, CROFT W B. Query expansion using local and global document analysis[J]. ACM SIGIR forum, 2017, 51(2): 168-175.

[26] AGIRRE E, ARREGI X, OTEGI A. Document expansion based on WordNet for robust IR[C]. Posters: In Proceedings of COLING 2010, 2010: 9-17.

[27] EFRON M, ORGANISCIAK P, FENLON K. Improving retrieval of short texts through document expansion[C]// Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2012: 911-920.

[28] LIU X Y, CROFT W B. Cluster-based retrieval using language models[C]// Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2004: 186-193.

[29] GAO J F, NIE J Y, WU G Y, et al. Dependence language model for information retrieval[C]// Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2004: 170-177.

[30] GUO J F, CAI Y Q, FAN Y X, et al. Semantic models for the first-stage retrieval: A comprehensive review[J]. ACM transactions on information systems, 40(4): 1-42.

[31] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching word vectors with subword information[J]. Transactions of the association for computational linguistics, 2017, 5: 135-146.

[32] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[J]. arXiv: 1310.4546, 2013.

[33] PENNINGTON J, SOCHER R, MANNING C. Glove: Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2014.

[34] DAI Z Y, CALLAN J. Context-aware sentence/passage term importance estimation for first stage retrieval[J]. arXiv: 1910.10687, 2019.

[35] NOGUEIRA R, YANG W, LIN J, et al. Document expansion by query prediction[J]. ArXiv: 1904.08375, 2019.

[36] GILLICK D, PRESTA A, TOMAR G S. End-to-end retrieval in continuous space[J]. arXiv: 1811.08008, 2018.

[37] JANG K R, KANG J M, HONG G, et al. UHD-BERT: Bucketed ultra-high dimensional sparse representations for full ranking[J]. arXiv: 2104.07198, 2021.

[38] KHATTAB O, ZAHARIA M. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT[C]// Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2020: 39-48.

[39] ZAMANI H, DEHGHANI M, CROFT W B, et al. From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing[C]// Proceedings of the 27th ACM International Conference on Information and Knowledge Management. New York: ACM, 2018: 497-506.

[40] BARONI M, DINU G, KRUSZEWSKI G. Don' t count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors[C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2014.

[41] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model[J]. J Mach learn res, 2003, 3: 1137-1155.

[42] QIU X P, SUN T X, XU Y G, et al. Pre-trained models for natural language processing: A survey[J]. Science China technological sciences, 2020, 63(10): 1872-1897.

[43] WANG D B, LIU C, ZHU Z H, et al. Construction and application of pre-trained models of siku Quanshu in orientation to digital humanities[J]. Library tribune, 2022, 42(6): 31-43.

[44] WANG P Y, REN Z C. The uncertainty-based retrieval framework for ancient Chinese CWS and POS[C]. Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages, 2022: 164-8.

[45] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv: 1810.04805, 2018.

[46] CHE W X, GUO J, CUI Y M. Natural language processing[M]. Beijing: Publishing House of Electronics Industry, 2021.

[47] SHAO H, LIU Y F. Pre-training language model[M]. Beijing: Publishing House of Electronics Industry, 2021.

[48] REIMERS N, GUREVYCH I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2019.

[49] GAO T Y, YAO X C, CHEN D Q. SimCSE: Simple contrastive learning of sentence embeddings[J]. arXiv: 2104.08821, 2021.

Author Biographies:

Liu Nanzhu (1991-), female, master' s degree, research direction: library and information science.

Wang Mo (1987-), male, Ph.D., associate researcher, research directions: agricultural information technology, agricultural knowledge management, data mining.

*Corresponding Author: Cui Yunpeng (1972-), male, Ph.D., researcher, research directions: agricultural information technology, agricultural knowledge management, data mining. Email: cuiyunpeng@caas.cn

Note: Figure translations are in progress. See original paper for figures.