Ensemble-Based Multi-Specialty Retrieval: Integrating Diverse Similarity Metrics for Enhanced Question Answering

Chi-Hsiang, Chao
National Central University
Taipei, Taiwan
samuelchao921221@gmail.com

Hsin-Fu, Chang
National Central University
Taipei, Taiwan
xuanchang@g.ncu.edu.tw

Po-Yuan, Teng
National Central University
Taipei, Taiwan
davidteng00000@gmail.com

Abstract

We propose an innovative approach for multi-specialty retrieval in question answering systems by integrating diverse similarity measures through ensemble learning. Traditional machine reading comprehension methods struggle to accurately capture specialty-specific terminology and semantic nuances due to their reliance on generic models. Our framework addresses this challenge by leveraging multiple pre-trained embedding models tailored specifically for Chinese, English, and financial texts, combined with various similarity metrics, including cosine similarity, modified Euclidean similarity, and modified Manhattan similarity. The core novelty of our framework lies in effectively combining these domain-specific embeddings and diverse similarity metrics through both unsupervised and supervised ensemble strategies, enabling robust relevance prediction across heterogeneous contexts. Extensive experiments on domain-specific and challenging cross-specialty datasets demonstrate significant improvements in accuracy, F1-score, and precision compared to single-embedding and single metric-metric baselines.

Keywords

MRC (Machine Reading Comprehension), RAC (Retrieval Augmented Classification), Multi-Specialty Retrieval

1 Introduction

Recent advances in deep learning and large language models (LLMs) have significantly enhanced question-answering (QA) systems. Traditional Machine Reading Comprehension(MRC)[1] methods often depend on arge, parameter-intensive models, yet they frequently struggle to capture nuanced semantic information unique to specialized fields such as finance, healthcare, and multilingual content. Retrieval-Augmented Generation (RAG)[2] has emerged as an effective alternative by supplementing generative models with externally retrieved information, thereby enhancing accuracy and comprehensiveness.

Despite these advancements, designing an efficient retrieval module for multi-specialty applications remains challenging. Each specialty possesses distinct terminologies, stylistic conventions, and data distributions, causing embedding models optimized for one specialty to sub-optimized when applied to another. Each specialty-specific embedding model inherently exhibits unique strengths and weaknesses, making it difficult to select a universally optimal solution.

To overcome these limitations, we propose an ensemble-based approach that strategically integrates multiple domain-specific embedding models with diverse similarity metrics—cosine similarity, modified Euclidean similarity, and modified Manhattan similarity. By combining these embeddings and metrics through both unsupervised and supervised ensemble techniques, our approach effectively leverages their complementary strengths, enhancing retrieval robustness even under resource-constrained scenarios.

2 Related Work

2.1 DIEM

Federico Tessari *et al.* [3] investigate the limitations of commonly used similarity metrics—namely, cosine similarity, Euclidean distance, and Manhattan distance—in high-dimensional spaces. Their analysis highlights that both Euclidean and Manhattan distances can suffer from deviations and instability due to the curse of dimensionality, which may negatively impact the accuracy of similarity computations. However, our preliminary experiments have shown that these metrics still hold promise in capturing diverse geometric properties within the embedding space. As a result, we include them in our experimental framework to better address semantic variations across multiple domains.

2.2 BGE M3-Embedding

The M3-Embedding framework, presented in [4], excels in multilingual, multifunctional, and multigranularity retrieval tasks. Using self-knowledge distillation, the approach demonstrates exceptional adaptability and robustness in diverse domains. The method effectively integrates semantic information from various sources, yielding improved performance in cross-domain retrieval tasks. Inspired by this work, our study adopts a multi-domain design that utilizes domain-specific embedding models along with multiple similarity metrics, aiming to improve performance in heterogeneous environments.

2.3 Multi-domain learning to rank

Maintaining multiple domain-specific information retrieval is costly. Deep domain adaptation (DDA) techniques address this by learning domain-invariant representations for knowledge transfer, but they may degrade performance in domain-specific tasks due to the loss of specialized features. Recent research [5] present Deep Domain Specialization (DDS), which consolidates multiple domains into a single ranking model while preserving domain-specific feature representations as a solution. By retaining specialized knowledge for each domain, DDS achieves superior performance without sacrificing domain effectiveness.

3 Dataset

This section outlines the datasets used for both training and testing, including Chinese, English, and financial question-responses.

3.1 Training Data

To capture specialty-specific characteristics across Chinese, English, and financial texts, we aggregated several publicly available QA datasets into a unified corpus. This unified approach allows our retrieval classification models to learn from diverse specialty-specific contexts. The following datasets were used for training.

3.1.1 Chinese specialty Question Answering Dataset. We employ three Chinese QA datasets to enhance our model's performance in Chinese comprehension:

- **DuReader[6]**: An open-domain machine reading comprehension dataset from Baidu, featuring real-world user queries and corresponding passages.
- **ChineseSquad**[1]: A Chinese adaptation of the original SQuAD dataset, generated via machine translation and manually refined to ensure high quality.
- **WebQA[7]**: A large-scale human-annotated dataset developed by Baidu based on web contexts.

For each dataset, we randomly selected a subset of QA pairs (as detailed in Table 1), resulting in a total of 10,000 pairs for the Chinese specialty.

3.1.2 English Question Answering Datasets. To improve our English comprehension capabilities, we integrate three English QA datasets:

- **SearchQA**[2]: This is a split derived from the MRQA 2019 dataset[8], a task focused on generalization in question answering, which has been reformatted and filtered specifically for question answering.
- **Disfl_qa[9]**: A dataset focused on contextual disfluencies in information-seeking scenarios from Google. We only utilize it's context and question-answering pairs.
- **Duorc[10]**: A dataset, composite by IBM, containing questions and answers collected from crowdsourced AMT workers based on Wikipedia and IMDB movie plots.

A random selection of QA pairs from each dataset (refer to Table 1) yields a total of 10,000 pairs for the English specialty.

3.1.3 Financial Question Answering Datasets. We also include two financial QA datasets to boost the model's capability in financial comprehension:

- **Finqa[11]**: A dataset designed for complex numerical reasoning over financial documents.
- **Sujet-Finance-QA-Vision-100k**[3]: A collection of financial document images with corresponding textual annotations and AI-generated QA pairs.

From these datasets, we randomly selected QA pairs (as shown in Table 1) to form a total of 10,000 pairs for the financial specialty.

3.1.4 Data Preprocessing. To comprehensively cover the Chinese, financial, and English specialties, we aggregated eight publicly available QA datasets. Each dataset contributes a specified number of correct QA pairs, selected from its total pool of correct and incorrect pairs. Incorrect pairs were generated by pairing each query with a mismatched context from the same dataset, ensuring a balanced distribution for the binary classification task. In total, the training corpus comprises 30,000 QA pairs (10,000 per specialty).

Table 1: Overview of the training datasets used in this study.

Source Dataset	Specialty	Selected QA Pairs
DuReader	Chinese	6666
ChineseSquad	Chinese	6666
WebQA	Chinese	6668
Finqa	Financial	10000
Sujet-Finance-QA-Vision-100k	Financial	10000
SearchQA	English	6666
Disfl_qa	English	6666
Duorc	English	6668

Table 2: Overview of the testing datasets used in this study.

Source Dataset	Specialty	QA Pairs Quantity
DRCD	Chinese	67906
NQ	English	15660
Financial-QA-10K	Financial	14000
BiPaR	Chinese&English	29336
AICUP2024	Chinese&Finance	2570

3.2 Testing Data

The testing set consists of five datasets (as shown in Table 2), each targeting different specialties and evaluation aspects:

3.2.1 DRCD (Delta Reading Comprehension Dataset)[12]. DRCD is a traditional Chinese machine reading comprehension dataset developed by Delta Electronics. Pair Chinese texts with the corresponding questions and answers, providing a challenging evaluation of the retrieval and quality control capabilities of the system.

3.2.2 Natural Questions[13]. Natural Questions (NQ) is a large-scale QA dataset. It includes real user queries from the Google search and corresponding answers extracted from Wikipedia passages, representing the English specialty.

3.2.3 Financial-QA-10K. Financial-QA-10K[4] comprises 10,000 QA pairs extracted from company financial reports. Covering topics such as financial analysis, company operations, and strategic insights, this dataset serves as a benchmark to evaluate the retrieval module in a specialty with specialized terminology and numerical reasoning.

3.2.4 BiPaR[14]. BiPaR is a manually annotated bilingual dataset originally developed for novel-style reading comprehension. In this study, we exclusively use its monolingual English subset, directly translated from Chinese, to evaluate how "Chinese-style" elements influence English contexts.

3.2.5 AICUP 2024 Dataset. AICUP[5] is a series of AI competitions in Taiwan challenging participants to develop machine learning

[1] https://github.com/pluto-junzeng/ChineseSquad
[2] https://huggingface.co/datasets/lucadiliello/searchqa
[3] https://huggingface.co/datasets/sujet-ai/Sujet-Finance-QA-Vision-100k
[4] https://www.kaggle.com/datasets/yousefsaeedian/financial-q-and-a-10k
[5] https://www.aicup.tw

solutions for real-world tasks. This dataset comes from one of the 2024 competition organized by E.SUN Bank, which focuses on applying RAG to financial question answering in Chinese based financial dataset.

4 Experiment

4.1 Similarity Method

In our experiments, we adopt three widely-used vector similarity and distance measures: cosine similarity, Euclidean distance, and Manhattan distance. Each method captures different aspects of vector relationships. However, while cosine similarity naturally yields a score in a bounded range (typically interpreted as 0 to 1), both Euclidean and Manhattan distances produce unbounded values in the range $[0, \infty)$. To facilitate a uniform comparison across all methods and to integrate them seamlessly into our downstream tasks, we transform the distance-based measures into normalized similarity scores using an exponential decay function.

Cosine Similarity: Cosine similarity evaluates the cosine of the angle between two vectors $\mathbf{a}$ and $\mathbf{b}$:

$$\frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\| + 10^{-10}} \tag{1}$$

where a small constant 10^{-10} is added to avoid division by zero. This metric inherently reflects the directional alignment between vectors.

Modified Euclidean Similarity: The conventional Euclidean distance, $\|\mathbf{a} - \mathbf{b}\|$, outputs values from 0 to infinity, with lower values indicating higher similarity. To map this distance to a similarity score between 0 and 1, we apply an exponential decay transformation:

$$e^{-0.1\|\mathbf{a}-\mathbf{b}\|} \tag{2}$$

This modification is motivated by several considerations:

- **Normalization:** The exponential function compresses the unbounded Euclidean distance into a bounded similarity score, making it directly comparable to cosine similarity.
- **Monotonic Decay:** The function decreases monotonically as the distance increases, ensuring that smaller distances correspond to higher similarity scores.
- **Sensitivity Control:** The decay factor (0.1) in our exponential decay function was empirically selected through preliminary experiments, regulating how quickly similarity decreases with increasing distance. This parameter effectively controls the sensitivity and ensures meaningful similarity interpretations across embedding spaces. While our empirical choice balances performance and interpretability effectively, future work could explore alternative normalization strategies such as inverse distance or softmax transformations, providing a deeper theoretical grounding and potential further performance improvements.

Modified Manhattan Similarity: Similarly, the Manhattan distance, given by $\sum |\mathbf{a} - \mathbf{b}|$, is unbounded. We convert it to a similarity score in the $[0, 1]$ range using a comparable exponential decay function, identical to Modified Euclidean Similarity:

$$e^{-0.1 \sum |\mathbf{a}-\mathbf{b}|} \tag{3}$$

The reasons for this modification are analogous to those for the Euclidean case.

In general, by converting the Euclidean and Manhattan distances into normalized similarity scores, we establish a consistent framework to compare vector similarities in different methods.

4.2 Embedding Models

We employ three pre-trained embedding models, each tailored for a specific specialty, to generate robust vector representations for our retrieval tasks.

Yuan-embedding-1.0 (Chinese Specialty)[6]: This model is specifically designed for Chinese context retrieval tasks and is available on Hugging Face. Although there is no dedicated paper for embedding, the underlying data is derived from the work presented by Inspur. In particular, Yuan-embedding-1.0 has demonstrated leading accuracy in the retrieval task of the legacy (before February 2025) MTEB benchmark, making it a strong choice for Chinese language applications.

all-MiniLM-L6-v2 (English Specialty)[7]: Developed by the sentence-transformers team and accessible through Hugging Face, the all-MiniLM-L6-v2 model is renowned for its high performance and compact model size. Given its robust representation capabilities and efficiency, we have selected this model as our primary embedding method for the English specialty.

finance-embeddings-investopedia (Finance Specialty)[8]: This embedding model, provided by FINLANG TRANSLATION SERVICES LIMITED, is fine-tuned on BAAI/bge-base-en v1.5. It is particularly well-suited for tasks such as clustering and semantic search in Retrieval Augmented Generation (RAG) applications. Its selection is motivated by its high download count in the relevant specialty and its strong performance with a reasonable parameter size, as observed in our preliminary tests.

4.3 Baseline 1: Embedding-Based Similarity Computation

In our first baseline experiment, we employed a single pre-trained embedding model to transform both context and query sentences into fixed-dimensional vector representations. Among various similarity metrics, Cosine Similarity was selected due to its widespread application in embedding-based retrieval tasks. For each context–query pair, a uniform threshold of 0.7 was applied, categorizing pairs as *relevant* if their similarity scores exceeded this threshold and as *irrelevant* otherwise. We evaluated the performance of this approach across multiple datasets from diverse specialties, using Accuracy, F1-score, and Precision as primary evaluation metrics. Detailed results obtained using Cosine Similarity are presented in Table 3.

4.4 Baseline 2: Multi-Embedding Similarity Ensemble via Averaging

To further explore the potential of embedding-based similarity, we extend our approach by leveraging multiple pre-trained embedding models. For each model, we extract the same three similarity scores

[6] https://huggingface.co/IEITYuan/Yuan-embedding-1.0
[7] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[8] https://huggingface.co/FinLang/finance-embeddings-investopedia

Table 3: Baseline 1 Performance

Specialty	Datasets	Cosine Similarity Performance		
		Accuracy	F1-score	Precision
Chinese	DRCD	0.9333	0.9357	0.9033
Chinese	financial-qa-10K	0.6159	0.7223	0.5656
Chinese	NQ	0.8626	0.8785	0.7874
Chinese	BiPaR	0.5683	0.6959	0.5372
Chinese	AICUP2024	0.5821	0.7027	0.5453
English	DRCD	0.5094	0.0372	0.9923
English	financial-qa-10K	0.8493	0.8228	0.9980
English	NQ	0.6126	0.3676	1.0000
English	BiPaR	0.5062	0.0246	0.9946
English	AICUP2024	0.5128	0.1448	0.5922
Finance	DRCD	0.5630	0.2352	0.9408
Finance	financial-qa-10K	0.8739	0.8560	0.9977
Finance	NQ	0.5430	0.1585	1.0000
Finance	BiPaR	0.5026	0.0103	1.0000
Finance	AICUP2024	0.5198	0.3830	0.5357

as in Baseline 1. We then aggregate these scores by computing their arithmetic mean. Specifically, we investigate the following ensemble configurations:

- **c**: Cosine similarity only.
- **e**: Modified Euclidean similarity only.
- **m**: Modified Manhattan similarity only.
- **c,e,m**: Mean of all three similarity scores.

A fixed threshold of 0.7 is applied to the aggregated score for relevance determination. Table 4 summarizes the performance of these ensemble configurations across various datasets.

In certain datasets, the *m* and *c,e,m* configurations exhibit an F1-score and precision of 0, coupled with an accuracy of 0.5. This pattern arises possibly due to the threshold-based classification mechanism employed. Because normalized Manhattan similarity scores (*m*) or their arithmetic mean with other metrics (*c,e,m*) seldom exceed the fixed threshold of 0.7, the classifier effectively labels all pairs as negative. In a balanced dataset with an equal number of positive and negative samples, always predicting the negative class leads to an accuracy of 0.5 (random guessing), while resulting in zero precision and recall for the positive class. Hence, the inherent low range of these similarity scores under the chosen threshold compels the model to make uniform (all-negative) predictions, yielding the observed metrics.

4.5 K-Means Clustering for Unsupervised Integration

Inspired by previous research[15], which found that while the Euclidean distance is effective in k-means clustering, its performance deteriorates in high-dimensional spaces, whereas cosine similarity remains robust, we propose an unsupervised integration strategy using k-means clustering. In our approach, we compute three similarity scores (cosine, modified Euclidean, and modified Manhattan) from each embedding model for every context–query pair and concatenate them into a single feature vector. We perform k-means clustering with $k = 2$ to partition the pairs into two clusters, corresponding to *relevant* and *irrelevant* relationships, with the cluster exhibiting higher average similarity values designated as *relevant*. New context–query pairs are assigned to the closest cluster based on the Euclidean distance. Table 5 summarizes the performance metrics of this approach across our datasets, demonstrating its effectiveness in uncovering latent similarity structures and serving as a valuable alternative to averaging or supervised methods when labeled data is limited.

Table 4: Baseline 2 Performance for Different Ensemble Combinations

Source Dataset	Union	Accuracy	F1-score	Precision
DRCD	c	0.5581	0.2083	0.9992
	e	0.5216	0.0829	1.0000
	m	0.5000	0.0000	0.0000
	c,e,m	0.5000	0.0001	1.0000
financial-qa-10K	c	0.9329	0.9282	0.9972
	e	0.7581	0.6812	1.0000
	m	0.5001	0.0003	1.0000
	c,e,m	0.5001	0.0006	1.0000
NQ	c	0.6677	0.5023	1.0000
	e	0.5404	0.1496	1.0000
	m	0.5000	0.0000	0.0000
	c,e,m	0.5000	0.0000	0.0000
BiPaR	c	0.5107	0.0418	0.9968
	e	0.5001	0.0004	1.0000
	m	0.5000	0.0000	0.0000
	c,e,m	0.5000	0.0000	0.0000
AICUP2024	c	0.5767	0.4499	0.6421
	e	0.5595	0.2215	0.9527
	m	0.5000	0.0000	0.0000
	c,e,m	0.5000	0.0000	0.0000

4.6 Multi-Embedding Similarity Ensemble with RandomForest

Building on the previous k-means approaches, we next incorporate a supervised learning method using a RandomForest classifier to fuse the similarity scores. For every context–query pair, the cosine, modified Euclidean, and modified Manhattan similarities from each embedding model are concatenated into a comprehensive feature vector. The RandomForest classifier is configured with 100 trees, a minimum samples split of 2, and a minimum samples leaf of 1. The classifier is trained using these feature vectors to predict relevance. A fixed threshold of 0.7 is applied to the output probability - pairs with probabilities greater than this value are classified as relevant. This method leverages the ensemble nature of RandomForest to capture nonlinear interactions among the similarity features, which

Table 5: Performance of the k-means-based Multi-Embedding Similarity Ensemble.

Source Dataset	Union	Accuracy	F1-score	Precision
DRCD	c	0.6227	0.7110	0.5762
	e	0.8324	0.7988	0.9984
	m	0.7213	0.6466	0.8831
	c,e,m	0.6419	0.7201	0.5911
financial-qa-10K	c	0.9393	0.9427	0.8927
	e	0.9303	0.9261	0.9848
	m	0.9784	0.9782	0.9835
	c,e,m	0.9444	0.9472	0.9011
NQ	c	0.9732	0.9726	0.9982
	e	0.8397	0.8092	0.9997
	m	0.8908	0.8774	0.9998
	c,e,m	0.9711	0.9703	0.9984
BiPaR	c	0.7457	0.6887	0.8877
	e	0.5545	0.1986	0.9876
	m	0.6056	0.3562	0.9691
	c,e,m	0.7411	0.6790	0.8934
AICUP2024	c	0.5642	0.6913	0.5352
	e	0.6969	0.6895	0.7066
	m	0.6000	0.6623	0.5731
	c,e,m	0.5708	0.6944	0.5392

often yields improved performance over simple averaging. Performance metrics for various embedding combinations are detailed in Table 6.

4.7 Multi-Embedding Similarity Ensemble with MLP

Finally, we investigate a neural network–based integration strategy using a Multi-Layer Perceptron (MLP). As before, the similarity scores extracted from multiple embedding models are concatenated to form the input feature vector. The MLP architecture is structured as follows:

- **Input Layer**: Accepts the concatenated similarity feature vector with 9 features.
- **Hidden Layers**: Three fully connected hidden layers with 64, 16, and 4 neurons, respectively, each followed by a ReLU activation function.
- **Dropout**: No explicit dropout mechanism is applied in this architecture.
- **Output Layer**: A single neuron with a sigmoid activation function outputs the probability of relevance.

The network is trained using binary cross-entropy loss and optimized with the Adam optimizer. The batch size is set to 32, and an L2 regularization term with $\alpha = 0.0001$ is applied to mitigate overfitting. Consistent with the other methods, a threshold of 0.7 is used to determine relevance. The MLP's ability to learn complex nonlinear interactions among the similarity features can lead to more refined decision boundaries. Experimental results for various embedding combinations are reported in Table 7.

Table 6: Performance of the RandomForest-based Multi-Embedding Similarity Ensemble.

Source Dataset	Union	Accuracy	F1-score	Precision
DRCD	c	0.9074	0.8995	0.9834
	e	0.9216	0.9185	0.9556
	m	0.9228	0.9200	0.9552
	c,e,m	0.9160	0.9094	0.9881
financial-qa-10K	c	0.9385	0.9415	0.8978
	e	0.9419	0.9431	0.924
	m	0.9416	0.9429	0.9222
	c,e,m	0.9483	0.9503	0.9151
NQ	c	0.9550	0.9529	0.9988
	e	0.9565	0.9546	0.9986
	m	0.9554	0.9533	0.9986
	c,e,m	0.9520	0.9497	0.9989
BiPaR	c	0.7429	0.6776	0.9079
	e	0.7240	0.6429	0.9107
	m	0.7258	0.6458	0.912
	c,e,m	0.7389	0.6681	0.9166
AICUP2024	c	0.6556	0.7273	0.6021
	e	0.6467	0.7223	0.5949
	m	0.6490	0.7245	0.5963
	c,e,m	0.6595	0.7304	0.6046

Table 7: Performance of the MLP-based Multi-Embedding Similarity Ensemble.

Source Dataset	Union	Accuracy	F1-score	Precision
DRCD	c	0.8962	0.8849	0.9938
	e	0.9353	0.9327	0.9729
	m	0.6606	0.5187	0.8912
	c,e,m	0.9301	0.9259	0.9849
financial-qa-10K	c	0.9537	0.9552	0.9245
	e	0.9456	0.9471	0.9226
	m	0.9590	0.9582	0.9782
	c,e,m	0.9466	0.9490	0.9087
NQ	c	0.9463	0.9433	0.9999
	e	0.9582	0.9565	0.9985
	m	0.8997	0.8885	0.9996
	c,e,m	0.9611	0.9596	0.999
BiPaR	c	0.7229	0.6348	0.9309
	e	0.7238	0.6398	0.9194
	m	0.6688	0.5226	0.9353
	c,e,m	0.7470	0.6816	0.9186
AICUP2024	c	0.6650	0.7317	0.6102
	e	0.6572	0.7298	0.6022
	m	0.5977	0.6190	0.5878
	c,e,m	0.6626	0.7327	0.6068

5 Results

Our experiments compared the performance of single similarity computation methods with machine-learning-based ensemble methods for retrieval across multiple specialties. The findings indicate that while a baseline using cosine similarity achieves decent performance on single-specialty datasets (such as DRCD, Financial-QA-10K, and NQ), it falls short on cross-specialty datasets like BiPaR and AICUP2024, where both accuracy and F1-scores are notably lower. This suggests that relying on a single similarity measure may be insufficient for capturing the nuanced semantic variations present in heterogeneous datasets.

5.1 Ensemble Strategy Improvement

By integrating multiple similarity metrics—cosine, modified Euclidean, and modified Manhattan—and applying supervised learning approaches (RandomForest and MLP) to fuse these features, we observed significant performance gains. For instance, on the Financial-QA-10K dataset, the baseline cosine similarity method achieved an accuracy of approximately 87.4%, whereas the RandomForest ensemble improved accuracy to 94.8%, and the MLP-based approach further boosted it to around 95.0%. Similar improvements were observed in the DRCD and NQ datasets, where supervised ensemble methods better captured the complementary information offered by the different embedding models and similarity measures.

5.2 Challenges and Observations on Cross-Specialty Datasets

The cross-specialty datasets, particularly BiPaR and AICUP2024, pose unique challenges due to the inherent diversity in language styles and specialty-specific terminology:

- **BiPaR Dataset**: The baseline cosine similarity method yielded accuracies in the range of 50–57%. However, when using an MLP-based ensemble with multiple similarity metrics, accuracy increased to approximately 74.7%. This improvement underscores the effectiveness of nonlinear feature integration in addressing semantic inconsistencies across languages and specialties.
- **AICUP2024 Dataset**: This dataset also presents cross-specialty challenges, combining elements of both Chinese and financial contexts. Here, the baseline accuracy of about 58.2% (for the Chinese specialty) improved to roughly 65.9% with the RandomForest ensemble. Although the gains are less pronounced than in single-specialty settings, the improvement highlights the relative stability of ensemble methods when dealing with complex, heterogeneous data.

It is noteworthy that the modified Manhattan similarity on its own consistently performed poorly across experiments. However, its inclusion in the ensemble—when combined with cosine and modified Euclidean similarity—helped enhance the overall model's ability to discern subtle semantic differences, particularly in the cross-specialty scenarios.

6 Interesting Phenomena and Future Directions

- **Nonlinear Complementarity**: The supervised ensemble methods (both RandomForest and MLP) are capable of capturing nonlinear interactions among the different similarity metrics. This synergy is especially beneficial in cross-specialty contexts, where semantic patterns are more diverse and complex.
- **Cross-Specialty Limitations**: Despite the improvements, the ensemble methods still show relatively lower performance gains on cross-specialty datasets (BiPaR and AICUP2024) compared to single-specialty datasets. This suggests that further enhancements may be achieved by incorporating additional specialty-specific embedding models or by developing more adaptive fusion strategies.
- **Selective Fusion Strategy**: In some cases, a single metric (e.g., cosine similarity or modified Euclidean similarity) performed comparably to or even slightly better than a naive fusion of all three metrics. This observation indicates that in certain scenarios, indiscriminate fusion might introduce noise, and a more selective or weighted fusion approach could yield better results.

In summary, our study demonstrates that combining multiple similarity measures through supervised machine learning ensembles can effectively boost retrieval performance, particularly in challenging cross-specialty environments. Future work will explore advanced fusion strategies and the integration of additional specialty-specific features to further enhance model robustness and accuracy.

7 Conclusion

This study presented a comprehensive evaluation of embedding models and similarity metrics for multi-specialty retrieval and classification tasks. Our experiments demonstrated that supervised ensemble methods—by integrating multiple similarity measures such as cosine, modified Euclidean, and modified Manhattan—consistently outperformed baseline approaches. In particular, these supervised techniques improved accuracy and F1-scores by approximately 5–10% compared to single-metric methods.

The combination of diverse similarity metrics enabled the models to capture richer semantic nuances, especially in challenging cross-specialty datasets like BiPaR and AICUP2024. Additionally, supervised methods were found to be more effective than unsupervised clustering approaches, highlighting the importance of labeled data and nonlinear feature integration in complex retrieval tasks.

Future work will focus on further optimizing these ensemble strategies, potentially incorporating additional specialty-specific embedding models and exploring advanced neural network architectures to develop even more robust multi-specialty retrieval systems.

8 Acknowledgments

We would like to express our gratitude to E.SUN Bank for providing the AICUP2024 dataset. Their support and the accessibility of the data were instrumental in making this research possible. We greatly appreciate their contribution and assistance throughout this project.

9 Appendices

The following is our model architecture figure.

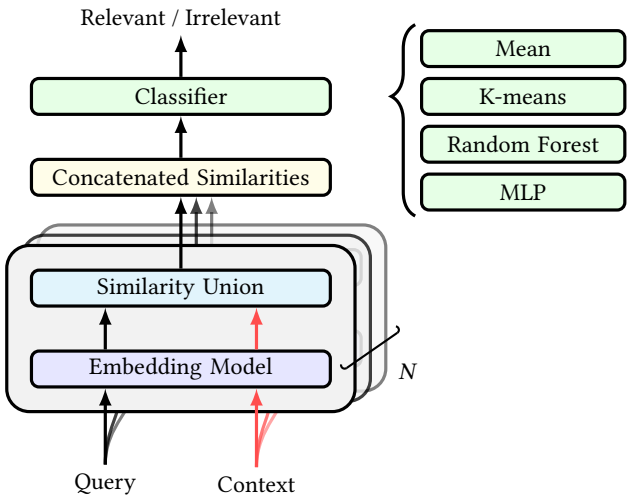

Figure 1: Relevance Prediction Architecture
N is the number of combinations of different similarity unions and embedding models

References

[1] Changchang Zeng, Shaobo Li, Qin Li, Jie Hu, and Jianjun Hu. A Survey on Machine Reading Comprehension—Tasks, Evaluation Metrics and Benchmark Datasets. *Applied Sciences*, 10(21):7640, January 2020. Number: 21 Publisher: Multidisciplinary Digital Publishing Institute.

[2] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-Augmented Generation for Large Language Models: A Survey. 2023. Publisher: arXiv Version Number: 5.

[3] Federico Tessari, Kunpeng Yao, and Neville Hogan. Surpassing Cosine Similarity for Multidimensional Comparisons: Dimension Insensitive Euclidean Metric (DIEM), December 2024. arXiv:2407.08623 [cs].

[4] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation, June 2024. arXiv:2402.03216 [cs].

[5] Abdelmaseeh Felfel and Paul Missault. Deep Domain Specialisation for single-model multi-domain learning to rank, 2024.

[6] Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications. In Eunsol Choi, Minjoon Seo, Danqi Chen, Robin Jia, and Jonathan Berant, editors, *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[7] Peng Li, Wei Li, Zhen He, Xuguang Wang, Ying Cao, Jie Zhou, and W. Xu. Dataset and Neural Recurrent Sequence Labeling Model for Open-Domain Factoid Question Answering. *ArXiv*, July 2016.

[8] Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. MRQA 2019 Shared Task: Evaluating Generalization in Reading Comprehension. In Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen, editors, *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China, January 2019. Association for Computational Linguistics.

[9] Aditya Gupta, Jiacheng Xu, Shyam Upadhyay, Diyi Yang, and Manaal Faruqui. Disfl-QA: A Benchmark Dataset for Understanding Disfluencies in Question Answering. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3309–3319, Online, August 2021. Association for Computational Linguistics.

[10] Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. DuoRC: Towards Complex Language Understanding with Paraphrased Reading Comprehension. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1693, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[11] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A Dataset of Numerical Reasoning over Financial Data. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic, January 2021. Association for Computational Linguistics.

[12] Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. DRCD: a Chinese Machine Reading Comprehension Dataset, May 2019. arXiv:1806.00920 [cs].

[13] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. Place: Cambridge, MA Publisher: MIT Press.

[14] Yimin Jing, Deyi Xiong, and Zhen Yan. BiPaR: A Bilingual Parallel Dataset for Multilingual and Cross-lingual Reading Comprehension on Novels. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2452–2462, Hong Kong, China, January 2019. Association for Computational Linguistics.

[15] Ruiyu Yang, Yuxiang Jiang, Scott Mathews, Elizabeth A. Housworth, Matthew W. Hahn, and Predrag Radivojac. A new class of metrics for learning on real-valued and structured data. *Data Mining and Knowledge Discovery*, 33(4):995–1016, July 2019.