A Chinese Short Text Similarity Method Integrating Sentence-level and Phrase-level Semantics

Zhenji Shen and Zhiyong Xiao *

Posted Date: 7 November 2024

doi: 10.20944/preprints202411.0453.v1

Keywords: Short text similarity; Chinese sentence pair classification; BERT; external knowledge integration

Article

A Chinese Short Text Similarity Method Integrating Sentence-Level and Phrase-Level Semantics

Zhenji Shen and Zhiyong Xiao

School of Artificial Intelligence and Computer Science, Jiangnan University
* Correspondence: zhiyong.xiao@jiangnan.edu.cn

Abstract: Short text similarity, as a pivotal research domain within Natural Language Processing (NLP), has been extensively utilized in intelligent search, recommendation systems, and question-answering systems. The majority of existing models for short text similarity concentrate on aligning the overall semantic content of entire sentences, frequently neglecting the semantic correlations between individual phrases within the sentences. This challenge is particularly acute in the Chinese language context, where synonyms and near-synonyms can introduce substantial interference in the computation of text similarity. In this paper, we introduce a short text similarity computation methodology that integrates both sentence-level and phrase-level semantics. By harnessing vector representations of Chinese words/phrases as external knowledge, our approach amalgamates global sentence characteristics with local phrase features to compute short text similarity from diverse perspectives, spanning from the global to the local level. Experimental findings substantiate that the proposed model surpasses previous approaches in Chinese short text similarity tasks. Specifically, it attains an accuracy of 90.16% on the LCQMC, marking an enhancement of 2.23% over ERNIE and 1.46% over the previously top-performing model, Glyce + BERT.

Keywords: Short text similarity; Chinese sentence pair classification; BERT; external knowledge integration

1. Introduction

Text similarity refers to the task of evaluating the extent of similarity between two sentences on the basis of their semantic, structural, and content-related attributes. It has substantial applications across diverse fields, including intelligent search, recommendation systems, and question-answering systems. As natural language processing (NLP) technologies continue to advance, enhancing the precision of text similarity calculations has emerged as a pivotal area of research focus, especially for short text similarity.

In recent years, deep learning algorithms, which have undergone training on extensive textual data, have exhibited the capacity to learn the semantic representations of texts. These algorithms possess the ability to represent words and sentences in the form of high-dimension vectors, better for natural language understanding. A notable example of such models is BERT[1], which is based on Transformers[2]. Through unsupervised pre-training on extensive text corpora, BERT effectively captures contextual representations of sentences, resulting in exceptional performance. However, despite the significant advancements, models based on Transformers still have several limitations.

Deep learning models are significantly dependent on the data they have been trained on and lack the ability to incorporate novel or extraneous knowledge beyond their training corpus[3]. Although models, such as BERT, can capture contextual relationships between words, they are unable to autonomously reason or leverage knowledge that exceeds the scope of their training data. In contrast to humans, deep learning models do not possess external world knowledge and are unable to infer or reason about information that transcends the text.

Furthermore, models face challenges in processing information related to temporal, spatial, or external factual updates. The data utilized during the pre-training phase of BERT is typically static,

 doi:10.20944/preprints202411.0453.v1

whereas real-world knowledge is continuously evolving. As time progresses and new knowledge emerges, these models fail to automatically update or integrate such new information[4]. Consequently, predictions based on outdated data may experience reduced accuracy.

These constraints also impact the effectiveness of models in handling tasks involving intricate reasoning, including common-sense reasoning or cross-domain inference, which require the system to engage in more elaborate reasoning processes that transcend basic text-based comprehension. These obstacles present opportunities for future advancements in Natural Language Processing (NLP), particularly regarding the integration of external knowledge with current deep learning models, a subject of growing interest among both academic researchers and industry professionals.

To overcome the limitations of deep learning models in reasoning with external knowledge, numerous researchers have achieved notable advancements by incorporating external knowledge into language models. Velickovic et al.[5], introduced knowledge graphs utilizing Graph Neural Networks (GNNs) to facilitate effective reasoning regarding complex relational structures. Lee et al.[6] investigated the possibility of treating language models as knowledge bases, thereby enhancing reasoning capabilities through the integration of structured knowledge bases. Petroni et al.[7] put forward a multi-task learning framework that integrates pre-trained language models with external knowledge bases, significantly boosting performance in open-domain question answering tasks. Wang et al.[8] systematically summarized methods for incorporating external knowledge into open-domain question answering through a combination of retrieval and reading strategies. Chen et al.[9] proposed a Retriever-Reader architecture that dynamically retrieves knowledge, resulting in substantial improvements in model accuracy. Ren et al.[10] improved model performance on domain-specific and common-sense reasoning tasks by embedding knowledge graphs into language model representations.

Furthermore, the work of Zhou et al.[11] introduced the KEBERT-GCN model, which incorporates external knowledge bases for computing sentence similarity. This model employs BERT to generate a novel adjacency matrix. This is achieved by performing a Hadamard product between the similarity matrix derived from external knowledge bases and the attention matrix, which is subsequently input into a Graph Convolutional Network (GCN) to capture intricate relationships between words. Notably, the model constructs the similarity matrix using semantic similarities between words from WordNet. This approach enhances the utilization of external knowledge within the attention layer. This work efforts have significantly contributed to the advancement of external knowledge integration in Natural Language Processing (NLP), resulting in substantial improvements in model performance for tasks involving complex reasoning and similarity calculations.

Models incorporating external knowledge have achieved remarkable success in English Natural Language Processing tasks[12], yet their application to Chinese poses distinct challenges. For example, prevalent methods such as the KEBERT-GCN model typically dissect sentences into individual words to construct similarity matrices. However, this technique frequently fails to adequately encapsulate the comprehensive semantics of Chinese sentences. In English, utilities like WordNet can be utilized to compute semantic distances between words, yielding semantic similarity matrices as inputs derived from external knowledge, thereby facilitating the understanding of word relationships. This methodology has proven efficacious, particularly in managing intricate semantic relationships. Conversely, in Chinese, the absence of a direct equivalent to WordNet restricts the applicability of this approach in Chinese text processing.

Moreover, these methods exhibits certain limitations even in English. It predominantly focuses on the relationships between individual words while often neglecting the actual contribution of each word to the overall semantic similarity computation. For instance, words possess varying degrees of importance within a sentence, yet existing similarity matrices may inadequately reflect this. Current methods[13,14] also tend to prioritize global semantic information, without adequately capturing intricate relationships within the local context. These issues are particularly acute in Chinese text processing, where the intricacy of word and character combinations necessitates that models address both global and local semantic attributes. As a result, new methods must be introduced to address

 doi:10.20944/preprints202411.0453.v1

the scarcity of external knowledge resources for Chinese and delve into integrate both global and local information for enhanced contextual comprehension in Chinese NLP.

To address the aforementioned challenges, this research introduces a methodology that incorporates external knowledge to enhance the performance of models in Chinese NLP tasks. This methodology utilizes vector representations derived from Tencent's 8-million-word corpus and Netease's embedding corpus as external knowledge resources for the BERT model, thereby allowing the model to access detailed, word-level external knowledge inputs in Chinese. The proposed methodology is highly flexible, as it allows for the substitution of the external knowledge base with domain-specific word embeddings tailored to the requirements of specific tasks, thus improving performance within those domains. Moreover, the structure of the proposed model is not limited to BERT and can be adapted to other pre-trained models, making it suitable for a variety of training tasks and data scenarios. Experimental results obtained on the LCQMC demonstrate that the proposed model significantly outperforms the traditional BERT model, especially in tasks involving semantic similarity calculation and contextual understanding. When compared to existing models, it achieves an accuracy of 90.16% on the LCQMC, representing an improvement of 2.23% over ERNIE and 1.46% over the previously top-performing model, Glyce + BERT, providing an effective solution for Chinese NLP tasks.

The key contributions of this paper are as follows:

1. Flexible Integration of External Knowledge: We introduce a methodology that facilitates the utilization of external knowledge,such as Tencent's 8-million-word corpus and Netease's embedding corpus, as inputs for Chinese word-level semantic understanding. This methodology offers the flexibility to incorporate domain-specific word embedding libraries based on task requirements.
2. Combination of Global and Local Information: By integrating external knowledge with internal attention mechanisms, the methodology effectively combines global semantic information with local contextual relationships. This addresses the limitations of previous methodologies that focused solely on global relationships.
3. Task Generalizability: The proposed methodology transcends the confines of short text similarity, extending its application to enhance semantic understanding and recognition accuracy across various NLP tasks. We conduct innovative experiments on Named Entity Recognition (NER), demonstrating remarkable results.
4. Extensive Experimental Validation: The proposed methodology undergoes extensive validation on datasets such as MSRA-NER. The results demonstrate not only superior performance in general tasks but also versatility across various language processing scenarios. These findings provide new insights for future NLP research.

The remainder of this paper is structured as follows: Section 2 presents the related works; Section 3 outlines the proposed method, discussing the overall model design, training details, and the integration of global and local information. Section 4 presents the experimental results, while Section 5 extends the application to other tasks. Finally, Section 6 concludes our work.

2. Related Work

The integration of external knowledge bases has become essential in natural language processing (NLP) for enhancing language model performance in tasks requiring deep semantic understanding. This section reviews related works.

2.1. Pre-Trained Models

Pre-trained models, especially BERT[1] introduced by Google in 2018, marked a breakthrough by using a bidirectional encoder that captures context from both directions within text, excelling in tasks like question answering and named entity recognition. RoBERTa[15] by Facebook in 2019 further optimized BERT's training, enhancing robustness through batch size adjustments and eliminating the Next Sentence Prediction task. ALBERT[16], also from Google, reduced model parameters for efficiency, while BART[17] combined BERT and GPT's strengths to improve tasks like

 doi:10.20944/preprints202411.0453.v1

summarization and machine translation. The GPT series*[18–20]*, known for its generative capacity, has been pivotal in text generation tasks, especially with GPT-3's scale and performance in areas such as translation and dialogue.

Despite these advancements, pre-trained models often struggle with tasks that require complex reasoning or domain-specific knowledge. Consequently, integrating external knowledge bases has emerged as a promising solution to enhance NLP performance beyond what pre-training alone can achieve.

2.2. Integration of External Knowledge Bases

Knowledge bases, particularly knowledge graphs (KGs) and semantic networks, enable NLP models to perform relationship-based reasoning. Knowledge graphs like Freebase[21] and Google's Knowledge Graph[22] organize entities and their relationships, supporting models in handling tasks that demand high levels of semantic understanding. Techniques for incorporating KGs include:

- Knowledge Graph Embeddings[23]: Converting KGs into low-dimensional vectors allows neural models to process entity relationships through vector arithmetic, enabling semantic relationship modeling.
- Graph Neural Networks (GNNs)[24]: GNNs aggregate information across graph nodes, with attention mechanisms like Graph Attention Network (GAT) enhancing node representation by focusing on the importance of connections.
- Knowledge-Enhanced Transformer Models[25]: Models like K-BERT incorporate KG embeddings directly, allowing the Transformer architecture to utilize structured external knowledge during both pre-training and fine-tuning, thereby improving tasks such as text classification.

These approaches demonstrate notable performance improvements, with models like KEPLER[26] and GraphBERT[27] achieving strong results in question answering and multi-hop reasoning tasks.

2.3. Applications of External Knowledge in Chinese NLP

External knowledge bases are particularly beneficial in Chinese NLP, especially for semantic similarity tasks, where unique linguistic challenges arise:

1. Polysemy and Homophones: Due to frequent ambiguity, models often require additional context to accurately interpret meaning.
2. Lexical Scarcity and Conciseness: Chinese language relies on idioms and phrases, making semantic analysis difficult without enhanced segmentation accuracy.

To address these, knowledge-enhanced lexical representations, knowledge graphs, and contextual information have been successfully applied:

1. Lexical Representations: thesaurus like the Tencent's 8-million-word corpus and NetEase bce embedding model enrich word embeddings for Chinese, which is especially helpful in handling synonyms and reducing semantic interference.
2. Knowledge Graph-Based Enhancement: Structured data from Chinese knowledge graphs provides hierarchical relationships, aiding semantic similarity by embedding this information into models for complex relationship handling.
3. Contextual Computations: External knowledge supports disambiguation, particularly for polysemy, thereby improving similarity computation accuracy.

Studies in this field have yielded several successful applications. For instance, Li et al. (2021) enhanced short-text similarity calculations by embedding a Chinese KG, which aided complex reasoning tasks. Zhang et al. (2022) leveraged multimodal knowledge bases combining text and image data to improve recommendation accuracy in e-commerce, while Wang et al. (2023) incorporated a legal knowledge base into a BERT-based model for Chinese legal document analysis, achieving superior document matching by addressing terminology complexity.

 doi:10.20944/preprints202411.0453.v1

2.4. Insights for This Study

The literature reveals that integrating external knowledge bases significantly enhances the performance of NLP models. This study builds on these insights by using the Tencent's 8-million-word corpus and NetEase bce embedding model alongside the BERT model to improve semantic understanding in Chinese short-text similarity tasks. Our approach enables a more adaptable model design with greater scalability, offering a foundation for future research on the deeper integration of external knowledge with pre-trained models for complex semantic tasks.

3. Methodology

This section provides a detailed explanation of the design and implementation of the proposed short-text similarity model, which integrates external knowledge.

3.1. Model Design

After the sentence is input into BERT, the global features are obtained from the first CLS token. These are then combined with the local features retrieved from the external knowledge base, and the similarity score is output after feature fusion. Our model archtecture shows in Figure 1.

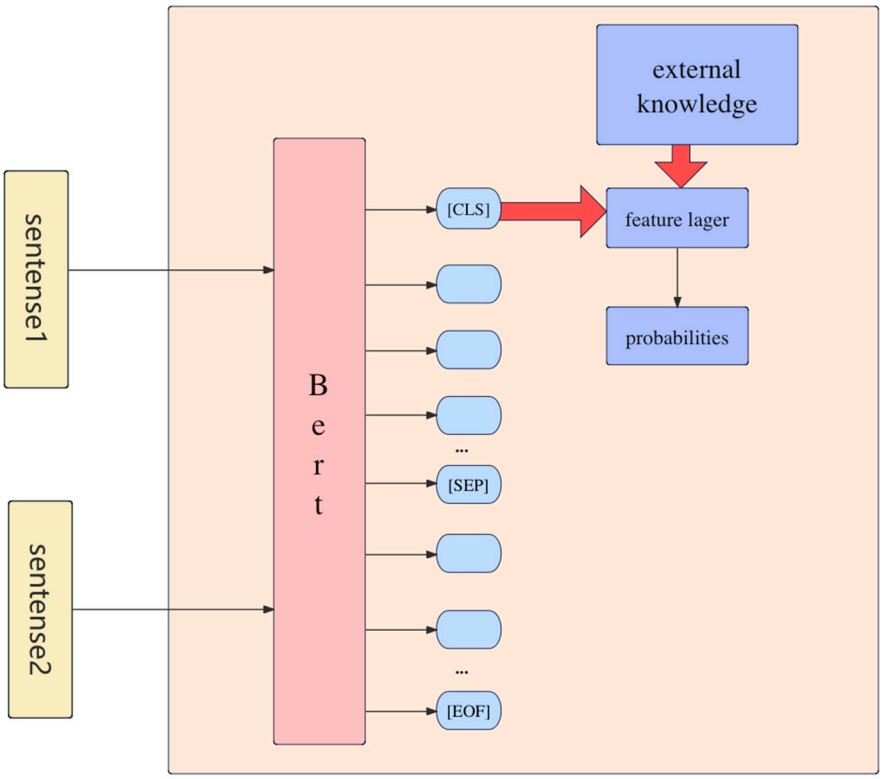

Figure 1. Overall Architecture Diagram.

3.1.1. Pre-Trained Language Models

The proposed model adopts BERT as the primary framework for language understanding. BERT, leveraging its multi-head attention mechanism, captures complex dependencies within sentences, offering strong support for obtaining global semantic representations of sentences. In this implementation, we utilize the base version of BERT-CHINESE-BASE, which comprises 12 Transformer layers, a hidden size of 768, and a total of 110 million parameters—sufficient to meet the requirements for semantic understanding of short Chinese texts.

To accommodate specific application scenarios, the model also supports the use of alternative pre-trained models, such as RoBERTa or ALBERT, providing flexibility for different corpus

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 November 2024 doi:10.20944/preprints202411.0453.v1

characteristics and performance demands. These models can be selected and adjusted according to various task-specific needs.

3.1.2. Integration of External Word Vector Databases

A key innovation of the model is the incorporation of external word vector databases to enhance the semantic understanding at the lexical level. In this study, the Tencent's 8-million-word corpus is used as the primary source of external word vectors. This extensive word corpus covers a wide range of vocabulary and provides high-quality word vectors, which is particularly useful for enhancing the model's understanding of out-of-vocabulary words that do not appear in the training data. Additionally, to meet domain-specific requirements, the model supports the integration of specialized word vector databases, such as those in the medical or legal fields, significantly improving the model's performance in domain-specific tasks.

Technically, we employ the Jieba segmentation system to segment the input sentences, and then retrieve the corresponding vector representation for each word from the word vector database. If a word is not found in the database, a zero vector is used as a substitute. The word vectors for each sentence are aggregated through mean pooling to form a single semantic vector, which is then fused with the sentence-level vector output from BERT.

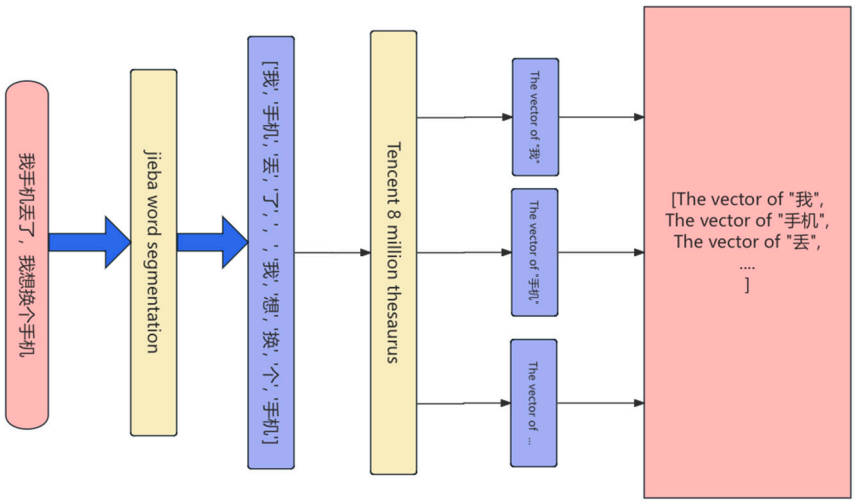

Figure 2. External Knowledge Processing.

3.1.3. Feature Fusion Strategy

To fully exploit the information derived from both the pre-trained language model and the external word vector database, a multi-stage feature fusion strategy is adopted. First, the CLS token vector output by BERT, representing the global semantic representation of the sentence, is passed through a linear transformation layer to reduce its dimensionality to 200 dimensions. Meanwhile, the word vectors of the two sentences, obtained via mean pooling, serve as local semantic vectors.

These three vectors (two local vectors and one global vector) are concatenated to form a 600-dimensional composite feature vector. This design not only retains global semantic information but also enhances the model's sensitivity to local lexical variations, enabling effective semantic analysis at both the global and local levels.

To prevent overfitting and improve the model's generalization to new data, Dropout regularization is applied after feature fusion. A dropout rate of 0.1 is used, which effectively "shuts off" a random subset of neurons, reducing the model's dependency on specific training samples.

 doi:10.20944/preprints202411.0453.v1

3.2. Similarity Computation and Optimization

The model's output layer is a linear layer that maps the 600-dimensional fused vector to a 1-dimensional similarity score. This score is converted into a probability value between 0 and 1 using the sigmoid activation function, indicating the degree of similarity between the two sentences. The model is trained using the binary cross-entropy loss function, and the optimization is performed using the Adam algorithm with an initial learning rate set to 1e-5 to ensure fast convergence during early training.

Through this design, the model not only improves the accuracy of short-text similarity computation but also enhances adaptability and practicality by allowing flexible integration of various external word vector databases. These advantages make the model particularly effective in real-world applications, especially in scenarios requiring the handling of specialized terminology or domain-specific language.

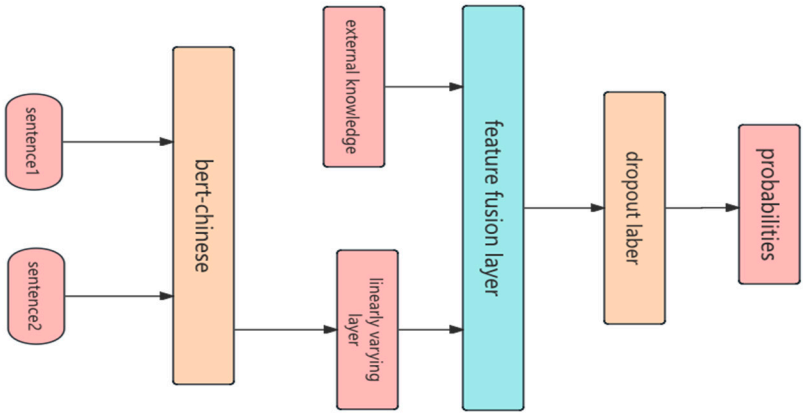

Figure 3. Feature Fusion Diagram.

4. Experiments

4.1. Experimental Setup

4.1.1. Dataset Overview

In this study, four standard datasets widely used for Chinese short-text similarity calculation were selected:

1. LCQMC(the dataset can be downloaded from https://www.modelscope.cn/datasets/DAMO_NLP/LCQMC/files)[28]: This is a large-scale Chinese question matching corpus designed to assess the semantic similarity between two questions. The dataset contains diverse short-text features and covers common question-answering scenarios in daily life, making it highly suitable for short-text similarity tasks.
2. BQ(the dataset can be downloaded from https://www.modelscope.cn/datasets/DAMO_NLP/BQ_Corpus)[29]: Derived from a banking question-answering system, this dataset focuses on short-text matching tasks in the financial domain. The BQ dataset has a complex structure involving financial terms and concepts, allowing it to test a model's adaptability in specialized domains.
3. OPPO-xiaobu(the dataset can be downloaded from https://github.com/CLUEbenchmark/FewCLUE/tree/main/datasets/bustm): This dataset, provided by OPPO, consists of real user queries from an intelligent assistant. It is used to evaluate short-text similarity performance in dialogue-based scenarios.
4. AFQMC(the dataset can be downloaded from https://www.modelscope.cn/datasets/modelscope/afqmc_small): A financial domain short-text similarity dataset containing 80,000 samples, primarily used to assess performance in matching financial-related queries.

 doi:10.20944/preprints202411.0453.v1

4.1.2. Experimental Environment

- Hardware: The experiments were conducted on a server equipped with an NVIDIA RTX 4090 GPU with 24 GB of memory and 500 GB of storage.
- Operating System: Ubuntu 20.04.
- Development Framework: Python was used for model implementation and training, with PyTorch serving as the deep learning framework.
- Dependencies: Jieba was utilized for Chinese word segmentation. External knowledge bases such as Tencent's 8-million-word corpus and NetEase's word vector library were integrated to enhance the model's lexical representation capabilities.

4.1.3. Model Selection and Configuration

To evaluate the impact of external knowledge bases, a comparative experiment was conducted using various pre-trained models and traditional models. The list of models used in the experiment is as follows:

1. BERT-Base-Chinese: A bidirectional encoder model based on the Transformer architecture, used as a baseline model without incorporating external knowledge.
2. RoBERTa-Base-Chinese: An optimized version of BERT, with the next-sentence prediction task removed and trained on a larger dataset for improved performance.
3. BART-Base-Chinese: A model combining features of BERT and GPT, suitable for generation and summarization tasks.
4. ERNIE 2.0: A knowledge-enhanced pre-trained model that integrates knowledge such as entities and concepts.
5. The proposed knowledge-enhanced model: Built upon pre-trained models, this model integrates Tencent's 8-million corpus and NetEase bce embedding model, fusing external lexical knowledge with pre-trained models to capture both global and local semantic features of sentences.

4.1.4. Feature Fusion Strategy

The proposed method adopts a multi-layer fusion strategy that combines pre-trained models with external knowledge bases. Specifically, the output vectors of pre-trained models (e.g., BERT, RoBERTa) are dimensionally reduced via a linear layer. The word vectors from external knowledge bases, obtained through Jieba segmentation, are then concatenated with the pre-trained model vectors. A multi-layer neural network is used to fuse these features, resulting in a comprehensive semantic representation vector for similarity calculation.

4.1.5. Experimental Parameters

- Optimization Algorithm: Adam optimizer with an initial learning rate of 1e-5, applying a linear decay strategy to a minimum of 1e-6.
- Batch Size: 64.
- Training Epochs: Maximum of 15 epochs, with early stopping applied if no performance improvement is observed on the validation set for 5 consecutive epochs.
- Loss Function: Binary cross-entropy loss.
- Activation Function: Sigmoid function, used to map the output similarity score to a probability between 0 and 1.

4.2. Experimental Design

4.2.1. Pre-Trained Model Comparison

This experiment compares the performance of different pre-trained models with and without the incorporation of external knowledge bases, focusing on accuracy across different datasets. The results are show in Table 1:

 doi:10.20944/preprints202411.0453.v1

Table 1. Performance Comparison of Different Pre-trained Models on Specific Datasets (With and Without External Knowledge).

Model	Dataset	Without External Knowledge	With External Knowledge
BERT-Base-Chinese	LCQMC	88.13%	88.66%
BERT-Base-Chinese	AFQMC	72.66%	74.50%
RoBERTa-Base-Chinese	LCQMC	88.80%	90.16%
RoBERTa-Base-Chinese	AFQMC	73.3%	74.56%
BART-large-Chinese	LCQMC	88.73%	90.11%
BART-large-Chinese	AFQMC	72.84%	74.68%

As shown in the table, the accuracy of different pre-trained models on the LCQMC and AFQMC datasets, comparing performance with and without incorporating external knowledge bases. The results indicate that adding external knowledge consistently enhances model accuracy across all tasks. For instance, the accuracy of the RoBERTa-Base-Chinese model on the LCQMC dataset increases from 88.80% to 90.16% with external knowledge. Similarly, the BERT and BART models show comparable improvements on the AFQMC dataset. These findings highlight the effectiveness of incorporating external knowledge in enhancing performance for Chinese NLP tasks across various models.

4.2.2. Cross-Dataset Performance Comparison

As shown in Table 2,the performance of the BERT-Base-Chinese model on different datasets (LCQMC, BQ, OPPO-xiaobu, AFQMC), comparing results with and without external knowledge. Incorporating external knowledge improves performance across all datasets, with a notable accuracy increase from 86.64% to 88.03% on the OPPO-xiaobu dataset. These findings demonstrate the cross-dataset applicability of external knowledge integration for enhancing performance in diverse Chinese NLP tasks.

Table 2. Cross-Dataset Performance Comparison (With and Without External Knowledge).

Model	Dataset	Without External Knowledge	With External Knowledge
BERT-Base-Chinese	LCQMC	88.13%	88.66%
	BQ	84.43%	85.32%
	OPPO-xiaobu	86.64%	88.03%
	AFQMC	72.66%	74.50%

4.2.3. Comparison of Different External Knowledge Bases

To evaluate the effect of different external knowledge bases on model performance, an experiment was conducted on the AFQMC, and shows in Table 3:

Table 3. Impact of External Knowledge on BERT Performance on the AFQMC Dataset.

Model	Dataset	External Knowledge	Accuracy
BERT-Base-Chinese	AFQMC	None	72.66%
	AFQMC	Tencent's 8-million-word corpus	73.4%
	AFQMC	NetEase bce embedding model	73.22%

As show in the table that introducing diverse external knowledge sources, such as Tencent's 8-million-word corpus(200 dimensional) and the NetEase bce embedding model(768 dimensional), effectively improves model performance. Compared to the base model without external knowledge, accuracy increases to 73.4% and 73.22%, respectively.

 doi:10.20944/preprints202411.0453.v1

4.2.4. Comparison with Other Advanced Models

To further validate the effectiveness of the proposed method, the performance of other advanced models was compared in Table 4.

Table 4. Comparison with Other Models.

Model	LCQMC	BQ
ERNIE 2.0 Base[30]	87.9%	85.0%
ERNIE 2.0 Large[30]	87.9%	85.2%
ZEN[31]	87.95%	-
KECM[32]	88.91%	87.62%
NEK[33]	88.15%	85.08%
Ours	90.16%	85.32%

As show in the table the performance of various models, highlighting that both KECM and NEK also incorporate external knowledge. The proposed method (RoBERTa-Base-Chinese with Tencent's 8-million-word corpus) achieves the highest accuracy on the LCQMC dataset, reaching 90.16% and surpassing all other models in performance.

4.3. Summary

This section presented the experimental setup and results of the proposed method, focusing on integrating external knowledge into pre-trained models to improve short-text similarity tasks in Chinese NLP. By utilizing standard datasets like LCQMC, BQ, OPPO-xiaobu, and AFQMC, the study assessed the efficacy of the model in various domains, including general question matching, finance, and intelligent dialogue systems.

The experimental results demonstrated that incorporating external knowledge sources, such as Tencent's 8-million-word corpus and NetEase's word vector library, significantly enhanced model accuracy across datasets. Specifically, the results showed consistent improvements across both general and domain-specific datasets, as well as across different pre-trained models (e.g., BERT, RoBERTa, and BART). Notably, the RoBERTa-Base-Chinese model with Tencent's corpus achieved the highest performance on the LCQMC dataset with an accuracy of 90.16%, surpassing other advanced models like KECM and NEK.

The comparison of various external knowledge sources further highlighted that both Tencent's corpus and NetEase bce embedding model contributed positively to model performance, with accuracy increases observed across different experimental settings. Additionally, the cross-dataset comparison underscored the generalizability of the proposed approach, as performance gains were noted across diverse domains.

Overall, the experiments provide strong evidence that integrating external knowledge sources into pre-trained models enhances their ability to capture semantic features more effectively, improving performance in short-text similarity tasks across different Chinese NLP scenarios. This approach not only boosts accuracy but also offers a scalable solution for enhancing language models with domain-specific knowledge, marking a promising direction for future NLP research and applications.

5. Extended Experiments

This paper proposes a short text matching model that integrates external knowledge bases, demonstrating superior performance in short text similarity computation. Beyond short text similarity tasks, this model also enhances semantic understanding and recognition accuracy in Named Entity Recognition (NER), showcasing its strong generalizability and applicability across various NLP tasks requiring precise semantic understanding and relationship modeling.

 doi:10.20944/preprints202411.0453.v1

5.1. NER

In the Named Entity Recognition (NER) task, the proposed model improves performance by integrating a pre-trained model with external knowledge bases in several ways: it uses the BERT-Base-Chinese model as the foundational language model, leveraging its multi-head attention mechanism to capture global semantic representations and using the CLS vector for overall context. The model incorporates external resources like Tencent's 8-million-word corpus and NetEase's embedding corpus, which enrich its lexical and entity information, significantly enhancing NER accuracy. Contextual vectors from BERT are concatenated with embeddings from the knowledge bases, followed by a linear transformation for feature fusion, while a Conditional Random Field (CRF) layer generates label sequences. The CRF layer combined with a cross-entropy loss function and Adam optimizer, with an initial learning rate of 2e-5, is used for optimization. The model's performance was validated on the MSRA-NER dataset, demonstrating superior effectiveness in standard Chinese NER tasks.

5.2. Experimental Results and Analysis

The results of the experiments conducted on a NER dataset show that the proposed model, which integrates external knowledge bases, outperforms traditional BERT models and other models in the NER task. Specific results on the MSRA-NER are presented in Table 5:

Table 5. Comparison with Other Models on MSRA-NER.

Model	Dataset	F1
ERNIE 2.0 Base	MSRA-NER	93.8%
ERNIE 2.0 Large	MSRA-NER	95.25%
ZEN	MSRA-NER	95.25%
Glyce + BERT[34]	MSRA-NER	95.54%
Bert	MSRA-NER	93.56%
Ours	MSRA-NER	95.74%

As the results show, the model with external knowledge base integration achieves higher F1 scores than baseline models, indicating that incorporating external knowledge significantly enhances the model's performance in the NER task.

5. Conclusion

This study addresses semantic challenges in Chinese short text matching by integrating external knowledge bases with pre-trained language models like BERT. By combining BERT's global features with local semantic features from external lexicons (e.g., Tencent's 8-million-word corpus), the model significantly improves accuracy and robustness in short text similarity tasks. Key contributions include the integration of external knowledge for enhanced semantic disambiguation, multi-layer feature fusion to capture nuanced meanings, validation on datasets like LCQMC and BQ where it outperformed traditional methods, and successful application to named entity recognition (NER), demonstrating cross-task utility.

future research could focus on dynamic knowledge integration, use of multimodal knowledge, expansion to domain-specific tasks, and optimization for resource efficiency, especially in low-resource settings. Addressing current limitations, such as noise from external knowledge bases, adaptability to dynamic updates, and constraints in specific domains, may involve strategies like noise filtering and dynamic knowledge retrieval.

References

1. Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
2. Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems.

 doi:10.20944/preprints202411.0453.v1

3. Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language models as knowledge bases?. arXiv preprint arXiv:1909.01066.
4. Roberts, A., Raffel, C., & Shazeer, N. (2020). How much knowledge can you pack into the parameters of a language model?. arXiv preprint arXiv:2002.08910.
5. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. stat, 1050(20), 10-48550.
6. Lee, K., Chang, M. W., & Toutanova, K. (2019). Latent retrieval for weakly supervised open domain question answering. arXiv preprint arXiv:1906.00300.
7. Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language models as knowledge bases?. arXiv preprint arXiv:1909.01066.
8. Wang, H., Liu, Y., Zhu, C., Shou, L., Gong, M., Xu, Y., & Zeng, M. (2021, August). Retrieval Enhanced Model for Commonsense Generation. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (pp. 3056-3062).
9. Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading Wikipedia to answer open-domain questions. In 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017 (pp. 1870-1879). Association for Computational Linguistics (ACL).
10. Lin, Y., Han, X., Xie, R., Liu, Z., & Sun, M. (2018). Knowledge representation learning: A quantitative review. arXiv preprint arXiv:1812.10901.
11. Zhou, Y., Li, C., Huang, G., Guo, Q., Li, H., & Wei, X. (2023). A short-text similarity model combining semantic and syntactic information. Electronics, 12(14), 3126.
12. Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. ERNIE: Enhanced Language Representation with Informative Entities
13. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.
14. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv preprint arXiv:1906.08237.
15. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach.
16. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019, September). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In International Conference on Learning Representations.
17. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (p. 7871). Association for Computational Linguistics.
18. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. Improving Language Understanding by Generative Pre-Training.
19. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.
20. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. Language Models are Few-Shot Learners.
21. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008, June). Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 1247-1250).
22. Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., & Taylor, J. (2019). Industry-scale Knowledge Graphs: Lessons and Challenges: Five diverse technology companies show how it's done. Queue, 17(2), 48-75.
23. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. Advances in neural information processing systems, 26.
24. Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.
25. Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., & Wang, P. (2020, April). K-bert: Enabling language representation with knowledge graph. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 03, pp. 2901-2908).

 doi:10.20944/preprints202411.0453.v1

26. Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., & Tang, J. (2021). KEPLER: A unified model for knowledge embedding and pre-trained language representation. Transactions of the Association for Computational Linguistics, 9, 176-194.
27. Zhang, J., Zhang, H., Xia, C., & Sun, L. (2020). Graph-bert: Only attention is needed for learning graph representations. arXiv preprint arXiv:2001.05140.
28. Liu, X., Chen, Q., Deng, C., Zeng, H., Chen, J., Li, D., & Tang, B. (2018, August). Lcqmc: A large-scale chinese question matching corpus. In Proceedings of the 27th international conference on computational linguistics (pp. 1952-1962).
29. Chen, J., Chen, Q., Liu, X., Yang, H., Lu, D., & Tang, B. (2018). The bq corpus: A large-scale domain-specific chinese corpus for sentence semantic equivalence identification. In Proceedings of the 2018 conference on empirical methods in natural language processing (pp. 4946-4951).
30. Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., & Wang, H. (2020, April). Ernie 2.0: A continual pre-training framework for language understanding. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 05, pp. 8968-8975).
31. Diao, S., Bai, J., Song, Y., Zhang, T., & Wang, Y. (2019). ZEN: Pre-training Chinese text encoder enhanced by n-gram representations. arXiv preprint arXiv:1911.00720.
32. Liu, R., Zhong, Q., Cui, M., Mai, H., Zhang, Q., Xu, S., ... & Du, Y. (2024, March). External Knowledge Enhanced Contrastive Learning for Chinese Short Text Matching. In 2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT) (pp. 1436-1440). IEEE.
33. Ma, H., Li, Z., & Guo, H. (2022, October). Using Noise and External Knowledge to Enhance Chinese Pre-trained Model. In 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 476-480). IEEE.
34. Meng, Y., Wu, W., Wang, F., Li, X., Nie, P., Yin, F., ... & Li, J. (2019). Glyce: Glyph-vectors for chinese character representations. Advances in Neural Information Processing Systems, 32.