information

Article

Robust Chinese Short Text Entity Disambiguation Method Based on Feature Fusion and Contrastive Learning

Qishun Mei [1] and Xuhui Li [1,2,*]

[1] School of Information Management, Wuhan University, Wuhan 430072, China; meiqishun@whu.edu.cn
[2] Big Data Institute, Wuhan University, Wuhan 430072, China
[*] Correspondence: lixuhui@whu.edu.cn

Abstract: To address the limitations of existing methods of short-text entity disambiguation, specifically in terms of their insufficient feature extraction and reliance on massive training samples, we propose an entity disambiguation model called COLBERT, which fuses LDA-based topic features and BERT-based semantic features, as well as using contrastive learning, to enhance the disambiguation process. Experiments on a publicly available Chinese short-text entity disambiguation dataset show that the proposed model achieves an F1-score of 84.0%, which outperforms the benchmark method by 0.6%. Moreover, our model achieves an F1-score of 74.5% with a limited number of training samples, which is 2.8% higher than the benchmark method. These results demonstrate that our model achieves better effectiveness and robustness and can reduce the burden of data annotation as well as training costs.

Keywords: short text; entity disambiguation; topic model; pre-trained model; feature fusion; contrastive learning

1. Introduction

Entity disambiguation is a prominent task in the field of Natural Language Processing (NLP), operating at the lexical semantic level. Its objective is to identify key entities within unstructured texts and determine their accurate meanings based on contextual cues; for example, the term “apple” can refer to a company or a fruit in different contexts. Entity disambiguation plays a crucial supporting role in higher-level NLP tasks such as sentiment analysis [1], event extraction [2], and knowledge graph construction [3].

In recent years, the rapid growth of the Internet, along with the emergence of social and online platforms, has empowered individuals to produce online content anywhere, anytime. This has resulted in an exponential surge in online texts, particularly short (fewer than 100 words) and non-standard texts, such as “the apple press conference is about to be held”, prevalent on platforms like Twitter and Weibo. These short texts have become a vehicle for diverse information and viewpoints that are rapidly disseminated across the Internet.

The proliferation of short texts on the Internet presents significant challenges for the entity disambiguation task. Existing entity disambiguation methods primarily focus on medium or long texts that contain enough information to easily distinguish entity mentions [4–9], making it more challenging to extract semantic information from short texts due to the reduced number of words and an increase in colloquial expressions. Consequently, these methods struggle to effectively capture text features in short-text scenarios, resulting in a substantial decline in disambiguation performance. Additionally, current entity disambiguation approaches, especially for short-text disambiguation [10,11], heavily rely on a large volume of training samples to ensure model generalization and disambiguation effectiveness. However, acquiring sufficient data and conducting model training for various tasks can be costly in real-world applications. Therefore, it is imperative to address

check for updates

Citation: Mei, Q.; Li, X. Robust Chinese Short Text Entity Disambiguation Method Based on Feature Fusion and Contrastive Learning. *Information* **2024**, *15*, 139. https://doi.org/10.3390/info15030139

Academic Editor: Katsuhide Fujita

Received: 25 January 2024
Revised: 20 February 2024
Accepted: 25 February 2024
Published: 29 February 2024

the pressing issues of extracting better text representations, reducing the reliance of entity disambiguation models on training samples, and enhancing model robustness to improve short-text disambiguation in scenarios with limited training data.

To address the aforementioned challenges, we propose COLBERT, which stands for combining Contrastive Learning with LDA and BERT for short-text entity disambiguation. COLBERT leverages the prevalent entity-linking-based disambiguation approach, where mentions of ambiguous entities are mapped to referent entities in an external knowledge base by matching algorithms [12]. The proposed method involves transforming the entity disambiguation task into a classification task by combining short texts with descriptions of referent entities from the knowledge base. Topic features and semantic features are extracted by the LDA and BERT models, respectively, and are fused to obtain comprehensive text representations. Furthermore, to further enhance the model's robustness and performance with limited samples, contrastive learning is introduced to improve the quality of text representations during training. The resulting text representations are then input into the classification layer for the matching decision, thereby completing the entity disambiguation task.

Experimental results on a publicly available Chinese dataset for short-text entity disambiguation demonstrate the effectiveness of the proposed method. The contributions of this paper are as follows:

(1) An end-to-end entity disambiguation training framework is proposed that combines topic and semantic features to address the issue of insufficient information extraction in short-text entity disambiguation tasks;

(2) Contrastive learning methods are introduced to the entity disambiguation task to further enhance text representation quality, improve performance in scenarios with limited training samples, and reduce the annotation workload and training costs;

(3) The model proposed in this paper demonstrates superior performance, with a 0.6% improvement in the F1-score on a full training set, and robustness, with a 2.8% improvement on a small training set, compared to the benchmark method, offering a novel approach to addressing short-text entity disambiguation tasks.

2. Related Work

The existing entity disambiguation methods are generally divided into three types: unsupervised-learning-based, supervised-learning-based, and graph-based collaborative disambiguation methods.

Unsupervised-learning-based entity disambiguation methods often employ vector space models (VSMs) [4], which represent text as a bag of words and establish context models for entity mentions and candidate entities. These models utilize vector representations of words, concepts, categories, etc., to measure similarity and select the most suitable entity to accomplish the goal of entity disambiguation. Fleischman et al. [5] utilized the maximum entropy model to calculate the probability of mapping two mentions to the same entity and then employed a bottom-up hierarchical clustering algorithm to disambiguate entity mentions. Pedersen et al. [6] constructed a co-occurrence matrix of entities, applied singular value decomposition to reduce dimensionality, and used repeated dichotomy to disambiguate entities.

Supervised learning methods for entity disambiguation typically frame the task as a classification task in which the input samples for the entity disambiguation model consist of an entity mention and its corresponding entities. These samples are then trained using machine learning or deep learning models to address entity ambiguity. Pilz et al. [7] employed the LDA topic model to extract the topic distribution of the entity mentions and the reference entities, calculating the distance between their topic distributions, then utilized the SVM model to construct a binary classifier. He et al. [8] proposed a deep neural network method for entity disambiguation that accomplishes the task by stacking automatic noise reduction encoders. Sun et al. [9] developed a Convolutional Neural Network (CNN) to encode the context of the entity mention and the candidate entities and then applied an additional position vector as auxiliary information.

The graph-based collaborative disambiguation methods apply graph models to capture dependencies among multiple entity mentions in texts [13], transforming the disambiguation task into a graph optimization problem, where nodes represent combinations of entity mentions and corresponding entities, and edges represent relationships between entities. The distance between two entities is calculated by some specific algorithms to determine the most suitable candidate entity. Minkov et al. [14] constructed an email network incorporating text content and a social network and then employed a re-sorting algorithm based on the graph walk similarity to accomplish the disambiguation task. Zhang et al. [15] proposed an entity disambiguation method that leverages link information from a collaborative network to aid the disambiguation process through candidate entity correlations. To address the sparsity issue in entity relation graphs, Phan et al. [16] iteratively selected pairs of entities with the highest confidence to discriminate between them at each step and used the weight of the minimum spanning tree to measure the consistency between the two entities.

Moreover, the development of extensive knowledge repositories such as Wikipedia and WordNet presents novel prospects for knowledge-driven approaches to solving entity disambiguation tasks with the help of their extensive semantic information. Han et al. [17] introduced a knowledge-based method named Structural Semantic Relatedness (SSR), which enhances the named entity disambiguation process by capturing and leveraging the structural semantic knowledge presented in multiple knowledge sources. Bouarroudj et al. [18] proposed WeLink, an entity recognition method based on WordNet for Question-Answering Systems that identifies distinct entities along with their types and contexts. Lommatzsch et al. [19] conducted an evaluation of various similarity measures and algorithms for extracting data to perform named entity disambiguation based on both German and English document corpora.

In recent years, scholars have also made extensive efforts to facilitate the task of short-text entity disambiguation. Zhang et al. [10] utilized features similar to those of twin networks to deeply analyze semantic relationships in texts and fully utilized the feature information of the texts to be disambiguated. Shi et al. [11] combined multiple embedding representations for entity linking in Chinese short texts to improve the performance of entity linking.

Existing entity disambiguation methods exhibit distinct characteristics and advantages in addressing entity disambiguation tasks, leading to significant contributions. However, these methods either are poor in fully extracting features from texts or heavily rely on extensive training data, so they have poor robustness. Consequently, their performance diminishes when confronted with short-text disambiguation tasks or insufficient training samples. Therefore, the development of a robust short-text entity disambiguation model has emerged as a crucial objective.

3. Method

To address the issues of inadequate feature extraction and poor robustness in short-text entity disambiguation, we propose a model based on feature fusion and contrastive learning called COLBERT. Our approach tackles the entity disambiguation task through entity linking. The task can be defined as follows: given a set of short texts, denoted by T, containing ambiguous entity mentions, and an external knowledge base, denoted by K, comprising a vast number of entity descriptions, our objective is to accurately link all entity mentions in T to their corresponding entity descriptions in K, thus completing the entity disambiguation task, as shown in Figure 1.

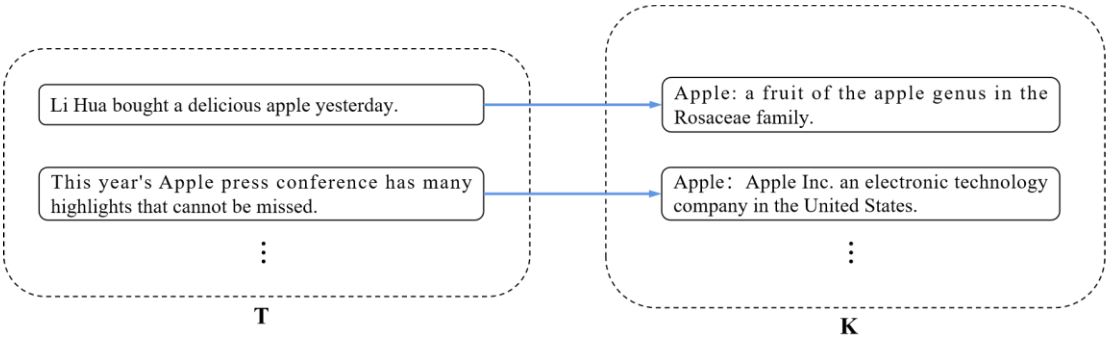
Li Hua bought a delicious apple yesterday.
This year's Apple press conference has many highlights that cannot be missed.
⋮
T
Apple: a fruit of the apple genus in the Rosaceae family.
Apple: Apple Inc. an electronic technology company in the United States.
⋮
K

Figure 1. The format of entity linking, using "apple" as an example.

The proposed model is depicted in Figure 2. Initially, we combine the short texts with the corresponding candidate entity descriptions from the knowledge base to create a binary classification task. For example, we can combine the short text "Li Hua bought a delicious apple yesterday" with the target entity description "Apple: a fruit of the apple genus in the Rosaceae family" to build a positive sample, and with another entity description, "Apple: Apple Inc. an electronic technology company in the United States", to build a negative sample. The input to the model consists of Context_m(A), which represents the m-th short text to be disambiguated with entity mention A, and EntiDes_M(A), which denotes the M-th description for entity A. Our model comprises five modules: the LDA model for topic feature extraction, the BERT model for semantic feature extraction, feature fusion, contrastive learning during training, and the classification layer.

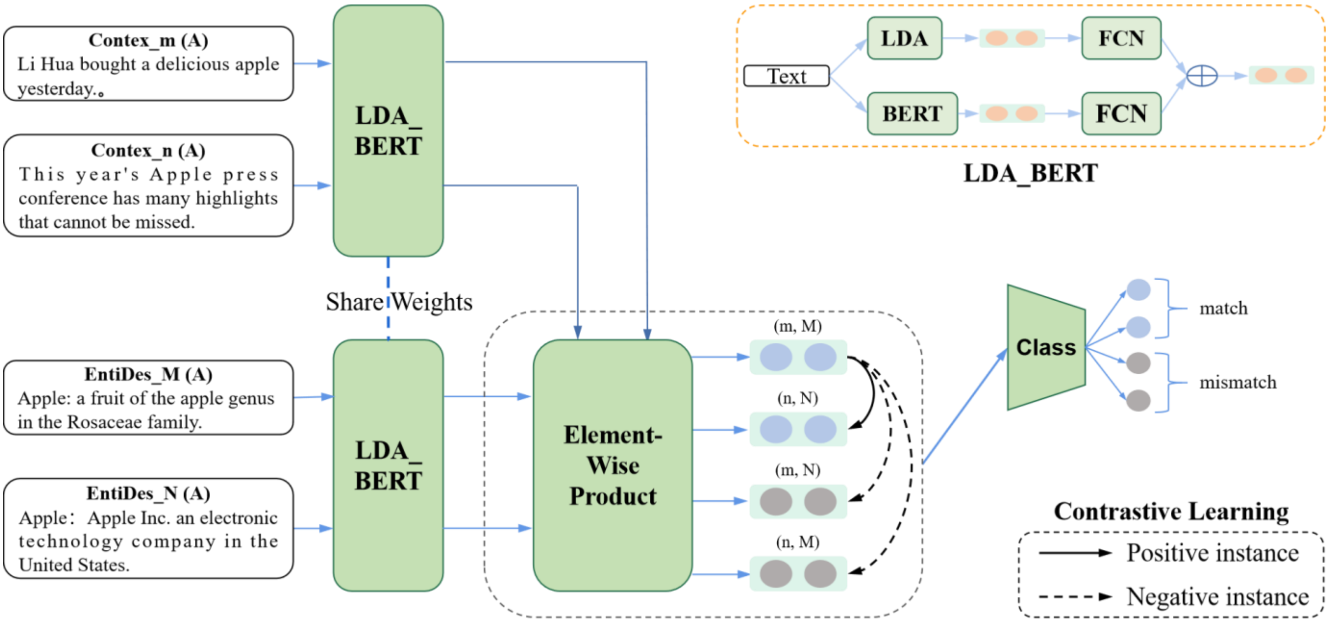
Contex_m (A)
Li Hua bought a delicious apple yesterday.
Contex_n (A)
This year's Apple press conference has many highlights that cannot be missed.
EntiDes_M (A)
Apple: a fruit of the apple genus in the Rosaceae family.
EntiDes_N (A)
Apple: Apple Inc. an electronic technology company in the United States.
LDA_ BERT
Share Weights
LDA_ BERT
Element-Wise Product
(m, M)
(n, N)
(m, N)
(n, M)
Text
LDA
BERT
FCN
FCN
LDA_BERT
Class
match
mismatch
Contrastive Learning
Positive instance
Negative instance

Figure 2. Short text entity disambiguation model: COLBERT.

The objective of the model is to determine whether the entity mention in the input short text matches the referent entity description, thereby linking the entity mention to the corresponding entity and completing the entity disambiguation task.

In the LDA_BERT module, each {short text, entity description} combination sample undergoes separate processing by the LDA model and BERT model to obtain vectors of topic features and semantic features, respectively. The two vectors are then mapped to the same dimension through a fully connected layer and summed together, resulting in a text representation that incorporates both topic and semantic information. The fusion vector is obtained by element-wise product of the text representations of the short text and the entity description. The classification layer, named Class, determines the disambiguation result based on this fusion vector.

Meanwhile, our model incorporates contrastive learning during the training process. We construct a contrastive loss based on the sample labels (i.e., match or not) and weigh the contrastive loss and classification loss to form a joint loss. The contrastive learning training process and the classification task learning process are carried out in coordination, facilitating parameter updates for the pre-trained BERT model and other layers. This means our model implements a complete "End-to-End" training process.

3.1. Feature Extraction

Feature extraction involves transforming an original text or document into a low-dimensional vector representation through which we can obtain fixed-length feature vectors by extracting features from entity descriptions and short texts. The task of entity disambiguation is then accomplished by calculating the similarity of these two vectors using a specific similarity algorithm. However, in practical applications, the effectiveness of feature extraction can be influenced by various factors, such as the feature extraction model and the semantic distribution of the corpus. In the context of short-text entity disambiguation, a single-feature extraction method often fails to yield satisfactory results.

To address this limitation and extract text information more comprehensively, we leverage both the LDA model and BERT model to extract topic and semantic features from texts, respectively. The fusion of these two feature extraction approaches enhances the effectiveness of short-text entity disambiguation. By combining the strengths of these models, we can capture information within the text more comprehensively, thereby improving the performance of the entity disambiguation task.

3.1.1. Topic Feature Extraction

The LDA (Latent Dirichlet Allocation) model [20] is a well-established and influential topic model in Natural Language Processing (NLP). It operates as a generative Bayesian probability model with a hierarchical structure encompassing words, topics, and the corpus. Researchers can train the LDA model on a specific corpus and utilize it to obtain the probability distribution of a new text across various topics, thereby representing the document or sentence as a topic vector.

In addition to the LDA model, other renowned topic models, such as the Non-Negative Matrix Factorization (NMF) model, have gained prominence. The LDA model, being extensively applied in various natural language tasks [21], particularly in short-text tasks [22], holds significance in NLP. Additionally, numerous studies have employed the LDA model for entity disambiguation tasks [23,24]. Hence, in this paper, we employ the LDA model not only due to its proven effectiveness in extracting topic features in numerous studies but also for the convenience of comparing it with existing methods for entity disambiguation.

Upon analyzing the nature of the entity disambiguation task, we observe that different interpretations of ambiguous entities often exhibit distinct topic distributions. Consequently, we employ the LDA model to extract topic features from the texts, aiming to enhance the quality of text representations.

Considering that we accomplish the entity disambiguation task through entity linking, the external knowledge base K assumes a critical role in determining the precise interpretation of entities within short texts. Therefore, to ensure the model's scalability in practical applications, we employed the entity description texts from the external knowledge base K as the training corpus for the LDA model. We trained the LDA model on the corpus after performing word segmentation and removing stop words. Subsequently, we utilized this model to obtain the topic feature vectors for both the entity descriptions and the short texts requiring disambiguation, as depicted in Figure 3. This approach allows us to fully leverage the information within the external knowledge base, thereby enhancing the model's performance and applicability.

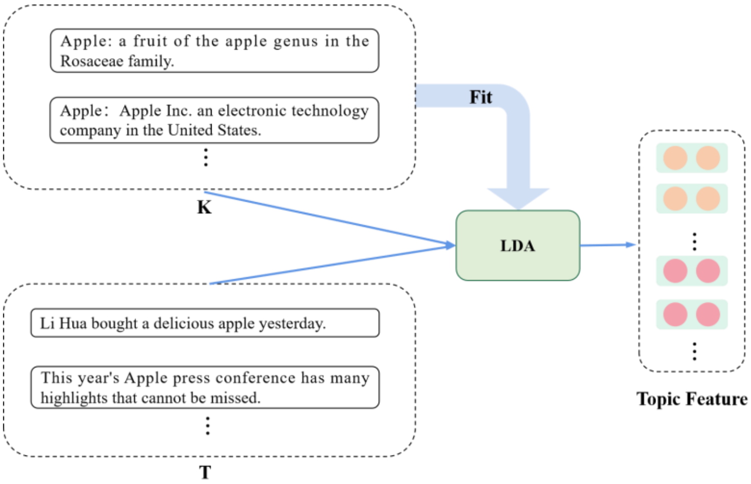
Apple: a fruit of the apple genus in the Rosaceae family.
Apple：Apple Inc. an electronic technology company in the United States.
K
Fit
LDA
Li Hua bought a delicious apple yesterday.
This year's Apple press conference has many highlights that cannot be missed.
T
Topic Feature

Figure 3. Training the LDA model and obtaining a vector of topic features. The yellow vectors represent the topic vectors of the entity description texts from K, and the red ones represent the topic vectors of the short texts from T.

3.1.2. Semantic Feature Extraction

The BERT (Bidirectional Encoder Representation from Transformers) model [25] is a renowned pre-trained model built on the transformer architecture [26]. It has introduced the pre-training + fine-tuning paradigm, which has become a prominent approach in NLP. Through pre-training tasks like Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) on extensive corpora, BERT demonstrates exceptional semantic understanding capabilities. Consequently, it can be effectively applied in downstream tasks to obtain high-quality semantic feature representations of texts.

In our study, we employed the BERT-base-Chinese (https://huggingface.co/bert-base-chinese, accessed on 1 October 2023) model to extract semantic features from both short texts and entity descriptions in the knowledge base, as depicted in Figure 4. To comprehensively capture the semantic information within the texts, we pool the embedded vectors of all tokens from the output of the final layer of the BERT model. This process generates feature vectors capable of representing the global semantic information of the texts. This method effectively captures the overall semantic representations of the texts.

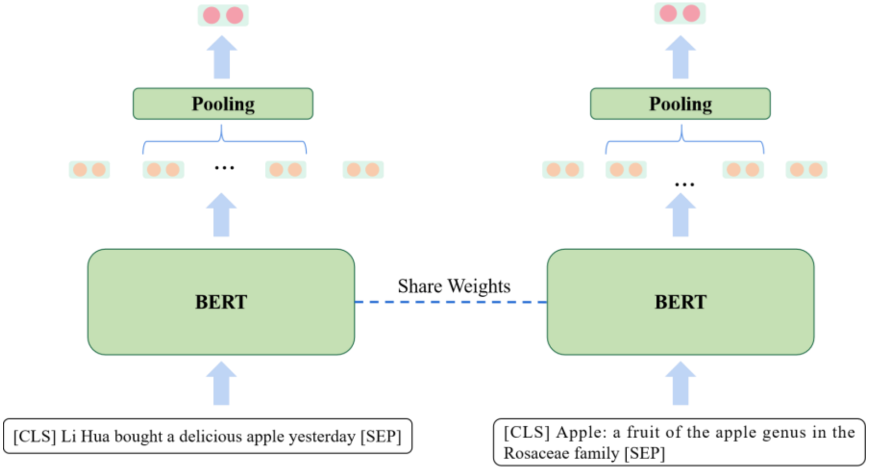
Pooling
Pooling
...
...
BERT
Share Weights
BERT
[CLS] Li Hua bought a delicious apple yesterday [SEP]
[CLS] Apple: a fruit of the apple genus in the Rosaceae family [SEP]

Figure 4. Generation of semantic feature vectors.

3.2. Feature Fusion and Disambiguation

Short texts often present challenges for computers to comprehend due to the limited number of words and informal language usage; meanwhile, they also contain extraneous information and unrelated words, which further complicates the task of entity disambiguation. These issues make it challenging for a single-feature extraction method to capture sufficient information to accomplish the task effectively.

Previous studies have shown that the fusion of topic features and semantic features can improve the performance of NLP models, such as text classification models [27]. To enhance the performance of entity disambiguation in short texts, we apply the integration of the topic feature and semantic feature. By fusing these two features, we obtain the final feature representations for both the short-text disambiguation and the referent entity description. This fusion process improves the quality of text representations. Subsequently, based on this feature vector, we determine whether the entity mentioned in the text matches the referent entity. Experimental results demonstrate that this feature fusion method effectively enhances the performance of the model.

Figure 5 illustrates the extraction of a topic feature vector (t_vector) and a semantic feature vector (s_vector) using the LDA model and the BERT model, respectively. The two vectors then undergo two linear layer transformations to ensure they share the same dimension and are added together after these transformations, resulting in the final fusion feature vector (m_vector) with a manually set dimension (m).

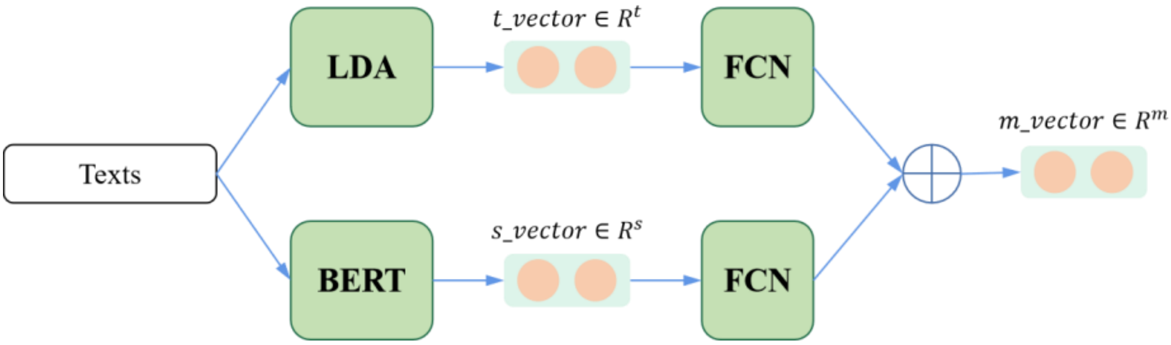

Figure 5. Feature fusion.

Meanwhile, we have formulated the task of entity disambiguation as a binary classification problem. The objective of our model is to determine whether the entity mention requiring disambiguation should be linked to a specific referent entity description. To achieve this, we constructed a positive sample by combining one short text with its target entity description and several negative samples by combining it with other referent entity descriptions. This transformation enabled us to convert the disambiguation task into a classification task.

In the feed-forward process of the model, the text requiring disambiguation and the corresponding entity descriptions undergo separate processing by the LDA_BERT module. The resulting fusion feature vectors obtained from the LDA_BERT module represent the two texts. The element-wise product of the vectors produces a measure vector, which quantifies the similarity between the two. Finally, the measure vector is passed through the classification layer to obtain the ultimate classification result, indicating whether there is a match or mismatch, as shown in Figure 6.

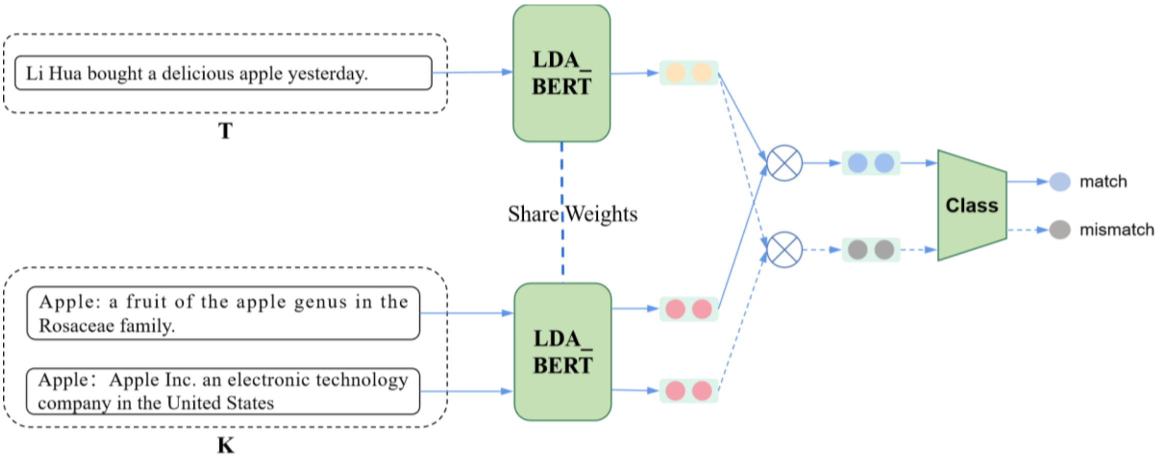
Li Hua bought a delicious apple yesterday.
T
LDA_
BERT
Share Weights
Apple: a fruit of the apple genus in the Rosaceae family.
Apple： Apple Inc. an electronic technology company in the United States
K
LDA_
BERT
Class
match
mismatch

Figure 6. Classification process.

The cross-entropy loss function is used to calculate the loss of the classification process for updating the model's parameters, as shown in Equation (1).

$$l_{clas} = -E_{t,k,y \sim P_{data}} log \left[p(\hat{y} = y | t, k, y \in (0,1)) \right] \tag{1}$$

where t is the short text to be disambiguated, k is the entity description from the knowledge base, and y is the label of the sample, whose value is 0 (mismatched) or 1 (matched).

3.3. Contrastive Learning

Contrastive learning [28] is a self-supervised learning method designed to enhance a model's performance on specific tasks by enabling it to learn improved data representations. In the semantic space, contrastive learning aims to enhance text embedding representations by grouping semantically similar texts while separating dissimilar ones, as depicted in Figure 7. The distance between embeddings can be computed using cosine similarity or the Euclidean distance. Previous studies [29] have demonstrated that contrastive learning can effectively enhance the performance of classification models, particularly in scenarios involving few-shot learning.

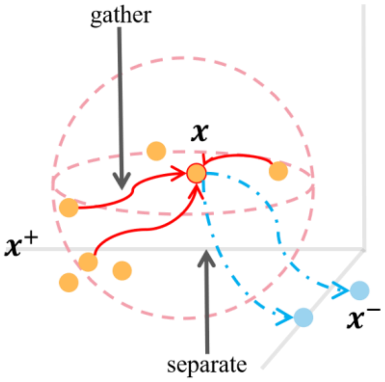

Figure 7. The process of contrastive learning. The nodes denote text embeddings, and the ones in the same color have similar semantics.

Given the real-world challenges of limited annotated samples, high annotation costs, and expensive model training associated with entity disambiguation, this study aimed to enhance the model's performance given insufficient training samples through the adoption of contrastive learning.

Specifically, we have transformed the entity disambiguation task into a binary classification problem, as described in Section 3.2. To enable effective discrimination between different categories in the classification layer, it is crucial for the feature extraction module to acquire robust feature representations in the vector space; this entails ensuring that sample features from different categories exhibit clear distinguishability in the vector space. To achieve this target, we introduced the contrastive learning training method after the LDA_BERT module. During training, for each sample in a batch, we randomly selected another sample belonging to the same category as a positive instance while treating the remaining samples in the batch as negative instances. The objective of contrastive learning is to minimize the distance between representations of samples from the same category and maximize the separation between samples from different categories. The contrastive loss function, as depicted in Equation (2), encapsulates this design.

$$l_{con} = \frac{1}{N}\sum_{i=1}^{N} -log \frac{e^{sim(h_i,h_i^+)/\tau}}{\sum_{j=1,h_j \neq h_i^+}^{N} e^{sim(h_i,h_j)/\tau}} \tag{2}$$

where N is batch_ Size, h is the feature extracted from the text by the LDA_BERT module, h_i^+ is a positive instance of h_i, and $sim(x,y)$ indicates cosine similarity, i.e., $\frac{x^T y}{\|x\|\|y\|}$.

3.4. Training Process

We employ the LDA model to extract topic features from texts. To obtain the core keywords and topic probabilities, the LDA model requires fitting on a specific corpus. In order to ensure practical scalability, we utilized the entity description texts from the knowledge base as the training corpus for the LDA model.

During the fitting process of the LDA model, the crucial parameter that requires manual configuration is the number of topics. To achieve a well-performing model, we employed two evaluation metrics, namely, confusion and consistency [30], to determine the optimal number of topics. Our objective is to identify the number of topics that minimizes confusion and maximizes consistency. Figure 8 illustrates the confusion and consistency of the LDA model fitted under various numbers of topics, facilitating the selection of the optimal number.

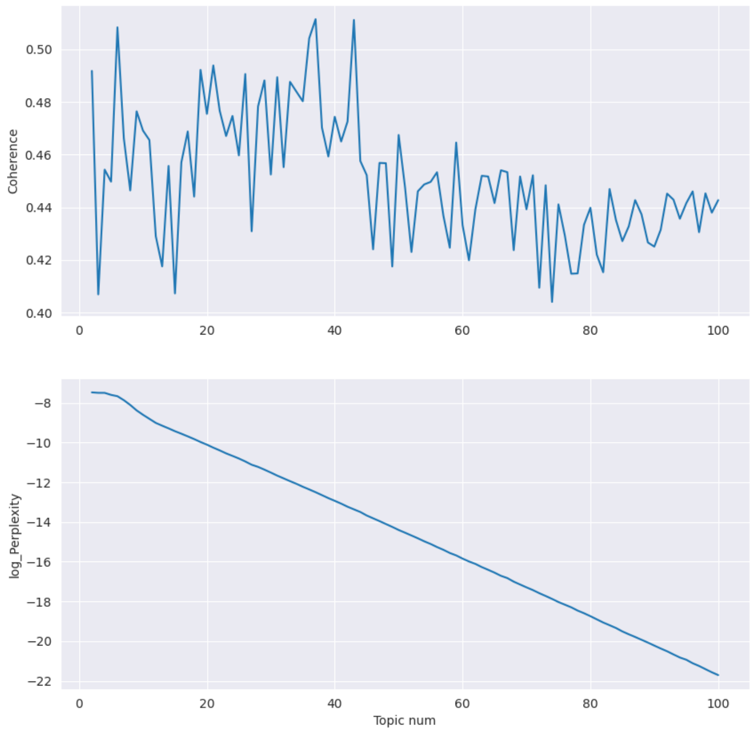

Figure 8. Searching for the optimal number of topics for the LDA model.

To ensure both the "end-to-end" process of the neural network modules and the model's robustness, we integrated the contrastive learning and classification training processes. This integration involves combining the contrastive loss and classification loss into a weighted sum, which serves as the final joint loss function, as shown in Equation (3).

$$l_{all} = \alpha l_{clas} + (1 - \alpha) l_{con} \tag{3}$$

where α is a weight parameter with a value ranging from 0 to 1.

4. Experiment and Analysis

4.1. Dataset

We used the public Chinese short-text entity disambiguation dataset named DUEL2 (https://www.luge.ai/#/luge/dataDetail?id=24, accessed on 26 September 2023) for experiments, which consists of two main components: short texts containing entity mentions slated for disambiguation and the associated knowledge base consisting of entity descriptions. The short texts are predominantly sourced from various resources, including Internet search queries, blogs, dialogues, and titles. Meanwhile, the knowledge base is derived from Baidu Encyclopedia (https://baike.baidu.com, accessed on 26 September 2023). Notably, all samples have undergone manual labeling.

The DUEL2 dataset comprises approximately 70,000 training samples, 10,000 validation samples, and 10,000 test samples. Furthermore, the knowledge base encompasses 324,000 entities. On average, a short text in the dataset contains 16.7 entity mentions, while an entity in the knowledge base possesses an average of 8.71 descriptions. Specific samples are provided in Table 1. Due to the limitations in computing resources and time cost, we randomly sampled 3000 validation set samples and 3000 test set samples from the original dataset for experiments. At the same time, 10,000 training set samples were sampled as the full training set scenario (with far more training samples than testing samples), and 3000 and 1000 training set samples were sampled as two small training set scenarios (with no more training samples than the testing samples) for experiments.

Table 1. Samples from experimental data.

Short texts	Chinese text	{"text_id": "1", "text": "小品◎战狼故事◎中, 吴京突破重重障碍解救爱人, 深情告白太感人", "mention_data": [{"kb_id": "159056", "mention": "吴京", "offset": "10"}]}
	English translation	{"text_id": "1", "text": "In the skit" Wolf Warrior Story ", Wu Jing breaks through numerous obstacles to rescue his lover, and his heartfelt confession is too touching." "mention_data": [{"kb_id": "159056", "mention": "Wu Jing", "offset": "10"}]}
Knowledge base	Chinese text	{"alias": [], "subject_id": "27429", "data": [{"predicate": "摘要", "object": "◎心魔◎是由张明师/张超南作词, 朱兴明作曲, 张雅静演唱的歌曲, 发行于2017年11月○"}, {"predicate": "义项描述", "object": "张雅静演唱的歌曲"}], "type": "Work", "subject": "心魔"}
	English translation	{"alias": [], "subject_id": "27429", "data": [{"predict": "abstract", "object": "Heart Demon" is a song written by Zhang Mingshi/Zhang Chaonan, composed by Zhu Xingming, and sung by Zhang Yajing. It was released in November 2017. "}, {"predict": "meaning description", "object": "Song sung by Zhang Yajing"}, "type": "Work", "subject": "Heart Demon"}

To align with the proposed model in this study, we conducted the following preprocessing of the dataset:

(1) For short texts containing multiple entities, we rebuilt individual samples by pairing each entity mention with the corresponding short text so that each sample is designed to disambiguate only one specific entity mention.

(2) We extracted and consolidated the structured entity information from the knowledge base, condensing it into a single text entity description. This consolidation facilitates the calculation of feature vectors.
(3) We combine the short texts to be disambiguated with their corresponding entity descriptions from the knowledge base, generating samples in the format of {short text, entity description} pairs.

4.2. Evaluation

We use Precision (P), Recall (R), and F1-score (F_1) as the evaluation metrics of the models, as shown in Equations (4)–(6).

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F_1 = \frac{2 \ast Precision \ast Recall}{Precision + Recall} \tag{6}$$

where TP represents the number of entity links correctly predicted by the model, FP represents the number of entity links incorrectly predicted by the model, and FN represents the number of entity links not detected by the model. The values of the three metrics range from 0 to 1. The larger the values, the better the effect of the models.

4.3. Experimental Environment and Parameter Settings

The experimental environment configuration in this work is shown in Table 2.

Table 2. Experimental environment.

Experimental Environment	Environment Configuration
Operating system	Ubuntu 20.04.6 LTS
CPU	Intel (R) Xeon (R) gold 6130 h
GPU	NVIDIA geforce RTX 3090 × 1
Memory	128 G
Python	3.8.11

The model parameters in this paper are shown in Table 3.

Table 3. Model parameters.

Parameter	Parameter Value
Topic num of LDA	43
Dim of m_vector	128
Epoch	3
Batch Size	128
Learning rate of BERT	5×10^{-5}
Learning rate of other nets	1×10^{-3}
Max sequence length of short texts	64
Max sequence length of entity descriptions	256
Optimizer	Adam
α of l_{all}	0.9

4.4. Results

We employed two widely used text feature extraction methods for NLP tasks, namely, BERT-CNN [31,32] and BERT-BiLSTM [33,34], along with the CHOLAN model [35], which has achieved state-of-the-art (SOTA) results in the entity disambiguation task, as benchmark methods in this paper. To evaluate the impact of different feature extraction methods on the model's performance, we applied the same training framework to BERT-CNN and BERT-BiLSTM for the entity disambiguation task. Specifically, these methods are employed to extract feature vectors from the short texts to be disambiguated and the entity descriptions, respectively. The entity disambiguation task is then accomplished by calculating the similarity between the two vectors using the element-wise product and a classification layer.

To assess the model's robustness and its performance with varying training samples, we randomly selected 10,000, 3000, and 1000 samples from the training dataset for model training. We evaluated the model's effectiveness on a separate test set containing 3000 test samples. To ensure that the experimental results are reliable and that the distribution of the experimental data is as close as possible to that of the original data, each group of experiments for a certain training set employed different random seeds for three random sampling experiments on the original training set. The mean values of the evaluation metrics were computed as the final results. Table 4 presents the experimental results for each benchmark method, as well as the proposed COLBERT model proposed in this paper.

Table 4. Experimental results (%). The↓ symbol with a percentage in parentheses refers to the degree of decrease in performance relative to the full training set (10,000 training samples), and the values in bold refer to the optimal results under the same conditions.

Model	10,000 Training Samples			3000 Training Samples			1000 Training Samples		
	P	***R***	***F*** $_1$	***P***	***R***	***F*** $_1$	***P***	***R***	***F*** $_1$
BERT-BiLSTM	84.3	77.3	80.6	72.6	85.6	78.6 (↓2.5%)	61.9	71.8	66.5 (↓17.5%)
BERT-CNN	83.1	80.6	81.8	**84.4**	63.8	72.7 (↓11.1%)	73.9	65.8	69.6 (↓14.9%)
CHOLAN	82.2	**84.6**	83.4	77.2	**80.9**	79.0 (↓5.3%)	69.9	73.7	71.7 (↓14.0%)
Our COLBERT	**84.6**	83.4	**84.0**	76.0	83.9	**79.8 (↓5%)**	**74.8**	**74.3**	**74.5 (↓11.3%)**

4.4.1. Full Training Set

Through a comparative analysis of the results obtained from BERT-BiLSTM, BERT-CNN, and COLBERT, it is evident that the model proposed in this paper outperforms the traditional text feature extraction and enhancement methods. Notably, the proposed model achieves notable increases in the F1-score of 3.4% and 2.2%, respectively, compared to BERT-BiLSTM and BERT-CNN. This demonstrates that although CNN and BiLSTM can enhance the semantic features extracted by the BERT model, they are still constrained to a single semantic feature, thereby exhibiting a significant gap when compared to the COLBERT model presented in this paper.

Furthermore, the Precision and Recall of COLBERT are also elevated, indicating a reduced likelihood of missing correct entity links and misjudging incorrect entity links for the entity disambiguation task. This further validates that the multi-feature fusion approach can extract more comprehensive text information compared to the single-feature enhancement method, resulting in superior performance in entity disambiguation tasks.

Moreover, when compared to the CHOLAN model, the proposed method in this paper also demonstrates an improvement, with a 0.6% increase in the F1-score. This confirms the superiority of the approach introduced in this paper.

4.4.2. Small Training Set

Based on the findings presented in Table 4, it is evident that the performance of each model declines as the amount of training data decreases. However, the impact on different models varies, indicating differences in their robustness. Compared to the effectiveness attained with 10,000 training samples, the F1-scores of BERT-BiLSTM, BERT-CNN, and CHOLAN decreased by 2.5%, 11.1%, and 5.3%, respectively, when the number of training samples was reduced to 3000. Moreover, with a further reduction to 1000 training samples, their F1-scores decreased by 17.5%, 14.9%, and 14.0%, respectively. Notably, while BERT-BiLSTM exhibits the smallest decrease in effectiveness with 3000 training samples, it demonstrates the most significant decline when the number of training samples is reduced to 1000, indicating relatively poor robustness.

In contrast, the proposed COLBERT model exhibits the best performance in terms of mitigating the decline in average effectiveness. The decline ratios of its F1-score are merely 5% and 11.3%, respectively. The proposed method in this paper exhibits a significantly smaller decrease in comparison to the benchmark methods. This observation suggests that the proposed method is less sensitive to variations in the training sample size, as the reduction in training samples has less of an impact on its performance. These results signify that the COLBERT model possesses better robustness. Consequently, in scenarios with limited training samples, the model introduced in this paper can enhance the effectiveness of entity disambiguation tasks. As a result, it holds the potential to alleviate the high costs associated with annotation and model training in practical applications.

4.5. Discussion

4.5.1. Effectiveness of Feature Fusion

In order to analyze the contribution of feature fusion in the COLBERT model to the entity disambiguation task, ablation experiments were applied, and the results are shown in Table 5.

Table 5. Results of ablation experiments (%). The values in bold refer to the optimal results under the same conditions.

Model	**10,000 Training Samples**			**3000 Training Samples**			**1000 Training Samples**		
	P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
COLBERT	**84.6**	83.4	**84.0**	**76.0**	83.9	**79.8**	**74.8**	74.3	**74.5**
-LDA	82.1	**85.8**	83.9	71.5	**88.5**	79.1	69.0	**80.2**	74.1
-BERT	60.1	68.3	64.0	56.2	65.1	60.3	50.3	60.8	55.1

Table 5 illustrates the entity disambiguation performance of the COLBERT model when topic features or semantic features are excluded. The findings indicate that both thematic and semantic features contribute to the effectiveness of entity disambiguation tasks, and the removal of either feature type diminishes the disambiguation effect. Notably, the absence of topic features leads to a slight decrease in model performance (average F1-score decreases by 0.4%). In contrast, the removal of semantic features extracted by the BERT model significantly impacts the model's effectiveness (average F1-score decreases by 19.6%). These results highlight the greater role of semantic features in the COLBERT model and emphasize the substantial impact of removing modules with a high number of parameters on the model's performance. Meanwhile, we can observe that Precision will decrease and Recall will improve when the LDA module is removed, which means that the LDA model tends to promote the accuracy of the model's judgment on negative samples.

4.5.2. The Role of Contrastive Learning

To evaluate the impact of the contrastive learning method in the COLBERT model, we conducted experiments by varying different α values of the joint loss function *l_all* and examined the model's performance. And the results are shown in Figure 9, where the contrastive learning training method is not introduced when α is 1. In this scenario, the training set consists of 10,000 samples.

Based on the observations from Figure 9, it can be inferred that the model's performance exhibits improvement to a certain degree with every value of α. Notably, there is no discernible linear relationship between the changing trends of the α value and the model's effectiveness. This indicates that introducing contrastive learning training enhances the model's performance in the entity disambiguation task. However, the value of α should be treated as a hyper-parameter and rigorously tested on the validation set to ascertain its optimal setting.

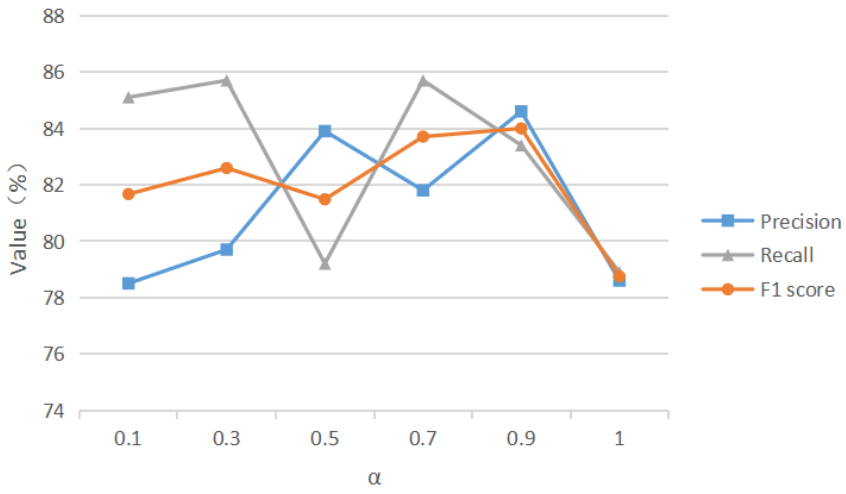

Figure 9. Exploring the effect of the contrastive learning method.

Furthermore, we investigated the impact of the contrastive learning method on the extracted feature representations of the model. Specifically, we employed two metrics, Alignment and Uniformity, as proposed by Wang and Isola [36], to evaluate the quality of text feature representations. The Alignment metric calculates the expected value of the distances between positive samples, assuming the vector has been normalized:

$$l_{align} \triangleq E_{(x,\, x^{+})\sim P_{pos}\|f(x)-f(x^{+})\|^{2}} \tag{7}$$

Meanwhile, Uniformity measures the uniform distribution of the feature representations of all samples:

$$l_{unif} \triangleq logE_{(x,\, y)\sim P_{data}}e^{-2\|f(x)-f(y)\|^{2}} \tag{8}$$

In classification tasks, lower values of Alignment and Uniformity indicate that the model is more proficient in discerning various types of text, suggesting the extraction of superior features. In this study, we employed these two metrics to assess the evolution of text features during training. The variation in the two metrics, along with the superimposed training rounds, is illustrated in Figure 10, with the arrow indicating the direction of change.

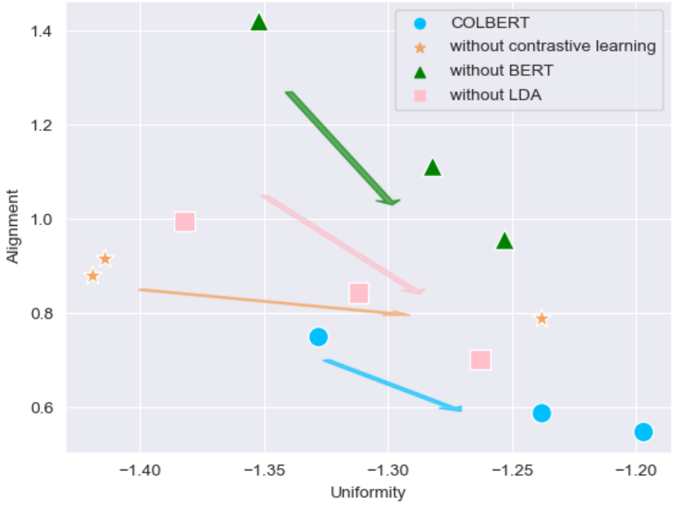

Figure 10. Development of text features' quality in training rounds.

According to Figure 10, we can draw the following conclusions:

(1) With the increase in training rounds, the value of the Alignment of features from each extractor continues to decrease, indicating that the text features of samples in the same category are clustered in the vector space, which helps the classification layer better distinguish the categories of samples. And it is notable that when the LDA model, BERT model, or contrastive learning method is removed, the quality of the features will decrease. This means that all of them can improve the text representations.

(2) The slope of the blue arrow is larger than that of the yellow one, indicating that the introduction of the contrastive learning method can accelerate this trend; that is, after adding contrastive learning, the model can reach a smaller Alignment value faster. Meanwhile, we observed that Uniformity increased during the training rounds, indicating that when the model separates different types of samples, the vector distribution inevitably becomes uneven. However, the introduction of the contrastive learning method has little effect on the Uniformity value, indicating that the contrastive learning method will not aggravate this trend.

In conclusion, the introduction of the contrastive learning method and feature fusion can make the model learn better text representations so as to improve the effect of the model in the task of entity disambiguation, especially when the training samples are not sufficient. And the contrastive learning method plays a very important role in this process by accelerating feature optimization processing and helping the model achieve better text representations.

5. Conclusions

We present COLBERT, a novel entity disambiguation model that leverages feature fusion and contrastive learning. The proposed model constructs an end-to-end disambiguation approach by combining LDA-based topic features and BERT-based semantic features. Additionally, the contrastive learning method is introduced during the training loop to enhance the fused features. This model effectively addresses the challenges of inadequate feature extraction and excessive reliance on training samples in short-text entity disambiguation tasks and outperforms benchmark methods in terms of performance and robustness, offering a novel approach to short-text entity disambiguation tasks. However, it is important to acknowledge certain limitations of the method. For instance, it requires prior efforts to determine the optimal number of topics and λ values, and the large number of model parameters may result in reduced computational efficiency during practical implementation.

In future work, we plan to enhance the model in the following aspects. Firstly, we aim to leverage more powerful pre-trained models and recently proposed topic models to extract more comprehensive features and further enhance the model's effectiveness. Secondly, we intend to integrate a broader knowledge base, such as HowNet [37], to enrich the descriptions of entities and improve the model's universality. Furthermore, our future work will aim to develop a multilingual entity disambiguation model that is applicable to various languages, including Chinese, English, and potentially others.

Author Contributions: Data collection, data processing, design and implementation of research programs, drafting of the paper, and revision of the paper: Q.M.; proposal of the research idea, improvement of the research scheme, proposal of several rounds of revision suggestions, and revision of the final paper: X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Key Research and Development Program sponsored by the Ministry of Science and Technology of the People's Republic of China (Grant No. 2022YFF0904300) and the National Social Science Fund of China (Grant No. 21&ZD334), it was also supported by National Demonstration Center for Experimental Library and Information Science Education, Wuhan University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets associated with the paper can be accessed at https://www.luge.ai/#/luge/dataDetail?id=24 (accessed on 26 September 2023).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Nemes, L.; Kiss, A. Information Extraction and Named Entity Recognition Supported Social Media Sentiment Analysis during the COVID-19 Pandemic. *Appl. Sci.* **2021**, *11*, 11017. [CrossRef]
2. Han, X.; Kim, J.; Kwoh, C. Active learning for ontological event extraction incorporating named entity recognition and unknown word handling. *J. Biomed. Semant.* **2016**, *7*, 22. [CrossRef] [PubMed]
3. Al-Moslmi, T.; Gallofré Ocaña, M.; LOpdahl, A.; Veres, C. Named Entity Extraction for Knowledge Graphs: A Literature Overview. *IEEE Access* **2020**, *8*, 32862–32881. [CrossRef]
4. Bagga, A.; Baldwin, B. Entity-based cross-document coreferencing using the vector space model. In Proceedings of the COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics, Stroudsburg, PA, USA, 10–14 August 1998.
5. Fleischman, M.; Hovy, E. Multi-document person name resolution. In Proceedings of the Conference on Reference Resolution and Its Applications, Barcelona, Spain, 25–26 July 2004; pp. 1–8.
6. Pedersen, T.; Purandare, A.; Kulkarni, A. Name discrimination by clustering similar contexts. In Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, Mexico, 13–19 February 2005; Springer: Berlin/Heidelberg, Germany, 2005; pp. 226–237.
7. Pilz, A.; Paaß, G. From names to entities using thematic context distance. In Proceedings of the 20th ACM international conference on Information and Knowledge Management, Glasgow, UK, 24–28 October 2011; pp. 857–866.
8. He, Z.; Liu, S.; Li, M.; Zhou, M.; Zhang, L.; Wang, H. Learning entity representation for entity disambiguation. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Sofia, Bulgaria, 4–9 August 2013; pp. 30–34.
9. Sun, Y.; Lin, L.; Tang, D.; Yangz, N.; Jiy, Z.; Wang, X. Modeling mention, context and entity with neural networks for entity disambiguation. In Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 25–31 July 2015.
10. Zhang, Y.; Liu, J.; Huang, B.; Chen, B. Entity Linking Method for Chinese Short Text Based on Siamese-Like Network. *Information* **2022**, *13*, 397. [CrossRef]
11. Shi, Y.; Yang, R.; Yin, C.; Lu, Y.; Yang, Y.; Tao, Y. Entity Linking Method for Chinese Short Texts with Multiple Embedded Representations. *Electronics* **2023**, *12*, 2692. [CrossRef]
12. Moller, C.; Lehmann, J.; Usbeck, R. Survey on English Entity Linking on Wikidata. *arXiv* 2021. [CrossRef]
13. De Bonis, M.; Falchi, F.; Manghi, P. Graph-based methods for Author Name Disambiguation: A survey. *PeerJ Comput. Sci.* **2023**, *9*, e1536. [CrossRef] [PubMed]
14. Minkov, E.; Cohen, W.W.; Ng, A. contextual search and name disambiguation in email using graphs. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA, USA, 6–11 August 2006; pp. 27–34.
15. Zhang, B.; Saha, T.K.; Al Hasan, M. Name disambiguation from link data in a collaboration graph. In Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), Beijing, China, 17–20 August 2014; pp. 81–84.
16. Phan, M.C.; Sun, A.; Tay, Y.; Han, J.; Li, C. Pair-linking for collective entity disambiguation: Two could be better than all. *IEEE Trans. Knowl. Data Eng.* **2019**, *31*, 1383–1396. [CrossRef]
17. Han, X.; Zhao, J. Structural Semantic Relatedness: A Knowledge-Based Method to Named Entity Disambiguation. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010.
18. Bouarroudj, W.; Boufaïda, Z.; Bellatreche, L. WeLink: A Named Entity Disambiguation Approach for a QAS over Knowledge Bases. In Proceedings of the International Conference on Flexible Query Answering Systems, Amantea, Italy, 17–19 June 2019.
19. Lommatzsch, A.; Ploch, D.; Luca, E.W.; Albayrak, S. Named Entity Disambiguation for German News Articles. *LWA* **2010**, *2*, 209–212.
20. Blei, D.M.; Ng, A.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2001**, *3*, 993–1022.
21. Jelodar, H.; Wang, Y.; Yuan, C.; Feng, X. Latent Dirichlet allocation (LDA) and topic modeling: Models, ap-plications, a survey. *Multimed. Tools Appl.* **2017**, *78*, 15169–15211. [CrossRef]
22. Chen, Q.; Yao, L.; Yang, J. Short text classification based on LDA topic model. In Proceedings of the 2016 International Conference on Audio, Language and Image Processing (ICALIP), Shanghai, China, 11–12 July 2016; pp. 749–753.
23. Jiang, S.; Xian, Y.; Wang, H.; Zhang, Z.; Li, H. Representation Learning with LDA Models for Entity Disam-biguation in Specific Domains. *J. Adv. Comput. Intell. Intell. Inform.* **2021**, *25*, 326–334. [CrossRef]
24. Zhang, W.; Su, J.; Tan, C.L. A Wikipedia-LDA Model for Entity Linking with Batch Size Changing Instance Selection. In Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, 8–13 November 2011.

25. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805. [CrossRef]
26. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
27. Zhang, P.; Zhao, H.; Wang, F.; Zeng, Q.; Amos, S. Fusing LDA Topic Features for BERT-based Text Classification. *Res. Sq.* **2022**. [CrossRef]
28. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G.E. A simple framework for contrastive learning of visual representations. *arXiv* **2020**, arXiv:2002.05709.
29. Majumder, O.; Ravichandran, A.; Maji, S.; Polito, M.; Bhotika, R.; Soatto, S. Revisiting Contrastive Learning for Few-Shot Classification. *arXiv* **2021**, arXiv:2101.11058.
30. Stevens, K.; Kegelmeyer, W.P.; Andrzejewski, D.; Buttler, D.J. Exploring Topic Coherence over Many Models and Many Topics. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Jeju Island, Republic of Korea, 12–14 July 2012.
31. Wan, C.; Li, B. Financial causal sentence recognition based on BERT-CNN text classification. *J. Supercomput.* **2021**, *78*, 6503–6527. [CrossRef]
32. Abas, A.R.; Elhenawy, I.; Zidan, M.; Othman, M. BERT-CNN: A Deep Learning Model for Detecting Emotions from Text. *Comput. Mater. Contin.* **2022**, *71*, 2943.
33. Dai, Z.; Wang, X.; Ni, P.; Li, Y.; Li, G.; Bai, X. Named Entity Recognition Using BERT BiLSTM CRF for Chinese Electronic Health Records. In Proceedings of the 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Suzhou, China, 19–21 October 2019; pp. 1–5.
34. Xia, L.; Ye, J.; Luo, D.; Guan, M.; Liu, J.; Cao, X. Short text automatic scoring system based on BERT-BiLSTM model. *J. Shenzhen Univ. Sci. Eng.* **2022**, *39*, 349. [CrossRef]
35. Ravi, M.P.; Singh, K.; Mulang, I.O.; Shekarpour, S.; Hoffart, J.; Lehmann, J. CHOLAN: A Modular Approach for Neural Entity Linking on Wikipedia and Wikidata. *arXiv* **2021**, arXiv:2101.09969.
36. Wang, T.; Isola, P. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020.
37. Dong, Z.; Dong, Q. HowNet—A hybrid language and knowledge resource. In Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China, 26–29 October 2003; pp. 820–824.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.