ISSN: 2664-9640

An Entity Linking Improvement Model Based on Sentence Semantic Embedding Enhancement

YuXin Luo [1,*], BaiLong Yang [1], DongHui Xu [1], LuoGeng Tian [1,2] and JingYuan He [1,3]

[1] Xi'an Reaseach Inst. Of Hi-Tech, Xi'an 710025, China

[2] National University of Defense, Wuhan 430030, China

[3] School of Mathematics and Computer Science, Yan'an University, Yan'an 716000, China

Abstract

Entity linking is one of the key technologies for knowledge graph applications, but the existing entity linking model has problems such as weak semantic expressiveness of generated sentence embedding, large errors in the calculation of semantic similarity features, and insufficient utilization of sentence-level entity features. In this paper, the existing entity linking model generates sentence embedding with weak semantic expressiveness, which brings errors to the calculation of relevant semantic similarity features, and in addition, entity descriptions as key information in knowledge graphs, and sentence-level features are not effectively utilized. An entity linking model based on sentence semantic embedding enhancement is proposed, which uses unsupervised contrastive learning to optimize the BERT semantic space, and the sentence embedding generated by it are semantically related to each other through an attention mechanism to enhance sentence semantic embedding, and sentence-level similarity features referring to context and entity descriptions are introduced as supplementary information to local item in the benchmark model mulrel-nel. The average F1 value of the proposed model on the five out-of-domain datasets is 86.28, which is 0.77 improvement compared to the benchmark model.

Keywords

Knowledge Graph; Entity Linking; Semantic Enhancement; Attention Mechanism; Contrastive Learning.

1. Introduction

Entity linking refers to linking the mentions in a text to the corresponding entities in the knowledge graph to solve the ambiguity problem in the text. Entity linking models play an important role in applications related to knowledge graphs, including information extraction [1], question answer [2], and semantic search [3], all of which are predicated on the exact semantics of the text.

Traditional entity linking models are based on statistical models and design many discriminative features, such as entity popularity [4], entity type [5], etc. Current scholars are devoted to constructing global models. the DeepED model proposed by The DeepED model proposed by Ganea and Hofmann [6] by constructing local and global terms that incorporating mentions of contextual information and entity consistency features, outperforms traditional methods in standard benchmark tests.. Le and Titovp [7] propose the mulrel-nel model based on DeepED [6], which incorporates potential relationship information between entities into global item, where relationships are considered as potential variables without additional supervision, and constructs relational embedding through representation learning.

ISSN: 2664-9640

Although these two models achieve excellent performance in the entity linking task, they suffer from two problems. On the one hand, no pre-trained language model is chosen but simply a few layers of neural networks to embedding representation of sentences, and on the other hand, entity descriptions are not fully utilized, which as important information can compensate for the sparsity of the knowledge graph. Chen et al [5] incorporated potential entity type information from entity descriptions into the local item of the DeepED model by pre-training the language model BERT [8], but did not consider entity description sentence-level features. Jia et al [9] constructed twin neural networks (Siamese network) based on BERT to semantically associate sentence embedding referring to context and entity descriptions, but only used the acquired sentence-level similarity as the only discriminative feature.Reimers and Gurevych [10] found that sentence embedding obtained directly with BERT have anisotropy and poor semantic expressiveness, and can even be weaker than sentence embedding generated by the Glove [11] model, and the problem persists even after fine-tuning BERT.

The above models fail to effectively utilize the sentence-level features of entity descriptions, and the generated sentence embedding have weak semantic expressiveness, which brings errors to the calculation of relevant semantic similarity features. For solving the problems in the current entity linking models, this paper has the following three main contributions:

1) In order to obtain high-quality sentence embedding, the BERT semantic space is optimized using unsupervised contrastive learning method, and the dataset consists of randomly selected texts of entity linking tasks. The experimental results show that the BERT semantic space optimization is able to obtain high-quality sentence embedding that are more applicable to this task

2) Association between sentence embedding based on attention mechanism to complement each other's semantics. Sentence-level similarity features referring to context and entity descriptions are aggregated to local item of the mulrel-nel model as complementary discriminative features.

(3) Since the current knowledge graph generally does not contain the description information of entities, this paper crawls the abstracts of all candidate entities in Wikipedia, which constitute a simple local document for experimental extraction. The proposed model in this paper performs validation experiments on in-domain and out-of-domain datasets, respectively. The results show that the proposed model has some improvements on the baseline and can effectively improve the quality of entity links.

2. Background and Related Work

2.1. Entity Linking Task

A text will contain several mentions $m_1, m_2, \ldots, m_n$. The goal of entity linking is to map each mention to the candidate entity that correctly corresponds to it in the knowledge graph, i.e., $m_i \to e_i$.

Entity linking is usually performed in two steps: candidate entity generation and entity disambiguation. A heuristic is generally used to obtain the set of candidate entities $C_i = (e_{i1}, \ldots, e_{il_i})$ and to disambiguate the unlikely options. The purpose of entity disambiguation is to find the entity that best fits the mention context of the statement from the set of candidate entities. In this paper, we focus on entity disambiguation. The current approach focuses on entity disambiguation jointly with local item, which correspond to the degree of entity fit to the mentioned context, and global item, which correspond to entity consistency.

ISSN: 2664-9640

2.2. Related Work

The work in this paper focuses on enhancing the semantic representation of sentence embedding by improving the semantic space of BERT and semantically associating sentence embedding that mention context and entity descriptions ,and the sentence-level similarity feature is introduced. The following two aspects are related to the previous approach in.

2.2.1 BERT Improvement

Gao et al [14] pointed out that the language modeling capability of BERT may be limited by the embedding space of each heterogeneous word.Ethayarajh [15] found that the sentence embedding generated by BERT are non-smooth in the semantic space, which makes it difficult to use sentence embedding by simple similarity measures (dot product or cosine similarity). Li et al [16] addressed how to fully utilize the semantic information of BERT-encoded sentences in an unsupervised situation by transforming the anisotropic sentence embedding distribution into a smooth isotropic Gaussian distribution through normalized flow, called "BERT-flow". Su et al [17] pointed out that the "BERT-flow" flow model has too large a parameter magnitude and produces limited effects. They used the whitening operation in machine learning instead of flow model to reduce the dimensionality of the vector distribution by PCA(Principal Component Analysis) to eliminate redundant information, called "BERT-whitening", and achieved comparable performance with BERT-flow. Gao et al [12] proposed SimCSE, which achieves SOTA for nonsupervised semantic similarity task by constructing positive samples for comparison learning with a simple "twice dropout". In this paper, we adopt the unsupervised contrastive learning method in SimCSE, randomly select certain mentioned contexts and entity descriptions as training data, and retrain BERT to make its semantic space more homogeneous.

2.2.2 Sentence Semantic Embedding

Scholars have continuously proposed improved methods on how to embed the semantic information of sentences more fully into a fixed-length vector. Ma et al [18] combined deep learning and language structure to propose a dependency-based convolutional framework for embedding representation of sentences. With the advent of BERT, which inputs individual sentences into BERT and produces fixed-size sentence embedding, subsequent NPL tasks using BERT to obtain sentence embedding have become the mainstream approach. Taking a sentence as an input to BERT, BERT outputs sentence embedding in two main ways, taking the output of text-tagged CLS or doing pooling operations on the output of all token. SBERT [14] improves on the network of BERT by introducing a triple network structure, achieving a significant improvement in sentence embedding methods. However, this improvement is there caused by high quality supervised training.IS-BERT [19] proposed a lightweight extension model of BERT using an unsupervised approach to derive meaningful sentence embedding and based on a mutual information maximization strategy for unsupervised tasks. Different from the above approaches, this paper uses BERT optimized by unsupervised contrastive learning to obtain sentence embedding and enhances the local semantic embedding of sentences by correlating between sentence embedding based on attention.

3. Model

The work in this paper focuses on enhancing the semantic representation of sentence embedding by improving the semantic space of BERT and semantically associating sentence embedding that mention context and entity descriptions ,and the sentence-level similarity feature is introduced. The following two aspects are related to the previous approach in.

ISSN: 2664-9640

3.1. Main Model

The entity linking model merges the local model with the global model by CRF (Conditional Random Field). A score function g is defined to calculate the joint total entity $m_1,...,m_n$ score after mapping the model to all mentions $e_1,...,e_n$ in the text. The function is as follows:

$$g(e_1,...,e_n \mid D) = \sum_{i=1}^{n} \psi(e_i, c_i) + \sum_{i \neq j} \varphi(e_i, e_j \mid D) \tag{1}$$

,where the first item is a local item and the second item is a global item.

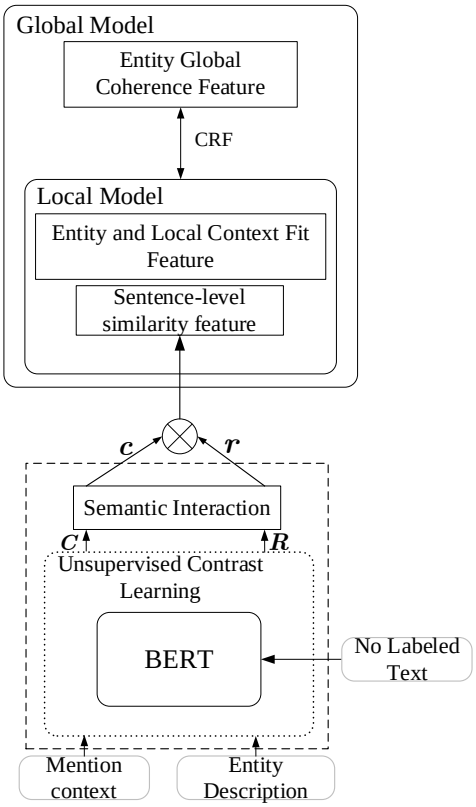


Figure. 1 Model Framework

3.1.1 Entity Linking Task

The local model calculates the score of the fit between the entity and the mentioned context, i.e., the local item. Let c_i be the local context of mentioning m_i, and e_i be the candidate entity after mapping, then the score function of the local model is :

$$\psi_{\text{entity}}(e_i, c_i) = \boldsymbol{e}_i^T \boldsymbol{B} f(c_i) \tag{2}$$

,where $\boldsymbol{e} \in \mathbb{Z}^d$ is the entity word embedding, $\boldsymbol{B} \in \mathbb{Z}^{d \times d}$ is the learnable diagonal matrix, and $f(c_i) \in \mathbb{Z}^d$ denotes the feature vector representation of the mentioned context obtained by neural network mapping. The local item selects the candidate entity with the highest score as the real entity corresponding to the mention:

$$e_i^* = \arg\max_{e_i \in C_i} \psi(e_i, c_i) \tag{3}$$

,where $i \in \{1,...,n\}$.

3.1.2 Global Model

The global model introduces the entity coherence, i.e., the global term, on top of the local model. Where the coherence score function of two entities is :

ISSN: 2664-9640

$$\varphi(e_i, e_j \mid D) = \sum_{k=1}^{K} \alpha_{ijk} \boldsymbol{e}_i^T \boldsymbol{R}_k \boldsymbol{e}_j \tag{4}$$

,where D is all mention contexts, $\boldsymbol{R}_k \in \mathbb{Z}^{d \times d}$ is the learnable diagonal matrix, k is the relationship between entities, and α_{ijk} is the normalized weight factor:

$$\alpha_{ijk} = \frac{1}{Z_{ijk}} \exp\left\{ \frac{f^{\top}(m_i, c_i) \boldsymbol{D}_k f(m_i, c_i)}{\sqrt{d}} \right\} \tag{5}$$

,where Z_{ijk} is the normalization factor, $f(m_i, c_i)$ is the mapping of mentions to their contexts into a feature vector $\mathbb{Z}^d$, and $\boldsymbol{D}_k \in \mathbb{Z}^{d \times d}$ is also a learnable diagonal matrix.

Then the global model is defined as:

$$q(E \mid D) \propto \exp\left\{ \sum_{i=1}^{n} \psi(e_i, c_i) + \sum_{i \neq j} \varphi(e_i, e_j \mid D) \right\} \tag{6}$$

Training and predicting the binary conditional random field of the global model is an NP-hard problem [13]. mulrel-nel uses a truncated fit LBP(loop belief propagation)algorithm, an approximate inference method based on message passing, to estimate the maximum edge probability for each mention:

$$\hat{q}_i(e_i \mid D) \approx \max_{\substack{e_1,\ldots,e_{i-1} \\ e_{i+1},\ldots,e_n}} q(E \mid D) \tag{7}$$

A mention that the final score function of m_i is:

$$\rho_i(e) = g(\hat{q}_i(e \mid D), p^*(e \mid m_i)) \tag{8}$$

where g is a two-layer neural network and $p^*(e \mid m_i)$ refers to the prior probability of selecting entity m_i conditional on mentioning e. This probability can be calculated from the hyperlinked statistics of mentions to entities in Wikipedia, large Web corpora and YAGO.

In this paper, we use a pre-trained language model, BERT, instead of a simple neural network to obtain a feature vector representation of the sentences. That is：

$$f(m_i, c_i) \rightarrow BERT(m_i, c_i) \tag{9}$$

3.2. Unsupervised Contrastive Learning for BERT

The idea of contrast learning is to aggregate similar samples and separate dissimilar ones [20]. The key to contrast learning is to construct positive example pairs, unlike images, natural languages with highly discrete structures are difficult to construct semantically consistent positive examples.

The Transformers [23] module in BERT has a dropout mask mechanism, which sets a smaller dropout (p=0.1) parameter value. Although the sentence embedding obtained from a sentence after two dropouts are not the same, the semantic expectations are basically the same and can be used as positive example pairs with each other and negative example pairs with other sentence embedding.

In this paper, a certain number of mention contexts and entity descriptions are randomly selected to form a training set $S = \{x_i\}_{i=1}^{m}$, and each sentence input BERT twice. Set $\boldsymbol{h}_i^z = f_\theta(x_i, z)$ to be the sentence embedding of sentence x_i, where z is the mask of random dropout. then $(\boldsymbol{h}_i^{z_i}, \boldsymbol{h}_i^{z_i'})$ is the positive example pair, and the sentence embedding of different sentences $(\boldsymbol{h}_i^{z_i}, \boldsymbol{h}_j^{z_j})$ is used as the negative example pair. In this paper, we follow the comparison framework of Chen et al [24], and the loss function is:

ISSN: 2664-9640

$$loss\ell_i = -\log \frac{e^{\mathrm{sim}(\boldsymbol{h}_i^{z_i}, \boldsymbol{h}_i^{z_i'})/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}(\boldsymbol{h}_i^{z_i}, \boldsymbol{h}_j^{z_j})/\tau}} \tag{10}$$

,where $\mathrm{sim}(\boldsymbol{h}_1, \boldsymbol{h}_2) = \frac{\boldsymbol{h}_1^{\top}\boldsymbol{h}_2}{\|\boldsymbol{h}_1\| \cdot \|\boldsymbol{h}_2\|}$, $\boldsymbol{h}_i$ is the coded representation of x_i and τ is the temperature coefficient.

3.3. Semantic Interaction

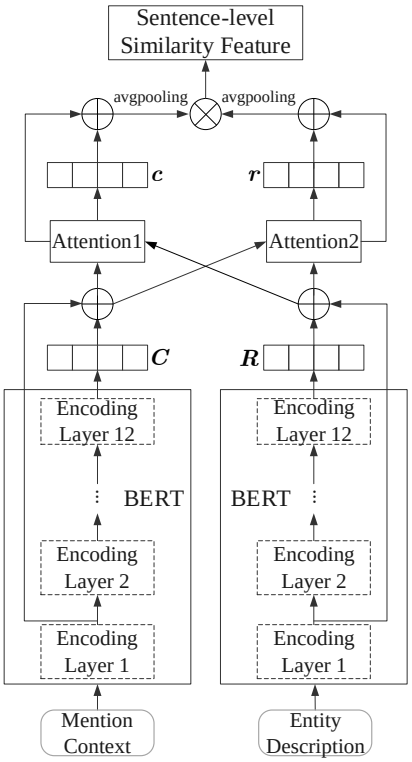


Figure. 2 Semantic interaction module

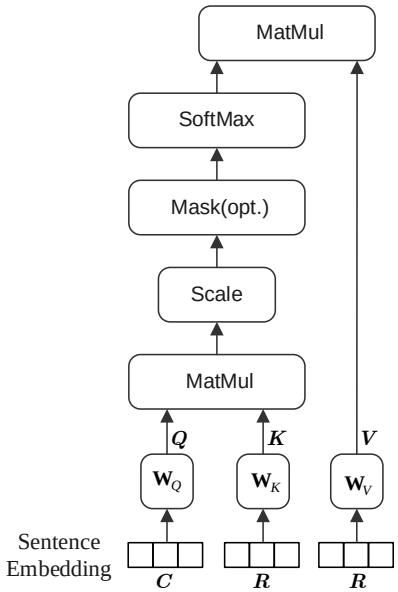


Figure. 3 Scaled Dot Product Attention Mechanism

ISSN: 2664-9640

Figure 2 shows the semantic interaction module, which takes the sentence input through the contrast-learning BERT and does the sum operation on all the corresponding position outputs of the first and last layer of the BERT encoding block. c denotes the context of the mention m, e denotes the candidate entity corresponding to the mention m, and r denotes the descriptive sentence of the candidate entity e.

$$\boldsymbol{C} = \mathrm{BERT}[c]_{\mathrm{layer1}} + \mathrm{BERT}[c]_{\mathrm{layer12}} \tag{11}$$

$$\boldsymbol{R} = \mathrm{BERT}[r]_{\mathrm{layer1}} + \mathrm{BERT}[r]_{\mathrm{layer12}} \tag{12}$$

,where $\boldsymbol{C}, \boldsymbol{R} \in \mathbb{Z}^{M*N}$ are the same as the sentence embedding matrix.

Considering that entity descriptions contain only partially useful information, while the attention mechanism can increase the weight of keywords in a sentence and reduce the interference of other irrelevant words, and can semantically relate two sentences. Therefore, sentence embedding that the mention context and entity descriptions after an additive-sum operation are interacted through the attention mechanism. In this paper, we follow the Scaled Dot Product attention mechanism in Transformer [23], as shown in Figure 3. the sentence embedding are first multiplied with the corresponding weight matrix for dimensional transformation, and then the attention mechanism related operations are performed. The formula is:

$$\mathrm{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{soft\,max}(\frac{\boldsymbol{Q}\boldsymbol{K}^{\top}}{\sqrt{d_k}})\boldsymbol{V} \tag{13}$$

,where d_k is the size after dimensional transformation. The updated sentence embedding matrix is:

$$\boldsymbol{C} = \boldsymbol{C} + \mathrm{Attention}(\boldsymbol{C}, \boldsymbol{R}, \boldsymbol{R}) \tag{14}$$

$$\boldsymbol{R} = \boldsymbol{R} + \mathrm{Attention}(\boldsymbol{R}, \boldsymbol{R}, \boldsymbol{C}) \tag{15}$$

Finally, the average pooling of the sentence embedding matrix is done to obtain the desired sentence embedding.

$$\boldsymbol{c} = \mathrm{avgpooling}(\boldsymbol{C}) \tag{16}$$

$$\boldsymbol{r} = \mathrm{avgpooling}(\boldsymbol{R}) \tag{17}$$

,where $\boldsymbol{c} \in \mathbb{Z}^{1*N}$, $\boldsymbol{r} \in \mathbb{Z}^{1*N}$.

Define a cosine similarity scoring function to calculate the sentence-level similarity feature scores referring to the mention context and entity descriptions.

$$\psi_{\mathrm{sentence}}(c, e) = \cos\mathrm{ine}(\boldsymbol{c}, \boldsymbol{r}) \tag{18}$$

3.4. Feature Fusion

To aggregate sentence-level similarity features into local item, the feature synthesis method of DeepED [6] is used to merge feature $\psi_{\mathrm{sentence}}(c, e)$ with the original feature $\psi_{\mathrm{entity}}(c, e)$ of the local item through two fully connected layers and a ReLU activation layer, where the feature spaces of $\psi_{\mathrm{entity}}(c, e)$ and $\psi_{\mathrm{sentence}}(c, e)$ are isomorphic.

$$\psi_{\mathrm{local}}(c, e) = f(\psi_{\mathrm{entity}}(c, e), \psi_{\mathrm{sentence}}(c, e)) \tag{19}$$

Define the conditional random global term from equation (20):

$$q(E \mid D) \propto \exp\left\{\sum_{i=1}^{n} \psi_{\mathrm{local}}(c_i, e_i) + \sum_{i \neq j} \varphi(e_i, e_j \mid D)\right\} \tag{20}$$

ISSN: 2664-9640

For the weights α_{ijk} in the global item $\varphi(e_i, e_j \mid D)$ (Equation 5), using the Mention-wise normalization in mulrel-nel, the normalization factor Z_{ijk} for α_{ijk} is:

$$Z_{ijk}=\sum_{\substack{j'=1 \\ j' \neq i}}^{n} \exp\left\{\frac{f^{\top}(m_i, c_i)\mathbf{D}_k f(m_{j'}, c_{j'})}{\sqrt{d}}\right\} \tag{21}$$

The LBP(belief propagation algorithm) [25] is used to estimate the maximum edge probability $\hat{q}_i(e_i \mid D)$ for each mention m_i. Then the score function of a mention m_i is $\rho_i(e)$, where g is another two-layer fully connected neural network used to combine the prior probability $p^*(e \mid m_i)$ and the maximum edge probability $\hat{q}_i(e_i \mid D)$:

$$\hat{q}_i(e_i \mid D) \approx \max_{\substack{e_1,\ldots,e_{i-1} \\ e_{i+1},\ldots,e_n}} q(E \mid D) \tag{22}$$

$$\rho_i(e) = g(\hat{q}_i(e \mid D), p^*(e \mid m_i)) \tag{23}$$

The goal of model training is to minimize the following loss function.

$$L(\theta) = \sum_{D \in \mathbf{E}} \sum_{m_i \in D} \sum_{e \in C_i} h(m_i, e) \tag{24}$$

$$h(m_i, e) = \max(0, \gamma - \rho_i(e_i^*) + \rho_i(e)) \tag{25}$$

where θ is the model parameter, E is the training data set, and e_i^* is the correctly linked entity.

4. Experiment

The model proposed in this paper is built on the Pytorch framework and trained on NVIDIA GeForce RTX 2080 Ti GPUs.

4.1. Datasets

To fully validate the reliability and generalization ability of the model, the model is first trained, evaluated and tested on the in-domain dataset AIDA-CoNLL [26]. The trained models are evaluated on the following five out-of-domain datasets: MSNBC, AQUAINT, ACE2004 maintained and updated by Guo and Barbosa [27], WNED-CWEB (CWEB), WNED-WIKI (WIKI) automatically extracted from ClueWeb and Wikipedia.

Most of the currently constructed knowledge graphs are sparse and may not contain all candidate entity nodes or have limited information available, and most existing methods add external information, such as Wikipedia.Chen et al [5] randomly sampled up to 100 entity descriptions in Wikipedia for each entity, and it is unlikely that the entity nodes in the knowledge graph contain so much textual information. In this paper, we crawl through Wikipedia and integrate the abstracts of all candidate entity descriptions to simulate a local document to provide entity descriptions to the model, where each entity has a corresponding ID number and description, as in Table 1. for unsupervised comparative learning of BERT, this paper randomly constitutes a training set consisting of a certain number of mentioned contexts and entity descriptions.

4.2. Parameter setting

Table1 Description of candidate entities crawled in wikipedia

id	entity	description
12	Anarchism	Anarchism is that advocates stateless societies……

ISSN: 2664-9640

25	Autism	Autism is a disorder of neural development characterized……
…	…	……
41534315	Cindy Griffin	Cindy Griffin is an American coach ……

In the candidate entity generation phase, for a mention, 30 top-ranked candidate entities are first selected based on $p^*(e|m_i)$. After that, the 4 entities with the largest $p^*(e|m_i)$ and the 3 entities with the highest degree of contextual fit are selected as the final candidate entities.

In this paper, the rest of the parameters of the model remain the same as the original parameters of mulrel-nel, except for the required parameters of the module for introducing sentence-level features. In this paper, we use HuggingFace's BERT-base-uncased pre-trained language model [28], which has 12 layers of coding blocks and 768 hidden layer neurons. setting the maximum sentence length to 64, the initial sentence embedding matrix size of a sentence output from BERT is 64*768. in this paper, we use three linear layers to represent the weight matrix $\mathbf{W}_Q$, $\mathbf{W}_K$, $\mathbf{W}_V$, and sentence embedding associated in the attention mechanism and the dimension is 1*768 after pooling by averaging for computing similarity features.

In training the model, the BERT learning rate is 1*10-5 and the rest of the network and statistical parameters learning rate is 2*10-3. mulrel-nel uses the Adam optimizer, but in this paper, we found no significant effect of Adam on the improvement of F1 value of BERT in our experiments, and to ensure the training quality of the model, the Adam improved AdamW [29] optimizer is used.

Usually BERT encodes the surface information of a sentence at the lower layer, captures syntactic information at the middle layer, and extracts semantic information at the higher layer. In order to determine which layer of BERT output as sentence embedding is more suitable for the task of this paper, this paper model performs entity linking experiments on dataset AIDA-B based on the sentence embedding of different encoding layers of BERT output. As shown in Table 2, the sentence embedding with the first and last layers summed achieves the best performance. Therefore, the model in this paper is based on the sentence embedding of the first and last layers of BERT added together.

4.3. Experimental Results

Table2 BERT different layer outputs on the AIDA-B dataset

Layer	AIDA-B
Layer1	89.58
Layer11	90.06
Layer12	91.22
Layer1+ Layer11	90.39
Layer1+ Layer12	**93.65**
Layer11+ Layer12	91.83
All layers	92.37

Table3 F1 scores on the AIDA-B dataset

Methods	AIDA-B
L2R.WNED-CONLL[27]	89.0
Globerson et al. [30]	91.0

ISSN: 2664-9640

Yamada et al.[31]	91.5
DeepED[6][6]	92.22
mulrel-nel[7]	93.07
BERT-Entity-Sim[5]	93.54
ELSR[9]	92.09
CSL-BERT- SE	91.26
BERT-SEAtt	93.15
CSL-BERT-SEAtt	**93.65**

Table 3 shows the F1 values of this paper and other advanced models on the AIDA-B dataset. In order to compare the effect of semantic interaction between contrast learning optimized BERT semantic space and sentence embedding via attention mechanism on the model performance, two additional control models are added: CSL-BERT-SE, a model without semantic interaction, and BERT-SEAtt, a model without contrast learning trained BERT. From Table 3, we can see that the experimental results of CSL-BERT-SE are poor and even inferior to most of the previous advanced models. the experimental results of BERT-SEAtt are better, only 0.39 lower than BERT-Entity-Sim. the model proposed in this paper, CSL-BERT-SEAtt, performs best compared to all previous models, where the F1 value is higher than mulrel-nel and BERT-Entity-Sim by 0.58 and 0.11, respectively.

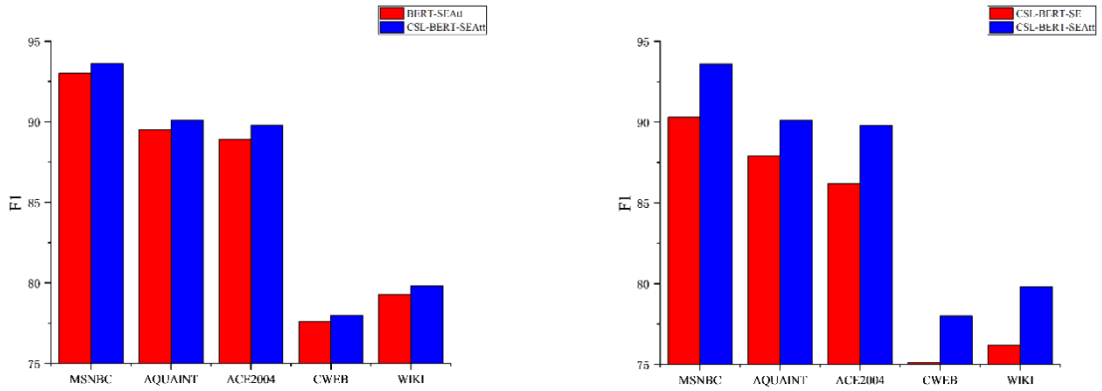
Table 4 validates the generalization ability and stability of the model, evaluated on five other out-of-domain datasets. On the AQUAINT and CWEB datasets, the model proposed in this paper achieved the highest F1 values, 1.8 and 0.5 higher than mulrel-nel, respectively, and the average F1 value is better than all advanced models, including 0.77 and 0.26 higher than mulrel-nel and BERT-Entity-Sim, respectively.

Table4 F1 scores on the out-of-domain dataset

Methods	MSNBC	AQUAINT	ACE2004	CWEB	WIKI	Avg
Cheng and Roth 错误！未找到引用源。	90	90	86	67.5	73.4	81.38
L2R.WNED-CONLL[27]	92	87	88	77	84.5	85.70
DeepED[6]	93.7±0.1	88.5±0.4	88.5±0.3	77.9±0.1	77.5±0.1	85.22
mulrel-nel[7]	**93.9**±0.2	88.3±0.6	**89.9**±0.8	77.5±0.1	78.0±0.1	85.51
BERT-Entity-Sim[5]	93.4±0.1	89.8±0.4	88.9±0.7	77.9±0.4	**80.1**±0.4	86.02
CSL-BERT-SE	90.3±0.3	87.9±0.7	86.2±0.5	75.1±0.4	76.2±0.1	83.14
BERT-SEAtt	93.0±0.4	89.5±0.4	88.9±0.5	77.6±0.2	79.3±0.5	85.66
CSL-BERT-SEAtt	93.6±0.2	**90.1**±0.4	89.8±0.4	**78.0**±0.1	79.8±0.2	**86.28**

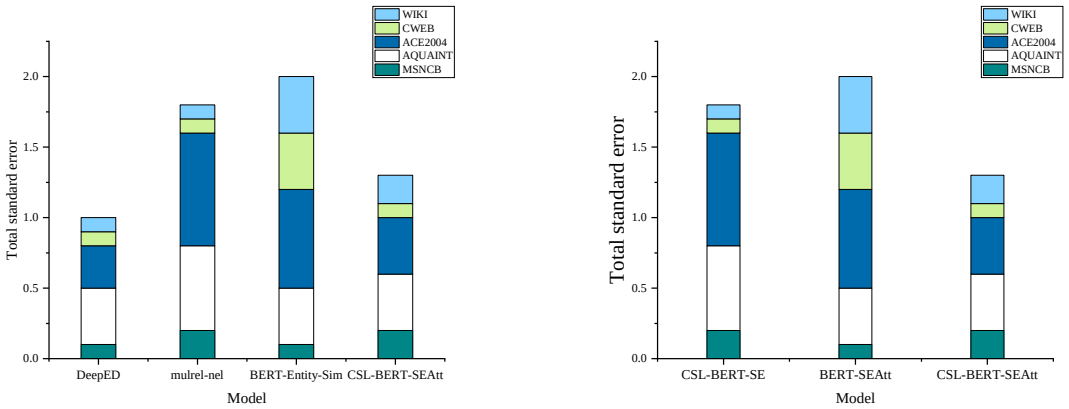
ISSN: 2664-9640

4.4. Analysis



(a) CSL-BERT-SEAtt and BERT-SEAtt (b) CSL-BERT-SEAtt and CSL-BERT-SE

Fig.4 Comparison of experimental results of CSL-BERT-SEAtt and control model



(a) CSL-BERT-SEAtt with advanced models (b) CSL-BERT-SEAtt with comparative model

Figure.5 Comparison of cumulative standard errors of CSL-BERT-SEAtt and other models

4.4.1 Unsupervised Contrastive Learning

The F1 scores of CSL-BERT-SEAtt are all higher than those of BERT-SEAtt on the out-of-domain test set, where the average F1 value is improved by 0.62, as shown in Figure 4(a). It is verified that unsupervised contrastive learning by using randomly selected mention context and entity description texts as training datasets can make the semantic space in BERT more balanced, and the sentence embedding generated by the optimized BERT are more suitable for the entity linking task in this paper, which is beneficial to the calculation of sentence-level similarity features of mention context and entity description.

4.4.2 Semantic Association

Directly calculating the similarity features between entity descriptions and mention contexts will bring more serious noise to the model. A sentence contains a large amount of information irrelevant to mentions and only a small part of useful information, and it is necessary to increase the semantic Interaction between entity descriptions and mention contexts based on the attention mechanism to give higher weight to useful information before calculating similarity features. The performance of the model is significantly improved after the semantic interaction module is introduced, as shown in Figure 4(b), the mean F1 of CSL-BERT-SEAtt is 3.14 higher than that without CSL-BERT-SE on the out-of-domain dataset.

4.4.3 Entity Description

ISSN: 2664-9640

Entity description as the most common information in the knowledge graph, previous approaches did not consider it at the sentence level. In order to be more in line with the objective conditions of knowledge graph, the model proposed in this paper introduces only summary information of entity description and extracts sentence-level features of entity description, and the mean F1 value on the out-of-domain dataset is 0.77 higher than mulrel-nel.

4.4.4 Stability

As shown in Figure 5, the sum of standard errors of CSL-BERT-SEAtt compared with all other models on the out-of-domain dataset. It can be seen that the model in this paper only lags behind the DeepED model in terms of stability, and combined with the F1 index, we are able to verify the advantages of the performance of the model in this paper.

4.4.5 Error Analysis

When analyzing the entity linking errors generated in the experiments, most of them are caused by too little mention of contextual information, or by the mention of appearing abbreviations. For example, the mention of "USS Cole" refers to the guided missile destroyer USS Cole, while its corresponding candidate entities include "USS Cole (DD-155)", "USS Cole (DDG-67)", and "USS Cole bombing", which correspond to different types of destroyers and specific events. "USS Cole (DDG-67)" as the correct entity, even though the entity description information was introduced, the mentioning context did not have enough information, resulting in a link error.

5. Conclusion and Prospect

1) In this paper, we improve the BERT semantic space using the unsupervised contrastive learning method in SIMCSE to output sentence embedding with better semantic quality. The semantic information of sentence embedding is further supplemented by semantic association between sentence embedding of mention context and entity descriptions. The sentence-level similarity features of both are introduced into the local items of the model. After experiments on different datasets, it is proved that the model proposed in this paper has certain advantages.

2) Due to the complex internal structure of BERT, it causes the model to be very time consuming for training evaluation. In the next step, a pre-trained model with better performance and faster training speed can be used according to the technological development. In addition the model proposed in this paper has more parameters and the parameter learning rate has only two fixed values. The learning rate of the parameters can be refined and the learning rate size can be dynamically adjusted to improve the training quality of the model.

References

[1] Ji H, Nothman J, Hachey B, et al. Overview of TAC-KBP2015 Tri-lingual Entity Discovery and Linking[C]//TAC. 2015.

[2] Yih S W, Chang M W, He X, et al. Semantic parsing via staged query graph generation: Question answering with knowledge base[C]//Proceedings of the Joint Conference of the 53rd Annual Meeting of the ACL and the 7th International Joint Conference on Natural Language Processing of the AFNLP. 2015.

[3] Bast H, Björn B, Haussmann E. Semantic search on text and knowledge bases[J]. Foundations and Trends in Information Retrieval, 2016, 10(2-3): 119-271.

[4] Milne D, Witten I H. Learning to link with wikipedia[C]//Proceedings of the 17th ACM conference on Information and knowledge management. 2008: 509-518.

[5] Chen S, Wang J, Jiang F, et al. Improving entity linking by modeling latent entity type information[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(05): 7529-7537.

[6] Ganea O E, Hofmann T. Deep joint entity disambiguation with local neural attention[J]. arXiv preprint arXiv:1704.04920, 2017.

[7] Le P, Titov I. Improving entity linking by modeling latent relations between mentions[J]. arXiv preprint arXiv:1804.10637, 2018.

[8] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[9] Jia B, Wu Z, Zhou P, et al. Entity Linking Based on Sentence Representation[J]. Complexity, 2021, 2021.

[10] Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks[J]. arXiv preprint arXiv:1908.10084, 2019.

[11] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation [C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.

[12] Gao T, Yao X, Chen D. Simcse: Simple contrastive learning of sentence embeddings[J]. arXiv preprint arXiv:2104.08821, 2021.

[13] Wainwright M J, Jordan M I. Graphical models, exponential families, and variational inference[J]. Foundations and Trends® in Machine Learning, 2008, 1(1–2): 1-305.

[14] Gao J, He D, Tan X, et al. Representation degeneration problem in training natural language generation models[J]. arXiv preprint arXiv:1907.12009, 2019.

[15] Ethayarajh K. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings[J]. arXiv preprint arXiv:1909.00512, 2019.

[16] Li B, Zhou H, He J, et al. On the sentence embeddings from pre-trained language models[J]. arXiv preprint arXiv:2011.05864, 2020.

[17] Su J, Cao J, Liu W, et al. Whitening sentence representations for better semantics and faster retrieval[J]. arXiv preprint arXiv:2103.15316, 2021.

[18] Ma M, Huang L, Xiang B, et al. Dependency-based convolutional neural networks for sentence embedding[J]. arXiv preprint arXiv:1507.01839, 2015.

[19] Zhang Y, He R, Liu Z, et al. An unsupervised sentence embedding method by mutual information maximization[J]. arXiv preprint arXiv:2009.12061, 2020.

[20] Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). IEEE, 2006, 2: 1735-1742.

[21] Wu Z, Wang S, Gu J, et al. Clear: Contrastive learning for sentence representation[J]. arXiv preprint arXiv:2012.15466, 2020.

[22] Meng Y, Xiong C, Bajaj P, et al. Coco-lm: Correcting and contrasting text sequences for language model pretraining[J]. Advances in Neural Information Processing Systems, 2021, 34.

[23] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[24] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020.A simple framework for contrastive learning of visual representations. In International Conference on Machine Learning(ICML), pages 1597–1607.

[25] Murphy K, Weiss Y, Jordan M I. Loopy belief propagation for approximate inference: An empirical study[J]. arXiv preprint arXiv:1301.6725, 2013.

[26] Hoffart J, Yosef M A, Bordino I, et al. Robust disambiguation of named entities in text[C]//Proceedings of the 2011 conference on empirical methods in natural language processing. 2011: 782-792.

[27] Guo Z, Barbosa D. Robust named entity disambiguation with random walks[J]. Semantic Web, 2018, 9(4): 459-479.

[28] https://huggingface.co/bert-base-uncased

ISSN: 2664-9640

[29] Loshchilov I, Hutter F. Decoupled weight decay regularization[J]. arXiv preprint arXiv:1711.05101, 2017.

[30] Globerson A, Lazic N, Chakrabarti S, et al. Collective entity resolution with multi-focal attention[J]. 2016.

[31] Yamada I, Shindo H, Takeda H, et al. Joint learning of the embedding of words and entities for named entity disambiguation[J]. arXiv preprint arXiv:1601.01343, 2016

[32] Cheng X, Roth D. Relational inference for wikification[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 1787-1796.