Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105 (Online)
Volume **8**, Number **1**, February **2023**

Entity Linking Improvement Model by Deep Modeling of Sentence Semantics

Yu-Xin Luo*

Xi'an Reaseach Inst. Of Hi-Tech, Xi'an 710025, China
iseef2@163.com

Bai-Long Yang

Xi'an Reaseach Inst. Of Hi-Tech, Xi'an 710025, China
xa_403@163.com

Dong-Hui Xu

Xi'an Reaseach Inst. Of Hi-Tech, Xi'an 710025, China
xdh_45@163.com

Luo-Geng Tian

Xi'an Reaseach Inst. Of Hi-Tech, Xi'an 710025, China
tianlg82@163.com

Jing-Yuan He

Xi'an Reaseach Inst. Of Hi-Tech, Xi'an 710025, China
yau_hejingyuan@163.com

*Corresponding author: Yu-Xin Luo

Received July 20, 2022, revised October 17, 2022, accepted December 1, 2022.

Abstract. *Entity linking is one of the critical technologies for knowledge graph applications. Still, the existing entity linking model has problems such as weak semantic expressiveness of the generated sentence embedding, significant errors in calculating semantic similarity features, and insufficient utilization of sentence-level entity features. The existing entity linking models develop weak semantic expressiveness of sentence embedding and bring mistakes to the computation of relevant semantic similarity features, in addition to entity descriptions, as crucial information in knowledge graphs and sentence-level elements are not effectively utilized. This paper proposes an entity linking model for deeply modeling sentence semantics, which uses unsupervised contrastive learning to optimize the BERT semantic space. The sentence embedding generated by it interacts semantically with each other through an attention mechanism to enhance sentence semantic embedding. It introduces sentence-level similarity features referring to context and entity descriptions as supplementary information to local terms 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 a 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 semantics sorting [1], question answering [2], and multi-modality learning [3], all of which are predicated on the exact semantics of the text.

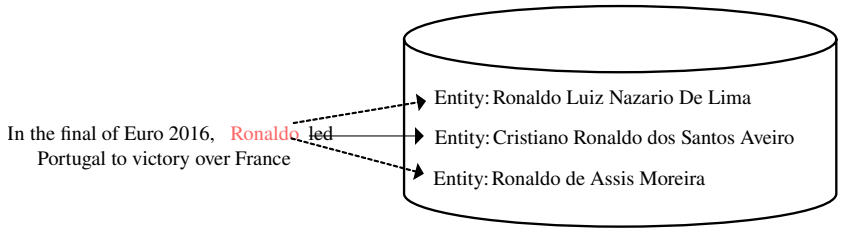


FIGURE 1. Example of entity linking task

Traditional entity linking models are based on statistical models and design many discriminative features, such as entity popularity [4], entity type [5], et al. Current scholars are devoted to constructing global models. The DeepED model proposed by Ganea and Hofmann [6] outperforms traditional methods in standard benchmark tests by constructing local and global terms that incorporate references to contextual information and entity consistency feature. 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 potential variables without additional supervision, and constructs relational embedding through representation learning.

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. 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]. Still, they 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. Sentence semantic embedding refers to encoding sentence semantics into a fixed-length embedded vector through deep learning, which can be used for the numerical calculation of related features between sentences. High-quality sentence embedding plays an essential role in enhancing the reliability of entity link discriminant features. The above methods mainly use BERT to obtain sentence embedding; Reimers and Gurevych [10] found that sentence embedding obtained directly with BERT has anisotropy and poor semantic expressiveness and can even be weaker than sentence embedding generated by the Glove [11] model. The problem persists even after fine-tuning BERT. It is shown that unsupervised contrast learning can improve the BERT semantic space and thus obtain high-quality sentence embeddings [12]. Existing entity linking methods construct similarity features between candidate entities and mention contexts based on different perspectives. However, the weak semantic representation of the sentence embeddings leads to significant errors in feature computation, and the entity description information is not effectively utilized as crucial information in the knowledge graph. For solving the problems in the current entity linking models, this paper has the following three main contributions:

1) To obtain high-quality sentence embeddings, the BERT semantic space is optimized using an unsupervised contrastive learning method. The dataset consists of randomly selected texts for the entity linking task. The experimental results show that the BERT semantic space optimization can obtain high-quality sentence embedding more suitable for this task, and its discriminant features are more reliable.
2) Interaction between sentence embeddings based on attention mechanism complements each other's semantics. Sentence-level similarity features referring to context and entity descriptions are aggregated to local terms of the mulrel-nel model as complementary discriminative features, further optimizing of entity-linking results.
3) Since current knowledge graphs generally do not contain descriptive information about entities, this paper crawls the abstracts of all candidate entities in Wikipedia to form a simple local document for experimental extraction. The proposed model in this paper performs validation experiments on the in-domain and out-of-domain datasets, respectively. The results show that the proposed model is somewhat improved in 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 $C_i = (e_{i1}, \ldots, e_{il})$ 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 the local item, which corresponds to the degree of entity fit to the mentioned context, and the global item, which compare to entity consistency.

2.2. **Related Work.** This paper focuses on enhancing the semantic representation of sentence embedding by improving the semantic space of BERT and semantically associating sentence embedding that mentions context and entity descriptions. The sentence-level similarity feature is introduced. The following two aspects are related to the previous approach.

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 is non-smooth in the semantic space, which makes it challenging 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 the 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 tasks 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 specific 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 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 the 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 tokens. SBERT [14] improves 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 based on a mutual information maximization strategy for unsupervised tasks. Unlike the above methods, this paper uses BERT optimized by unsupervised contrastive learning to obtain sentence embedding and enhances the local semantic embedding of sentences by correlating sentence embedding based on attention.

3. **Model.** The proposed model optimizes the BERT semantic space by unsupervised contrastive learning. The sentence embeddings generated by it referring to the context and entity descriptions interact on an attention mechanism to enhance the sentence semantic embedding. The sentence-level similarity features are introduced into the mulrel-nel [7] local model as complementary discriminative features. And the global and local models have selected the optimally linked entities by introducing different discriminative features and evaluating the scores of the linked entities using a score function, respectively. Figure 2 shows the model framework.

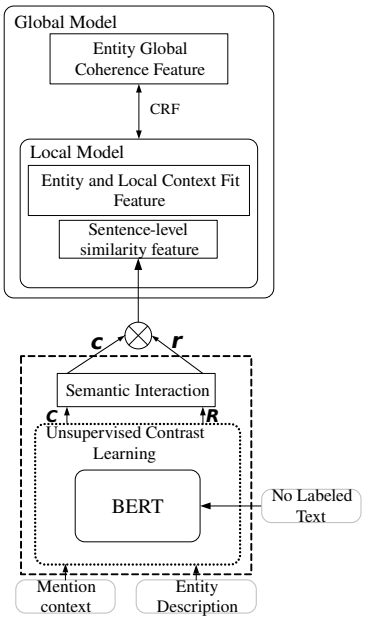


FIGURE 2. Model Framework

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 $m_1, \ldots, m_n$ score after mapping the model to all mentions $e_1, \ldots, e_n$ in the text. The function is as follows:

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

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

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:

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

where $e \in \mathbb{Z}^d$ is the entity word embedding, $B \in \mathbb{Z}^{d \times d}$ is the learnable diagonal matrix, and $f\left(c_i\right) \in \mathbb{Z}^d$ 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:

$$e_i^* = \underset{e_i \in C_i}{\arg\max} \psi\left(e_i, c_i\right) \tag{3}$$

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

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

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

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:

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

where Z_{ijk} is the normalization factor, $f\left(m_i, c_i\right)$ 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:

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

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:

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

A mention that the final score function of m_i is:

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

where g is a two-layer neural network and $p^*\left(e \mid m_i\right)$ 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:

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

3.2. **Unsupervised Contrastive Learning for BERT.** The idea of contrastive learning is to aggregate similar and separate dissimilar samples [20], and the key to contrastive learning is to construct positive example pairs. The BERT Transformers [23] module has a dropout mask mechanism when a small dropout ($p = 0.1$) parameter value is set. Although the sentence obtained by two dropouts of a sentence embedding is not the same, the semantic expectation is the same, and Gao et al. [12] verified this method on SIMCSE.

Therefore, this paper randomly selects a certain number of mentioned contexts and entity descriptions to form a training set $S = \{x_i\}_{i=1}^{m}$, and each sentence inputs BERT twice. Set $\boldsymbol{h}_i^z = f_\theta\left(x_i, z\right)$ to be the sentence embedding of sentence x_i, where z is the mask of random dropout. Then $\left(\boldsymbol{h}_i^{z_i}, \boldsymbol{h}_i^{z_i'}\right)$ is the positive example pair, and the sentence embedding of different sentences $\left(\boldsymbol{h}_i^{z_i}, \boldsymbol{h}_j^{z_j}\right)$ 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:

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

where $\left(\boldsymbol{h}_i^{z_i}, \boldsymbol{h}_j^{z_j}\right), \boldsymbol{h}_i$ is the coded representation of x_i and τ is the temperature coefficient.

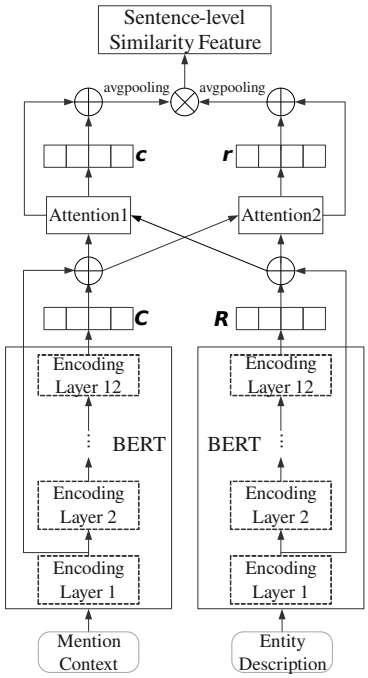


FIGURE 3. Semantic interaction module

3.3. **Semantic Interaction.** Figure 3 shows the semantic interaction module, which takes the sentence input through the Comparative Learning BERT and does the sum operation on all the corresponding position outputs of the first and last layers of the BERT encoding block. c denotes the context of the mention m, e represents the candidate entity

corresponding to the mention m, and r means the descriptive sentence of the candidate entity e.

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

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

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

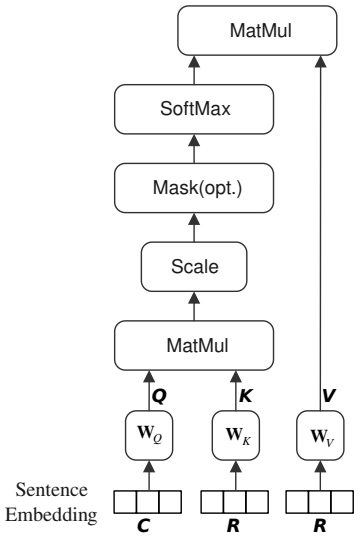


FIGURE 4. Scaled Dot Product Attention Mechanism

Considering that entity descriptions contain only partially helpful information, the attention mechanism can increase the weight of keywords in a sentence, 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 have interacted through the attention mechanism. This paper follows the Scaled Dot Product attention mechanism in Transformer [23], as shown in Figure 4. The sentence embedding is first multiplied with the corresponding weight matrix for dimensional transformation. Then the attention mechanism-related operations are performed. The formula is:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{soft}\max\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\cdot}}{\sqrt{d_k}}\right)V \tag{13}$$

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

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

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

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

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

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

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.

$$\Psi_{\text{sentence}}(c, e) = \text{cosine}(\boldsymbol{c}, \boldsymbol{r}) \tag{18}$$

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

$$\Psi_{\text{local}}(c, e) = f\left(\Psi_{\text{entity}}(c, e), \Psi_{\text{sentence}}(c, e)\right) \tag{19}$$

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

$$\varphi\left(e_i, e_j \mid D\right) \tag{20}$$

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

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

The LBP [25] is used to estimate the maximum edge probability $\hat{q}_i\left(e_i \mid D\right)$ 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 $p^*\left(e \mid m_i\right)$ and the maximum edge probability $\hat{q}_i\left(e_i \mid D\right)$:

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

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

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

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

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

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.** 4.1.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 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. The entity nodes in the knowledge graph are unlikely to 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.** In the candidate entity generation phase, for a mention, 30 top-ranked candidate entities are first selected based on $p^*(e \mid m_i)$. After that, the 4 entities with the largest $p^*(e \mid m_i)$ and the 3 entities with the highest degree of contextual fit are selected as the final candidate entities.

Table 1. Description of candidate entities crawled in wikipedia

id	entity	description
12	Anarchism	Anarchism is that advocates stateless societies .....
25	Autism	Autism is a disorder of neural development characterized .....
...	...	...
41534315	Cindy Griffin	Cindy Griffin is an American coach .....

In this paper, the rest of the model's parameters 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. This paper uses three linear layers to represent the weight matrix $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$, and sentence embedding associated with the attention mechanism. 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 the F1 value of BERT in our experiments, and to ensure the training quality of the model, the Adam improved AdamW [29] optimizer is used.

Table 2. 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

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. 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. **Comparison methods.** 4.3.The performance of the model in this paper is experimentally compared with the following four models: (1) DeepED (2) mulrel-nel (3) BERT-Entity-Sim (4) ELSR.

4.4. **Experimental Results.** 4.4.Table 3 shows the F1 values of this paper and other advanced models on the AIDA-B dataset. To compare the effect of semantic interaction between contrastive 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 contrastive 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 3. F1 scores on the AIDA-B dataset

Methods	AIDA-B
L2R.WNED-CONLL [27]	89.0
Globerson et al. [30]	91.0
Yamada et al. [31]	91.5
DeepED [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

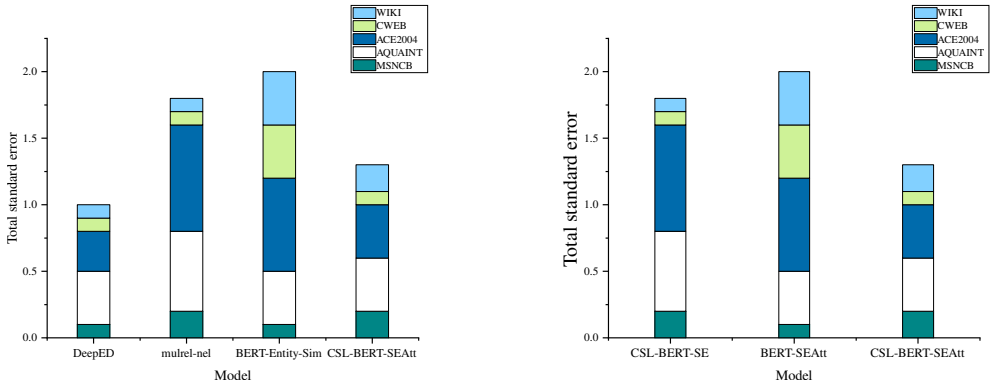
Table 4. F1 scores on the out-of-domain dataset

Methods	MSNBC	AQUAINT	ACE2004	CWEB	WIKI	Avg
Cheng and Roth [32]	90	90	86	67.5	73.4	81.38
L2R.WNED-CONLL [27]	92	87	88	77	84.5	85.70
DeepED	93.7±0.1	88.5±0.4	88.5±0.3	77.9±0.1	77.5±0.1	85.22
mulrel-nel	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	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

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.

4.5. **Analysis.**

4.5.1. *Unsupervised Contrastive Learning.* 4.5.1.The F1 values of CSL-BERT-SEAtt on the out-of-domain test set are all higher than those of BERT-SEAtt, where the average



(A) CSL-BERT-SEAtt with advanced models

(B) CSL-BERT-SEAtt with comparative model

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

F1 value is improved by 0.62, as shown in Table 4. verifies that unsupervised comparative 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 BERT generated by the optimized sentence embeddings, which are more suitable for the entity linking task in this paper, facilitate the calculation of sentence-level similarity features for mentioning contexts and entity descriptions.

4.5.2. *Semantic Association.* Calculating the similarity features between entity descriptions and mention context bring more noise to the model. A sentence contains a large amount of information irrelevant to mentions and only a tiny portion of helpful information. It is necessary to increase the semantic interaction between entity descriptions and mention contexts based on the attention mechanism to give more weight to useful information before calculating similarity features. The model's performance is significantly improved after introducing the semantic interaction module. As shown in Table 4, the mean F1 of CSL-BERT-SEAtt is 3.14 higher than that without CSL-BERT-SE on the out-of-domain dataset.

4.5.3. *Entity Description.* Entity description is the most common information in the knowledge graph, and previous approaches did not consider it at the sentence level. To be more consistent with the objective conditions of the knowledge graph, the model proposed in this paper only introduces the summary information of entity description and extracts the sentence-level features of entity description. The mean F1 value is 0.77 higher than mulrel-nel on the out-of-domain dataset. Chen et al. [5] mainly pull the potential entity type information in entity description, and although this paper considers the entity description information from a different perspective from theirs and introduces this paper considers entity description information from a different perspective than they do and presents less entity description information, the F1 value of CSL-BERT-SEAtt is still 0.26 higher than that of BERT-Entity-Sim, indicating that the entity description summary sentence-level information contains most of the information of entities and is sufficient for entity disambiguation.

4.5.4. *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, which,

combined with the F1 index, can verify the advantages of the performance of the model in this paper.

4.5.5. *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. At the same time, 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)" is 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.** In this paper, we improve the BERT semantic space using the unsupervised contrastive learning method in SIMCSE to output sentence embeddings with better semantic quality. The semantic information of the sentence embeddings is further supplemented by semantic interactions between sentence embeddings referring to the context and entity descriptions. The sentence-level similarity features are introduced into the local terms of the model. After experiments on different datasets, it is demonstrated that the model proposed in this paper has certain advantages. The complexity of the internal structure of BERT causes the model to be very time-consuming to train and evaluate. In the next step, a pre-trained model with better performance and faster training speed can be used according to technological development. In addition, the model proposed in this paper has more parameters, and the learning rate of the parameters 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] K. Wang, C.-M. Chen, M.-S. Obaidat, S. Kumari, S. Kumar, and J Long, "Deep Semantics Sorting of Voice Interaction-Enabled Industrial Control System," *IEEE Internet of Things Journal*, 2021. [Online]. Available: https://doi.org/10.1109/JIOT.2021.3093496

[2] S.-W.-T. Yih, M.-W. Chang, X. He, and J. Gao, "Semantic parsing via staged query graph generation: Question answering with knowledge base," *in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (IJCNLP)*. ACL, 2015, pp. 1321-1331.

[3] E.-K. Wang, X. Zhang, F. Wang, T.-Y. Wu, and C.-M. Chen, "Multilayer dense attention model for image caption," *IEEE Access*, vol. 7, pp. 66358-66368, 2019.

[4] D. Milne, and I.-H. Witten, "Learning to link with Wikipedia," *in Proceedings of the 17th ACM conference on Information and knowledge management (CIKM)*. ACM, 2008, pp. 509-518.

[5] S. Chen, J.-P. Wang, F. Jiang, and C.-Y. Lin, "Improving entity linking by modeling latent entity type information," *in Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. AAAI, 2020, pp. 7529-7537.

[6] O.-E. Ganea, and T. Hofmann, "Deep joint entity disambiguation with local neural attention," *Arxiv Preprint ArXiv*: 1704.04920, 2017.

[7] P. Le, and I. Titov, "Improving entity linking by modeling latent relations between mentions," *Arxiv Preprint ArXiv*: 1804.10637, 2018.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *Arxiv Preprint ArXiv*: 1810.04805, 2018.

[9] B.-J. Jia, Z.-L. Wu, P.-P. Zhou, and B. Wu, "Entity Linking Based on Sentence Representation," *Complexity*, vol. 2021, Article ID 8895742, 9 pages, 2021.

[10] N. Reimers, and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *Arxiv Preprint ArXiv*: 1908.10084, 2019.

[11] J. Pennington, R. Socher, and C.-D. Manning, "Glove: Global vectors for word representation," *in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2014, pp. 1532-1543.
[12] T.-Y. Gao, X.-C. Yao, and D.-Q. Chen, "Simcse: Simple contrastive learning of sentence embeddings," *Arxiv Preprint ArXiv*: 2104.08821, 2021.
[13] M.-J. Wainwright, and M.-I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends® in Machine Learning*, vol. 1, pp. 1-305, 2008.
[14] J. Gao, D. He, X. Tan, T. Qin, L.-W. Wang, and T.-Y. Liu, "Representation degeneration problem in training natural language generation models," *Arxiv Preprint ArXiv*: 1907.12009, 2019.
[15] K. Ethayarajh, "How contextual are contextualized word representations. Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings," *ArXiv Preprint ArXiv*: 1909.00512, 2019.
[16] B.-H. Li, H. Zhou, J.-X. He, M.-X. Wang, Y.-M. Yang, and L. Li, "On the sentence embeddings from pre-trained language models," *Arxiv Preprint ArXiv*: 2011.05864, 2020.
[17] J.-L. Su, J.-R. Cao, W.-J. Liu, and Y.-Y.-W. Ou, "Whitening sentence representations for better semantics and faster retrieval," *Arxiv Preprint ArXiv*: 2103.15316, 2021.
[18] M.-B. Ma, L. Huang, B. Xiang, and B.-W. Zhou, "Dependency-based convolutional neural networks for sentence embedding," *Arxiv Preprint ArXiv*: 1507.01839, 2015.
[19] Y. Zhang, R.-D. He, Z.-Z. Liu, K.-H. Lim, and L.-D. Bing, "An unsupervised sentence embedding method by mutual information maximization," *Arxiv Preprint ArXiv*: 2009.12061, 2020.
[20] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," *in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2006, pp. 1735-1742.
[21] Z.-F. Wu, S.-N. Wang, J.-T. Gu, M. Khabsa, F. Sun, and H. Ma, "Clear: Contrastive learning for sentence representation," *Arxiv Preprint ArXiv*: 2012.15466, 2020.
[22] Y. Meng, C.-Y. Xiong, P. Bajaj, P. Bennett, J.-W. Han, and X. Song, "Coco-lm: Correcting and contrasting text sequences for language model pretraining," *Advances in Neural Information Processing Systems (NeurIPS)*. MIT Press, 2021, pp. 23102-23114.
[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.-N. Gomez, ... and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems (NeurIPS)*. MIT Press, 2017, pp. 6000-6010.
[24] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *in International Conference on Machine Learning (PMLR)*. MLR, 2020, pp. 1597-1607.
[25] K. Murphy, Y. Weiss, and M.-I. Jordan, "Loopy belief propagation for approximate inference: An empirical study," *Arxiv Preprint ArXiv*: 1301.6725, 2013.
[26] J. Hoffart, M.-A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, ... and G. Weikum, "Robust disambiguation of named entities in text," *in Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2011, pp. 782-792.
[27] Z.-C. Guo, and D. Barbosa, "Robust named entity disambiguation with random walks," *Semantic Web*, vol. 9, no. 4, pp. 459-479, 2018.
[28] https://huggingface.co/bert-base-uncased
[29] I. Loshchilov, and F. Hutter, "Decoupled weight decay regularization," *Arxiv Preprint ArXiv*: 1711.05101, 2017.
[30] A. Globerson, N. Lazic, S. Chakrabarti, A. Subramanya, M. Ringaard, and F. Pereira, "Collective entity resolution with multi-focal attention," *in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL, 2016, pp. 621-631.
[31] I. Yamada, H. Shindo, H. Takeda, and Y. Takefuji, "Joint learning of the embedding of words and entities for named entity disambiguation," *Arxiv Preprint ArXiv*: 1601.01343, 2016.
[32] X. Cheng, and D. Roth, "Relational inference for wikification," *in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2013, pp. 1787-1796.