Knowledge-enhanced Representation based-on Contrastive Learning and Informative Entities

Pingchuang Ma[1], Jianhua Miao[2*], Chunyang Ruan[3]

[1]Department of information and intelligent engineering, Shanghai Publishing and Printing College, Shanghai, 200093, China
[2]Group of Computer Shanghai Caoyang Vocational School, Shanghai, 200333, China
[3]Algorithm, Shanghai Enflame Technology Company, Shanghai, 201306
**Corresponding author: Jianhua Miao: 62163346@163.com, luckycat336@163.com*

Abstract

In the field of NLP, sentence representation model is a popular task. The emergence of pre-trained representation model based on transformer structure yields significant results for downstream tasks. Besides, since the introduction of contrastive learning based on the pre-train representation model two years ago, there has been great interest in its notable benefits. For the sake of achieving better training results for sentence vector representation, we propose to use a training framework of contrastive learning and bring about knowledge graph information to improve language representation. This method enables the model to learn more linguistic information in sentence presentation and will improve the effect of sentences in tasks like semantic matching and classification.

Keywords: *sentence representation; contrastive learning; knowledge enhancement; Transformer; knowledge graph*

1 INTRODUCTION

In recent years, sentence vector representation has become a popular topic in the NLP field. Through self-supervised learning, the sentence vector representation model mines the ability to learn general information from massive unable texts and then achieves remarkable results in some tasks by supervised training of a small amount of data in downstream tasks. For example, the pre-trained language models using large-scale data trained by BERT [5] are widely used in industry and academia, especially in tasks such as semantic textual similarity and dense text retrieval. The pre-trained model calculates the encoding representation of two sentences and uses spatial distance similarity to measure whether the two sentences are semantically related. It performs well.

However, although BERT-based models have shown good performance on many NLP tasks, the model has their anisotropy problem in which the derived sentence vectors occupy a narrow conical space in the vector space. This problem causes most sentences to have high similarity scores when the sentence vector output by BERT is employed for similarity calculation, even those that are semantically irrelevant. That means the representation of the sentence has "collapse" [3] [7] [10] phenomena.

For the problem of vector representation anisotropy, some mainstream solutions, such as BERT-flow [3] and BERT-whitening [6], apply linear transformations of vector-matrix space to alleviate this problem. In addition, there are also some methods to solve the problems through contrastive learning and training---utilizing operations such as translation, insertion, deletion, sequencing, etc., to construct positive and negative samples. This model narrows the distance between positive sample pairs and widens those between negative sample pairs to achieve a uniform effect of the sentence vector in the spatial distribution. Furthermore, in 2021, Tianyu Gao et al. proposed the SimCSE [15] framework, which adopted dropout [13] in unsupervised contrastive learning to solve the problem of vector anisotropy, and refreshed the SOTA results of supervised and unsupervised semantic similarity. Moreover, the unsupervised SimCSE model far outperforms all supervised models, including SBERT [12] in the STS task.

Despite its simplicity and effectiveness, using dropout as a minimal method of “data argumentation” in

B. Fox et al. (Eds.): IC-ICAIE 2022, AHCS 9, pp. 1280-1286, 2023.
https://doi.org/10.2991/978-94-6463-040-4_190

SimCSE has its shortcomings. In unsupervised training, the same sentence is used to obtain two vectors as positive samples via a pre-trained language model, and the sentence vectors derived from the same sentence are of the same length. During the learning process, the model tends to consider that sentences of identical lengths are semantically similar. Based on SimCSE, the author then proposed the ESimCSE [18] model, which has made several improvements: using batch processing to perform word-repetition operations, so that the positive pairs no longer have consistent lengths; and using momentum contrast in the selection of the negative sample, so that fuller comparison for negative samples can be achieved. Even though these improvements can make up for the problem of constructing false positive and negative samples in contrastive learning training to some extent, the noise will still be introduced into the construction of positive and negative samples by the simple word repetition. This paper proposes a positive sample construction method for knowledge graph data enhancement based on the framework of SimCSE. By adding some entity-associated knowledge graph information to the sentence, it introduces additional prior knowledge information without changing the semantics of the sentence. It can not only solve the problem of identical sentence lengths in the construction of positive samples, but also enable the construction of positive samples to obtain richer semantic features. Furthermore, this paper also offers a method of using knowledge graph retrieval to assist in the retrieval of negative samples, so that when training the model, the selected negative samples are more similar in literal or entity-relationship, but not in semantics.

2 BACKGROUND

2.1 Contrastive Learning

Contrastive learning is a method of self-supervised learning which is popular in the CV field first. Facebook and Google successively proposed the contrastive learning framework of MoCo [8] and SimCLR [16]. Moreover, contrastive learning's application in NLP is inspired by frameworks like SimCLR. The idea behind contrastive learning is to pull similar samples closer and dissimilar samples further, with the goal of learning a good semantic representation space from the samples [14]. The definition of contrastive learning is as follows. Assuming a dataset containing similar pairs is shown as the equation (1):

$$\mathcal{D} = \{(\mathcal{X}_i\ , \mathcal{X}_i^+)\}_{i=1}^{m} \tag{1}$$

Cross-entropy is used as the loss function [11] [17], the objective of the contrastive learning is shown as the equation (2):

$$\ell = \log \frac{e^{\mathrm{sim}\left(h_i\ , h_i^+\right)/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}\left(h_i\ , h_j^+\right)/\tau}} \tag{2}$$

where τ is the temperature hyperparameter coefficient, $\mathrm{sim}(h_1\ , h_2)$ is defined as the similarity of two vectors, and is measured by the spatial distance. Equation (3) is the distance measurement method:

$$L\ =\ \frac{h_1^T h_2}{\|h_1\| \cdot \|h_2\|} \tag{3}$$

The vector representation of the sentences in this paper is derived from an open-source pre-trained model.

2.1.1 Positive Sample Construction

The most critical issue in contrastive learning is the construction of the positive sample $(\mathcal{X}_i\ , \mathcal{X}_i^+)$. In the CV field, the positive samples are usually constructed by cropping, flipping, twisting, and rotating the images [2][4][9]. In the NLP field, some research adopts deletion, synonym replacement, mutual translation, etc. [19] to realize this target. Akbar Karimi et al. proposed a method of randomly inserting punctuation marks into the original text [1] in 2021, which avoids the problem of adding noise in the traditional way and is relatively simple to implement. In the past two years, a contrastive learning framework for NLP tasks has emerged. The representative frameworks such as ConSERT [20] and SimCSE directly use the sentence matrix that have been pre-trained in the pre-trained representation model and random dropout to achieve the goal of data argumentation.

2.1.2 Alignment and Uniformity

Since the representation of contrastive learning generally uses regularization, the vector representation is concentrated on a hypersphere. A good representation space should meet two conditions: one is alignment, which means that the representation of similar samples should be as close as possible; and the other is uniformity, which means that the representations of non-similar samples should be distributed on the hypersphere evenly. If a representation space meets such a condition, it is linearly separable. A linear is enough for classification.

2.2 SimCSE Framework

The SimCSE framework mainly proposes the use of self-supervised contrastive learning to enhance the sentence representation ability. Two methods of contrastive learning training are given: unsupervised contrastive learning training and supervised training. The former uses the sentence itself as a positive sample and non-sentence as a negative sample and then obtains two different embedding representations of the same sentence with the same semantics by random dropout.

This paper is improved on the basis of the framework of the SimCSE. Given a pair of sentences $\{\mathcal{X}_i, \mathcal{X}_i^+\}$, which will be treated as a positive sample pair. $\mathcal{X}_i$ and $\mathcal{X}_i^+$ are semantically related. The central idea of the unsupervised SimCSE is to use the same sentence in constructing positive pairs, i.e., $\mathcal{X}_i^+ = \mathcal{X}_i$. A dropout mask is placed between the Feed Forword Network fully connected layer and Multi-head Attention in the Transformer construe of BERT. By applying different dropout masks z_i and z_i^+, the same input $\mathcal{X}_i$ is input twice into the Transformer layer while two separate sentence embedding h_i and h_i^+ ($h_i^+$for positive samples) are output to construct a positive pair. The construction process is shown as the equation (4):

$$h_i = f_\theta(\mathcal{X}_i, z_i), h_i^+ = f_\theta(\mathcal{X}_i, z_i^+) \qquad (4)$$

3 KNOWLEDGE-ENHANCED CONTRASTIVE LEARNING

In the SimCSE-based contrastive learning framework, even though using dropout as the minimum data argumentation is a simple and effective way, the characteristics of the pre-trained language model like BERT will cause deviation of the results. Such model is based on the Transformer structure that the length information of the sentence is encoded by the position representation. As a result, pairs of positive samples generated by BERT from the same sentence will have the same length, while pairs of negative samples from different sentences will typically contain different lengths. This makes the positive pairs and negative pairs contains different length information, which can be used as a feature to distinguish the positive one and the negative one. Specifically, such differences can lead to a bias in model training, allowing the model to automatically consider two sentences of the same or similar length to be more semantically similar. To alleviate this problem, the author proposed a simple but effective method for sample engagement that adopting random insertion and deletion for each positive sample pair to change the length of the sentences. Whereas, although this can change the sentence length, randomly inserting words in sentences may introduce additional noise. Besides, removing keywords from a sentence can also cause a drastic change in its semantics. For example, “The author of *The Dream of the Red Chamber* is Xueqin Cao”, operations such as random insertion, deletion and repetition are adopted, as shown below (Table 1):

Table 1: Insertion, deletion, duplication similarity score

Method	Text	Whether the semantics are similar
Original sentence	The author of *The Dream of the Red Chamber* is Xueqin Cao.	--
Random insertion	Xueqin Cao is the author of *The Dream of the Red Chamber.*	YES
Random deletion	The author of *The Dream of the Red Chamber* is ~~Xueqin Cao.~~	NO
Random word repetition	The author of *The Dream Dream of the Red Red Chamber* is Xueqin Cao.	NO

From the example above, it is evident that operations such as insertion, deletion and repetition may bring a certain degree of noise. Therefore, based on this framework, this paper proposes a method that utilizing knowledge graphs as “word annotation” to resolve the problem different length information of positive and negative samples. At the same time, adding additional knowledge graph information to enhance sentence semantics.

3.1 Positive Sample Pair Construction Method

In terms of positive sample construction, entity recognition and entity linking will be carried out for each query. The linked entity attribute information is annotated to the entire sentence for the entity word. As in the above instance, the “Word annotation” of the entity is shown in Figure 1:

Original sentence: The author of *The Dreame of the Red Chamber* is Xueqin Cao.

Word annotation: The author of *The Dreame of the Red Chamber* is Xueqin Cao. (Xueqin Cao Literature the Dreame of the Red Chamber)

Figure 1: Entity “word annotation”

In the training process, a batch of the positive sample construction process will query the entity knowledge base and will make “word annotation” for the linked entity. The sentences without links can be left out. The whole training process is shown in the Figure 2:

3.2 Negative Sample Pair Construction Method

In addition to the above construction methods for positive pairs, the selection of negatives is also crucial in contrastive learning. Theoretically, the more negative samples, the better compared with the previous one. One way to increase the amount of the negative samples is to expand the Batch size of training. However, the setting of the Batch size will degrade the performance of the model. Thus, how to select effective negative sample pairs is also critical in contrastive learning. In this paper, we still use the prior knowledge of the knowledge graph to filter negative samples. The premise of the method is that the model is better trained after the negative samples with high literal coincidence and different semantics with the positive samples are selected. As a result, when selecting a negative sample, we will first select N link results $\mathrm{Query_E} = \mathrm{KG}\,(\mathrm{E_1}\,,\mathrm{E_2}\,,\mathrm{E_3}\ldots\mathrm{E_n})$ related to the entity from the results of the positive sample entity-linking, and then take the N linked results to calculate entity correlation one by one. The query with a high correlation is reserved as a negative sample. In this way, negative samples that are literally coincident with the positive samples and semantically different are constructed. The update process of negative samples in the overall training is shown in Figure 2.

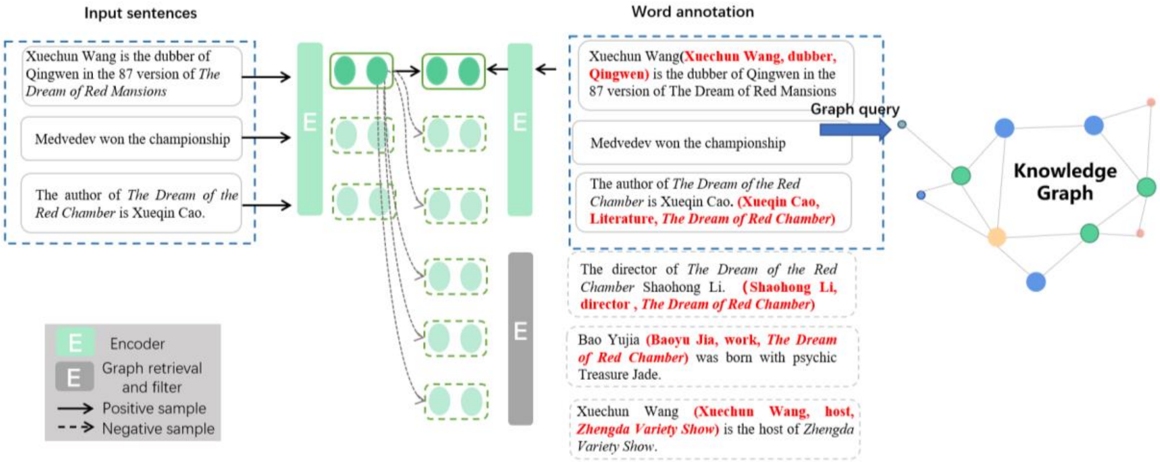
Input sentences
Xuechun Wang is the dubber of Qingwen in the 87 version of The Dream of Red Mansions
Medvedev won the championship
The author of The Dream of the Red Chamber is Xueqin Cao.
E
E
Word annotation
Xuechun Wang(Xuechun Wang, dubber, Qingwen) is the dubber of Qingwen in the 87 version of The Dream of Red Mansions
Medvedev won the championship
The author of The Dream of the Red Chamber is Xueqin Cao. (Xueqin Cao, Literature, The Dream of Red Chamber)
Graph query
Knowledge Graph
The director of The Dream of the Red Chamber Shaohong Li. (Shaohong Li, director , The Dream of Red Chamber)
Bao Yujia (Baoyu Jia, work, The Dream of Red Chamber) was born with psychic Treasure Jade.
Xuechun Wang (Xuechun Wang, host, Zhengda Variety Show) is the host of Zhengda Variety Show.
E
E Encoder
E Graph retrieval and filter
→ Positive sample
--> Negative sample

Figure 2 Knowledge-enhanced contrastive learning for positive example construction and negative example retrieval

4 EXPERIMENTAL RESULTS AND ANALYSIS

This paper uses the knowledge-enhanced contrastive learning method to train the representation model of the sentence. On the entity-linking task, take the trained representation model to make the category prediction of the entity mention and verify the task of matching the link in the knowledge base and entity mention.

4.1 Experimental Data

This paper uses the open source dataset of the entity-linking of the Chinese short text. The dataset is extracted from natural web page tags, multi-modal titles, and search queries. And the data annotation is completed by manual crowd-sourcing, of which the accuracy rate is 95.32%. The short text contains the category of each entity mention and the mapped id in the knowledge base. The amount of the category is 21. The knowledge base comes from the Baidu Encyclopedia knowledge base. Table 2 shows the relevant properties of the dataset.

Table 2 Chinese short text entity-linking task dataset

Dataset	Training set	Validation set	Test set
Chinese short text entity-linking	7W	1W	1W

4.2 Experimental Setup

The experiment uses the pre-trained model ERNIE1.0 [21] to obtain the embedding representation of the sentence's cold start. And the whole experiment is divided into two parts: basic experiment and knowledge-enhanced contrastive learning experiment.

In the basic experiment part, we first use the pre-trained model to represent the prediction process on the entity-linking task of the model. On this basis, a step-by-step training classification model is adopted for those inaccurately determined categories. In the step-by-step training classification model, above all, 2 epochs of training are performed on the entities whose prediction results are non-category in the training set, and then 5 epochs of training are performed on the data of the classified entities until the model reaches the optimum. The position features introducing model mainly aim at the problem of the incorrect mapping of the entity knowledge base. Based on the baseline model, the output layer directly splices the text with the description of the candidate entity, and adds "#" to mark the position of the entity. After the sentence input layer is encoded, the correlation degree will be calculated directly through the fully connected layer and tanh. The main purpose is to emphasize the role of entities, use the positional encoding of entities as features, and then splice it with the sentence encoding before going through the fully connected layer and tanh, and introducing the results as features into the correlation calculation.

In the knowledge-enhanced contrastive learning experiment, the overall framework adopts the contrastive learning framework of SimCSE. Firstly, take the query and entity mention information in the training set of the entity-linking task for the knowledge-enhanced contrastive learning training. After the training, the representation model is used to fine-tune the entity-mapping task and the category determination of the entity-linking task, to verify the effect of the representation model trained by the knowledge-enhanced contrastive learning on the classification and matching tasks.

In terms of experiment training settings, the Batch Size is set to 64, the learning rate is set to 0.00005, and the model is saved every 1000 steps. And models will be selected in 3w steps of the final training. The results of the experiment are shown in Table 3.

It can be seen in the experiment that after the introduction of the contrastive learning, the general F1 value is improved in the entity-mapping and the category determination of the entity-linking compared with the baseline representation model. Besides, after the introduction of knowledge-enhanced contrastive learning, the overall F1 value is largely improved compared to the baseline representation model. Moreover, it can be seen from the experimental results that after the introduction of knowledge-enhanced contrastive learning, the category determination of entity mention, or the classification model, is not significantly improved its performance, compared with knowledge-enhanced contrastive learning. But in the knowledge base mapping of entity mention, the improvement is noticeable. These results also verify that using contrastive learning to learn a representation model better can bring a more significant improvement to the representation and discrimination of sentence semantics.

Table 3 Experimental results

Scheme	Dev F1	Dev Precision	Dev recall	Test F1
ERNIE1.0	0.802	0.802	0.802	--
ERNIE1.0 + Category multi-step training	0.864	0.864	0.864	0.873
ERNIE1.0 + Position encoding feature + Category multi-step training	0.874	0.874	0.874	0.878
ERNIE1.0 + contrastive learning	0.833	0.833	0.833	0.853
ERNIE1.0 + Knowledge-enhanced contrastive learning	0.844	0.844	0.844	0.861
ERNIE1.0 + Knowledge-enhanced contrastive learning + Category multi-step training	0.880	0.880	0.880	0.879
ERNIE1.0 + Knowledge-enhanced contrastive learning + Position encoding feature + Category multi-step training	0.882	0.882	0.882	0.881

5 CONCLUSIONS

This paper focuses on the collapse phenomenon in the representation of the sentences of the BERT-based pre-trained representation model and improves on the representative contrastive learning framework SimCSE. The purpose is to achieve the goal of knowledge enhancement by introducing the entity knowledge in the knowledge graph in contrastive learning, so as to enhance the semantic representative ability of the learned sentence representation.

This paper respectively verified that by introducing the graph entity information for knowledge-enhanced contrastive learning, the learning ability can be enhanced, the sentences representing semantic information can be enriched, and the classification and matching tasks can be improved to a certain extent, through the effect verification on the short text entity-linking task, discrimination of the entity type in the entity-linking task, and the matching mapping task the entity knowledge base.

REFERENCES

[1] Akbar Karimi, Leonardo Rossi, and Andrea Prati. AEDA: An Easier Data Augmentation Technique for Text Classification[C]//In Findings of the Association for Computational Linguistics: EMNLP 2021:pages 2748–2754.

[2] Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning representations by maximizing mutual information across views[J]//In Advances in Neural Information Processing Systems, pp. 15509–15519, 2019.

[3] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models[C] //In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020: pages 9119–9130, Online. Association for Computational Linguistics.

[4] Hénaff, O. J., Razavi, A., Doersch, C., Eslami, S., and Oord, A. v. d. Data-efficient image recognition with contrastive predictive coding[J]//arXiv preprint arXiv:1905.09272, 2019.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding [C] //In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019: pages 4171–4186.

[6] Jianlin Su, Jiarun Cao, Weijie Liu, Yangyiwen Ou. Whitening Sentence Representations for Better Semantics and Faster Retrieval[J]//arXiv preprint arXiv: 2103.15316, 2021.

[7] Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and TieYan Liu. Representation degeneration problem in training natural language generation models [J]//arXiv preprint arXiv:1907.1,2019.

[8] Kaiming He and Haoqi Fan and Yuxin Wu and Saining Xie and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning [J]//arXiv preprint arXiv:1911.05722, 2019.

[9] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks[J]//In Advances in neural information processing systems, 2012: pp. 1097–1105.

[10] Lingxiao Wang, Jing Huang, Kevin Huang, Ziniu Hu, Guangtao Wang, and Quanquan Gu. Improving neural language generation with spectrum control [C]//In International Conference on Learning Representations.2019.

[11] Matthew Henderson, Rami Al-Rfou, Brian Strope, YunHsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. Effificient natural language response suggestion for smart reply[J]// arXiv preprint arXiv:1705.00652, 2017.

[12] Nils Reimers and Iryna Gurevych. 2019. Sentencebert: Sentence embeddings using siamese bertnetworks[C]// In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019:pages 3973–3983.

[13] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfifitting[J]//The Journal of Machine Learning Research (JMLR), 15(1), 2014:1929–1958.

[14] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping[J]//In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2006: volume 2, pages 1735–1742. IEEE.

[15] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple Contrastive Learning of Sentence Embeddings[C]//In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing,2021: pages 6894–6910.

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

[17] Ting Chen, Yizhou Sun, Yue Shi, and Liangjie Hong. On sampling strategies for neural network based collaborative fifiltering[J]//In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017:pages 767–776.

[18] Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, Songlin Hu.ESimCSE: Enhanced Sample Building Method for Contrastive Learning of Unsupervised Sentence Embedding [J]//arXiv preprint arXiv:2109.04380, 2021.

[19] Yanru Qu, Dinghan Shen, Yelong Shen, Sandra Sajeev, Jiawei Han,Weizhu Chen. CoDA: Contrast-enhanced and Diversity-promoting Data Augmentation for Natural Language Understanding [J]//arXiv preprint arXiv:2010.08670, 2020.

[20] Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, Weiran Xu. ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer[J]//arXiv preprint arXiv:2105.11741, 2021.

[21] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, Hua Wu. ERNIE: Enhanced Representation through Knowledge Integration [J]//arXiv preprint arXiv:1904.09223, 2019.