Unsupervised Keyphrase Extraction Approach for Hazard Records by Contrastive Learning

Wenbin Luo Xinbo Ai Yanjun Guo Guangsheng Liu Ange Li

School of Artificial Intelligence Beijing University of Posts and Telecommunications

Abstract

In this study, we introduce a KPE method called Domain-oriented Joint Scores Rank, which evaluates and ranks candidate phrases by combining scores from multiple dimensions. Further, we have developed a security domain-oriented BERT model by proposing a lightweight comparative learning fine-tuning method. The HKBERT-based DJSRank outperforms the SOTA general domain model on the KPE task for hazard records, with an average F1 improvement of 4.3.

DJSRANK: DOMAIN-ORIENTED JOINTLY SCORES RANK

In this section, we will introduce an unsupervised KPE method called Domain-Oriented Jointly Scores Rank.The structure of DJSRank is shown in Fig. 1.

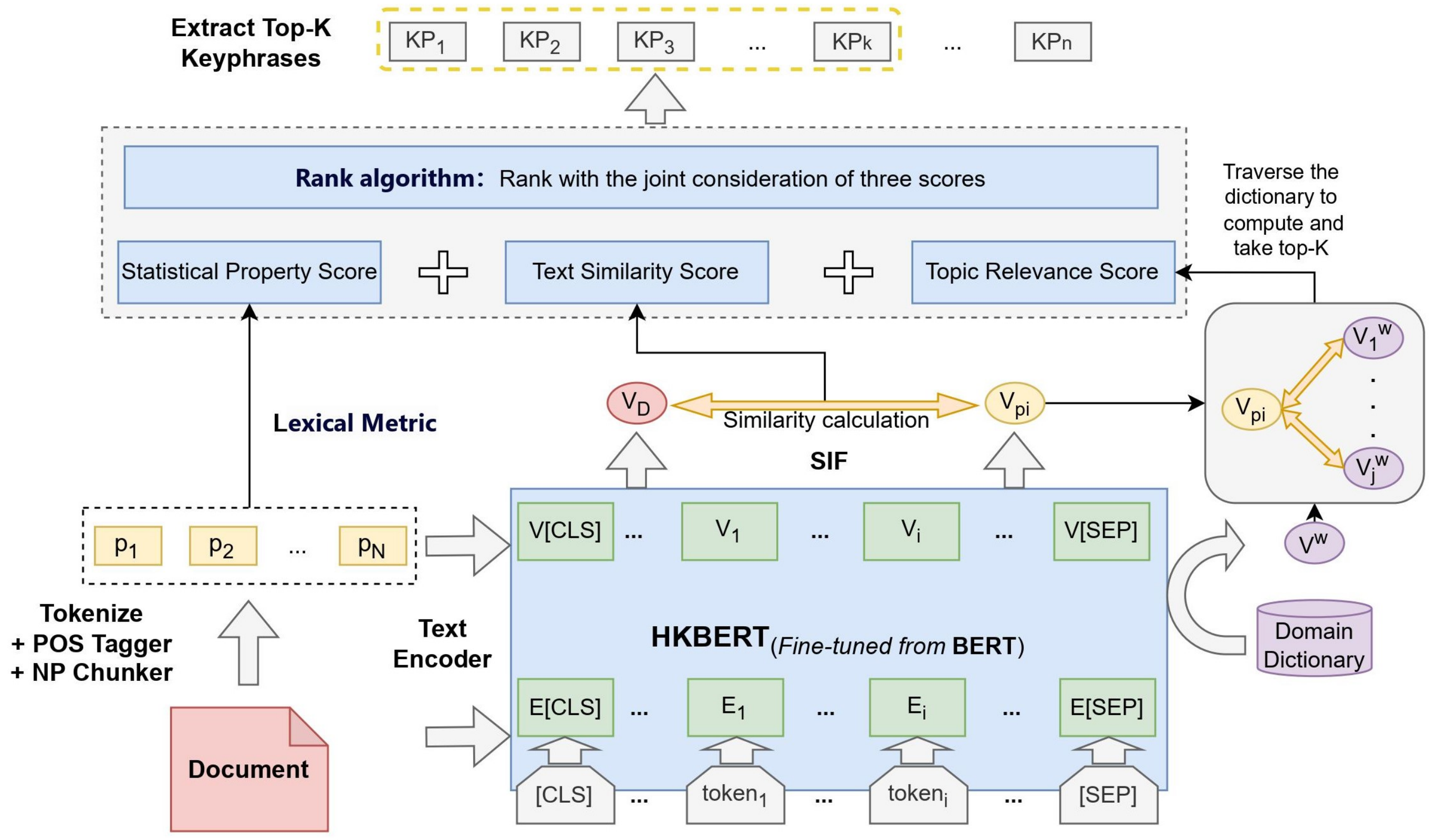

Fig. 1: DJSRank

To incorporate Text Similarity scores, Statistical Property scores , and Topic Relevance scores, we aggregate them by addition, resulting in the final scores for the candidate phrases, as shown below:

$$\text{FinalScore}\ (p_i) = \text{SimScore} + \text{StaScore} + \text{TopScore}$$

Text Similarity scores is shown that:

$$\mathrm{SimScore}\,(p_i) = \cos\left(v_D, v_{p_i}\right) = \frac{\vec{v}_D \cdot \vec{v}_{p_i}}{\|\vec{v}_D\|\,\|\vec{v}_{p_i}\|}$$

Statistical Property scores is shown that:

$$l\_\ \mathrm{score}\left(v_{p_i}\right) = \begin{cases} 0 & \text{if } L\left(v_{p_i}\right) = 0 \\ \log_{10}\left(L\left(v_{p_i}\right)\right) & \text{if } L\left(v_{p_i}\right) \neq 0 \end{cases}$$
$$L\left(v_{p_i}\right) = \min\left[0, \left|\mathrm{len}\left(p_i\right) - \sup(T)\right|, \left|\mathrm{len}\left(p_i\right) - \inf(T)\right|\right]$$
$$t\_score\left(v_{p_i}\right) = \log_{10}\left(TF\left(v_{p_i}, D\right)\right)$$
$$StaScore\left(v_{p_i}\right) = -\alpha \cdot l\_score\left(v_{p_i}\right) + \beta \cdot t\_score\left(v_{p_i}\right)$$

Topic Relevance scores is shown that:

$$TopScore(p_i) = \gamma\cos\left(v_H, v_{p_i}\right) + \delta\frac{\sum_{j=1}^{U}\cos\left(v_j^W, v_{p_i}\right)}{U}$$

HKBERTDOMAIN-ORIENTED CONTRASTIVE LEARNING

According to the analysis, the BERT-based model has issues with vector representation. Therefore, when calculating the similarity at the sentence level, if both sentences are composed of high-frequency words, the similarity may be very high. Conversely, if both sentences are composed of low-frequency words, the similarity obtained may be relatively low. This leads to interference from common words in hazard records KPE tasks. This paper proposes a fine-tuning method using contrastive learning to train a new PLM, Hazard-oriented KPE BERT (HKBERT). This new fine-tuning method is called Domain-oriented Contrastive Learning, as shown in Fig. 2. .

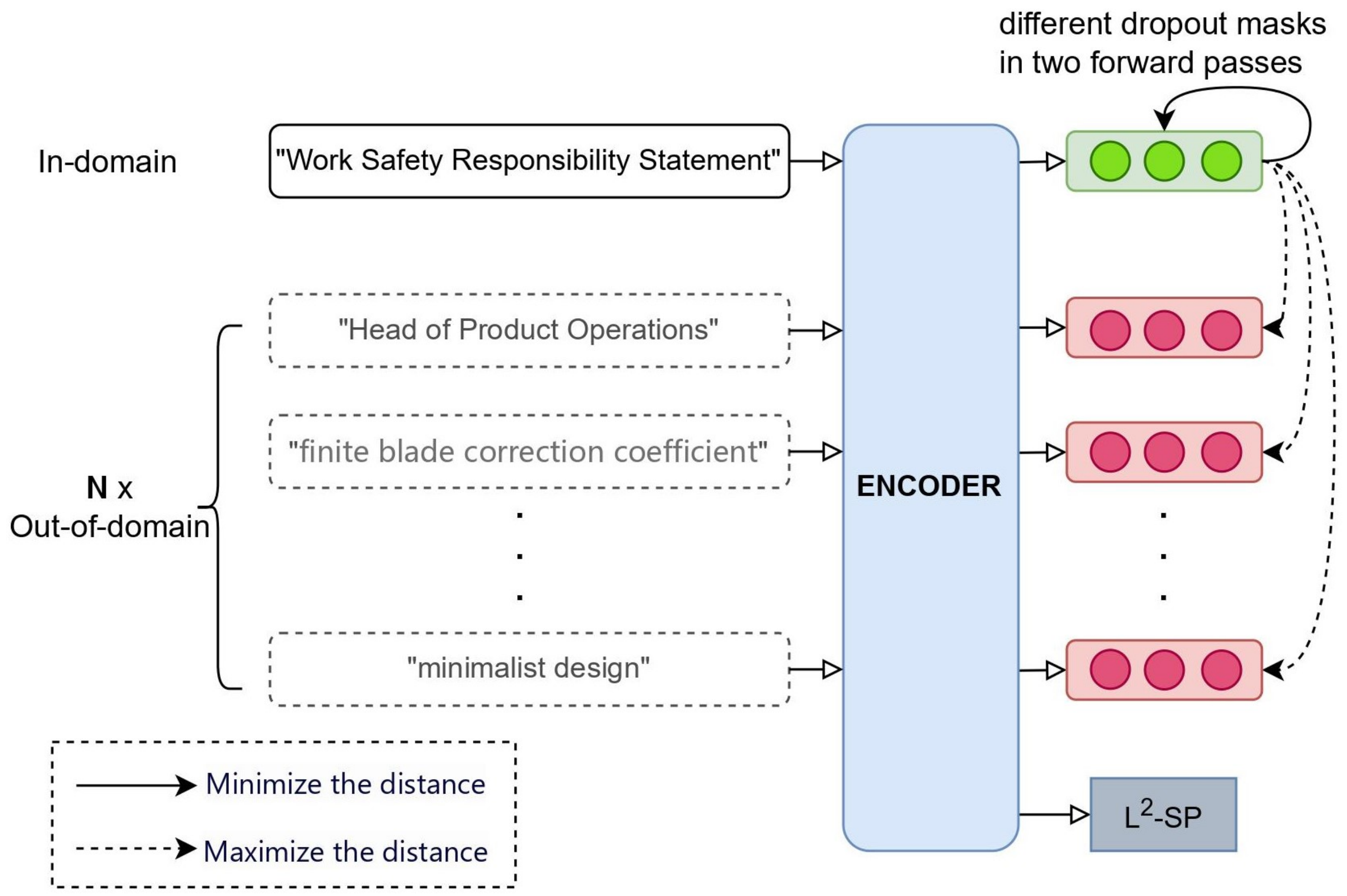
different dropout masks in two forward passes
In-domain
"Work Safety Responsibility Statement"
N x Out-of-domain
"Head of Product Operations"
"finite blade correction coefficient"
"minimalist design"
ENCODER
Minimize the distance
Maximize the distance
L2-SP

Fig. 2: HKBERT

We adopt the SimCSE method to create a positive example by adjusting the dropout rate of the transformers framework (as set in this paper). We input a specialized domain text into the PLM twice to generate two different embeddings for anchor and positive examples. For creating negative examples, we use embeddings from text in general domains. Within a mini-batch, one text from a specialized domain is included, while the remaining data comprises texts from general domains. We define the ternary loss as follows:

$$TopScore(p_i) = \gamma\cos\left(v_H, v_{p_i}\right) + \delta\frac{\sum_{j=1}^{U}\cos\left(v_j^W, v_{p_i}\right)}{U}$$

We also incorporate the $L^2 - SP$ regularization loss to prevent catastrophic forgetting, negative transfer, and overfitting issues that may occur during model fine-tuning. The regularization loss is defined as follows:

$$\ell_{L^2-SP}(\omega) = \frac{\alpha}{2}\left\|\omega_s - \omega_s^0\right\|_2^2$$

Therefore, final training loss is set as follows:

$$\ell = \ell_{tri} + \lambda\ell_{L^2-SP}$$

EXPERIMENTS

We present the results of a comparative experiment in Table1. Where (B) denotes that the method uses the BERT model as the embedding model, and similarly, (D) denotes that the Doc2vec model is used as the embedding model, and (H) denotes that the HKBERT model is used as the embedding model. We observe that the BERT model performs best when the ranking models are consistent. We believe this is because the BERT model is more capable of learning semantic information. When all the embedded models are BERT models, EmbedRank(BERT) performs poorly. This is similar to the performance of the general domains. The MDERank method is slightly less effective than the SIFRank (BERT) method, possibly because hazard records are generally short, while MDERank performs better on long texts. Finally, our model shows a more significant performance improvement for evaluation phrases N=3, 7, and 10 compared to the general domain model. The improvement is more noticeable at N=3 than at N=10.

TABLE I. Comparative Experiment

Method	N=3			N=5			N=10		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
EmbedRank(B)	59.20	38.85	46.92	51.97	62.68	56.83	48.68	72.75	58.33
EmbedRank(D)	51.33	33.68	40.67	46.62	56.22	50.97	44.79	66.93	53.66
SIFRank	43.77	29.22	35.04	39.56	49.72	44.06	37.68	59.55	46.15
SIFRank+	41.80	27.91	33.47	38.40	48.26	42.77	36.68	57.97	44.93
SIFRank(B)	**66.66**	**43.77**	**52.84**	**55.01**	**66.39**	**60.17**	**51.05**	**76.36**	**61.19**
MDERank	65.14	43.12	51.89	53.88	66.37	59.48	49.31	76.07	59.83
DJSRank(H)	**72.65**	**47.71**	**57.59**	**59.38**	**71.65**	**64.94**	**54.46**	**81.46**	**65.27**

Fig. 3: HKBERT

CONCLUSION

We introduce DJSRank, a novel multi-score joint ranking algorithm designed for unsupervised KPE in security, surpassing the state-of-the-art method. We propose a domain-specific fine-tuning approach using contrastive learning, introduce Hazard-oriented KPE BERT (HKBERT), and enhance DJSRank performance in hazard records KPE task with HKBERT. Further research is needed to investigate remaining issues. Improving the domain dictionary's quality will be a long-term effort. This process involves improving dictionary coverage and domain specificity, and validating the impact of enhancing dictionary quality on the Topic Relevance Score. And future research should consider the importance of local information for KPE tasks. Integrating the assessment of local and global information poses a challenging task.