An Entity Relationship Extraction Model Based on BERT-BLSTM-CRF Algorithm for Cosmetics Domain

Authors: Min ZUO, Wei GE, Wei DONG, Qingchuan ZHANG
Adress: Beijing Technology and Business University, Beijing, China

Introduction

In view of the lack of timeliness in the regulation of the cosmetics field, we establish a model for extracting the relationship between entities in the cosmetics field, assist the relevant departments in building an intelligent regulatory system, and rely on data to achieve scientific decision-making and effective regulation. In this paper, we use the BERT (Bidirectional Encoder Representations from Transformers) network model to train the word vectors, and combine the BLSTM (bidirectional long short-term memory) network with the fused positional attention mechanism to perform entity relationship extraction, and the extracted relationship features extracted by the BERT neural network are incorporated into the word dimension text vector, and then the entity pairs are extracted by the BLSTM with the fused positional attention mechanism, and finally the predicted labels are decoded by CRF (Conditional Random Field). The experimental results show that the BERT-BLSTM-CRF-based entity relationship extraction model in cosmetics domain constructed in this paper exhibits good feature extraction and classification performance.

BERT-BLSTM-CRF model elaboration

In this paper, a Bert-BLSTM-CRF information extraction model is used, and the model generates word vectors by pre-training the BERT model on the character features of the cosmetic domain corpus.
The model structure is shown in Figure 1.

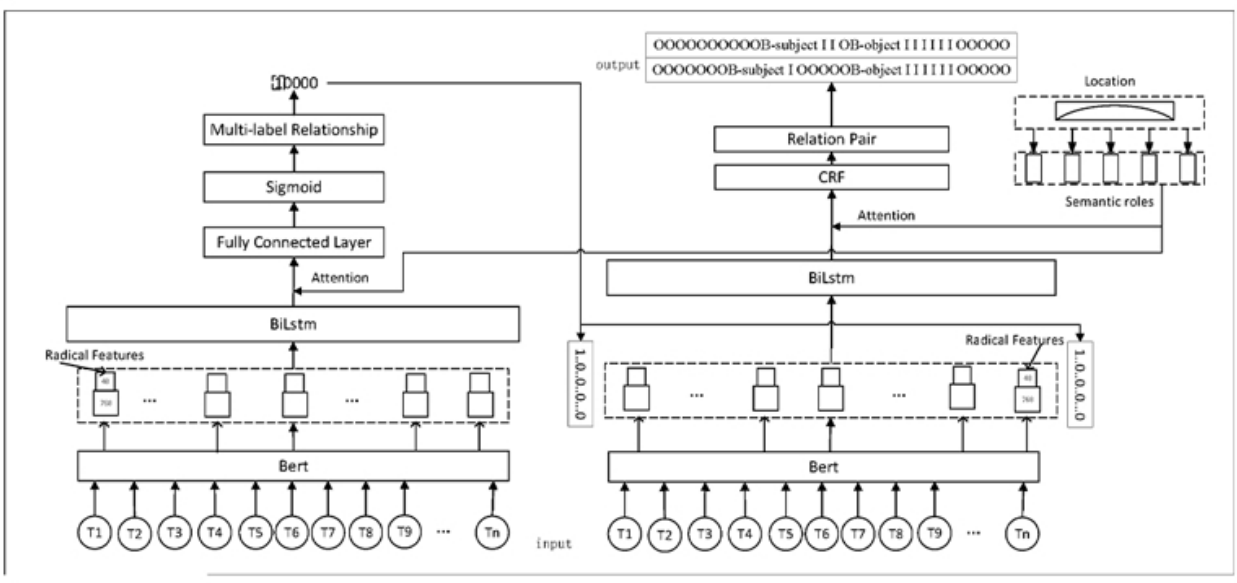

Figure 1. Structure of BERT-BLSTM-CRF model

The word dimensional text vectorized representation of the fused Chinese radical features obtained in the lower side BERT network in the network model structure diagram first.

In the left half of the model structure diagram, on the basis of the cosmetic domain lexicon, the model first determines the domain keyword semantic roles and locations to generate location-aware attention-based attention. Then the multi-categorization relations are computed by a BLSTM network incorporating the location-aware attention mechanism.

In the right half of the model structure diagram, the multiclassification relationship results are added to the text vector of the BERT neural network on the upper side, and again after BLSTM computation by fusing the location-aware attention mechanism, the best output information tagging sequence is finally obtained by decoding the predicted tags through CRF, and

Experimental results and analysis

In this paper, the dataset was obtained from the literature on the Internet and the Baidu encyclopaedia, and the relationship pairs were labelled according to five types of relationships, and a total of 5320 pairs were obtained.
The data set is divided into training set, validation set and test set according to the division. For the part of relevant indexes used in the experiments, accuracy, recall and F-value are used as the criteria for judging the experimental results in this paper.
In terms of parameter settings for the experiments, BERT uses a 12-layer 768-dimensional hidden layer and the BLSTM model uses a 128-dimensional hidden layer dimension.

Table 1. Comparison of the experimental results

Dataset	Precision	Recall	F1
CNN-ATT	0.8416	0.8378	0.8397
BLSTM	0.8576	0.8495	0.8535
BLSTM-ATT	0.8962	0.9048	0.9005
BERT-BLSTM-CRF	0.9187	0.9213	0.9199

As can be seen from Table 1, the traditional BLSTM network model has better precision but slightly worse recall in the corpus of the cosmetics domain. In this paper, the attention mechanism is introduced into the BLSTM model, and its precision, recall rate and F1 are improved to some extent. From the comparison of the experimental results of the third model and the first model, we can see that for the overall semantics of sentences, the CNN model with the learning mechanism introduced is still lacking compared with the BLSTM model with the learning mechanism introduced. In addition, Chinese radical features are introduced in this paper, and from the experimental results, it can be seen that the BERT-BLSTM-CRF model with the introduction of attention mechanism proposed in this paper achieves 91.99% in the F1 score, which exceeds all other baseline models.

Conclusion

In this paper, we propose an entity relationship extraction model for the cosmetics domain based on the real-world needs of cosmetics safety regulators. The model uses character embedding level feature representation and incorporates a location-aware attention mechanism in order to automatically discriminate entity relationships in cosmetic domain corpus, which in turn provides a decision reference for risk assessment in the cosmetic domain. The experiments show that the model constructed in this paper exhibits better performance on experimental data compared to the baseline neural network model.
Nevertheless, there is still much room for improvement in the model of this paper. In the future, in addition to the problems of continuing to optimise the hyperparameters of the network structure and improving the training efficiency, it is also necessary to increase the construction of the corpus and actively explore more methods that are less dependent on the corpus.

References

[1] Wang, H., Qin, K., Zakari, R.Y. et al. Deep neural network-based relation extraction: an overview. Neural Comput & Applic 34, 4781- Neural Comput & Applic 34, 4781- 4801 (2022).
[2] DAI Jianhua, PENG Ruoyao, XU Lu, JIANG Chao, ZENG Daojian, LI Yangding. A Review of Information Extraction Research Based on Deep Neural Networks[J]. Journal of Southwest Normal University (Natural Science Edition), 2022, 47(4): 1-11.
[3] Z. Geng, Y. Zhang, and Y. Han, "Joint entity and relation extraction model based on rich semantics," Neurocomputing, vol. 429, pp. 132-140, 2021.
[4] Gormley MR, Yu M, Dredze M. Improved Relation Extraction with Feature-Rich Compositional Embedding Models. arXiv 2015:1505.02419.
[5] LI X Y, ZHANG H, ZHOU X H. Chinese Clinical Named Entity Recognition with Variant Neural Structures Based on BERT Methods[J]. Journal of Biomedical Informatics, 2020, 107:

E-mail:dongwei2019@btbu.edu.cn