Bi-GRU Relation Extraction Model Based on Keyword Attention Postprint

Authors: Yuanyuan, Zhang, Yu, Chen, Shengkang, Yu, Xiaoqin Gu, Mengqiong, Song, Yu, Peng, I am ready to translate your academic paper. Please provide the full text in Simplified Chinese, and I will:

- Preserve all LaTeX commands and mathematical formulas exactly as they appear
- Keep all citation commands ([], **??**, (**??**), etc.) unchanged
- Maintain all … paragraph wrapper tags with their original IDs
- Use precise technical terminology and formal academic tone
- Translate all content completely without omission

You can paste the entire document, and I will return the English translation following your specified requirements., Qi, Liu, Yuanyuan Zhang

Date: 2022-11-28T00:00:00+00:00

Abstract

Relational extraction plays an important role in the field of natural language processing to predict semantic

relationships between entities in a sentence. Currently, most models have typically utilized the natural

language processing tools to capture high-level features with an attention mechanism to mitigate the adverse

effects of noise in sentences for the prediction results. However, in the task of relational classification, these

attention mechanisms do not take full advantage of the semantic information of some keywords which have

information on relational expressions in the sentences. Therefore, we propose a novel relation extraction model

based on the attention mechanism with keywords, named Relation Extraction Based on Keywords Attention (REKA).

In particular, the proposed model makes use of bi-directional GRU (Bi-GRU) to reduce computation, obtain the

representation of sentences , and extracts prior knowledge of entity pair without any NLP tools. Besides the calculation

of the entity-pair similarity, Keywords attention in the REKA model also utilizes a linear-chain conditional random

field (CRF) combining entity-pair features, similarity features between entity-pair features, and its hidden

vectors, to obtain the attention weight resulting from the marginal distribution of each word. Experiments

demonstrate that the proposed approach can utilize keywords incorporating relational expression semantics

in sentences without the assistance of any high-level features and achieve better performance than traditional

methods.

Full Text

Preamble

Bi-GRU Relation Extraction Model Based on Keywords Attention

Yuanyuan Zhang[1]†, Yu Chen[2], Shengkang Yu[1], Xiaoqin Gu[1], Mengqiong Song[1], Yu Peng[1], Jianxia Chen[2] & Qi Liu[2]
[1]Technical Training Center of State Grid Hubei Electric Power Co., Ltd., Wuhan 430070, China
[2]Hubei University of Technology, School of Computer Science, Wuhan 430068, China

Keywords: Relation extraction; Bi-GRU; CRF keywords attention; Hidden similarity

Citation: Zhang, Y.Y. et al.: Bi-GRU Relation Extraction Model Based on Keywords Attention. *Data Intelligence* 4(3), 552-572 (2022). DOI: 10.1162/dint_a_{00147}
Received: Oct. 11, 2021; Revised: Jan. 15, 2022; Accepted: Feb. 10, 2022

ABSTRACT

Relational extraction plays a crucial role in natural language processing by predicting semantic relationships between entities in sentences. While most current models utilize natural language processing tools to capture high-level features

with attention mechanisms to mitigate noise, these approaches do not fully exploit the semantic information carried by keywords that express relationships. To address this limitation, we propose a novel relation extraction model based on keyword-aware attention, named Relation Extraction Based on Keywords Attention (REKA).

Specifically, our model employs bidirectional GRU (Bi-GRU) to efficiently obtain sentence representations while extracting entity pair prior knowledge without relying on NLP tools. The keyword attention mechanism in REKA utilizes a linear-chain conditional random field (CRF) that combines entity-pair features, similarity features between entity-pair features, and hidden vectors to compute attention weights derived from the marginal distribution of each word. Experiments demonstrate that our approach effectively leverages keywords incorporating relational expression semantics without high-level feature engineering, achieving superior performance compared to traditional methods.

Corresponding author: Yuanyuan Zhang (E-mail: 16823650@qq.com; ORCID: 0000-0002-5353-2989)

1. INTRODUCTION

The Web generates and shares abundant data daily, with relational facts between subjects (entities) in text often used to capture associations among data. These relationships are typically represented as triples (e_1, r, e_2), indicating that entity e_1 has relation r with entity e_2. Knowledge graphs such as FreeBase [1] and DBpedia [2] exemplify this triple-based representation.

Relation extraction is an NLP sub-task that discovers relationships between entity pairs in unstructured text. Early work relied heavily on kernel and feature-based methods [3], while recent research employs data-driven deep neural networks (DNNs) to eliminate dependence on conventional NLP approaches. DNN-based methods [4–6] automatically learn features instead of requiring manually designed features from various NLP toolkits, surpassing traditional methods and achieving excellent results. Among these, supervised and distant supervision methods are the most popular solutions, with supervised methods performing better in specific domains and distant supervision excelling in generic domains.

Given this context, we focus on DNN-based supervised methods in this research. These are typically classified by architecture into CNN [6–10], RNN [5, 11, 12], or mixed structures, with RNN variants such as LSTM [13–15] and GRU [16] offering particular advantages. While CNNs efficiently process local and structural information through parallelization, they struggle to capture global features and sequential information—limitations that RNNs, LSTMs, and GRUs address through their sequence modeling capabilities.

However, RNN-based methods share a common drawback: they introduce many external artificial features without effective filtering mechanisms [17]. Semantic-

oriented approaches and attention mechanisms have been developed to improve representation by capturing internal text associations and reducing word-level noise [18–21]. Recent state-of-the-art attention-based methods [19, 22, 23] have shown particular promise.

While attention mechanisms can mitigate intra-sentence noise by independently weighting words, important information often resides in continuous phrases. Yu et al. [24] proposed a CRF-based attention mechanism that incorporates such keyword information into neural relation extractors, demonstrating its importance for constructing better attention weights compared to feature-based classifiers and baseline neural models.

Building on this analysis, we propose REKA, a novel relation extraction model with keyword-aware attention that uses Bi-GRU to reduce computational requirements without NLP tools. Our CRF attention mechanism comprises two components: entity pair attention and segment attention. Entity pair attention adds weight to entity mentions to increase their influence, while segment attention assumes each sentence has a binary state sequence where each state indicates whether a word is relevant to relation extraction (0 or 1). Using a linear-chain CRF, we obtain the marginal distribution of each state variable as an attention weight.

Our contributions are: (1) a novel Bi-GRU model with keyword attention for relation extraction; (2) incorporation of both entity pair similarity and segment features in the attention mechanism; (3) state-of-the-art performance without NLP tool assistance; and (4) improved interpretability over standard Bi-GRU models.

2.1 RNN-Based Relation Extraction Models

Recent relation extraction research focuses on extracting relational features using neural networks [25–27]. Zhang et al. [28] argued that RNN-based models outperform CNN-based approaches because CNNs only capture local features while RNNs learn long-distance dependencies between entities. LSTM [15] subsequently addressed RNN gradient explosion problems through gating mechanisms. Xu et al. [5] proposed the SDP-LSTM model, which uses LSTM along the shortest dependency path (SDP) between entities, incorporating word vectors, POS tags, grammatical relations, and WordNet hypernyms as external information. To address shallow architecture limitations, Xu et al. [29] developed methods to obtain abstract features along SDP sub-paths.

Since dependency trees are directed graphs, SPD must be divided into two sub-paths directed from each entity toward their common ancestor. However, unidirectional LSTM models cannot represent complete sequential information, prompting Zhang et al. [30] to utilize bidirectional LSTM (BiLSTM) for sentence-level representation with lexical features. Their results demonstrated

that word embeddings alone achieve excellent performance. To address SDP's lack of feature extraction capability, attention mechanisms were subsequently introduced for BiLSTM-based relation extraction [31].

2.2 Attention Mechanisms for Relation Extraction

Since useful information can appear anywhere in a sentence, researchers have developed attention-based models to capture important semantic information. Zhou et al. [31] introduced attention mechanisms in BiLSTM that automatically extract important features from raw text. Similarly, Xiao et al. [32] proposed a two-level BiLSTM architecture with a two-level attention mechanism for high-level sentence representation.

While attention mechanisms capture important features, Zhou et al. [31] used random weights without considering prior knowledge. Qin et al. [33] addressed this with EAtt-BiGRU, which leverages entity pairs as prior knowledge for attention weights, using Bi-GRU to reduce computation and capture sentence representations. Zhang et al. [34] proposed a Bi-GRU model with SDP-based attention for prior knowledge extraction, while Nguyen et al. [35] introduced dependency analysis in attention mechanisms to consider interconnections between potential features.

With BERT achieving excellent performance across NLP tasks, many studies have applied it to relation extraction [36]. However, BERT suffers from high training costs and long prediction times. Our model is inspired by Lee et al. [22] but differs by using Bi-GRU instead of BiLSTM for efficiency and by combining entity pair attention with segment attention via CRF, similar to Yu et al. [24], which learns phrase-like features for relational expressions.

While these methods provide a solid foundation, limitations remain, particularly insufficient training corpus size for supervised RE. Distant supervision methods [37–40] address this but introduce significant noise. Given these trade-offs, this paper focuses on supervised methods.

---

3. METHODOLOGY

The REKA model comprises four components, as shown in Figure 1 [Figure 1: see original paper]: (1) an input layer containing word vector and location information; (2) a self-attention layer that processes word vectors to obtain representations; (3) a Bi-GRU layer that captures contextual information for each word; and (4) a keyword-attention layer that extracts key information for the final classification layer.

3.1 Input Layer

The input layer transforms sentences into embedding vectors with rich feature information. Input sentences are denoted as $\{w_1, w_2, \ldots, w\}$, with each word's relative position features to entity pair e (where j {1,2}) represented as vectors.

To enhance semantic capture, we employ ELMo [43] for word embeddings, which better handles polysemy compared to static vectors from word2vec [41] or GloVe [42]. ELMo is a trained model that infers word vectors based on contextual sentences or paragraphs, allowing dynamic understanding of words in context. After embedding, $\{x_1, x_2, \ldots, x\}$ represents d -dimensional vectors passed to the next layer as position feature vectors.

3.2 Multi-Head Attention Layer

While ELMo provides non-fixed word vectors, we use Multi-Head Attention (MHA) to further process these outputs, helping the model understand deep semantic information and address long-term dependencies. MHA is a self-attention mechanism [17, 19] that constructs a symmetric similarity matrix from the input word vector sequence.

As shown in Figure 2 [Figure 2: see original paper], given keys K, queries Q, and values V, the multi-head attention module executes attention h times. The calculation process follows equations (1–3):

[Figure 2: see original paper] shows a sample of Multi-Head Attention [17]. The inputs Q, K, V all equal the word embedding vector $\{x_1, x_2, \ldots, x\}$ in the multi-head attention [17]. The output is a sequence of features containing contextual information about the input sentences.

3.3 Bi-GRU Network

The Bi-GRU network layer obtains semantic information from the MHA self-attentive output sequence. As shown in Figure 3 [Figure 3: see original paper], GRU optimizes LSTM by retaining only two gate operations (update and reset gates), resulting in fewer parameters and faster convergence.

For simplicity, we denote GRU unit processing of m as GRU(m). The contextualized word representation is calculated via equations (4–6):

The input M from the MHA layer is fed into the Bi-GRU network step-by-step. To simultaneously use past and future feature information at each time step, we connect the hidden state of the forward GRU network with the hidden state of the backward GRU network at each step, where d denotes the hidden state dimension and $\{h_1, h_2, \ldots, h\}$ represents each word's hidden state vector.

3.4 Keywords Attention based on CRF

While attention mechanisms achieve state-of-the-art results, most fail to fully exploit keyword information crucial for relation extraction. Our keyword attention mechanism assigns more reasonable weights to hidden layer vectors as linear combinations of scalars, with all weights between 0 and 1.

Unlike traditional attention, we define a state variable z for each word: z = 0 indicates irrelevance to relation classification, while z = 1 indicates relevance. Each input sentence has a corresponding sequence of z variables. The expected value of a hidden state N is calculated as the probability of its word being selected, shown in equation (7):

To compute p(z = 1|H), we introduce CRF to calculate weight sequences for hidden vectors H = {h_1, h_2, ..., h }, where H is the input sequence and h is the GRU hidden output for the i-th word. CRF provides transition probabilities for conditional probability computation between sequences.

The linear-chain CRF defines conditional probability p(z = 1|H) via equations (8–9):

where is the set of state sequences, Z(H) is the normalization constant, and Zc is the subset of z given by individual clique c. The potential function y(Zc, H) is defined by equation (10):

The feature extractor uses two feature functions: vertex feature function y_1(z , H) and edge feature function y_2(z , z $_1$). y_1 maps GRU output h to state variable z, while y_2 models transitions between adjacent state variables, defined in equations (11–13):

where W and W are trainable parameters and b is a trainable bias term. These calculate contextual information as feature scores using entity location features and keyword embedding vectors (entity pair hidden similarity features t_1, t_2, and entity pair features).

The CRF keyword attention mechanism performs soft selection by assigning higher weights to words more relevant for classification. Figure 4 [Figure 4: see original paper] illustrates this with the example sentence "The boy ran into the school cafeteria," where "into" receives higher weight alongside entity words "boy" and "cafeteria" due to its relational significance.

Entity position feature: Our keyword attention incorporates both word embeddings and position embeddings. To represent contextual and relative location information, we connect position features p with corresponding hidden layer outputs h , as shown in F_1 of equation (12). Positional vectors transform relative positional scalars into feature embedding vectors through an embedding matrix, where L is the maximum sentence length and d is the position vector dimension.

Entity hidden similarity features: We extract entity hidden similarity features to replace traditional entity feature extraction methods, avoiding NLP tools. The calculation is defined in equations (14–15):

Entities are categorized by hidden vector similarity, where c denotes a potential vector representing similar entity classes and K is a hyperparameter for the number of similarity-based entity classes. The j-th entity hidden similarity feature t is calculated by weighting similarity between c and hidden layer output h based on the j-th entity. Entity features are constructed by cascading hidden states at entity locations with the entity pair's potential type representation, shown as F_2 in equation (12).

3.5 Classification Layer

A softmax layer after the keyword attention layer computes the output distribution probability p, shown in equation (16):

where |R| is the number of relationship categories, b ^{|R|} is a bias term, and W maps the expected value of hidden state N to relational label feature scores.

3.6 Training

The keyword attention is trained using cross-entropy loss for relation extraction, defined in equation (17):

where |D| is the training dataset size and ($S^{( )}$, $y^{( )}$) is the i-th sample. We use the AdaDelta optimizer to minimize loss parameters. To prevent overfitting, L2 regularization is added, where λ_1 and λ_2 are regularization hyperparameters. The second regularizer encourages the model to focus on significant words, producing sparse weight distributions. The final objective function L is shown in equation (18):

4.1 Dataset and Metric

We evaluate our model on the SemEval-2010 Task 8 dataset, a benchmark widely used in relation extraction research. The dataset contains 19 relationship types, including nine directional relationships and an "Other" category, as shown in Table 1 .

The dataset includes 10,717 sentences (8,000 training samples and 2,717 test samples). Evaluation uses macro-averaged F1 score, the official metric for this dataset.

4.2 Implementation Details

We initialize word embeddings using a publicly available pre-trained ELMo model, with other weights initialized randomly from a zero-mean Gaussian distribution. Hyperparameters are listed in Table 2 , with grid search selecting regularization coefficients λ_1 and λ_2 from 0 to 0.2.

4.3 Comparison Models

We compare REKA against several benchmark models:

- **SVM** [44]: A non-neural model achieving top SemEval-2010 results using handcrafted features (WordNet, PropBank, FrameNet, etc.)
- **MV-RNN** [45]: An SDP-based model iterating along the shortest dependency path between entities
- **CNN** [4]: An end-to-end model building convolutional networks for sentence-level feature learning
- **BiLSTM** [30]: A classic RNN-based model using bidirectional LSTM for sentence representations
- **DepNN** [46]: Combines RNN for subtrees and CNN for shortest path features
- **FCM** [45]: Decomposes sentences into sub-structures, extracts features separately, then merges them
- **SDP-LSTM** [5]: Uses LSTM along the shortest dependency path with pairwise ranking loss
- **Purely self-attention** [47]: Uses only self-attentive encoding with position-aware encoders
- **CASREL BERT** [36]: A cascade binary tagging framework with BERT
- **Entity-Aware BERT** [48]: Builds on BERT with structured predictions and entity-aware self-attention

4.4 Experimental Results

For deeper evaluation, we compare REKA with RNN-based models using Precision-Recall (PR) curves and complexity analysis, shown in Figure 5 [Figure 5: see original paper] for different dataset proportions (1%, 20%, 100%).

Table 3 shows comparison results on the SemEval-2010 Task 8 test set, while Table 4 presents average precision (AP) scores compared to RNN methods.

Results demonstrate that REKA outperforms conventional models with fewer features, though it scores slightly lower than Entity-Aware BERT and CASREL

BERT. However, BERT's large pre-trained files require longer training times and higher hardware specifications.

Ablation experiments on the development dataset (Table 5) explore component contributions. Removing position embeddings decreased F1 by 0.2, while removing MHA, pre-trained ELMo embeddings, and entity hidden similarity features reduced F1 by 0.5, 1.2, and 0.8 respectively. The keyword-aware attention mechanism itself provides a 2.3% F1 improvement. These components work complementarily, achieving 84.6 F1 when combined.

5. CONCLUSION

We propose REKA, a novel Bi-GRU network with keyword attention for relation extraction on SemEval-2010. The model effectively extracts available dataset features through keyword attention, achieving 84.8 F1 without NLP tools. Our CRF keyword attention mechanism uses entity hidden vector similarities and relative position features to compute marginal distributions as attention weights. Future work will explore attention mechanisms for better key information extraction and extend the approach to multi-entity relationship identification.

ACKNOWLEDGMENTS

This work is supported by the Science and Technology Project of Hubei Electric Power Co., LTD., State Grid (149).2020

AUTHOR CONTRIBUTIONS

Yuanyuan Zhang (E-mail: 16823650@qq.com, ORCID: 0000-0002-5353-2989): Model design and manuscript writing
Yu Chen (E-mail: 1148848330@qq.com, ORCID: 0000-0001-7316-3570): Coding, experiments, analysis, manuscript writing
Shengkang Yu (E-mail: 12052033@qq.com, ORCID: 0000-0001-6374-3395): Experiments and analysis
Xiaoqin Gu (E-mail: 1564785699@qq.com, ORCID: 0000-0001-6308-8474): Experiments and analysis
Mengqiong Song (E-mail: 297365728@qq.com, ORCID: 0000-0002-2816-5670): Experiments and analysis
Yu Peng (E-mail: 1039079148@qq.com, ORCID: 0000-0002-5353-2989): Manuscript revision
Jianxia Chen (E-mail: 1607447166@qq.com, ORCID: 0000-0001-6662-1895): Model design, problem analysis, writing and revision

Qi Liu (E-mail: 260129443@qq.com, ORCID: 0000-0003-1066-898X): Writing and revision

---

REFERENCES

[1] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pp. 1247–1250 (2008, June)
[2] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z.: Dbpedia: A nucleus for a web of open data. In *The semantic web*, pp. 722–735 (2007). Springer, Berlin, Heidelberg
[3] Pawar, S., Palshikar, G.K., & Bhattacharyya, P.: Relation extraction: A survey. arXiv preprint arXiv:1712.05191 (2017)
[4] Zeng, D., Liu, K., Lai, S., Zhou, G., & Zhao, J.: Relation classification via convolutional deep neural network. In: Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers, pp. 2335–2344 (2014, August)
[5] Xu, Y., Mou, L., Li, G., Chen, Y., Peng, H., & Jin, Z.: Classifying relations via long short term memory networks along shortest dependency paths. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp. 1785–1794 (2015, September)
[6] Liu, C., Sun, W., Chao, W., & Che, W.: Convolution neural network for relation extraction. In: International conference on advanced data mining and applications, pp. 231–242 (2013, December). Springer, Berlin, Heidelberg
[7] Nguyen, T.H., & Grishman, R.: Relation extraction: Perspective from convolutional neural networks. In: Proceedings of the 1st workshop on vector space modeling for natural language processing, pp. 39–48 (2015, June)
[8] Santos, C.N.D., Xiang, B., & Zhou, B.: Classifying relations by ranking with convolutional neural networks. arXiv preprint arXiv:1504.06580 (2015)
[9] Chen, Y.: Convolutional neural network for sentence classification (Master's thesis, University of Waterloo) (2015)
[10] Kalchbrenner, N., Grefenstette, E., & Blunsom, P.: A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188 (2014)
[11] Elman, J.L. Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning* 7(2), 195–225 (1991)
[12] Zhang, D., & Wang, D.: Relation classification via recurrent neural network. arXiv preprint arXiv:1508.01006 (2015)
[13] Zhang, S., Zheng, D., Hu, X., & Yang, M.: Bidirectional long short-term memory networks for relation classification. In: Proceedings of the 29th Pacific Asia conference on language, information and computation, pp. 73–78 (2015, October)
[14] Sundermeyer, M., Schlüter, R., & Ney, H.: LSTM neural networks for language modeling. In *Thirteenth annual conference of the international speech*

communication association (2012)
[15] Hochreiter, S., & Schmidhuber, J.: Long short-term memory. *Neural computation* 9(8), 1735–1780 (1997)
[16] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
[17] Wang, H., Qin, K., Zakari, R. Y., Lu, G., & Yin, J.: Deep neural network-based relation extraction: an overview. *Neural Computing and Applications*, 1–21 (2022)
[18] Xu, Y., Mou, L., Li, G., Chen, Y., Peng, H., & Jin, Z.: Classifying relations via long short term memory networks along shortest dependency paths. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp. 1785–1794 (2015, September)
[19] Zhang, Y., Zhong, V., Chen, D., Angeli, G., & Manning, C.D.: Position-aware attention and supervised data improve slot filling. In: Conference on Empirical Methods in Natural Language Processing (2017)
[20] Zhang, Y., Qi, P., & Manning, C.D.: Graph convolution over pruned dependency trees improves relation extraction. arXiv preprint arXiv:1809.10185 (2018)
[21] Liu, T., Zhang, X., Zhou, W., & Jia, W.: Neural relation extraction via inner-sentence noise reduction and transfer learning. arXiv preprint arXiv:1808.06738 (2018)
[22] Lee, J., Seo, S., & Choi, Y.S.: Semantic relation classification via bidirectional lstm networks with entity-aware attention using latent entity typing. *Symmetry* 11(6), 785 (2019)
[23] Wang, H., Qin, K., Lu, G., Luo, G., & Liu, G.: Direction-sensitive relation extraction using bi-sdp attention model. *Knowledge-Based Systems* 198, 105928 (2020)
[24] Yu, B., Zhang, Z., Liu, T., Wang, B., Li, S., & Li, Q.: Beyond Word Attention: Using Segment Attention in Neural Relation Extraction. In: *IJCAI*, pp. 5401–5407 (2019, August)
[25] Aydar, M., Bozal, O., & Ozbay, F.: Neural relation extraction: a survey. arXiv e-prints, arXiv-2007 (2020)
[26] Socher, R., Huval, B., Manning, C.D., & Ng, A.Y.: Semantic compositionality through recursive matrix-vector spaces. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, pp. 1201–1211 (2012, July)
[27] Zeng, D., Liu, K., Chen, Y., & Zhao, J.: Distant supervision for relation extraction via piecewise convolutional neural networks. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp. 1753–1762 (2015, September)
[28] Zhang, D., & Wang, D.: Relation classification via recurrent neural network. arXiv preprint arXiv:1508.01006 (2015)
[29] Xu, Y., Jia, R., Mou, L., Li, G., Chen, Y., Lu, Y., & Jin, Z.: Improved relation classification by deep recurrent neural networks with data augmentation. arXiv preprint arXiv:1601.03651 (2016)

[30] Zhang, S., Zheng, D., Hu, X., & Yang, M.: Bidirectional long short-term memory networks for relation classification. In: Proceedings of the 29th Pacific Asia conference on language, information and computation, pp. 73–78 (2015, October)
[31] Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B.: Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers), pp. 207–212 (2016, August)
[32] Xiao, M., & Liu, C.: Semantic relation classification via hierarchical recurrent neural network with attention. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 1254–1263 (2016, December)
[33] Qin, P., Xu, W., & Guo, J.: Designing an adaptive attention mechanism for relation classification. In: 2017 International Joint Conference on Neural Networks (IJCNN), pp. 4356–4362 (2017, May). IEEE
[34] Zhang, C., Cui, C., Gao, S., Nie, X., Xu, W., Yang, L., … & Yin, Y.: Multi-gram CNN-based self-attention model for relation classification. *IEEE Access* 7, 5343–5357 (2018)
[35] Zhang, C., Cui, C., Gao, S., Nie, X., Xu, W., Yang, L., … & Yin, Y.: Multi-gram CNN-based self-attention model for relation classification. *IEEE Access* 7, 5343–5357 (2018)
[36] Wei, Z., Su, J., Wang, Y., Tian, Y., & Chang, Y.: A novel cascade binary tagging framework for relational triple extraction. arXiv preprint arXiv:1909.03227 (2019)
[37] Mintz, M., Bills, S., Snow, R., & Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 1003–1011 (2009, August)
[38] He, Y., Li, Z., Yang, Q., Chen, Z., Liu, A., Zhao, L., & Zhou, X.: End-to-end relation extraction based on bootstrapped multi-level distant supervision. *World Wide Web* 23(5), 2933–2956 (2020)
[39] Han, X., Liu, Z., & Sun, M.: Neural knowledge acquisition via mutual attention between knowledge graph and text. In: Proceedings of the AAAI Conference on Artificial Intelligence 32(1) (2018, April)
[40] Wang, G., Zhang, W., Wang, R., Zhou, Y., Chen, X., Zhang, W., … & Chen, H.: Label-free distant supervision for relation extraction via knowledge graph embedding. In: Proceedings of the 2018 conference on empirical methods in natural language processing, pp. 2246–2255 (2018)
[41] Mikolov, T., Chen, K., Corrado, G., & Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
[42] Pennington, J., Socher, R., & Manning, C.D.: Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543 (2014, October)
[43] Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak,

I., Jarkiewicz, M., & Okruszek, L.: Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research* 304, 114135 (2021)
[44] Rink, B., & Harabagiu, S.: Utd: Classifying semantic relations by combining lexical and semantic resources. In: Proceedings of the 5th international workshop on semantic evaluation, pp. 256–259 (2010, July)
[45] Socher, R., Huval, B., Manning, C.D., & Ng, A.Y.: Semantic compositionality through recursive matrix-vector spaces. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, pp. 1201–1211 (2012, July)
[46] Liu, Y., Wei, F., Li, S., Ji, H., Zhou, M., & Wang, H.: A dependency-based neural network for relation classification. arXiv preprint arXiv:1507.04646 (2015)
[47] Bilan, I., & Roth, B.: Position-aware self-attention with relative positional encodings for slot filling. arXiv preprint arXiv:1807.03052 (2018)
[48] Wang, H., Tan, M., Yu, M., Chang, S., Wang, D., Xu, K., … & Potdar, S.: Extracting multiple-relations in one-pass with pre-trained transformers. arXiv preprint arXiv:1902.01030 (2019)

AUTHOR BIOGRAPHY

Yuanyuan Zhang (1979–), male, Ph.D. from Wuhan University, associate professor and senior engineer at Technical Training Center of State Grid Hubei Electric Power Co., Ltd. Research: intelligent substation technology, intelligent power grid operation and inspection. E-mail: 16823650@qq.com, ORCID: 0000-0002-5353-2989

Yu Chen (1996–), male, graduate student at Hubei University of Technology. Research: Artificial Intelligence, NLP. E-mail: 1148848330@qq.com, ORCID: 0000-0001-7316-3570

Shengkang Yu (1993–), male, M.S. from Huazhong University of Science and Technology, lecturer and intermediate engineer at Technical Training Center of State Grid Hubei Electric Power Co., Ltd. Research: fault diagnosis of electrical equipment. E-mail: 120520338@qq.com, ORCID: 0000-0001-6374-3395

Xiaoqin Gu (1973–), female, M.S. from Hubei University, lecturer at Technical Training Center of State Grid Hubei Electric Power Co., Ltd. Research: power grid operation technology. E-mail: 1564785699@qq.com, ORCID: 0000-0001-6308-8474

Mengqiong Song (1991–), female, M.S. from Wuhan University, intermediate engineer at Technical Training Center of State Grid Hubei Electric Power Co., Ltd. Research: power grid operation technology. E-mail: 297365728@qq.com, ORCID: 0000-0002-2816-5670

Yu Peng, female, graduate of Wuhan University. Research: grid power electronics. E-mail: 1039079148@qq.com, ORCID: 0000-0002-5353-2989

Jianxia Chen is an associate professor in School of Computer Science at Hubei University of Technology. She obtained her M.S. at Huazhong University of Science & Technology and has worked as a research fellow at CCF in China and ACM in USA. Her research focuses on knowledge graphs and recommendation systems. ORCID: 0000-0001-6662-1895

Qi Liu, female, graduate student at Hubei University of Technology. Research: Artificial Intelligence, NLP. E-mail: 260129443@qq.com, ORCID: 0000-0003-1066-898X

Note: Figure translations are in progress. See original paper for figures.