UKT: A Unified Knowledgeable Tuning Framework for Chinese Information Extraction

Jiyong Zhou[1,2], Chengyu Wang[2(✉)], Junbing Yan[2,3], Jianing Wang[3], Yukang Xie[1,2], Jun Huang[2], and Ying Gao[1(✉)]

[1] South China University of Technology, Guangzhou, Guangdong, China
{csjyzhou,aukangyuxie}@mail.scut.edu.cn, gaoying@scut.edu.cn
[2] Alibaba Group, Hangzhou, Zhejiang, China
{chengyu.wcy,huangjun.hj}@alibaba-inc.com
[3] East China Normal University, Shanghai, China
51215901034@stu.ecnu.edu.cn

Abstract. Large Language Models (LLMs) have significantly improved the performance of various NLP tasks. Yet, for Chinese Information Extraction (IE), LLMs can perform poorly due to the lack of fine-grained linguistic and semantic knowledge. In this paper, we propose Unified Knowledgeable Tuning (UKT), a lightweight yet effective framework that is applicable to several recently proposed Chinese IE models based on Transformer. In UKT, both linguistic and semantic knowledge is incorporated into word representations. We further propose the relational knowledge validation technique in UKT to force model to learn the injected knowledge to increase its generalization ability. We evaluate our UKT on five public datasets related to two major Chinese IE tasks. Experiments confirm the effectiveness and universality of our approach, which achieves consistent improvement over state-of-the-art models.

Keywords: Chinese information extraction · knowledge injection · knowledge validation

1 Introduction

Recently, Large Language Models (LLMs) have significantly improved the performance of various NLP tasks [15,26]. For example, ChatGPT[1] has shown remarkable capabilities of understanding user intention and providing complete and well-organized responses. Yet, it is infeasible to apply LLMs to all NLP tasks and scenarios. Consider Information Extraction (IE), which aims to extract key information from raw texts, including tasks such as Relation Extraction (RE) and Named Entity Recognition (NER) [10,11,31]. Some examples of ChatGPT for Chinese IE are presented in Fig. 1. We can see that ChatGPT can perform

[1] https://openai.com/blog/chatgpt.

F. Liu et al. (Eds.): NLPCC 2023, LNAI 14303, pp. 210–222, 2023.
https://doi.org/10.1007/978-3-031-44696-2_17

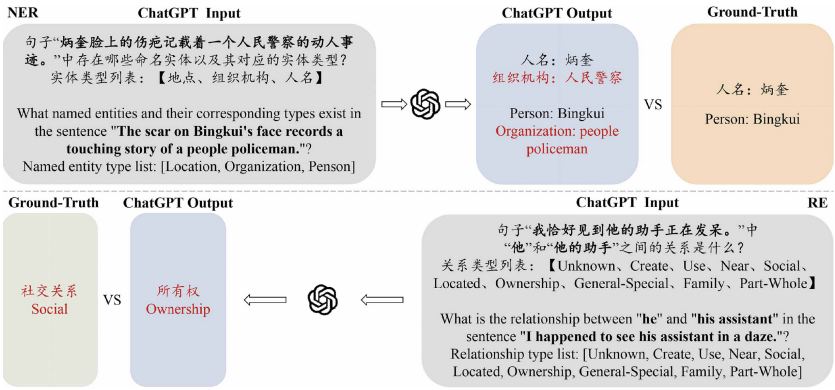


Fig. 1. Some examples with predictions generated by ChatGPT for Chinese IE. The red indicates errors. (Color figure online)

poorly, with reasons stated below: i) LLMs with decoder-only structures are more suitable for generative tasks such as machine translation and text generation, rather than text understanding and extraction [13]. ii) The training process of LLMs primarily is based on statistical characteristics of texts and pays little attention to fine-grained linguistic and semantic knowledge, which plays a vital role to improve the performance of IE [18,21]. In addition, LLMs suffer from high training and deployment costs, leading a low Return on Investment (ROI) ratio for real-world applications. Hence, it is more desirable to design task-specific moderate-size models that digest fine-grained linguistic and semantic knowledge to address Chinese IE with constrained computational resources.

In the literature, some works focus on moderate-size models for several IE tasks [7,9,10,17], which still suffer from three issues: i) Most existing approaches do not simultaneously incorporate a variety of linguistic and semantic knowledge [10,18]. ii) The majority of these approaches which incorporate linguistic or semantic knowledge, specifically structural knowledge, suffer from the lack of the distinction between different types of relationships or the lack of meaningful relationship representations, resulting in the loss of key information when the model digests the input knowledge [9,17]. iii) Previous approaches based on knowledge injection mostly focus on "exploiting" the knowledge only, without validating whether the model truly "captures" the knowledge and uses it as the guidance for IE [28,29].

In response to the aforementioned drawbacks, we propose a lightweight yet effective Chinese IE framework utilizing the Unified Knowledgeable Tuning (UKT) methodology. This framework operates in a plug-and-play fashion and can be utilized in various Chinese IE models built upon the Transformer architecture. Moreover, its coherence lies in its ability to address multiple Chinese IE tasks, including NER and RE, in a highly analogous manner. In UKT, we propose a Multi-relational, Multi-head Knowledge Fusion (MMKF) module to address the problems of missing and misrepresentation of knowledge mentioned

above, which incorporates both linguistic and semantic knowledge into representations. We further learn one-to-one meaningful relationship representations based on BERT [2] to represent different relationships between text fragments. In addition, UKT integrates a novel Relational Knowledge Validation (RKV) module to address the issue of knowledge usability. It explicitly forces the model to learn the injected relational knowledge through back propagation, making the model more knowledgeable and easier to generalize to previously unseen cases.

In experiments, we apply UKT to popular Chinese IE models (i.e., FLAT [10] and ATSSA [7]) on five public datasets related to two major Chinese IE tasks (i.e., NER and RE). Experiments confirm the effectiveness and universality of our approach, achieving consistent improvement over state-of-the-art models. In addition, we show that UKT is more capable of tackling Chinese IE, surpassing prevalent LLMs. The major contributions of this paper can be summarized as:

- We propose a lightweight yet effective framework for Chinese IE based on UKT which works in a plug-and-play fashion and is applicable to popular Chinese IE models based on the Transformer architecture. It can address several Chinese IE tasks in a highly analogous manner.[2]
- We propose MMKF in UKT to incorporate both linguistic and semantic knowledge into Chinese word representations rightly. We further propose RKV to force model to learn the injected relational knowledge.
- We apply UKT to two popular Chinese IE models on five public datasets related to two major Chinese IE tasks. Experimental results confirm the effectiveness and universality of our approach.

2 Related Work

Chinese Information Extraction. Lattice structure has received significant attention from the Chinese IE community by incorporating word information and boundary information. There have been several attempts at designing a model to optimize performance with lattice input, such as Lattice LSTM [30], MG Lattice [11] and LR-CNN [4]. However, both RNN-based and CNN-based models struggle with modeling long-distance dependencies and limited computational efficiency. Another common approach is encoding lattice graph by graph neural network (GNN), such as LGN [5] and CGN [19]. These approaches still rely on LSTMs as the bottom encoder due to the lack of sequential structure, thus increasing the complexity of the model. Recently, Transformer [20] has shown promising results in many downstream NLP tasks, owing to its good parallelism and ability to capture long-distance dependencies. Many scholars have explored the combination of Transformer and lattice structure. LAN [31] achieves it through lattice-aligned attention. FLAT [10] converts lattice structure into flat structure via ingenious span relative position encoding. Based on FLAT, ATSSA [7] has been improved by activating keys selectively. In this work,

[2] Source codes will be released in the EasyNLP framework [22].

we further propose a novel UKT framework for Chinese IE and apply it to the Transformer architecture.

Incorporating Linguistic and Semantic Knowledge for IE. Linguistic and semantic knowledge plays a vital role for various NLP tasks [18,21]. Specifically, structural knowledge, such as dependency parse tree (DEPT) and semantic dependency parse tree (SDPT), serves as a key component for IE. There is long-standing history digesting DEPT or SDPT for IE [3,25]. More recently, several studies try to encode structural knowledge graph by GNN [9,17]. Hence, these approaches usually need to combine with sequential structure models, thus increasing the complexity of the model. With the development of transformer-based approaches, researchers pay more attention on exploiting structural knowledge on the Transformer structure for IE [1,18]. For example, Sachan et al. [18] investigate strategy of applying GNN on the output of transformer and strategy of infusing syntax structure into the transformer attention layers for incorporating structural knowledge. Chen et al. [1] propose a type-aware map memory module based on the Transformer architecture to encode dependency information. However, most approaches may suffer from the knowledge usability issue and the missing or misrepresentation of knowledge issue. To address these limitations, we propose a novel UKT framework to correctly incorporate both DEPT and SDPT for Chinese IE.

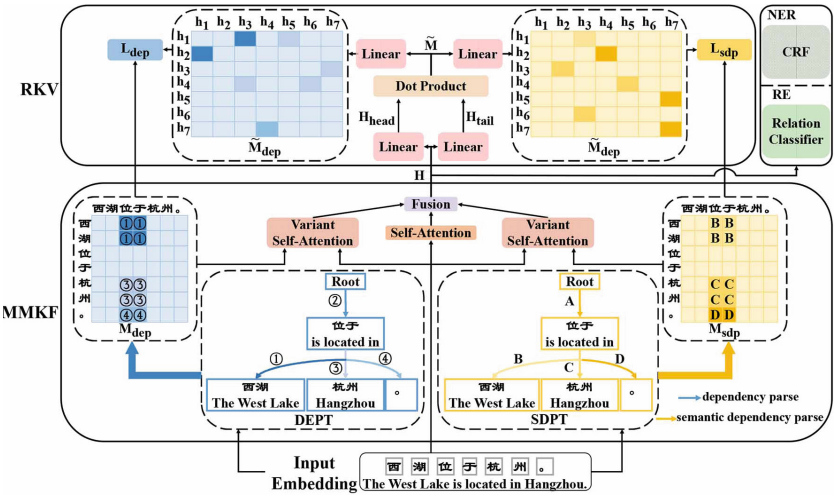


Fig. 2. Overall architecture. ①②③④ represent "subject-verb", "head", "verb-object" and "punctuation" dependency relationships, respectively. A,B,C,D represent "head", "experience", "location" and "punctuation" semantic relationships, respectively.

3 UKT: The Proposed Method

3.1 A Brief Overview of UKT

The overall architecture of Unified Knowledgeable Tuning (UKT) is shown in Fig. 2. Consider the example token sentence "The West Lake is located in Hangzhou.". We can obtain the structural knowledge (i.e. dependency parse tree (DEPT) and semantic dependency parse tree (SDPT)) by an off-the-shelf tool[3], as shown in Fig. 2. Here, each node in the tree is a text fragment, such as "the West Lake". Each edge in the tree represents the relationship between text fragments. For example, there is a "verb-object" dependency relationship and is a "location" semantic dependency relationship between "is located in" and "Hangzhou". We represent relationship knowledge as shown by M_{dep} and M_{sdp} separately in Fig. 2 and inject it into the model through a Multi-relational, Multi-head Knowledge Fusion (MMKF) module, detailed described in Subsect. 3.2. In order to validate whether the model truly "captures" the knowledge, we propose a Relational Knowledge Validation (RKV) module to calculate and back propagate the gap between the knowledge embedded in the model output sequence and the ground-truth knowledge, detailed in Subsect. 3.3. Take DEPT as an example. Specifically, we calculate and back propagate the gap between the dependency relationship between "is located in" and "Hangzhou" learned by the model and the ground-truth "verb-object" relationship.

We need to further clarify that UKT can be integrated into several Chinese IE models based on the Transformer architecture non-intrusively, i.e., FLAT [10] and ATSSA [7]. It addresses several Chinese IE tasks in a highly analogous manner. We will explain the details in Subsect. 3.4.

3.2 Multi-relational Multi-head Knowledge Fusion

Structural knowledge such as DEPT and SDPT indicates the existence and types of relationship between text fragments in the sentence, and provides crucial prior knowledge for IE [3,21]. As shown in Fig. 2, noun-based elements such as objects, subjects and complements are more likely to be named entities (such as "the West Lake" and "Hangzhou"). According to DEPT, there may be a relationship between "the West Lake" and "Hangzhou" through the intermediary jump of "is located in". In addition, different types of relationships reflect different key information for IE. For example, the semantic dependency relationship "location" between "is located in" and "Hangzhou" in the figure helps the model to infer that "Hangzhou" is a named entity of the location type. However, existing approaches pay little attention to DEPT and SDPT simultaneously, or lack the distinction of different types of relationships or lack one-to-one meaningful relationship representations. In UKT, We propose MMKF to learn one-to-one meaningful relationship knowledge representations based on BERT [2] and exploit them in a variant of self-attention to address the above limitations.

[3] http://ltp.ai/.

Relationship Knowledge Representation. Take DEPT as an example. After obtaining DEPT of a token sentence $C = [c_1, c_2, \cdots, c_n]$ (n is length of C), as shown in Fig. 2, $T_{dep} = \{Root_{dep}, Node_{dep}, Edge_{dep}\}$ is used to define the tree abstractly, where $Root_{dep}$ is the root of the tree, $Node_{dep}$ and $Edge_{dep}$ are the node set and the edge set respectively. Each node in the tree is a text fragment ent of C. Note that a text fragment may consist of more than one token. Each edge can be represented as $\{ent_p, ent_q, r_{dep}\}$, which means that there is an r_{dep} relationship between ent_p and ent_q. $r_{dep} \in \mathcal{R}_{dep}$, where $\mathcal{R}_{dep}$ indicates the set of all possible dependency relationships in total.

Consider the edge $\{ent_p, ent_q, r_{dep}\}$, we can feed the prompt sequence C_{prompt}, i.e., $C_{prompt} = \texttt{[CLS]} + C + \texttt{[SEP]} + ent_p + \texttt{[SEP]} + ent_q$ into BERT [2] to obtain the corresponding sentence-level relationship knowledge representation $\boldsymbol{r}_{dep}(C_{prompt})$ of the relationship r_{dep}. The one-to-one meaningful relationship knowledge representation for r_{dep}, denoted as $\boldsymbol{r}_{dep}$, is the averaged result of all such sentence-level representations (i.e., all $\boldsymbol{r}_{dep}(C_{prompt})$).

We represent T_{dep} as dependency relationship knowledge representation matrix M_{dep}. Let $M_{dep}^{(i,j)}$ be the (i,j)-th element of M_{dep}, initialized as $\mathbf{0}$. For each edge $\{ent_p, ent_q, r_{dep}\}$ of T_{dep}, we have $M_{dep}^{(i,j)} = \boldsymbol{r}_{dep}$ when $ent_p^l \le i \le ent_p^r$ and $ent_q^l \le j \le ent_q^r$. ent_p^l and ent_p^r respectively represent the starting and ending token position index of the text fragment ent_p in the token sentence. Relationship knowledge extends from the text fragment level to the token level. Thus, the knowledge of $\{ent_p, ent_q, r_{dep}\}$ that there is a dependency relationship r_{dep} between ent_p and ent_q can be encoded in M_{dep}. Similarly, we also obtain semantic dependency relationship knowledge representation matrix M_{sdp}.

Knowledge Injection. In UKT, we propose a variant of the self-attention mechanism which implement multiple attention heads following common practice to inject structural knowledge into the model. Let Q, K and V as the query, key and value in the vanilla self-attention mechanism. The vanilla self-attentive representation is calculated by: $Attn = \text{softmax}(Q \cdot K^T) \cdot V$.

Next, we consider the process of knowledge injection. Take DEPT as an example. We replace the key K with the dependency relationship knowledge representation matrix M_{dep} to obtain the weighted self-attentive representation $Attn_{dep}$ related to DEPT, shown as follows:

$$A_{dep} = (Q + v_{dep}) \cdot M_{dep}^T \quad , \quad Attn_{dep} = \text{softmax}(A_{dep}) \cdot V, \tag{1}$$

where v_{dep} is a learnable parameter for DEPT. Similarly, for SDPT, we have:

$$A_{sdp} = (Q + v_{sdp}) \cdot M_{sdp}^T \quad , \quad Attn_{sdp} = \text{softmax}(A_{sdp}) \cdot V. \tag{2}$$

In addition, different self-attentive representations, i.e., $Attn$, $Attn_{dep}$ and $Attn_{sdp}$, hold different levels of importance in different tasks and datasets. We choose to use a learnable weight W to solve this problem to compute the final self-attentive representation $Attn^*$. The formula can be expressed as follows:

$$Attn^* = W \cdot [Attn, Attn_{sdp}, Attn_{dep}]^T. \tag{3}$$

3.3 Relational Knowledge Validation

In order to force the model to learn injected relational knowledge, RKV is proposed to examine whether the model truly captures the knowledge. Denote H as the model's output representation, which may store the injected relational knowledge. We obtain H_{head} and H_{tail} by implementing linear transformations on H. The former represents the semantic and syntax information of the first text fragment in the relationship, while the latter represents those of the second text fragment. Let $\tilde{M}$ be a reproduced relationship knowledge matrix learned by the model, which can be obtained according to the following formula: $\tilde{M} = H_{head} \cdot H_{tail}^{T}$. We again take DEPT as an example. We obtain the dependency reproduced relationship knowledge matrix $\tilde{M}_{dep}$ by implementing linear transformations on $\tilde{M}$. $\tilde{M}_{dep}^{(i,j)}$ is the (i,j)-th element of $\tilde{M}_{dep}$, which represents the dependency relationship knowledge learned by the model between the i-th and the j-th token. We assume that there is a real dependency relationship r_{dep}. After that, $\tilde{M}_{dep}^{(i,j)}$ is fed into a softmax classifier to obtain $\Pr^{(i,j)}(r_{dep}|C)$, which is the estimated probability for i-th and j-th token having the relation r_{dep}.

During model training, as a regularizer, we back propagate the gap between the predicted relationship probabilistic distribution and the ground-truth to force the model to capture the true dependency relationship type. The sample-wise loss function related to DEPT (denoted as $L_{dep}(\Theta)$) is defined as:

$$L_{dep}(\Theta) = -\sum_{i}^{n}\sum_{j}^{n} \log Pr^{(i,j)}(r_{dep}|C), \tag{4}$$

where Θ indicates all parameters of our model. The loss function w.r.t. SDPT is similar to $L_{dep}(\Theta)$, i.e., $L_{sdp}(\Theta) = -\sum_{i}^{n}\sum_{j}^{n} \log Pr^{(i,j)}(r_{sdp}|C)$. The final loss function of model is $L(\Theta)$, defined as follows:

$$L(\Theta) = L_{task}(\Theta) + \lambda_{sdp} \cdot L_{sdp}(\Theta) + \lambda_{dep} \cdot L_{dep}(\Theta), \tag{5}$$

where λ_{sdp} and λ_{dep} are trade-off between multiple terms. $L_{task}(\Theta)$ is the original loss function of the IE task (either NER or RE).

3.4 UKT for Chinese IE

UKT works in a plug-and-play mode and is applicable to several popular Chinese IE models based on the Transformer architecture. Its universality also lies in its ability to address multiple Chinese IE tasks, including NER and RE, in a highly similar manner. In the following, we describe how to apply UKT to FLAT [10] for Chinese IE. For the NER task, we obtain the span embeddings E_{NER} directly from FLAT's span representations of the input. For the RE task, because many studies [11,27] have proven that positional embeddings are highly important, we combine them with FLAT's span representations to obtain the span embeddings for RE, denoted as E_{RE}. After that, E_{NER} or E_{RE} are passed to MMKF for

self-attention computation, with structural knowledge injected. The formula for obtaining the final representations of UKT-enhanced FLAT is:

$$Attn^* = W \cdot [Attn_{FLAT}, Attn_{sdp}, Attn_{dep}]^T, \tag{6}$$

where $Attn_{FLAT}$ is the self-attentive representation of FLAT.

For NER, a standard Conditional Random Field (CRF) module is built upon the modified transformer block. For RE, a simple relation classifier is added to the last transformer layer. The total loss for UKT-enhanced FLAT is as follows:

$$L(\Theta) = L_{FLAT}(\Theta) + \lambda_{sdp} \cdot L_{sdp}(\Theta) + \lambda_{dep} \cdot L_{dep}(\Theta), \tag{7}$$

where $L_{FLAT}(\Theta)$ is the original loss function of FLAT related to different tasks. Similarly, we can integrate UKT into ATSSA [7] in a highly similar manner, with two aspects modified: i) knowledge-injected self-attention by MMKF, and ii) additional loss functions for RKV. Due to space limitation, we omit the details.

4 Experiments

In this section, we conduct extensive experiments on five public Chinese IE datasets. We evaluate UKT using F1-score, with a comparison of state-of-the-art approaches. All experiments are conducted on NVIDIA V100 GPU (32GB).

4.1 Experiments Setup

Datasets. Three mainstream Chinese NER benchmark datasets (Weibo [6,16], Resume [30], MSRA [8]) and two mainstream Chinese RE benchmark datasets (SanWen [24] and FinRE [11]) are used for evaluating the performance of UKT. We show detailed statistics in Table 1, following the settings of LAN [31].

Table 1. Statistics of five Chinese IE datasets.

Dataset	NER			RE	
	Weibo	Resume	MSRA	SanWen	FinRE
# Train	73.8K	124.1K	2,169.9K	515K	727K
# Dev	14.5K	139K	–	55K	81K
# Test	14.8K	15.1K	172.6K	68K	203K

Baselines. ATSSA [7] and FLAT [10] are two recent popular Chinese NER models which can be adapted for Chinese RE with simple modifications as mentioned above. We treat FLAT and ATSSA as state-of-the-art baselines. In addition, several popular methods for Chinese NER and RE are also compared.

Implementation Details. We use the pre-trained character embeddings and bigram embeddings trained with word2vec [14]. BERT in the experiments is "BERT-wwm" released in [2]. The above three embeddings respectively have size of 50, 50, 768. Different baseline models use different word embedding dictionaries including YJ[4], LS[5], TX[6] and we conduct experiments using corresponding dictionaries along different baseline models suggested by their original papers. All of above four embeddings are fine-tuned during training. We use the stochastic gradient descent (SGD) optimizer with a momentum of 0.9 on all datasets for optimization. To avoid overfitting, the dropout technique is applied. For RKV, λ_{sdp} and λ_{dep} in the loss function are set to 1e-5 in default, which are also tuned over the development sets.

4.2 Overall Performance

Table 2 shows the overall performance on five public datasets. Generally, our approach UKT consistently outperforms all baseline models on both tasks, which demonstrates the effectiveness and universality of the proposed approach.

Table 2. Overall results in terms of F1 (%). * denotes models for RE only. ** denotes models for NER only. # denotes our re-production results based on open-source codes.

Model	Lexicon	NER			RE	
		MSRA	Resume	Weibo	FinRE	SanWen
MG Lattice* [11]	–	–	–	–	49.26	65.61
Lattice LSTM** [30]	YJ	93.18	94.46	58.79	–	–
CGN** [19]	LS	93.47	–	63.09	–	–
SoftLexicon** [12]	YJ	95.42	96.11	70.50	–	–
MECT** [23]	YJ	96.24	95.98	70.43	–	–
LAN [31]	TX	96.41	96.67	71.27	51.35	69.85
FLAT [10]	YJ	96.09	95.86	68.55	50.07#	72.84#
FLAT w/. UKT (**ours**)	YJ	**96.36**	**96.66**	**69.55**	**51.00**	**74.48**
ATSSA [7]	LS	96.45	96.73	72.53	52.28#	73.86#
ATSSA w/. UKT (**ours**)	LS	**96.49**	**96.81**	**73.18**	**52.81**	**74.84**
ChatGPT (for reference only)	–	50.46	54.60	23.10	28.47	42.81

In detail, on three NER datasets, we obtain large performance improvement over Weibo (+1.00%), Resume (+0.8%) and MSRA (+0.27%) compared with vanilla FLAT. For ATSSA, our UKT-based enhancement also achieves good performance. In addition, injecting linguistic and semantic knowledge into vanilla

[4] https://github.com/jiesutd/RichWordSegmentor.
[5] https://github.com/Embedding/Chinese-Word-Vectors.
[6] https://ai.tencent.com/ailab/nlp/en/embedding.html(v0.1.0).

FLAT improves the RE performance on FinRE and SanWen datasets by 0.93% and 1.64% in F1, respectively. For ATSSA, the corresponding improvement scores are 0.53% and 0.98%. We also present the F1 scores of ChatGPT (gpt-3.5-turbo) over zero-shot Chinese IE for reference. We can see that ChatGPT performs poorly. By analyzing examples, we can intuitively find that relationship or entity types generated by ChatGPT may not be in the optional list provided in input prompts. For Chinese NER task, ChatGPT sometimes does not generate the correct numbers of named entities corresponding to inputs. In a few cases, ChatGPT does not even follow instructions and generates answers irrelevant to the questions. Hence, our approach has practical values in the LLM era.

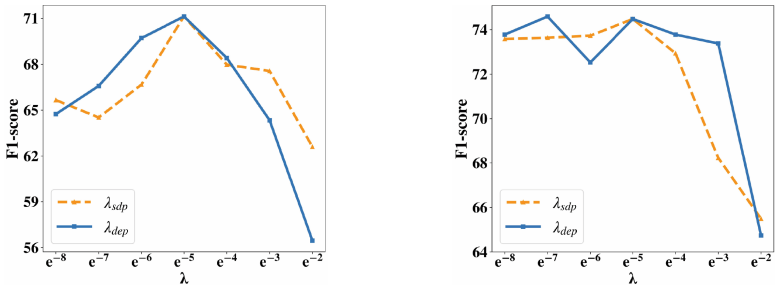
4.3 Model Analysis

Ablation Study. To further investigate the effectiveness of all the major components of our proposed approach, using FLAT as our base model, we conduct ablation experiments, with results reported in Table 3. We have the following findings. i) The degradation in performance with the removal of two types of knowledge proves that injecting it into models brings substantial improvement for Chinese IE. ii) We replace all relationship knowledge representation matrices M (i.e. M_{dep} and M_{sdp}) with all-zero matrices. The decline in performance indicates the validity of one-to-one meaningful relationship representations. iii) Removing RKV leads to significantly worse results on all datasets, which suggests that RKV can truly force the model to learn injected relational knowledge, thus improving the effectiveness of knowledge injection for Chinese IE.

Influence of Hyperparameters. To analyze the influence of λ (i.e. λ_{dep} and λ_{sdp}), we conduct experiments on UKT-enhanced FLAT over one NRE dataset (i.e. Weibo) and one RE dataset (i.e. SanWen), in which one hyperparameter is kept as 1e-5 while the other is varied. The F1-score results are summarized in Fig. 3. As the value of λ increases, the scores generally exhibit a trend of initially increasing followed by decreasing. This is because when λ is small, the RKV technique fails to exert its effect. When λ is large, the model tends to allocate more attention to the learning of the injected knowledge, deviating from the required IE task and leading to a decrease in F1 score.

Table 3. Ablation results in terms of F1 (%).

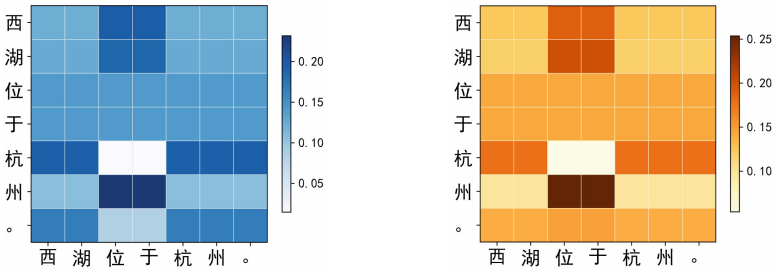
Model	Resume	Weibo	MSRA	SanWen	FinRE
Full implementation	**96.66**	**69.55**	**96.36**	**74.48**	**51.00**
w/o. DEPT	96.56	69.18	96.16	72.99	50.77
w/o. SDPT	96.52	69.13	96.29	74.06	50.68
w/o. M_{dep} and M_{sdp}	96.44	69.46	96.13	73.40	50.53
w/o. RKV	96.49	68.71	96.26	73.57	50.28



(a) Weibo (b) SanWen

Fig. 3. Parameter analysis of λ_{dep} and λ_{sdp}.

Visualizations. In order to vividly demonstrate the impact of knowledge for model learning, we visualize the attention distributions related to DEPT and SDPT of the example "The West Lake is located in Hangzhou.", as shown in Fig. 4. In the figure, the vertical axis represents queries and the horizontal axis indicates keys. We observe that there is differential attention given to tokens where structural knowledge exists compared to other tokens. Take DEPT as an example, the attention of queries in "The West Lake" is more likely to be assigned to keys in "is located in", which exactly corresponds to the "subject-verb" dependency relationship between "The West Lake" and "is located in".



(a) DEPT (b) SDPT

Fig. 4. Attention distributions related to knowledge derived from DEPT and SDPT.

5 Conclusion

In this paper, we propose UKT to inject linguistic and semantic knowledge for Chinese IE. It plays in a plug-and-play fashion and is applicable to popular Chinese IE models based on the Transformer architecture. To achieve our goal, we propose MMKF in UKT to incorporate knowledge and further introduce RKV to force model to learn the injected relational knowledge. We evaluate the proposed approach on five public datasets related to two Chinese IE tasks. Experimental results demonstrates the effectiveness and universality of the proposed

approach, which achieves consistent improvement over state-of-the-art models. In the future, we will extend our study to a broader range of Chinese IE tasks.

Acknowledgments. This work is supported by the Guangzhou Science and Technology Program key projects (202103010005), the National Natural Science Foundation of China (61876066) and Alibaba Cloud Group through the Research Talent Program with South China University of Technology.

References

1. Chen, G., Tian, Y., Song, Y., Wan, X.: Relation extraction with type-aware map memories of word dependencies. In: ACL-IJCNLP, pp. 2501–2512 (2021)
2. Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z.: Pre-training with whole word masking for Chinese bert. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 3504–3514 (2021)
3. Fundel, K., Küffner, R., Zimmer, R.: Relex—relation extraction using dependency parse trees. Bioinformatics **23**(3), 365–371 (2007)
4. Gui, T., Ma, R., Zhang, Q., Zhao, L., Jiang, Y.G., Huang, X.: CNN-based Chinese NER with lexicon rethinking. In: IJCAI, pp. 4982–4988 (2019)
5. Gui, T., et al.: A lexicon-based graph neural network for Chinese NER. In: EMNLP-IJCNLP, pp. 1040–1050 (2019)
6. He, H., Sun, X.: F-score driven max margin neural network for named entity recognition in Chinese social media. In: EACL, pp. 713–718 (2017)
7. Hu, B., Huang, Z., Hu, M., Zhang, Z., Dou, Y.: Adaptive threshold selective self-attention for Chinese NER. In: COLING, pp. 1823–1833 (2022)
8. Levow, G.A.: The third international Chinese language processing bakeoff: word segmentation and named entity recognition. In: SIGHAN, pp. 108–117 (2006)
9. Li, F., Lin, Z., Zhang, M., Ji, D.: A span-based model for joint overlapped and discontinuous named entity recognition. In: ACL/IJCNLP, pp. 4814–4828 (2021)
10. Li, X., Yan, H., Qiu, X., Huang, X.J.: Flat: Chinese NER using flat-lattice transformer. In: ACL, pp. 6836–6842 (2020)
11. Li, Z., Ding, N., Liu, Z., Zheng, H., Shen, Y.: Chinese relation extraction with multi-grained information and external linguistic knowledge. In: ACL, pp. 4377–4386 (2019)
12. Ma, R., Peng, M., Zhang, Q., Wei, Z., Huang, X.J.: Simplify the usage of lexicon in Chinese NER. In: ACL, pp. 5951–5960 (2020)
13. Ma, Y., Cao, Y., Hong, Y., Sun, A.: Large language model is not a good few-shot information extractor, but a good reranker for hard samples! CoRR abs/2303.08559 (2023)
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS, pp. 3111–3119 (2013)
15. Ouyang, L., et al.: Training language models to follow instructions with human feedback. In: NIPS, pp. 27730–27744 (2022)
16. Peng, N., Dredze, M.: Named entity recognition for Chinese social media with jointly trained embeddings. In: EMNLP, pp. 548–554 (2015)
17. Qin, H., Tian, Y., Song, Y.: Relation extraction with word graphs from N-grams. In: EMNLP, pp. 2860–2868 (2021)

18. Sachan, D., Zhang, Y., Qi, P., Hamilton, W.L.: Do syntax trees help pre-trained transformers extract information? In: EACL, pp. 2647–2661 (2021)
19. Sui, D., Chen, Y., Liu, K., Zhao, J., Liu, S.: Leverage lexical knowledge for Chinese named entity recognition via collaborative graph network. In: EMNLP-IJCNLP, pp. 3830–3840 (2019)
20. Vaswani, A., et al.: Attention is all you need. In: NIPS, pp. 5998–6008 (2017)
21. Wan, Q., Wan, C., Hu, R., Liu, D.: Chinese financial event extraction based on syntactic and semantic dependency parsing. Chin. J. Comput. **44**(3), 508–530 (2021)
22. Wang, C., et al.: EasyNLP: a comprehensive and easy-to-use toolkit for natural language processing. In: EMNLP, pp. 22–29 (2022)
23. Wu, S., Song, X., FENG, Z.: Mect: multi-metadata embedding based cross-transformer for Chinese named entity recognition. In: ACL-IJCNLP, pp. 1529–1539 (2021)
24. Xu, J., Wen, J., Sun, X., Su, Q.: A discourse-level named entity recognition and relation extraction dataset for Chinese literature text. CoRR abs/1711.07010 (2017)
25. 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: EMNLP, pp. 1785–1794 (2015)
26. Zeng, A., et al.: GLM-130B: an open bilingual pre-trained model. CoRR abs/2210.02414 (2022)
27. Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J.: Relation classification via convolutional deep neural network. In: COLING, pp. 2335–2344 (2014)
28. Zhang, T., et al.: HORNET: enriching pre-trained language representations with heterogeneous knowledge sources. In: CIKM, pp. 2608–2617 (2021)
29. Zhang, T., et al.: DKPLM: decomposable knowledge-enhanced pre-trained language model for natural language understanding. In: AAAI, pp. 11703–11711 (2022)
30. Zhang, Y., Yang, J.: Chinese NER using lattice LSTM. In: ACL, pp. 1554–1564 (2018)
31. Zhao, S., Hu, M., Cai, Z., Zhang, Z., Zhou, T., Liu, F.: Enhancing Chinese character representation with lattice-aligned attention. IEEE Trans. Neural Netw. Learn. Syst. **34**(7), 3727–3736 (2023). https://doi.org/10.1109/TNNLS.2021.3114378