Improving Relation Extraction with Knowledge-attention

Pengfei Li[1], Kezhi Mao[1]*, Xuefeng Yang[3], and Qi Li[2]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[2]Interdisciplinary Graduate School, Nanyang Technological University, Singapore
[3]ZhuiYi Technology, Shenzhen, China
{pli006,ekzmao,liqi0024}@ntu.edu.sg, ryan@wezhuiyi.com

Abstract

While attention mechanisms have been proven to be effective in many NLP tasks, majority of them are data-driven. We propose a novel knowledge-attention encoder which incorporates prior knowledge from external lexical resources into deep neural networks for relation extraction task. Furthermore, we present three effective ways of integrating knowledge-attention with self-attention to maximize the utilization of both knowledge and data. The proposed relation extraction system is end-to-end and fully attention-based. Experiment results show that the proposed knowledge-attention mechanism has complementary strengths with self-attention, and our integrated models outperform existing CNN, RNN, and self-attention based models. State-of-the-art performance is achieved on TACRED, a complex and large-scale relation extraction dataset.

1 Introduction

Relation extraction aims to detect the semantic relationship between two entities in a sentence. For example, given the sentence: "*James Dobson* has resigned as chairman of *Focus On The Family*, which he founded thirty years ago.", the goal is to recognize the organization-founder relation held between "*Focus On The Family*" and "*James Dobson*". The various relations between entities extracted from large-scale unstructured texts can be used for ontology and knowledge base population (Chen et al., 2018a; Fossati et al., 2018), as well as facilitating downstream tasks that requires relational understanding of texts such as question answering (Yu et al., 2017) and dialogue systems (Young et al., 2018).

Traditional feature-based and kernel-based approaches require extensive feature engineering (Suchanek et al., 2006; Qian et al., 2008; Rink and Harabagiu, 2010). Deep neural networks such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have the ability of exploring more complex semantics and extracting features automatically from raw texts for relation extraction tasks (Xu et al., 2016; Vu et al., 2016; Lee et al., 2017). Recently, attention mechanisms have been introduced to deep neural networks to improve their performance (Zhou et al., 2016; Wang et al., 2016; Zhang et al., 2017). Especially, the Transformer proposed by Vaswani et al. (2017) is based solely on self-attention and has demonstrated better performance than traditional RNNs (Bilan and Roth, 2018; Verga et al., 2018). However, deep neural networks normally require sufficient labeled data to train their numerous model parameters. The scarcity or low quality of training data will limit the model's ability to recognize complex relations and also cause overfitting issue.

A recent study (Li and Mao, 2019) shows that incorporating prior knowledge from external lexical resources into deep neural network can reduce the reliance on training data and improve relation extraction performance. Motivated by this, we propose a novel knowledge-attention mechanism, which transforms texts from word semantic space into relational semantic space by attending to relation indicators that are useful in recognizing different relations. The relation indicators are automatically generated from lexical knowledge bases which represent keywords and cue phrases of different relation expressions. While the existing self-attention encoder learns internal semantic features by attending to the input texts themselves, the proposed knowledge-attention encoder captures the linguistic clues of different relations based on external knowledge. Since the two attention mechanisms complement each other, we integrate them into a single model to maximize the uti-

*Corresponding author.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 229–239, Hong Kong, China, November 3–7, 2019.

lization of both knowledge and data, and achieve optimal performance for relation extraction.

In summary, the main contributions of the paper are: (1) We propose knowledge-attention encoder, a novel attention mechanism which incorporates prior knowledge from external lexical resources to effectively capture the informative linguistic clues for relation extraction. (2) To take the advantages of both knowledge-attention and self-attention, we propose three integration strategies: multi-channel attention, softmax interpolation, and knowledge-informed self-attention. Our final models are fully attention-based and can be easily set up for end-to-end training. (3) We present detailed analysis on knowledge-attention encoder. Results show that it has complementary strengths with self-attention encoder, and the integrated models achieve start-of-the-art results for relation extraction.

2 Related Works

We focus here on deep neural networks for relation extraction since they have demonstrated better performance than traditional feature-based and kernel-based approaches.

Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are the earliest and commonly used approaches for relation extraction. Zeng et al. (2014) showed that CNN with position embeddings is effective for relation extraction. Similarly, CNN with multiple filter sizes (Nguyen and Grishman, 2015), pairwise ranking loss function (dos Santos et al., 2015) and auxiliary embeddings (Lee et al., 2017) were proposed to improve performance. Zhang and Wang (2015) proposed bi-directional RNN with max pooling to model the sequential relations. Instead of modeling the whole sentence, performing RNN on sub-dependency trees (e.g. shortest dependency path between two entities) has demonstrated to be effective in capturing long-distance relation patterns (Xu et al., 2016; Miwa and Bansal, 2016). Zhang et al. (2018) proposed graph convolution over dependency trees and achieved state-of-the-art results on TACRED dataset.

Recently, attention mechanisms have been widely applied to CNNs (Wang et al., 2016; Han et al., 2018) and RNNs (Zhou et al., 2016; Zhang et al., 2017; Du et al., 2018). The improved performance demonstrated the effectiveness of attention mechanisms in deep neural networks. Particularly, Vaswani et al. (2017) proposed a solely self-attention-based model called Transformer, which is more effective than RNNs in capturing long-distance features since it is able to draw global dependencies without regard to their distances in the sequences. Bilan and Roth (2018) first applied self-attention encoder to relation extraction task and achieved competitive results on TACRED dataset. Verga et al. (2018) used self-attention to encode long contexts spanning multiple sentences for biological relation extraction. However, more attention heads and layers are required for self-attention encoder to capture complex semantic and syntactic information since learning is solely based on training data. Hence, more high quality training data and computational power are needed. Our work utilizes the knowledge from external lexical resources to improve deep neural network's ability of capturing informative linguistic clues.

External knowledge has shown to be effective in neural networks for many NLP tasks. Existing works focus on utilizing external knowledge to improve embedding representations (Chen et al., 2015; Liu et al., 2015; Sinoara et al., 2019), CNNs (Toutanova et al., 2015; Wang et al., 2017; Li and Mao, 2019), and RNNs (Ahn et al., 2016; Chen et al., 2016, 2018b; Shen et al., 2018). Our work is the first to incorporate knowledge into Transformer through a novel knowledge-attention mechanism to improve its performance on relation extraction task.

3 Knowledge-attention Encoder

We present the proposed knowledge-attention encoder in this section. Relation indicators are first generated from external lexical resources (Section 3.1); Then the input texts are transformed from word semantic space into relational semantic space by attending to the relation indicators using knowledge-attention mechanism (Section 3.2); Finally, position-aware attention is used to summarize the input sequence by taking both relation semantics and relative positions into consideration (Section 3.3).

3.1 Relation Indicators Generation

Relation indicators represent the keywords or cue phrases of various relation types, which are essential for knowledge-attention encoder to capture the linguistic clues of certain relation from texts.

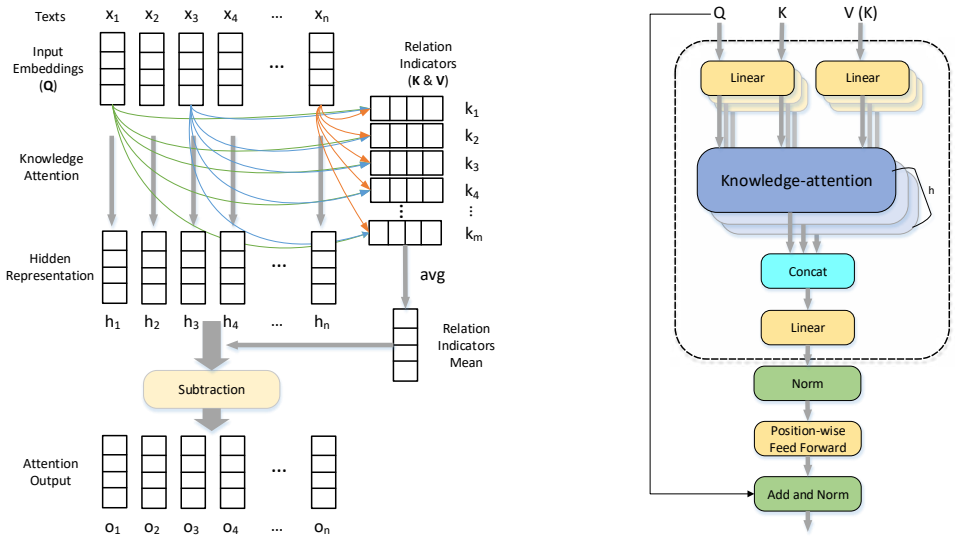

Figure 1: Knowledge-attention process (left) and multi-head structure (right) of knowledge-attention encoder.

We utilize two publicly available lexical resources including FrameNet[1] and Thesaurus.com[2] to find such lexical units.

FrameNet is a large lexical knowledge base which categorizes English words and sentences into higher level semantic frames (Ruppenhofer et al., 2006). Each frame is a conceptual structure describing a type of event, object or relation. FrameNet contains over 1200 semantic frames, many of which represent various semantic relations. For each relation type in our relation extraction task, we first find all the relevant semantic frames by searching from FrameNet (refer Appendix for detailed semantic frames used). Then we extract all the lexical units involved in these frames, which are exactly the keywords or phrases that often used to express such relation. Thesaurus.com is the largest online thesaurus which has over 3 million synonyms and antonyms. It also has the flexibility to filter search results by relevance, POS tag, word length, and complexity. To broaden the coverage of relation indicators, we utilize the synonyms in Thesaurus.com to extend the lexical units extracted from FrameNet. To reduce noise, only the most relevant synonyms with the same POS tag are selected.

Relation indicators are generated based on the word embeddings and POS tags of lexical units. Formally, given a word in a lexical unit, we find its word embedding $\mathbf{w}_i \in \mathbb{R}^{d_w}$ and POS embedding $\mathbf{t}_i \in \mathbb{R}^{d_t}$ by looking up the word embedding matrix $\mathbf{W}^{wrd} \in \mathbb{R}^{d_w \times V^{wrd}}$ and POS embedding matrix $\mathbf{W}^{pos} \in \mathbb{R}^{d_t \times V^{pos}}$ respectively, where d_w and d_t are the dimensions of word and POS embeddings, V^{wrd} is vocabulary size[3] and V^{pos} is total number of POS tags. The corresponding relation indicator is formed by concatenating word embedding and POS embedding, $\mathbf{k}_i = [\mathbf{w}_i, \mathbf{t}_i]$. If a lexical unit contains multiple words (i.e. phrase), the corresponding relation indicator is formed by averaging the embeddings of all words. Eventually, around 3000 relation indicators (including 2000 synonyms) are generated: $\mathbf{K} = \{\mathbf{k}_1, \mathbf{k}_2, ..., \mathbf{k}_m\}$.

3.2 Knowledge Attention

3.2.1 Knowledge-attention process

In a typical attention mechanism, a query (q) is compared with the keys (K) in a set of key-value pairs and the corresponding attention weights are calculated. The attention output is weighted sum of values (V) using the attention weights. In our proposed knowledge-attention encoder, the queries are input texts and the key-value pairs are both relation indicators. The detailed process of knowledge-attention is shown in Figure 1 (left).

Formally, given text input $x = \{x_1, x_2, ..., x_n\}$, the input embeddings $\mathbf{Q} = \{\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_n\}$ are generated by concatenating each word's word embedding and POS embedding in the same way as relation indicator generation in Section 3.1. The

[1] https://framenet.icsi.berkeley.edu/fndrupal

[2] https://www.thesaurus.com

[3] Same word embedding matrix is used for relation indicators and input texts, hence the vocabulary also includes all the words in the training corpus.

hidden representations $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n\}$ are obtained by attending to the knowledge indicators $\mathbf{K}$, as shown in Equation 1. The final knowledge-attention outputs are obtained by subtracting the hidden representations with the relation indicators mean, as shown in Equation 2.

$$\mathbf{H} = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V} \quad (1)$$

$$knwl(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{H} - \sum \mathbf{K}/m \quad (2)$$

where $knwl$ indicates knowledge-attention process, m is the number of relation indicators, and d_k is dimension of key/query vectors which is a scaling factor same as in Vaswani et al. (2017).

The subtraction of relation indicators mean will result in small outputs for irrelevant words. More importantly, the resulted output will be close to the related relation indicators and further apart from other relation indicators in relational semantic space. Therefore, the proposed knowledge-attention mechanism is effective in capturing the linguistic clues of relations represented by relation indicators in the relational semantic space.

3.2.2 Multi-head knowledge-attention

Inspired by the multi-head attention in Transformer (Vaswani et al., 2017), we also have multi-head knowledge-attention which first linearly transforms $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ h times, and then perform h knowledge-attentions simultaneously, as shown in Figure 1 (right).

Different from the Transformer encoder, we use the same linear transformation for $\mathbf{Q}$ and $\mathbf{K}$ in each head to keep the correspondence between queries and keys.

$$head_i = knwl(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^Q, \mathbf{V}\mathbf{W}_i^V) \quad (3)$$

where $\mathbf{W}_i^Q, \mathbf{W}_i^V \in \mathbb{R}^{d_k \times (d_k/h)}$ and $i \in [1, 2, ...h]$. Besides, only one residual connection from input embeddings to outputs of position-wise feed forward networks is used. We also mask the outputs of padding tokens using zero vectors.

The multi-head structure in knowledge-attention allows the model to jointly attend inputs to different relational semantic subspaces with different contributions of relation indicators. This is beneficial in recognizing complex relations where various compositions of relation indicators are needed.

3.3 Position-aware Attention

It has been proven that the relative position information of each token with respective to the two target entities is beneficial for relation extraction task (Zeng et al., 2014). We modify the position-aware attention originally proposed by Zhang et al. (2017) to incorporate such relative position information and find the importance of each token to the final sentence representation.

Assume the relative position of token x_i to target entity is $\hat{p}_i$. We apply position binning function (Equation 4) to make it easier for the model to distinguish long and short relative distances.

$$p_i = \begin{cases} \hat{p}_i & |\hat{p}_i| \leq 2 \\ \frac{\hat{p}_i}{|\hat{p}_i|} \lceil log_2 |\hat{p}_i| + 1 \rceil & |\hat{p}_i| > 2 \end{cases} \quad (4)$$

After getting the relative positions p_i^s and p_i^o to the two entities of interest (subject and object respectively), we map them to position embeddings base on a shared position embedding matrix $\mathbf{W}^p$. The two embeddings are concatenated to form the final position embedding for token x_i: $\mathbf{p}_i = [\mathbf{p}_i^s, \mathbf{p}_i^o]$.

Position-aware attention is performed on the outputs of knowledge-attention $\mathbf{O} \in \mathbb{R}^{n \times d_k}$, taking the corresponding relative position embeddings $\mathbf{P} \in \mathbb{R}^{n \times d_p}$ into consideration:

$$\mathbf{f} = O^T softmax(tanh(\mathbf{O}\mathbf{W}_o + \mathbf{P}\mathbf{W}_p)\mathbf{c}) \quad (5)$$

where $\mathbf{W}_o \in \mathbb{R}^{d_k \times d_a}$, $\mathbf{W}_p \in \mathbb{R}^{d_p \times d_a}$, d_a is attention dimension, and $\mathbf{c} \in \mathbb{R}^{d_a}$ is a context vector learned by the neural network.

4 Integrate Knowledge-attention with Self-attention

The self-attention encoder proposed by Vaswani et al. (2017) learns internal semantic features by modeling pair-wise interactions within the texts themselves, which is effective in capturing long-distance dependencies. Our proposed knowledge-attention encoder has complementary strengths of capturing the linguistic clues of relations precisely based on external knowledge. Therefore, it is beneficial to integrate the two models to maximize the utilization of both external knowledge and training data. In this section, we propose three integration approaches as shown in Figure 2, and each approach has its own advantages.

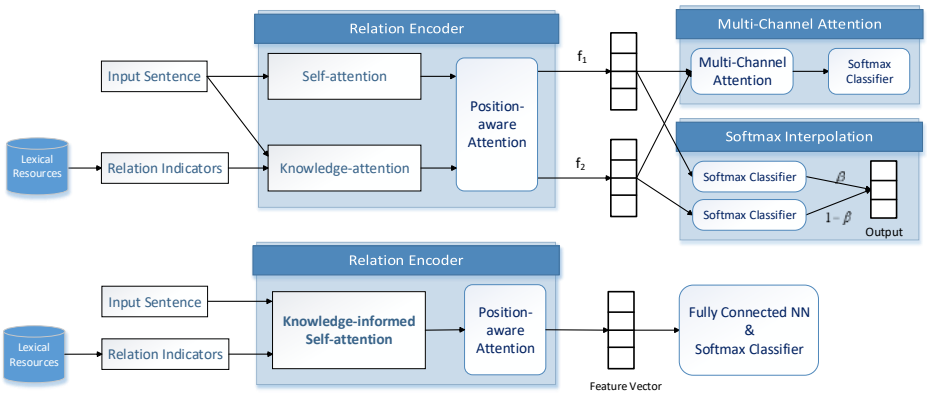

Figure 2: Three ways of integrating knowledge-attention with self-attention: multi-channel attention and softmax interpolation (top), as well as knowledge-informed self-attention (bottom).

4.1 Multi-channel Attention

In this approach, self-attention and knowledge-attention are treated as two separate channels to model sentence from different perspectives. After applying position-aware attention, two feature vectors $\mathbf{f}_1$ and $\mathbf{f}_2$ are obtained from self-attention and knowledge-attention respectively. We apply another attention mechanism called multi-channel attention to integrate the feature vectors.

In multi-channel attention, feature vectors are first fed into a fully connected neural network to get their hidden representations $\mathbf{h}_i$. Then attention weights are calculated using a learnable context vector $\mathbf{c}$, which reflects the importance of each feature vector to final relation classification. Finally, the feature vectors are integrated based on attention weights, as shown in Equation 6.

$$\mathbf{r} = \sum_i softmax(\mathbf{h}_i^T \mathbf{c})\mathbf{h}_i \tag{6}$$

After obtaining the integrated feature vector $\mathbf{r}$, we pass it to a softmax classifier to determine the relation class. The model is trained using stochastic gradient descent with momentum and learning rate decay to minimize the cross-entropy loss.

The main advantage of this approach is flexibility. Since the two channels process information independently, the input components are not necessary to be the same. Besides, we can add more features from other sources (e.g. subject and object categories) to multi-channel attention to make final decision based on all the information sources.

4.2 Softmax Interpolation

Similar as multi-channel attention, we also use two independent channels for self-attention and knowledge-attention in softmax interpolation. Instead of integrating the feature vectors, we make two independent predictions using two softmax classifiers based on the feature vectors from the two channels. The loss function is defined as total cross-entropy loss of the two classifiers. The final prediction is obtained using an interpolation function of the two softmax distributions:

$$\mathbf{p} = \beta \cdot \mathbf{p}_1 + (1 - \beta) \cdot \mathbf{p}_2 \tag{7}$$

where $\mathbf{p}_1$, $\mathbf{p}_2$ are the softmax distributions obtained form self-attention and knowledge-attention respectively, and β is the priority weight assigned to self-attention.

Since knowledge-attention focuses on capturing the keywords and cue phrases of relations, the precision will be higher than self-attention while the recall is lower. The proposed softmax interpolation approach is able to take the advantages of both attention mechanisms and balance the precision and recall by adjusting the priority weight β.

4.3 Knowledge-informed Self-attention

Since knowledge-attention and self-attention share similar structures, it is also possible to integrate them into a single channel. We propose knowledge-informed self-attention encoder which incorporates knowledge-attention into every self-attention head to jointly model the semantic relations based on both knowledge and data.

The structure of knowledge-informed self-attention is shown in Figure 3. Formally, given texts input matrix $\mathbf{Q} \in \mathbb{R}^{n \times d_k}$ and knowledge indicators $\mathbf{K} \in \mathbb{R}^{m \times d_k}$. The output of each attention head is calculated as follows:

$$\begin{aligned} head_i = \; & knwl(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^V) + \\ & self(\mathbf{Q}\mathbf{W}_i^{Q_s}, \mathbf{Q}\mathbf{W}_i^{K_s}, \mathbf{Q}\mathbf{W}_i^{V_s}) \end{aligned} \tag{8}$$

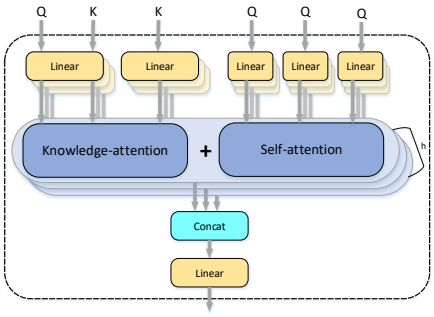

Figure 3: Knowledge-informed self-attention structure. Q, K represent input matrix and knowledge indicators respectively, h is the number of attention heads.

where $knwl$ and $self$ indicate knowledge-attention and self-attention respectively, and all the linear transformation weight matrices have the dimensionality of $\mathbf{W} \in \mathbb{R}^{d_k \times (d_k/h)}$.

Since each self-attention head is aided with prior knowledge in knowledge-attention, the knowledge-informed self-attention encoder is able to capture more lexical and semantic information than single attention encoder.

5 Experiment and Analysis

5.1 Baseline Models

To study the performance of our proposed models, the following baseline models are used for comparison:

CNN-based models including: (1) CNN: the classical convolutional neural network for sentence classification (Kim, 2014). (2) CNN-PE: CNN with position embeddings dedicated for relation classification (Nguyen and Grishman, 2015). (3) GCN: a graph convolutional network over the pruned dependency trees of the sentence (Zhang et al., 2018).

RNN-based models including: (1) LSTM: long short-term memory network to sequentially model the texts. Classification is based on the last hidden output. (2) PA-LSTM: Similar position-aware attention mechanism as our work is used to summarize the LSTM outputs (Zhang et al., 2017).

CNN-RNN hybrid model including contextualized GCN (C-GCN) where the input vectors are obtained using bi-directional LSTM network (Zhang et al., 2018).

Self-attention-based model (Self-attn) which uses self-attention encoder to model the input sentence. Our implementation is based on Bilan and Roth (2018) where several modifications are made on the original Transformer encoder, including the use of relative positional encodings instead of absolute sinusoidal encodings, as well as other configurations such as residual connection, activation function and normalization.

For our model, we evaluate both the proposed knowledge-attention encoder (Knwl-attn) as well as the integrated models with self-attention including multi-channel attention (MCA), softmax interpolation (SI) and knowledge-informed self-attention (KISA).

5.2 Experiment Settings

We conduct our main experiments on TACRED, a large-scale relation extraction dataset introduced by Zhang et al. (2017). TACRED contains over 106k sentences with hand-annotated subject and object entities as well as the relations between them. It is a very complex relation extraction dataset with 41 relation types and a *no_relation* class when no relation is hold between entities. The dataset is suited for real-word relation extraction since it is unbalanced with 79.5% *no_relation* samples, and multiple relations between different entity pairs can be exist in one sentence. Besides, the samples are normally long sentences with an average of 36.2 words.

Since the dataset is already partitioned into train (68124 samples), dev (22631 samples) and test (15509 samples) sets, we tune model hyperparameters using dev set and evaluate model using test set. The evaluation metrics are micro-averaged precision, recall and F_1 score. For fair comparison, we select the model with median F_1 score on dev set from 5 independent runs, same as Zhang et al. (2017). The same "entity mask" strategy is used which replaces subject (or object) entity with special $\langle$NER$\rangle$-SUBJ (or $\langle$NER$\rangle$-OBJ) tokens to avoid overfittting on specific entities and provide entity type information.

Besides TACRED, another dataset called SemEval2010-Task8 (Hendrickx et al., 2009) is used to evaluate the generalization ability of our proposed model. The dataset is significantly smaller and simpler than TACRED, which has 8000 training samples and 2717 testing samples. It contains 9 directed relations and 1 other relation (19 relation classes in total). We use the official macro-averaged F_1 score as evaluation metric.

We use one layer encoder with 6 attention heads for both knowledge-attention and self-attention

since further increasing the number of layers and attention heads will degrade the performance. For softmax interpolation, we choose $\beta = 0.8$ to balance precision and recall. Word embeddings are fine-tuned based on pre-trained GloVe (Pennington et al., 2014) with dimensionality of 300. Dropout (Srivastava et al., 2014) is used during trianing to alleviate overfitting. Other model hyperparameters and training details are described in Appendix due to space limitations.

5.3 Results and Analysis

5.3.1 Results on TACRED dataset

Table 1 shows the results of baseline as well as our proposed models on TACRED dataset. It is observed that our proposed knowledge-attention encoder outperforms all CNN-based and RNN-based models by at least 1.3 F_1. Meanwhile, it achieves comparable results with C-GCN and self-attention encoder, which are the current start-of-the-art single-model systems.

Comparing with self-attention encoder, it is observed that knowledge-attention encoder results in higher precision but lower recall. This is reasonable since knowledge-attention encoder focuses on capturing the significant linguistic clues of relations based on external knowledge, it will result in high precision for the predicted relations similar to rule-based systems. Self-attention encoder is able to capture more long-distance dependency features by learning from data, resulting in better recall. By integrating self-attention and knowledge-attention using the proposed approaches, a more balanced precision and recall can be obtained, suggesting the complementary effects of self-attention and knowledge-attention mechanisms. The integrated models improve performance by at least 0.9 F_1 score and achieve new state-of-the-art results among all the single end-to-end models.

Comparing the three integrated models, softmax interpolation (SI) achieves the best performance. More interestingly, we found that the precision and recall can be controlled by adjusting the priority weight β. Figure 4 shows impact of β on precision, recall and F_1 score. As β increases, precision decreases and recall increases. Therefore, we can choose a small β for relation extraction system which requires high precision, and a large β for the system requiring better recall. F_1 score reaches the highest value when precision and recall are balanced ($\beta = 0.8$).

Model	P	R	F_1
CNN†	**72.1**	50.3	59.2
CNN-PE†	68.2	55.4	61.1
GCN‡	69.8	59.0	64.0
LSTM†	61.4	61.7	61.5
PA-LSTM†	65.7	64.5	65.1
C-GCN‡	69.9	63.3	66.4
Self-attn††	64.6	**68.6**	66.5
Knwl-attn	70.0	63.1	66.4
Knwl+Self (MCA)	68.4	66.1	67.3*
Knwl+Self (SI)	67.1	68.4	**67.8***
Know+Self (KISA)	69.4	66.0	67.7*

Table 1: Micro-averaged precision (P), recall (R) and F_1 score on TACRED dataset. †, ‡ and †† mark the results reported in (Zhang et al., 2017), (Zhang et al., 2018) and (Bilan and Roth, 2018) respectively. $*$ marks statistically significant improvements over Self-attn with $p < 0.01$ under one-tailed t-test.

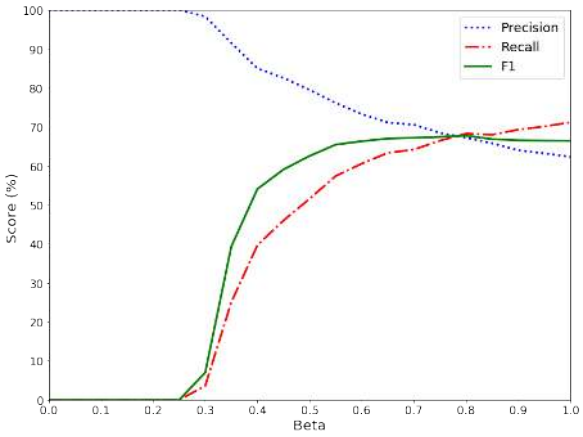

Figure 4: Change of precision, recall and F_1 score on dev set as the priority weight β in softmax interpolation changes.

Knowledge-informed self-attention (KISA) has comparable performance with softmax interpolation, and without the need of hyper-parameter tuning since knowledge-attention and self-attention are integrated into a single channel. The performance gain over self-attention encoder is 1.2 F_1 with much improved precision, demonstrating the effectiveness of incorporating knowledge-attention into self-attention to jointly model the sentence based on both knowledge and data.

Performance gain is the lowest for multi-channel attention (MCA). However, the model is more flexible in the way that features from other information sources can be easily added to the

Model	P	R	F_1
MCA	68.4	66.1	67.3
+ NER	68.7	**66.2**	67.5
+ Entity category	70.0	65.8	67.8
+ Both	**70.1**	66.0	**67.9**

Table 2: Results of adding NER embeddings and entity categorical embeddings to the multi-channel attention (MCA) integrated model.

model to further improve its performance. Table 2 shows the results of adding NER embeddings of each token to self-attention channel, and entity (subject and object) categorical embeddings to multi-channel attention as additional feature vectors. We use dimensionality of 30 and 60 for NER and entity categorical embeddings respectively, and the two embedding matrixes are learned by the neural network. Results show that adding NER and entity categorical information to MCA integrated model improves F_1 score by 0.2 and 0.5 respectively, and adding both improves precision significantly, resulting a new best F_1 score.

5.3.2 Results on SemEval2010-Task8 dataset

We use SemEval2010-Task8 dataset to evaluate the generalization ability of our proposed model. Experiments are conducted in two manners: mask or keep the entities of interest. Results in Table 3 show that the "entity mask" strategy degrades the performance, indicating that there exist strong correlations between entities of interest and relation classes in SemEval2010-Task8 dataset. Although the results of keeping the entities are better, the model tends to remember these entities instead of focusing on learning the linguistic clues of relations. This will result in bad generalization for sentences with unseen entities.

Regardless of whether the entity mask is used, by incorporating knowledge-attention mechanism, our model improves the performance of self-attention by a statistically significant margin, especially the softmax interpolation integrated model. The results on SemEval2010-Task8 are consistent with that of TACRED, demonstrating the effectiveness and robustness of our proposed method.

5.3.3 Ablation study

To study the contributions of specific components of knowledge-attention encoder, we perform ablation experiments on the dev set of TACRED. The results of knowledge-attention encoder with and without certain components are shown in Table 4.

Model	mask entity	keep entity
Self-attn	76.8	83.1
Knwl-attn	76.1	82.3
Knwl+Self (MCA)	77.4*	84.0*
Knwl+Self (SI)	**78.0***	**84.3***
Know+Self (KISA)	77.5*	84.0*

Table 3: Macro-averaged F_1 score on SemEval2010-Task8 dataset. $*$ marks statistically significant improvements over Self-attn with $p < 0.01$ under one-tailed t-test.

Model	Dev F_1
Knwl-attn Encoder	66.5
1. − Multi-head structure	64.6
2. − Synonym relation indicators	64.7
3. − Relation indicators mean	65.0
4. − Output masking	65.8
5. − Entity masking	65.4
6. − Relative positions	63.0

Table 4: Ablation study on knowledge-attention encoder. Results are the median F_1 scores of 5 independent runs on dev set of TACRED.

It is observed that: (1) The proposed multi-head knowledge-attention structure outperforms single-head significantly. This demonstrates the effectiveness of jointly attending texts to different relational semantic subspaces in the multi-head structure. (2) The synonyms improve the performance of knowledge-attention since they are able to broaden the coverage of relation indicators and form a robust relational semantic space. (3) The subtraction of relation indicators mean vector from attention hidden representations helps to suppress the activation of irrelevant words and results in a better representation for each word to capture the linguistic clues of relations. (4-5) The two masking strategies are helpful for our model: the output masking eliminates the effects of the padding tokens and the entity masking avoids entity overfitting while providing entity type information. (6) The relative position embedding term in position-aware attention contributes a significant amount of F_1 score. This shows that positional information is particularly important for relation extraction task.

5.3.4 Attention visualization

To verify the complementary effects of knowledge-attention encoder and self-attention

Sample Sentences	True Relation	Predict
SUBJ-PERSON graduated in 1992 from the OBJ-ORGANIZATION OBJ-ORGANIZATION OBJ-ORGANIZATION with a degree in computer science and had worked as a systems analyst at a Pittsburgh law firm since 1999 .	per:schools_attended	correct
SUBJ-PERSON graduated in 1992 from the OBJ-ORGANIZATION OBJ-ORGANIZATION OBJ-ORGANIZATION with a degree in computer science and had worked as a systems analyst at a Pittsburgh law firm since 1999 .		correct
OBJ-PERSON OBJ-PERSON , a public affairs and government relations strategist , was executive director of the SUBJ-ORGANIZATION SUBJ-ORGANIZATION Policy Institute from 2005 to 2010 .	org:top_members/employees	correct
OBJ-PERSON OBJ-PERSON , a public affairs and government relations strategist , was executive director of the SUBJ-ORGANIZATION SUBJ-ORGANIZATION Policy Institute from 2005 to 2010 .		wrong
Founded in 1992 in Schaumburg , Illinois , the SUBJ-ORGANIZATION is one of the largest Chinese - American associations of professionals in the OBJ-COUNTRY OBJ-COUNTRY .	org:country_of_headquarters	wrong
Founded in 1992 in Schaumburg , Illinois , the SUBJ-ORGANIZATION is one of the largest Chinese - American associations of professionals in the OBJ-COUNTRY OBJ-COUNTRY .		correct

Table 5: Attention visualization for knowledge-attention encoder (first) and self-attention encoder (second). Words are highlighted based on the attention weights assigned to them. Best viewed in color.

encoder, we compare the attention weights assigned to words from the two encoders. Table 5 presents the attention visualization results on sample sentences. For each sample sentence, attention weights from knowledge-attention encoder are visualized first, followed by self-attention encoder. It is observed that knowledge-attention encoder focuses more on the specific keywords or cue phrases of certain relations, such as "graduated", "executive director" and "founded"; while self-attention encoder attends to a wide range of words in the sentence and pays more attention to the surrounding words of target entities especially the words indicating the syntactic structure, such as "is", "in" and "of". Therefore, knowledge-attention encoder and self-attention encoder have complementary strengths that focus on different perspectives for relation extraction.

5.3.5 Error analysis

To investigate the limitations of our proposed model and provide insights for future research, we analyze the errors produced by the system on the test set of TACRED. For knowledge-attention encoder, 58% errors are false negative (FN) due to the limited ability in capturing long-distance dependencies and some unseen linguistic clues during training. For our integrated model[4] that takes the benefits of both self-attention and knowledge-attention, FN is reduced by 10%. However, false positive (FP) is not improved due to overfitting that leads to wrong predictions. Many errors are caused by multiple entities with different relations co-occurred in one sentence. Our model may mistake irrelevant entities as a relation pair. We also observed that many FP errors are due to the confusions between related relations such as "city_of_death"and "city_of_residence". More data or knowledge is needed to distinguish "death" and "residence". Besides, some errors are caused by imperfect annotations.

6 Conclusion and Future Work

We introduce knowledge-attention encoder which effectively incorporates prior knowledge from external lexical resources for relation extraction. The proposed knowledge-attention mechanism transforms texts from word space into relational semantic space and captures the informative linguistic clues of relations effectively. Furthermore, we show the complementary strengths of knowledge-attention and self-attention, and propose three different ways of integrating them to maximize the utilization of both knowledge and data. The proposed models are fully attention-based end-to-end systems and achieve state-of-the-art results on TACRED dataset, outperforming existing CNN, RNN, and self-attention based models.

In future work, besides lexical knowledge, we will incorporate conceptual knowledge from encyclopedic knowledge bases into knowledge-attention encoder to capture the high-level semantics of texts. We will also apply knowledge-attention in other tasks such as text classification, sentiment analysis and question answering.

[4]We observed similar error behaviors of the three proposed integrated models.

References

Sungjin Ahn, Heeyoul Choi, Tanel Pärnamaa, and Yoshua Bengio. 2016. A neural knowledge language model. *arXiv preprint arXiv:1608.00318*.

Ivan Bilan and Benjamin Roth. 2018. Position-aware self-attention with relative positional encodings for slot filling. *arXiv preprint arXiv:1807.03052*.

Muhao Chen, Yingtao Tian, Xuelu Chen, Zijun Xue, and Carlo Zaniolo. 2018a. On2vec: Embedding-based relation prediction for ontology population. In *Proceedings of the 2018 SIAM International Conference on Data Mining*, pages 315–323. SIAM.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2018b. Neural natural language inference models enhanced with external knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2406–2417.

Yun-Nung Chen, Dilek Hakkani-Tur, Gokhan Tur, Asli Celikyilmaz, Jianfeng Gao, and Li Deng. 2016. Knowledge as a teacher: Knowledge-guided structural attention networks. *arXiv preprint arXiv:1609.03286*.

Zhigang Chen, Wei Lin, Qian Chen, Xiaoping Chen, Si Wei, Hui Jiang, and Xiaodan Zhu. 2015. Revisiting word embedding for contrasting meaning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 106–115.

Jinhua Du, Jingguang Han, Andy Way, and Dadong Wan. 2018. Multi-level structured self-attentions for distantly supervised relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2216–2225.

Marco Fossati, Emilio Dorigatti, and Claudio Giuliano. 2018. N-ary relation extraction for simultaneous t-box and a-box knowledge base augmentation. *Semantic Web*, 9(4):413–439.

Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun, and Peng Li. 2018. Hierarchical relation extraction with coarse-to-fine grained attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2245.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 94–99. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2017. Mit at semeval-2017 task 10: Relation extraction with convolutional neural networks. *arXiv preprint arXiv:1704.01523*.

Pengfei Li and Kezhi Mao. 2019. Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. *Expert Systems with Applications*, 115:512–523.

Quan Liu, Hui Jiang, Si Wei, Zhen-Hua Ling, and Yu Hu. 2015. Learning semantic word embeddings based on ordinal knowledge constraints. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1501–1511.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1105–1116.

Thien Huu Nguyen and Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 39–48.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Longhua Qian, Guodong Zhou, Fang Kong, Qiaoming Zhu, and Peide Qian. 2008. Exploiting constituent dependencies for tree kernel-based semantic relation extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 697–704. Association for Computational Linguistics.

Bryan Rink and Sanda Harabagiu. 2010. Utd: Classifying semantic relations by combining lexical and semantic resources. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 256–259. Association for Computational Linguistics.

Josef Ruppenhofer, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher R Johnson, and Jan Scheffczyk. 2006. Framenet ii: Extended theory and practice.

Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying relations by ranking with convolutional neural networks. In *Proceedings of the*

53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 626–634.

Ying Shen, Yang Deng, Min Yang, Yaliang Li, Nan Du, Wei Fan, and Kai Lei. 2018. Knowledge-aware attentive neural network for ranking question answer pairs. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 901–904. ACM.

Roberta A Sinoara, Jose Camacho-Collados, Rafael G Rossi, Roberto Navigli, and Solange O Rezende. 2019. Knowledge-enhanced document embeddings for text classification. *Knowledge-Based Systems*, 163:955–971.

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

Fabian M Suchanek, Georgiana Ifrim, and Gerhard Weikum. 2006. Combining linguistic and statistical analysis to extract relations from web documents. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 712–717. ACM.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. 2015. Representing text for joint embedding of text and knowledge bases. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 872–884.

Ngoc Thang Vu, Heike Adel, Pankaj Gupta, et al. 2016. Combining recurrent and convolutional neural networks for relation classification. In *Proceedings of NAACL-HLT*, pages 534–539.

Jin Wang, Zhongyuan Wang, Dawei Zhang, and Jun Yan. 2017. Combining knowledge with deep convolutional neural networks for short text classification. In *IJCAI*, pages 2915–2921.

Linlin Wang, Zhu Cao, Gerard De Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention cnns.

Yan Xu, Ran Jia, Lili Mou, Ge Li, Yunchuan Chen, Yangyang Lu, and Zhi Jin. 2016. Improved relation classification by deep recurrent neural networks with data augmentation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1461–1470.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Mo Yu, Wenpeng Yin, Kazi Saidul Hasan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2017. Improved neural relation detection for knowledge base question answering. *arXiv preprint arXiv:1704.06194*.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, et al. 2014. Relation classification via convolutional deep neural network.

Dongxu Zhang and Dong Wang. 2015. Relation classification via recurrent neural network. *arXiv preprint arXiv:1508.01006*.

Yuhao Zhang, Peng Qi, and Christopher D Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. 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)*, volume 2, pages 207–212.