Graph Enhanced Dual Attention Network for Document-Level Relation Extraction

Bo Li[1,2], **Wei Ye**[1]*, **Zhonghao Sheng**[1,2], **Rui Xie**[1,2], **Xiangyu Xi**[1,2], **Shikun Zhang**[1]
[1]National Engineering Research Center for Software Engineering, Peking University
[2]School of Software and Microelectronics, Peking University
deepblue.lb@stu.pku.edu.cn,
{wye, zhonghao.sheng, ruixie, xixy, zhangsk}@pku.edu.cn

Abstract

Document-level relation extraction requires inter-sentence reasoning capabilities to capture local and global contextual information for multiple relational facts. To improve inter-sentence reasoning, we propose to characterize the complex interaction between sentences and potential relation instances via a Graph Enhanced Dual Attention network (GEDA). In GEDA, sentence representation generated by the sentence-to-relation (S2R) attention is refined and synthesized by a Heterogeneous Graph Convolutional Network before being fed into the relation-to-sentence (R2S) attention . We further design a simple yet effective regularizer based on the natural duality of the S2R and R2S attention, whose weights are also supervised by the supporting evidence of relation instances during training. An extensive set of experiments on an existing large-scale dataset show that our model achieves competitive performance, especially for the inter-sentence relation extraction, while the neural predictions can also be interpretable and easily observed.

1 Introduction

Relation extraction (RE) is an important research topic in natural language processing with plenty of applications. The task of RE is to detect the relationship from a given context and target entities. Depending on the given context, RE can be divided into two types: sentence-level RE (Zeng et al., 2015; Zhou et al., 2016; Ji et al., 2017; He et al., 2018; Fu et al., 2019; Guo et al., 2019; Ye et al., 2019) and document-level RE (Sahu et al., 2019; Gupta et al., 2019; Christopoulou et al., 2019; Wang et al., 2019). Compared with widely-studied sentence-level RE, document-level RE is a more challenging task that requires more investigations.

One of main challenges in document-level RE is how to perform inter-sentence reasoning over a long document to synthesize local and global information for potential relation facts. According to the statistics of DocRed (Yao et al., 2019), a large-scale human-annotated document-level RE dataset, there are 40.7% relational facts can only be extracted from multiple sentences in this dataset. That means a desirable neural model for document-level RE requires sophisticated inter-sentence reasoning capabilities.

In this paper, we propose to improve inter-sentence reasoning from a perspective of better characterizing the complex interaction between multiple sentences and multiple potential relation instances in a document. Since one relation instance may be expressed by multiple sentences and one sentence may reveal multiple relational facts (or parts of relational facts), it is natural and straightforward to use attention between sentences and potential relation instances to capture the complex many-to-many interaction. To this end, we introduce a bi-directional attention mechanism consisting of the attention paid by a sentence to relation instances (sentence-to-relation, S2R) and the attention paid by a relation instance to sentences (relation-to-sentence, R2S). Though bi-directional attention has been widely used in other NLP tasks (e.g., machine comprehension (Seo et al., 2017) and sentiment analysis (Zhao et al., 2020), what makes

* Corresponding author.

Proceedings of the 28th International Conference on Computational Linguistics, pages 1551–1560
Barcelona, Spain (Online), December 8-13, 2020

our architecture unique and innovative is the following three designs based on the classic bi-directional attention.

- **Graph-enhancing operation**. Sentences that express a specific relational fact may be located in different parts of a document, e.g., with a long distance. Cross-sentence information synthesizing with many noisy sentences in-between may not be accurate enough if we generate sentence representation with a classic attention mechanism. Since sentences and entities in a document naturally form a graph with rich semantic, we refine and synthesize the sentence representation generated by S2R attention via a Heterogeneous Graph Convolutional Network before feeding them into the R2S attention layer, to generate more accurate representation of potential relation instances. As demonstrated in the experiments, this graph-enhancing operation profits inter-sentence reasoning significantly.

- **Regularizer of attention duality**. Intuitively, the more attention a sentence pays to a relation instance, the more supporting evidence the sentence contains for the relation instance. Conversely, the relation instance should also pay more attention to the sentence to obtain a more accurate representation. This observation inspires us that there is a duality between S2R and R2S attention. The natural duality can provide our architecture with a useful induction bias as a simple and effective regularizer.

- **Attention supervision from supporting evidence**. We have achieved an architecture with graph-enhanced dual attention by the above two novel designs. Normally, the attention weights are trained implicitly with the signal from the ground truth of relation instances. We further leverage the supporting evidence for relation instance as a supervised signal for the weights of R2S attention, which also provides our model with more interpretability.

To evaluate our approach, We carried out extensive experiments on the DocRED dataset (Yao et al., 2019). From the experiment result, we found that characterizing the interaction between sentences and relation instances by our graph-enhanced dual attention network could significantly improve the performance of document-level RE. Our main contributions are:

- We proposed a Graph Enhanced Dual Attention network (GEDA) for document-level relation extraction, which is capable of improving inter-sentence reasoning by better characterizing the complex interaction between sentences and potential relation instances.

- The novelty of GEDA lies in its three well-designed components consisting of a graph-enhancing operation, a regularizer of attention duality, and the attention supervision from support evidence, which are proved to be effective on improving the performance of document-level RE and providing sound interpretability.

2 Proposed Approach

In this section, we present the GEDA in detail. The overall architecture is shown in Fig.1. Given a document D contains n words, m sentences, and k different entities, there will be $k \cdot (k-1)$ potential relation instances, our goal is to extract all the potential relation instances in a parallel way. GEDA mainly consists of 5 components: 1) the encoding layer, which generates the preliminary representations of sentences, entities and relation instances; 2) the graph-enhanced bi-directional attention layer, which synthesizes the intra-sentence and inter-sentence information to generate a refined representation of relation instances; 3) the constraint of attention duality as a regularizer; 4) the evidence-supervision loss; 5) and the final classification layer, which projects the representation of a potential relation instance to the probabilities for each relation type.

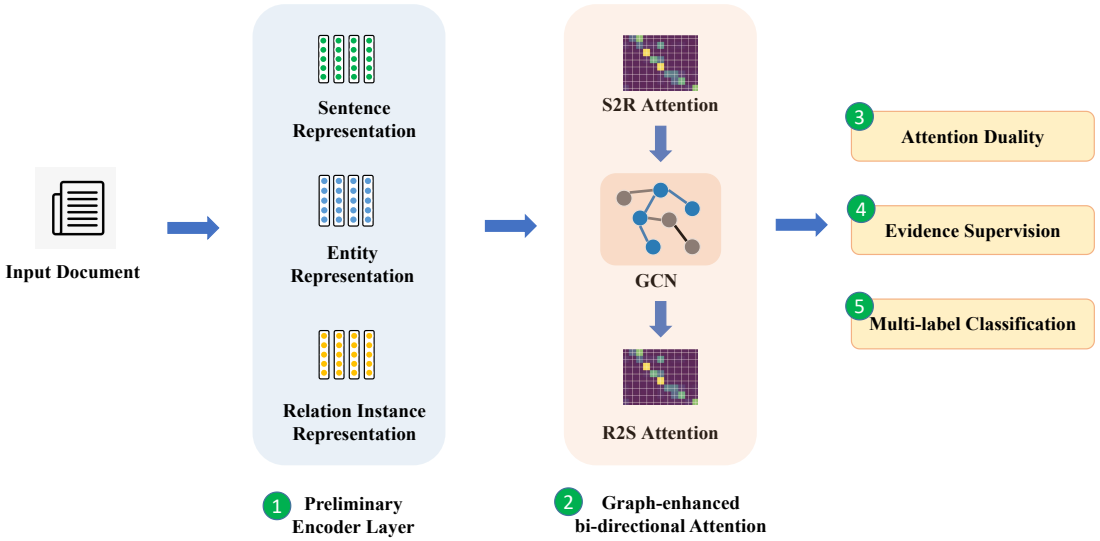


Figure 1: The overview architecture of GEDA. S2R attention means attention paid by a sentence to potential relation instances and R2S attention means attention paid by a relation instance to sentences in a document.

2.1 Encoding Layer

The encoding layer first converts the input document into a real-valued matrix, which contains three types of embeddings: 1) the word embedding; 2) the entity type embedding, which indicates the entity type information of each word; 3) and the entity order embedding, which represents the order of its first appearance in the document (Yao et al., 2019). An BiLSTM layer with h hidden units is used as an encoder to extract the semantic information, and the output of BiLSTM is denoted as the semantic representation H of a document, where $H \in \mathbb{R}^{n \times 2h}$. We then generate preliminary representations of sentences, entities and relation instances based on H.

Preliminary representations of sentences. We use max-pooling to obtain the preliminary representation for each sentence. Here we use l_i to denotes the preliminary representation of i-th sentence and $l_i \in \mathbb{R}^{1 \times 2h}$.

Preliminary representations of entities. Since there may be several entity mentions existing in the document for a given entity, to obtain the preliminary entity representation, we first extracts entity mention representations from H. For an entity mention ranging from the a-th to b-th word, the current entity mention representation for entity t_j is calculated as $\hat{t_j} = \frac{1}{b-a+1} \sum_{loc=a}^{b} H_{loc}$. And e_j is the average of all entity mention vectors of j-th entity, where $e_j \in \mathbb{R}^{1 \times 2h}$.

Preliminary representations of relation instances.We generates the preliminary relation representation for every entity pair $< e_p, e_q >$ using a bilinear function, where $p, q \in [1, k]$ and $p \neq q$. For the $k \cdot (k-1)$ potential relation instances, there will be $k \cdot (k-1)$ vectors generated. we stack them together as an preliminary relation instance representation $\mathbf{T}$, where $\mathbf{T} \in \mathbb{R}^{k \cdot (k-1) \times d}$.

2.2 Graph-Enhanced Bi-directional Attention

The graph-enhanced bi-directional attention layer aims to model the complex interactions between sentences and relation instances, which generates refined representation of relation instance by synthesizing both intra-sentence and inter-sentence information. This component consists of the S2R layer, the GCN layer, and the R2S layer.

2.2.1 S2R Layer

The S2R layer outputs the relation-oriented representation of sentence, where the query vector is the preliminary sentence representation l_i, and the key vector v_j is each row of tensor $\mathbf{T}$. The weight of the

attention paid by the sentence i to the relation instance j is denoted as α_{ij} and computed as follows:

$$\alpha_{ij} = \frac{exp(w_{ij})}{\sum_{j=1}^{k\cdot(k-1)} exp(w_{ij})}. \tag{1}$$

where $w_{ij} = l_i \cdot W_1 \cdot v_j$, $l_i \in \mathbb{R}^{1\times 2h}$. We get the weighted combination all the attention vectors calculated over each row in $\mathbf{T}$ as relation-oriented representation $\bar{l}_i$ of sentence i , where $\overline{l_i} = \sum_{j=1}^{k(k-1)} \alpha_{ij} \cdot v_j$. Via S2R layer, we obtain an attention weight matrix $\mathbf{W}_{S2R} \in \mathbb{R}^{m\times k\cdot(k-1)}$.

2.2.2 GCN Layer

In GEDA, we build a heterogeneous GCN with two types of nodes: entity nodes and sentence nodes. There are three different edges: 1) sentence-sentence edges, which link two sentence nodes if the two sentences contain the same entity; 2) entity-entity edges, which link two entity nodes if the two entities are co-occurrent in a sentence; 3) entity-sentence edges, which link an entity node and a sentence node if the entity resides in the sentence.

Since the entity representation e_i has different dimension with sentence representation $\bar{l}_j$, a matrix $W_2 \in \mathbb{R}^{2h\times d}$ is used to transform the e_i into $\overline{e}_i \in \mathbb{R}^{1\times d}$. Then the feature matrix $\mathbf{X}$ of GCN is computed as $\mathbf{X} = [\overline{e}_1; \overline{e}_2; ..., \overline{e}_k; \bar{l}_1; \bar{l}_2; ...; \bar{l}_m]$, where $\mathbf{X} \in \mathbb{R}^{(k+m)\times d}$. As for the adjacency matrix $\mathbf{A} \in \mathbb{R}^{(k+m)\times(k+m)}$, we set the diagonal elements to 1 since the self-loops. The weight of the edge is set to 1 if the edge exists between two nodes else 0. For a one-layer GCN, we can get the new node feature matrix $\mathbf{L} \in \mathbb{R}^{(m+k)\times s}$ by the following equation:

$$\mathbf{L} = \rho(\hat{\mathbf{A}}\mathbf{X}W_3), \hat{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{\frac{1}{2}}. \tag{2}$$

where $\hat{\mathbf{A}}$ is the normalized symmetric adjacency matrix and $W_3 \in \mathbb{R}^{d\times s}$ is a weight matrix. Output of GCN can be interpreted as two parts: 1) the refined entity representations as the first k rows in $\mathbf{L}$, denoted as $(\hat{e}_1, \hat{e}_2, ..., \hat{e}_k)$; 2) the refined sentence representation ranging from the $(k+1)$-th row to the $(k+m)$-th row in $\mathbf{L}$, denoted as $(\hat{l}_1, \hat{l}_2, ..., \hat{l}_m)$. Then the bilinear function is applied on refined entity representation to obtain refined relation instance representation $\hat{\mathbf{T}}$.

2.2.3 R2S Layer

Similar to S2R layer, R2S layer is used to obtain the sentence-oriented representation of relation instances, and the differences are in two aspects: 1) the query vector is $\hat{v}_i$ in $\hat{\mathbf{T}}$, which is the representation of each relation instance, and 2) the key vector is the representation $\hat{l}_j$ of each sentence. Finally, R2S layer outputs the sentence-oriented representation matrix $\tilde{\mathbf{T}}$ for all the potential relation instances, in which $\tilde{v}_i$ is the i-th row corresponding to the i-th relation instance. We also obtain a weight matrix $\mathbf{W}_{R2S} \in \mathbb{R}^{k\cdot(k-1)\times m}$.

2.3 Regularizer of Attention Duality

The attention paid by a sentence to a relation instance is generally consistent with the attention in the opposite direction paid by the relation instance to the sentence, which means there exists a natural duality between the two weight matrices $\mathbf{W}_{S2R}$ and $\mathbf{W}_{R2S}$. We leverage this duality to design a simple regularization term to introduce this useful induction bias. The mathematical expression of this regularizer is shown in the following, where $\|\cdot\|_2$ is the L2-regularization:

$$r_{duality} = \frac{1}{m \times k\cdot(k-1)} \|\mathbf{W}_{S2R} - \mathbf{W}_{R2S}^T\|_2. \tag{3}$$

2.4 Evidence Supervision

Supporting evidence information identifies which sentences contribute to a specific relation instance. We can transform this information into a real-valued vector. For example, given a document that has m

sentences, for the i-th relation instance, if the first two sentences are the supporting evidence, then the evidence vector is:

$$c_i = \underbrace{[0.5, 0.5, 0, ..., 0]}_{m}. \tag{4}$$

If a given relation instance can not be assigned to any relation type, the values in c_i are then all set to $1/m$.

Note that the i-th row in $\mathbf{W}_{R2S}$, termed as w_i, is the attention weights that i-th relation instance paid to all sentences. Intuitively, w_i should be close to the evidence vector to focus on most relevant sentences. Thus, we use Kullback-Leibler divergence (Kullback and Leibler, 1951) to measure the distribution differences between c_i and w_i as an extra loss. The loss of all potential relation instances in a document are as follows:

$$r_{evd} = \frac{1}{k \cdot (k-1)} \sum_{i=1}^{k \cdot (k-1)} D_{KL}(c_i | w_i). \tag{5}$$

where $D_{KL}(p|q) = \sum_x p(x) log \frac{p(x)}{q(x)}$.

2.5 Classification Layer

Since the relation prediction in our scenario is a multi-label problem, we use $\tilde{v}_i$, which is the i-th row in $\tilde{\mathbf{T}}$, to predict whether the i-th relation instance has the relation type r:

$$\hat{y}_r = \sigma(W_4 \tilde{v}_i + b_4). \tag{6}$$

where σ is the sigmoid function, W_4 and b_4 are the trainable parameters. Finally, for a given document contains m sentences, k different entities and t pre-defined relation types, the loss function is defined as:

$$loss = \sum_{i=1}^{k \cdot (k-1)} \sum_{r=1}^{t} y_r^i log(\hat{y}_r^i) + \alpha r_{duality} + \beta r_{evd} + \lambda \|\theta\|_2. \tag{7}$$

where y_r^i is a binary label of the i-th relation instance for the relation type r, θ is the parameters that need to be regularized, and $\|\theta\|_2$ is the L2-regularization, α, β, and λ are the coefficients.

3 Experiments

3.1 Dataset and Evaluation Metrics

The dataset we used is DocRED (Yao et al., 2019), which is a large-scale document-level RE dataset. DocRED has 3053 training documents, 1000 development documents and 1000 test documents, with 96 relation types. Note that an entity pair may be assigned by one or more relations in DocRED, we formulate RE as a multi-label classification problem in the experiment.

Following prior work(Yao et al., 2019), we use F1 and IgnF1 as the evaluation metrics, in which IgnF1 is calculated after removing the entity pairs that have appeared in the training set. Besides, to evaluate inter-sentence reasoning capabilities, we split the development set into two parts based on whether an entity pair exist in the same sentence. F1 on both splits are reported, named intra-F1 and inter-F1 respectively.

3.2 Baseline Models

In this paper, we compare GEDA with three types of models: 1) the vanilla models, including **CNN** (Zeng et al., 2014), **BiLSTM** (Cai et al., 2016) and **Context-Aware** (Sorokin and Gurevych, 2017); 2) graph-based models, such as **Graph LSTM** (Peng et al., 2017), **GCNN** (Sahu et al., 2019) and **EoG** (Christopoulou et al., 2019); 3) BERT-based models, including **BERT model** and **BERT-Two-Step model** (Wang et al., 2019). For comparison with BERT-based models, we build a variant of our model named BERT-GEDA, which uses Uncased BERT-Base (Devlin et al., 2019) as the encoder instead of the word embedding layer.

3.3 Experiment Settings

Most of the experiment settings are the same as (Yao et al., 2019). Specifically, 1) words initialized with 100 dimension Glove Embeddings (Pennington et al., 2014), and are fixed during training procedure; 2) the dimension of entity order embedding and entity type embedding are 20; 3) the optimizer is Adam (Kingma and Ba, 2015) with learning rate of 0.001; 4) the hidden size of LSTM is 128; 5) all coefficients are 1e-3; 6) as for a BERT based GEDA model (named BERT-GEDA), we use a transformation layer to project the BERT embedding of each word into a low-dimensional space of size 100, which is the same as the word embedding. The learning rate of the BERT-base model is 10^{-5}; 7) the batch size is 20.

Methods	Dev		Test		Dev	
	F1	IgnF1	F1	IgnF1	Intra-F1	Inter-F1
CNN	43.45*	37.99*	42.33*	36.44*	51.07	36.31
BiLSTM	50.95*	45.12*	51.06*	44.73*	56.81	43.36
Context-Aware	51.10*	44.84*	50.64*	43.93*	56.97	43.46
GraphLSTM	50.52	47.35	49.57	47.53	56.24	43.98
GCNN	51.45	47.42	50.44	47.82	57.75	44.52
EoG	51.03	48.11	51.25	48.48	57.33	44.31
GEDA w/o GCN (Ours)	52.51	48.28	52.01	48.59	58.05	45.34
GEDA (Ours)	**53.60**	**51.03**	**52.97**	**51.22**	**58.83**	**47.72**
BERT	54.16*	-	53.20*	-	60.12	47.53
BERT-Two-Step	54.42*	-	53.92*	-	60.32	47.60
BERT-GEDA (Ours)	**56.16**	**54.52**	**55.74**	**53.71**	**61.85**	**49.46**

Table 1: Performance of different models and GEDA on the DocRED. $*$ means the results are reported from (Yao et al., 2019) or (Wang et al., 2019), others are reproduced by ourselves. Besides, we also present the intra-F1 and inter-F1 for further analysis. The significance tests are conducted for testing the robustness of approaches.

3.4 Main Results

The experimental results for all models are shown in Table 1, from which we can observe that:

1. GEDA (Ours) and BERT-GEDA (Ours) outperform other proposed models significantly, showing the effectiveness of our graph enhanced dual attention network. For example, compared with the highest score among all previous none-BERT models, GEDA (Ours) enhances F1 for 2.15% and IgnF1 for 2.92% on the development set, and F1 for 1.72% and IgnF1 for 2.74% on the test dataset. Note that BERT-based methods leverage external knowledge from large-scale corpus and greatly improve RE. Our proposed BERT-GEDA (Ours) can enhance BERT-based methods with reasoning capacities, thus further gains significant improvement over BERT-based methods. This observation verifies that GEDA can bring consistent and robust improvement, even over very competitive baseline models.

2. Context-Aware has a similar performance with BiLSTM, though it employs an attention mechanism based on BiLSTM. Compared with Context-Aware, however, GEDA (Ours) improves F1-score by over 2.5%. This indicates that document-level RE requires a more sophisticated attention mechanism to handle the complex interaction between sentences and relation instances, and our technique is more suitable for document-level RE. We attribute this superiority to the well-designed graph-enhancing operation and effective regularizers over a classic bi-directional attention mechanism.

3. GNN-based methods achieve better performance over vanilla methods especially on Inter-F1, which shows that GNN is capable of characterizing the document-level context and thus learn the latent patterns of relations across sentences better.

3.5 Analysis of Inter-Sentence Reasoning

To investigate the reasoning ability of different methods, we report their performance on both intra-sentence and inter-sentence instances on development set, as the last two columns of Table 1 shows.

Taking none-BERT models as an example, GEDA (Ours) outperforms other models in both scenarios, and the relative improvement in inter-sentence is more significant than intra-sentence cases (3.20% of inter-F1 v.s. 1.08% of intra-F1). The experimental results of BERT-based model follow a similar trend. The results verifies that modeling the complex interaction between sentences and relation instances by our graph-enhanced dual attention can enrich inter-sentence reasoning skills. Besides, the improvements in intra-sentence cases are also notable, which reveals that contextual information is also useful to identify intra-sentence relational facts and GEDA (Ours) can well capture it.

3.6 Effects of Core Components

Setting	F1	IgnF1	Inter-F1
GEDA	53.60	51.03	47.72
w/o Attention Duality	52.62	49.71	46.54
w/o Evident Supervision	53.14	50.50	47.09
w/o GCN	52.51	48.28	45.34

Table 2: Ablation study of core components.

Setting	Average Weight
GEDA	0.83
w/o Evident Supervision	0.65
w/o Attention Duality	0.58

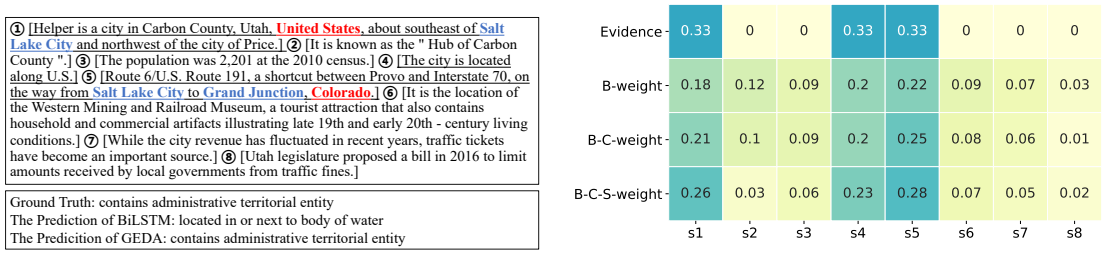
Table 3: Analysis of regularizers on development set.

To evaluate the effect of three core components, we test three GEDA variants by remove attention duality, evident supervision, and GCN, respectively. Due to space limitations, we only report the results of F1, IngF1, and Inter-F1 on the developing set. Results in Table 2 shows that three components all yield significant enhancements, verifying the effectiveness of the three novel designs based on the classic bi-directional attention mechanism. Meanwhile, the performance decrements of all GEDA variants on IngF1 and inter-F1 are much more notable than that on F1. This observation demonstrated that all three components not only improve inner-sentence reasoning but also provide GEDA a better generalization ability to predict the unseen potential relation instances more accurately.

3.7 Analysis of Attention Weights

We further investigate the impact of the two regularizers on attention weights. For each relation instance, we calculate the sum of R2S attention weights on the relevant sentences (supporting evidence), then get their average value over all the relation instances. As Table 3 shows, average weight value of GEDA (Ours) is larger than value of other two scenarios. The results indicate the two regularizers can both help GEDA pay more attention to the relevant sentences.

3.8 Case Study and Visualization



(a) Case study. (b) Visualization

Figure 2: (a) Case study. The document contains 8 sentences. Words in red are target entities, words in blue are non-target entities that are useful for inter-sentence inference, and the evidence sentences are underlined. (b) Visualization of different attention weights and sentence evidence information. The deeper color means the higher weight.

To promote understanding of the neural predictions, we use a document with 8 sentences from DocRED for case study, as shown in Fig.2 (a). There are 3 relevant sentences for the entity pair {United States, Colorado}. Due to lack of reasoning ability over multiple sentences, BiLSTM mislabels the instance, while GEDA can perform inter-sentence reasoning thus predicts a correct relation for the entity pair. Meanwhile, we visualize the R2S attention weights generated by GEDA. As shown in Fig.2(b), the

"Evidence" row shows the ground-truth attention weights for sentences within the document. B-weight is generated by bi-directional attention, B-C-weight is generated by bi-directional attention with duality constraint and B-C-S-weight is the attention weights generated by further adding supporting evidence. As shown in Fig.2(b), by incorporating duality constraints and supervised attention, correct sentences are assigned with more weights. The visualization results not only verifies the effectiveness of the regularizers, but also reveals the interpretability of our proposed GEDA.

4 Related Work

For document-level RE, (Peng et al., 2017) used graph-LSTM networks to cross-sentence n-ary RE, (Gupta et al., 2019) models inter-sentential dependency through a similar graph. (Nguyen and Verspoor, 2018) applied CNN to cross-sentence RE with improved character encoding. (Verga et al., 2018) used a modified Transformer(Vaswani et al., 2017) with CNN, followed by a bi-affine pairwise scoring prediction, and applying distant supervision and multi-task learning. (Jia et al., 2019) addressed document-level n-ary RE with the design of multi-scale representation learning, which aims to learn the representation of entity tuples at both mention-level and entity-level. (Sahu et al., 2019) used GCN as encoder and a multi-instance based classifier. (Christopoulou et al., 2019) applied GCN with a graph consisted of different types of nodes and edges, aiming to infer those representations from other edges. Unlike previous work, we focus on inter-sentence reasoning and model the interaction between relation instances and sentences.

5 Conclusion

We have introduced our neural architecture GEDA for document-level relation extraction. The intuition behind GEDA is to characterize the interaction between sentences and relation instances better to improve inter-sentence reasoning over the whole document. The novelty of GEDA mainly lies in the graph-based refinement of sentence representation and two simple yet effective regularizers based on attention duality and supporting evidence respectively. Experiments verified the superiority of the graph-enhanced dual attention mechanism, especially for the inter-sentence relation extraction. In the future, we will investigate more sophisticated reasoning techniques targeting more specific scenarios of inter-sentence relation extraction, e.g., involving common-sense reasoning.

6 Acknowledgements

This research was supported by the National Key Research And Development Program of China (No. 2019YFB1405802).

References

Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. Bidirectional recurrent convolutional neural network for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4924–4935. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. 2019. Graphrel: Modeling text as relational graphs for joint entity and relation extraction. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the*

57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 1409–1418. Association for Computational Linguistics.

Zhijiang Guo, Yan Zhang, and Wei Lu. 2019. Attention guided graph convolutional networks for relation extraction. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 241–251. Association for Computational Linguistics.

Pankaj Gupta, Subburam Rajaram, Hinrich Schütze, and Thomas A. Runkler. 2019. Neural relation extraction within and across sentence boundaries. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6513–6520. AAAI Press.

Zhengqiu He, Wenliang Chen, Zhenghua Li, Meishan Zhang, Wei Zhang, and Min Zhang. 2018. SEE: syntax-aware entity embedding for neural relation extraction. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5795–5802. AAAI Press.

Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In Satinder P. Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3060–3066. AAAI Press.

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multiscale representation learning. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3693–3704. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*.

Dat Quoc Nguyen and Karin Verspoor. 2018. Convolutional neural networks for chemical-disease relation extraction are improved with character-based word embeddings. In Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors, *Proceedings of the BioNLP 2018 workshop, Melbourne, Australia, July 19, 2018*, pages 129–136. Association for Computational Linguistics.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms. *Trans. Assoc. Comput. Linguistics*, 5:101–115.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Sunil Kumar Sahu, Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Inter-sentence relation extraction with document-level graph convolutional neural network. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4309–4316. Association for Computational Linguistics.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Daniil Sorokin and Iryna Gurevych. 2017. Context-aware representations for knowledge base relation extraction. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1784–1789. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 872–884. Association for Computational Linguistics.

Hong Wang, Christfried Focke, Rob Sylvester, Nilesh Mishra, and William Wang. 2019. Fine-tune bert for docred with two-step process. *CoRR*, abs/1909.11898.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 764–777. Association for Computational Linguistics.

Wei Ye, Bo Li, Rui Xie, Zhonghao Sheng, Long Chen, and Shikun Zhang. 2019. Exploiting entity BIO tag embeddings and multi-task learning for relation extraction with imbalanced data. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1351–1360. Association for Computational Linguistics.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In Jan Hajic and Junichi Tsujii, editors, *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 2335–2344. ACL.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1753–1762. The Association for Computational Linguistics.

Pinlong Zhao, Linlin Hou, and Ou Wu. 2020. Modeling sentiment dependencies with graph convolutional networks for aspect-level sentiment classification. *Knowl. Based Syst.*, 193:105443.

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, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. The Association for Computer Linguistics.