Key Mention Pairs Guided Document-Level Relation Extraction

Feng Jiang[1], **Jianwei Niu**[1]*, **Shasha Mo**[2]*, **Shengda Fan**[2]
[1] State Key Laboratory of Virtual Reality Technology and Systems,
School of Computer Science and Engineering, Beihang University, Beijing 100191, China
[2] School of Cyber Science and Technology, Beihang University, Beijing 100191, China
{jfeng, niujianwei, moshasha, fanshengda}@buaa.edu.cn

Abstract

Document-level Relation Extraction (DocRE) aims at extracting relations between entities in a given document. Since different mention pairs may express different relations or even no relation, it is crucial to identify key mention pairs responsible for the entity-level relation labels. However, most recent studies treat different mentions equally while predicting the relations between entities, leading to sub-optimal performance. To this end, we propose a novel DocRE model called **K**ey **M**ention pairs **G**uided **R**elation **E**xtractor (KMGRE) to directly model mention-level relations, containing two modules: a mention-level relation extractor and a key instance classifier. These two modules could be iteratively optimized with an EM-based algorithm to enhance each other. We also propose a new method to solve the multi-label problem in optimizing the mention-level relation extractor. Experimental results on two public DocRE datasets demonstrate that the proposed model is effective and outperforms previous state-of-the-art models.

1 Introduction

Relation Extraction (RE), which aims to identify the relations between entities in a given text, has been explored at the sentence level for decades (Culotta and Sorensen, 2004; Zeng et al., 2014, 2015). However, according to Yao et al. (2019), a large amount of relations can only be identified across multiple sentences in the real-world scenarios. Therefore, researchers have recently turned to extracting relations directly in documents (Zeng et al., 2020; Zhou et al., 2021; Huang et al., 2021; Ru et al., 2021).

Document-level Relation Extraction (DocRE) encounters many new challenges compared to its sentence-level counterpart. A document may include numerous entities, and the same entity may appear multiple times in different sentences. It requires the DocRE models to recognize and focus on the part of the document that has relevant context for a particular entity pair. Many previous works solve the above problems by obtaining stronger context-aware entity pair representations. There are two main ways to achieve this: the graph-based methods (Guo et al., 2019; Nan et al., 2020; Zeng et al., 2020) and the sequence-based methods (Yao et al., 2019; Zhou et al., 2021). The graph-based methods construct a document graph and then use Graph Neural Networks (GNNs) to aggregate information across nodes. Besides, as Transformer (Vaswani et al., 2017) could be regarded as a fully connected GNN, the sequence-based methods attempt to directly use Transformer-based Pre-trained Language Models (PLMs) for DocRE without graph structure. The sequence-based methods generally use strong PLMs (e.g., BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019)) to model the input text and use different strategies to get entity pair representations, e.g., average pooling (Yao et al., 2019) and attentive pooling (Zhou et al., 2021).

However, despite these successful efforts, most existing methods still ignore the critical issue of treating different mentions equally in an entity pair. And it is at odds with the actual situation, as different mention pairs may express different relations or even no relation. For the example in Figure 1, multiple relations exist between *Genc Ruli* and *University of Tirana*, e.g., *employer* and *educated at*. These two relations can be inferred by different mention pairs of them, and at the same time, there are also several mention pairs don't express any relation. The multi-mention property of DocRE makes it difficult to establish context to relation mapping at the entity level directly. Therefore, equal treating all mentions ignores the difference between different mentions' contexts and may introduce irrelevant information to mislead model training.

*Corresponding author.

Proceedings of the 29th International Conference on Computational Linguistics, pages 1904–1914
October 12–17, 2022.

[s1] Genc Ruli (born April 11, 1958) is an Albanian politician. … [s3] Ruli holds a bachelor's degree in Economics and a bachelor's degree in Law from the University of Tirana. [s4] He holds a PhD in Economics from the Faculty of Economics, University of Tirana. [s5] Ruli is given the title Professor from the Faculty of Economics, University of Tirana. [s6] He has served as a Professor of Finance and Accounting in the Faculty of Economics, at the University of Tirana. [s7] Ruli has an extensive experience as the Minister of Finance and Economy in early 90's and as the Minister of Economy, Trade and Energy during 2005 - 2009. [s8] Ruli resigned from his position as Finance Minister on 9 November 1993, following allegations of corruption. … [s14] Ruli has written several publications in the areas of economics and public policies.

Subject: Genc Ruli **Object**: University of Tirana

Relation: employer, educated at

Figure 1: An example of multi-mention and multi-label problems from DocRED (Yao et al., 2019). Head entity *Genc Ruli* and tail entity *University of Tirana* express relations *employer* and *educated at*. This entity pair contains multiple mention pairs, and only several of them express relations. Other entities in the document is highlighted in grey.

To handle the multi-mention problem, we reformulate DocRE task as a Multiple Instance Learning (MIL) problem (Carbonneau et al., 2018) and propose a novel model called **K**ey **M**ention pairs **G**uided **R**elation **E**xtractor (KMGRE). Our approach consists of a mention-level relation extractor and a key instance classifier, which are iteratively trained to enhance each other. The relation extractor provides mention-level relation pseudo labels to help train the key instance classifier, and the key instance classifier distinguishes key mention pairs to improve relation extractor training. Those two modules can be efficiently optimized with the Expectation-Maximization (EM) algorithm (Neal and Hinton, 1998). By introducing key instances, KMGRE can effectively filter out mention pairs that do not express any relation to reduce the impact of redundant information.

Such a mention-level relation extractor suffers from the multi-label problem. It could be difficult to distinguish what kind of relation each mention pair expresses in multi-label situations, making generating the mention-level relation pseudo labels challenging. To alleviate the multi-label problem in optimizing the mention-level relation extractor, we propose to generate entity-level relation predictions by fusing mention-level predictions. Then we optimize our model's parameters with the entity-level relation labels. The contributions of this paper are summarized as follows:

- We regard the multi-mention problem in DocRE as a particular case of MIL and exptend a novel framework to directly model mention-level relations.
- We propose a new method to fuse the mention-level predictions. It could avoid the wrong guide to the model caused by false labeling mention pairs in the multi-label case.
- Experiments on two public DocRE datasets demonstrate that the proposed model is effective and outperforms previous state-of-the-art models.

2 Related Work

Sentence-level relation extraction has been explored for decades (Culotta and Sorensen, 2004; Zeng et al., 2014, 2015), but the relational facts that can only be extracted through multiple sentences cannot be handled well with traditional sentence-level relation extraction methods (Yao et al., 2019). For this reason, DocRE has attracted significant attention from researchers.

Most previous DocRE approaches focus on obtaining a strong contextual representation for each entity or entity pair. There are two main ways to achieve this: graph-based and sequence-based methods. The graph-based methods first construct a document graph and then use GNNs to model the interaction between different words and sentences. Guo et al. (2019) propose attention-guided GNNs to model full dependency trees of input documents and selectively attend to the useful dependencies. Nan et al. (2020) use a novel procedure to induce the latent document-level graph and perform multi-hop inference on the document graph. Zeng et al. (2020) construct two different levels of document graphs to aggregate information and combine the comprehensive inferential path information to infer relations.

As Transformer (Vaswani et al., 2017) could be regarded as a fully connected graph neural network, the sequence-based methods directly use Transformer-based PLMs (Devlin et al., 2019; Liu et al., 2019) to model the given text and get entity pair representations by different strategies, e.g., average pooling (Yao et al., 2019), max pooling (Li et al., 2021), and attentive pooling (Zhou et al.,

2021). However, most existing methods treat different mentions of each entity equally, which is counterintuitive, as different mention pairs may express different relations in a given document.

Some methods also consider the effect of different mentions. For instance, Christopoulou et al. (2019) put mention nodes into the document graph and use GNNs to gather different mentions' information. Li et al. (2021) propose to use convolution neural networks to capture the local mention-to-mention interactions. Eberts and Ulges (2021) propose to regard DocRE as a MIL problem and obtain entity pair representations by aggregating information from different mention pairs. However, the above methods don't consider that different mention pairs may express different relations or even no relation. And they treat different mention pairs equally in constructing the entity-level representation. Unlike previous methods, we directly model mention-level relations and further design a key instance classifier to distinguish those key mention pairs.

3 Methodology

This section introduces the proposed model KMGRE which directly extracts relations at the mention level. We first introduce the task formulation of DocRE. With a document $\mathcal{D}$ that contains a set of entities $\mathcal{E} = \{e_i\}_{i=1}^{n}$, the task is to extract relations between each entity pair (e_h, e_t), where $e_h, e_t \in \mathcal{E}$ are the head entity and the tail entity, respectively. An entity e_i may occur multiple times in the document, which could be defined as $\{m_j^i\}_{j=1}^{N_{e_i}}$, and therefore the mention pairs of (e_h, e_t) could be defined as $\boldsymbol{X} = \{(m_i, m_j) | m_i \in \{m_i^h\}_{i=1}^{N_{e_h}}, m_j \in \{m_j^t\}_{j=1}^{N_{e_t}}\}$. We denote m_i^h and m_j^t as m_i and m_j below for convenience. $\mathcal{C}$ is the set of pre-defined relation types. There exists a relation $c \in \mathcal{C}$ between e_h and e_t only if any pair of their mentions could express it.

It is intuitive to train a mention-level relation extractor and fuse its results to generate the entity-level relation label. However, instead of one mention pair being matched to one relation label, we only have the relation labels of entity pairs. Therefore, it could be challenging to train a mention-level relation extractor, directly. We propose to regard DocRE as a MIL problem and extend a novel probabilistic model to handle this issue as shown in Figure 2.

To identify the key mention pairs, we assign a binary variable $z \in \{0, 1\}$ to each mention pair, denoting whether it is responsible for the relation label of the entity pair. Inspired by EM-MIL (Luo et al., 2020), the relation label of (e_h, e_t) is generated with probability:

$$p(y_c = 1|\boldsymbol{X}, \boldsymbol{z}) = \max \{p(y_c = 1|m_i, m_j) \cdot \boldsymbol{I}\left(z_{(i,j)} = 1\right)\} \quad (1)$$

where m_i and m_j are mentions of e_h and e_t, respectively. $I(\cdot)$ is the indicator function. $y_c = 1$ if this entity pair (or mention pair) contains relation c, otherwise $y_c = 0$. We then design two modules, i.e., the mention-level relation extractor and the key instance classifier. These two modules are parameterized by θ and ω, and used to estimate the distribution $p_\theta(y_c = 1|m_i, m_j)$ and $p_\omega\left(z_{(i,j)} = 1|m_i, m_j\right)$, respectively.

The goal is to jointly train the relation extractor and the key instance classifier to maximize the likelihood of the training data. Formally, the objective function is presented as below:

$$\begin{aligned}\mathcal{O}(\theta, \omega) &= \mathbb{E}[\log p_{\theta,\omega}(y_c|\boldsymbol{X})] \\ &= \mathbb{E}[\log p_{\theta,\omega}(\boldsymbol{z}, y_c|\boldsymbol{X}) - \log p(\boldsymbol{z}|\boldsymbol{X}, y_c)]\end{aligned} \quad (2)$$

Since we do not know the true distribution of $\boldsymbol{z}$, it is difficult to directly optimize Equation 2. Following previous work (Luo et al., 2020), we optimize the above objective function by maximizing its variational lower bound:

$$\begin{aligned}&\log p_{\theta,\omega}(y_c|\boldsymbol{X}) \\ =&KL(p_\omega(\boldsymbol{z}|\boldsymbol{X})||p_\theta(\boldsymbol{z}|\boldsymbol{X}, y_c)) \\ &+ \int p_\omega(\boldsymbol{z}|\boldsymbol{X}) \log \frac{p_\theta(\boldsymbol{z}, y_c|\boldsymbol{X})}{p_\omega(\boldsymbol{z}|\boldsymbol{X})} d\boldsymbol{z} \\ \geq &\int p_\omega(\boldsymbol{z}|\boldsymbol{X}) \log p_\theta(\boldsymbol{z}, y_c|\boldsymbol{X}) d\boldsymbol{z} + H(p_\omega(\boldsymbol{z}|\boldsymbol{X}))\end{aligned} \quad (3)$$

where $H(p_\omega(\boldsymbol{z}|\boldsymbol{X}))$ is the entropy of p_ω. Therefore, we use an EM-based algorithm to optimize the objective function iteratively. In the E-step, we update ω by minimizing the KL divergence between $p_\omega(\boldsymbol{z}|\boldsymbol{X})$ and $p_\theta(\boldsymbol{z}|\boldsymbol{X}, y_c)$ to obtain a tighter lower bound. In the M-step, we update θ by maximizing the lower bound. Notably, unlike the previous work (Luo et al., 2020) that directly assigns the bag's label to each instance, we further propose a new optimization method to alleviate its limitations in the case of multi-label.

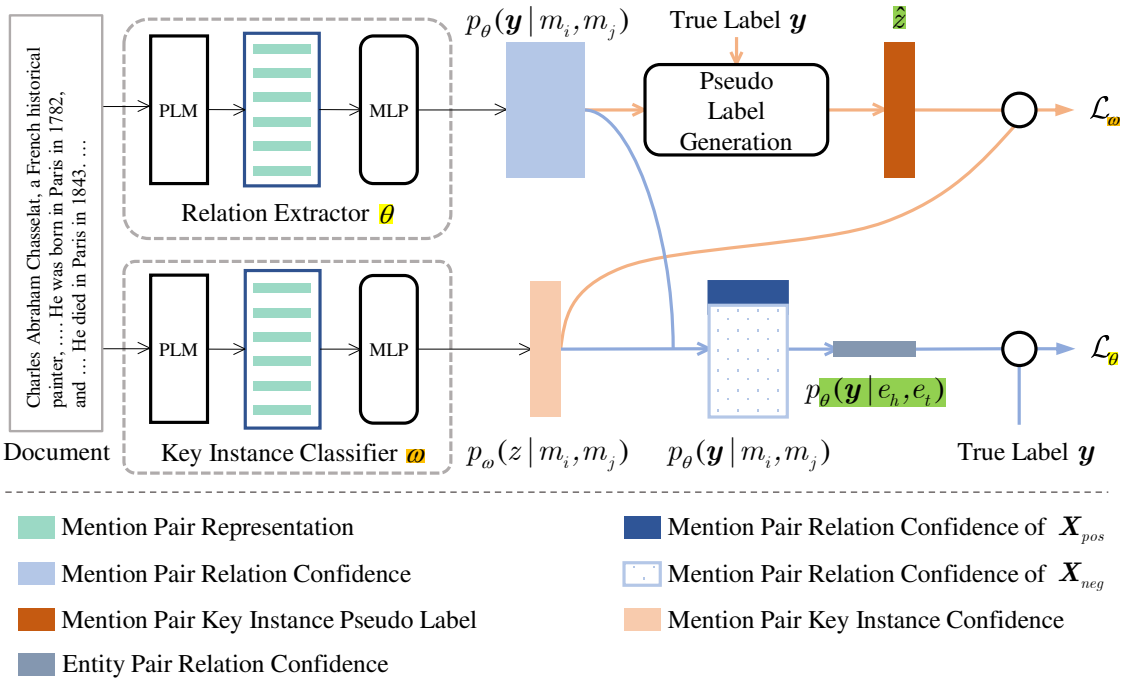


Figure 2: The overall architecture of KMGRE. We use two PLMs that do not share parameters to provide contextual embedding for the relation extractor and the key instance classifier. In the E-step, we update the parameters ω of the key instance classifier with $\mathcal{L}_\omega$; In the M-step, we update the parameters θ of the relation extractor with $\mathcal{L}_\theta$.

3.1 Parameterization

We use neural networks to parameterize the relation extractor and the key instance classifier. Specific details are described as follows.

Relation Extractor. Given an entity pair (e_h, e_t), the relation extractor generates relation probability distribution $p_\theta(y_c|m_i, m_j)$ for its mention pairs.

For a document of length ℓ, we first insert a special token "*" into every mention's start and end position. It is then fed into a PLM to obtain the contextual representation $\boldsymbol{H} \in \mathbb{R}^{\ell \times d}$ of each word, where d is the hidden dimension of the PLM. For a mention m_i, we take the representation of "*" at the start position as its embedding $\boldsymbol{h}_{m_i}$ and get its self-attention weight $\boldsymbol{A}_{m_i} \in \mathbb{R}^{H \times l}$ in H attention heads. m_j is similar to m_i. The contextual representation of mention pair (m_i, m_j) is calculated as:

$$\boldsymbol{c}^{(i,j)} = \boldsymbol{H}^\top \sum_{k=0}^{H} \frac{\boldsymbol{A}^k_{m_i} \cdot \boldsymbol{A}^k_{m_j}}{\mathbf{1}^\top \left(\boldsymbol{A}^k_{m_i} \cdot \boldsymbol{A}^k_{m_j}\right)}. \quad (4)$$

Then $\boldsymbol{c}^{(i,j)}$ is concatenated with the embedding of m_i and m_j to get the representation $\boldsymbol{x}^{(i,j)}$:

$$\boldsymbol{x}^{(i,j)} = [\boldsymbol{h}_{m_i}; \boldsymbol{h}_{m_j}; \boldsymbol{c}^{(i,j)}]. \quad (5)$$

We calculate the probability of relation c by a linear function and sigmoid activation:

$$p_\theta(y_c|m_i, m_j) = \sigma(\boldsymbol{w}_c \boldsymbol{x}^{(i,j)} + b_c) \quad (6)$$

where $\boldsymbol{w}_c \in \mathbb{R}^{3d}$ and $b_c \in \mathbb{R}$ are model parameters.

Key Instance Classifier. Since we only have the entity-level relation annotation, it is against intuition to directly train the above relation extractor. Therefore, we design this key instance classifier to generate the probability distribution $p_\omega(z_{(i,j)}|m_i, m_j)$, and assume the independence between different mention pairs. Moreover, we use this module to help train the relation extractor.

Like the above relation extractor, we use the same method to get the contextual embedding of (m_i, m_j) and concatenate it with $\boldsymbol{h}'_{m_i}$ and $\boldsymbol{h}'_{m_j}$:

$$x^{(i,j)'} = [\boldsymbol{h}'_{m_i}; \boldsymbol{h}'_{m_j}; \boldsymbol{c}^{(i,j)'}] \quad (7)$$

where the superscript $'$ means we use another PLM to get this embedding. We use two PLMs that do not share parameters to provide contextual embedding for the relation extractor and the key instance classifier, respectively, to avoid mutual interference during training.

We calculate the probability of (m_i, m_j) being a key instance by a linear function and sigmoid activation:

$$p_\omega(z_{(i,j)}|m_i, m_j) = \sigma(\boldsymbol{w}_k x^{(i,j)'} + b_k) \quad (8)$$

Algorithm 1 EM Optimization for $\mathcal{O}(\theta, \omega)$

Input: θ and ω, learning rate β, threshold control hyperparameter τ;

while not converged **do**
  for $(\boldsymbol{X}, \boldsymbol{y})$ in train set **do**
    Calculate the mention-level relation probability $p_\theta(y_c|m_i, m_j)$.
    Generate key instance pseudo label $\widehat{z}_{(i,j)}$ for all the mention pairs as Equation 9.
    Calculate the distribution of key instances $p_\omega(z_{(i,j)}|m_i, m_j)$.
    Calculate the E-step loss function $\mathcal{L}_\omega$ as Equation 11 and Equation 10.
    $\omega \leftarrow \omega - \beta \cdot \nabla_\omega \mathcal{L}_\omega$.
    Update the threshold control hyperparameter τ in Equation 13.
  end for ▷ E-step
  for $(\boldsymbol{X}, \boldsymbol{y})$ in train set **do**
    Calculate the distribution of key instances $p_\omega(z_{(i,j)}|m_i, m_j)$.
    Calculate the threshold $\widetilde{p}_\omega(z)$ as Equation 12 and Equation 13.
    Divide the mention pairs set $\boldsymbol{X}$ into $\boldsymbol{X}_{pos}$ and $\boldsymbol{X}_{neg}$ as Equation 14 and 15.
    Get the entity-level relation logit l_c as Equation 16.
    Calculate the M-step loss function $\mathcal{L}_\theta$ as Equation 19.
    $\theta \leftarrow \theta - \beta \cdot \nabla_\theta \mathcal{L}_\theta$.
  end for ▷ M-step
end while

where $\boldsymbol{w}_k \in \mathbb{R}^{3d}$ and $b_k \in \mathbb{R}$ are model parameters.

3.2 Optimization

Next, we introduce how we optimize the relation extractor and the key instance classifier to maximize the objective in Equation 2. We first train the relation extractor and the key instance classifier for several epochs before using the EM algorithm. Then at each iteration, the mention-level relation predictions and gold relation labels are first used to generate the key instance pseudo labels. After that, we update ω to minimize the KL divergence between $p_\omega(\boldsymbol{z}|\boldsymbol{X})$ and $p_\theta(\boldsymbol{z}|\boldsymbol{X}, y_c)$. Furthermore, we use the key instance predictions and gold relation labels to update θ and maximize the lower bound in Equation 3. The complete algorithm of KMGRE is shown in Algorithm 1, and the specifics are detailed below.

E-step. In the E-step, we first use the mention-level relation predictions and gold relation labels to generate the key instance pseudo labels $\widehat{\boldsymbol{z}}$ as below:

$$\widehat{z}_{(i,j)} = \begin{cases} 1, & \text{if } \exists\, c \in \mathcal{C}, \text{s.t. } y_c = 1 \wedge \\ & \quad p_\theta(y_c|m_i, m_j) \geq \overline{p}_\theta(y_c|e_h, e_t) \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where $\overline{p}_\theta(y_c|e_h, e_t) = \sum_{i,j} p_\theta(y_c|m_i, m_j)/(N_{e_h} \cdot N_{e_t})$ and y_c is the gold relation label of (e_h, e_t).

We update ω using binary focal loss (FC, (Lin et al., 2017)) as below:

$$\mathcal{L}_\omega = -\alpha_\omega (1 - p_\omega(z_{(i,j)}))^{\gamma_\omega} \log(p_\omega(z_{(i,j)})) \tag{10}$$

where α_ω and γ_ω are pre-defined hyperparameters. $p_\omega(z_{(i,j)})$ is defined below:

$$p_\omega(z_{(i,j)}) = \begin{cases} p_\omega(z_{(i,j)}|m_i, m_j), & \text{if } \widehat{z}_{(i,j)} = 1 \\ 1 - p_\omega(z_{(i,j)}|m_i, m_j), & \text{otherwise} \end{cases} \tag{11}$$

M-step. Unlike previous methods that directly label key mention pairs with the same label as entity pairs, we propose a new optimization method to alleviate the multi-label problem (e.g., the example in Figure 1 that the same entity pair may contain multiple relations). We fuse the mention-level relation results of key mention pairs to obtain the entity-level relation predictions and update θ by the entity-level relation extraction loss.

We first divide $\boldsymbol{X}$ into two different subsets $\boldsymbol{X}_{pos}$ and $\boldsymbol{X}_{neg}$ as below:

$$\overline{p}_\omega(z) = \sum_{i,j} \frac{p_\omega(z_{(i,j)})}{N_{e_h} \cdot N_{e_t}} \tag{12}$$

$$\widetilde{p}_\omega(z) = \min\big(\overline{p}_\omega(z) + \xi \cdot (\max\{p_\omega(z_{(i,j)})\} - \min\{p_\omega(z_{(i,j)})\}), \tau\big) \tag{13}$$

$$\boldsymbol{X}_{pos} = \{(m_i, m_j)|p_\omega(z_{(i,j)}) \geq \widetilde{p}_\omega(z)\} \tag{14}$$

$$\boldsymbol{X}_{neg} = \{(m_i, m_j)|p_\omega(z_{(i,j)}) < \widetilde{p}_\omega(z)\} \tag{15}$$

where $\xi > 0$ is set to control the degree of relaxation, $p_\omega(z_{(i,j)})$ means $p_\omega(z_{(i,j)}|m_i, m_j)$, and τ is a hyperparameter that increases gradually with the training process. The entity-level output logit of relation c is calculated as below:

$$l_c = \log \sum_{\boldsymbol{X}_{pos}} \exp(\boldsymbol{w}_c \boldsymbol{x}^{(i,j)} + b_c). \tag{16}$$

Following previous work (Zhou et al., 2021), we introduce a special relation class TH as the adaptive threshold and use the following loss function to update θ:

$$\mathcal{L}'_\theta = -\sum_{r \in \mathcal{P}_\mathcal{T}} \log \left(\frac{\exp(l_r)}{\sum_{r' \in \mathcal{P}_\mathcal{T} \cup TH} \exp(l_{r'})} \right) \quad (17)$$

$$\mathcal{L}''_\theta = -\log \left(\frac{\exp(l_{TH})}{\sum_{r' \in \mathcal{N}_\mathcal{T} \cup TH} \exp(l_{r'})} \right) \quad (18)$$

$$\mathcal{L}_\theta = \mathcal{L}'_\theta + \mathcal{L}''_\theta \quad (19)$$

where $\mathcal{P}_\mathcal{T}$ is the set of relations contained in (e_h, e_t) and $\mathcal{N}_\mathcal{T} = \mathcal{C} \setminus \mathcal{N}_\mathcal{T}$.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate our approach on two public DocRE datasets.

DWIE[1] (Zaporojets et al., 2021) is an entity-centric multi-task dataset containing 602/98/99 documents for training, validation, and testing, respectively. In the DWIE dataset, on average each entity pair contains 3.97 mention pairs. And about 26% of its entity pairs that express relations have more than one relation label.

DocRED (Yao et al., 2019) is a large scare human-annotated DocRE dataset containing 5053 documents from Wikipedia and Wikidata. As the original DocRED has a considerable amount of false-negative samples, we conduct experiments on two re-annotated versions of it, i.e., **Revisit-DocRED**[2] (Huang et al., 2022) and **Re-DocRED**[3] (Tan et al., 2022).

Following previous works, we use micro F1 and micro Ign F1 as the evaluation metrics for DocRE tasks. Ign F1 is proposed in Yao et al. (2019) with the relational facts shared by training and test sets excluded.

4.2 Baseline Models

We compare KMGRE with several RE models, e.g. CNN, LSTM, BiLSTM and Context-Aware (Sorokin and Gurevych, 2017). We also select several state-of-the-art DocRE models for comparison.

GAIN (Zeng et al., 2020) is a state-of-the-art graph-based DocRE model, which constructs two diagrams of mention level and entity level to aggregate the dependencies at different levels.

SSAN (Xu et al., 2021) takes the structural dependencies into account in the self-attention mechanism.

ATLOP (Zhou et al., 2021) proposes an adaptive threshold mechanism and optimizes it with a specific objective function and our method has a similar structure with it in implementation.

4.3 Implementation Details

Our model is implemented in PyTorch and HuggingFace's Transformers (Wolf et al., 2019)[4]. We use the uncased BERT-base (Devlin et al., 2019) as the base encoder to get contextual representation and attention weights.

For optimization, we use AdamW (Loshchilov and Hutter, 2019) with a learning rate of 5e-5 and a weight decay of 1e-5 to optimize our model. We apply a linear warmup on the first 6% steps. The focusing hyperparameters γ_ω and α_ω are set to 2 and 0.3, respectively. The threshold control hyperparameter ξ is set to 0.15 for Revisit-DocRED and 0.1 for DWIE.

We noticed in our experiments that if τ is set to a fixed high value, the model may misclassify some key mention pairs in the initial stage, which would mislead the relation extractor. Therefore, we introduce a warm-up process by calculating τ based on the steps as $\tau = 0.5 \cdot (1 - 0.999^{step})$.

4.4 Main Results

Results on DWIE. Our main results on the DWIE dataset are shown in Table 1. We can observe that our model has significant improvements in both development and test sets. In particular, KMGRE already achieves a state-of-the-art F1 score of 76.71% on the test set.

Results on DocRED. We also report the Ign F1 and F1 metrics on the Revisit-DocRED and Re-DocRED in Table 2. As seen, in the test set of Revisit-DocRED and Re-DocRED, KMGRE consistently outperforms previous methods. Notably, the performance of these models in the test set of Revisit-DocRED is much lower than reported in their original papers. This phenomenon is caused by the occurrence of many false-negative samples in the origin DocRED dataset (Huang et al., 2022).

[1] https://github.com/klimzaporojets/DWIE
[2] https://github.com/AndrewZhe/Revisit-DocRED
[3] https://github.com/tonytan48/Re-DocRED
[4] The code and training scripts will be released at https://github.com/toyfana/KMGRE.

Model	Dev		Test	
	Ign F1	F1	Ign F1	F1
CNN*	37.65	47.73	34.65	46.14
LSTM*	40.86	51.77	40.81	52.60
BiLSTM*	40.46	51.92	42.03	54.47
Context-Aware*	42.06	53.05	45.37	56.58
GAIN*	58.63	62.55	62.37	67.57
SSAN†	58.62	64.49	62.58	69.39
ATLOP	63.57	69.96	67.56	74.36
KMGRE	**65.56 ± 0.77**	**71.40 ± 0.37**	**69.94**	**76.71**

Table 1: Performance (%) on the development and test set of DWIE. We report the mean and standard deviation of F1 on the development set and test set by conducting 5 runs of training using different random seeds. The results with * are reported in Ru et al. (2021). The result with † is reported in Yu et al. (2022).

Model	Revisit-DocRED		Re-DocRED			
	Test		Dev		Test	
	Ign F1	F1	Ign F1	F1	Ign F1	F1
CNN	29.70	30.04	53.95	55.60	52.80	54.88
LSTM	31.32	31.77	56.40	58.30	56.31	57.83
BiLSTM	32.50	32.91	58.20	60.04	57.84	59.93
GAIN	41.27	41.64	71.99	73.49	71.88	73.44
SSAN	41.64	41.92	-	-	-	-
ATLOP	41.62	41.90	**73.35**	74.22	73.22	74.02
KMGRE	**42.78**	**43.16**	73.33	**74.44**	**73.39**	**74.46**

Table 2: Performance (%) on the dev/test set of Revisit-DocRED and Re-DocRED. The SSAN here uses the officially provided checkpoint based on RoBERTa-base.

Nevertheless, our model can still achieve large improvement on the test set compared to previous methods, demonstrating the effectiveness of modeling the mention-level relations.

Efficiency Comparison. We also benchmark the time and memory usage of KMGRE on a Tesla V100 GPU. Table 3 shows that our model incurs ∼22% training time and ∼63% GPU memory overhead.

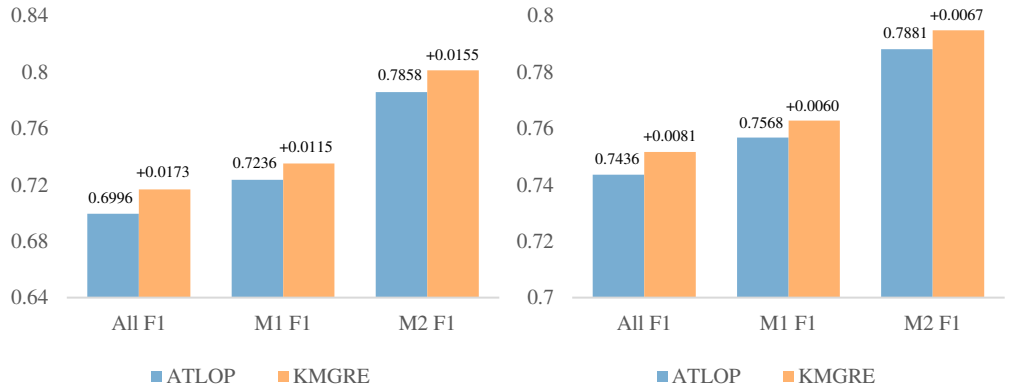
Model	Memory	Training time
ATLOP-BERT-base	4849 MB	4.21 it/s
KMGRE-BERT-base	7891 MB	3.45 it/s

Table 3: Training time and memory usage on Re-DocRED.

4.5 Ablation Studies

To better understand the impact of different components of our methods, we evaluate our model by removing each component. The results are shown in Table 4.

Effectiveness of the Key Instance Classifier. To evaluate the effectiveness of the key instance classifier, we directly train a model that only contains the mention-level relation extractor in KMGRE. By turning off the key instance classifier, KMGRE could be regarded as an instance-level approach of MIL (Ilse et al., 2018). As shown in Table 4, KMGRE performs better than without the key instance classifier. It means that our key instance classifier could effectively filter out mention pairs that do not express any relation to reduce the impact of redundant information. At the same time, our KMGRE can still achieve a better classification performance than ATLOP even without the key instance classifier, which means that directly modeling the mention-level relations is more reasonable.



(a) Results in development set.

(b) Results in test set.

Figure 3: The results of different mention numbers in DWIE. The **M1** subset denotes those entity pairs in which head or tail entity has multiple mentions. The **M2** subset denotes those entity pairs in which both head and tail entities contain multiple mentions.

Components	Ign F1	F1
ATLOP	63.57	69.96
KMGRE	65.56	71.40
-Key Instance Classifier	64.87	70.38
-Fusion of Mention-Level Results	59.67	63.93

Table 4: An ablation study of KMGRE on DWIE.

Effectiveness of the Mention-Level Results' Fusion. We further explore the effectiveness of the mention-level results' fusion by using the same pseudo-label generation procedure as the E-step. As shown in Table 4, we could observe a significant performance decay without the fusion of mention-level results. Since direct assigning the labels of entity pairs to key mention pairs will produce a large number of wrong labeled mention pairs, it seriously misleads the mention-level relation extractor. As about 26% of the positive entity pairs have more than one relation label, this phenomenon is particularly prominent in DWIE.

4.6 Effect Analysis for Mention Number

To explore the effect of mention number in DocRE, we compare our model's relation extraction performance in different cases. Following previous work (Yu et al., 2022), we divide the DWIE dataset into several subsets according to the mention number of head/tail entity, e.g., the **M1** subset denotes those entity pairs in which head or tail entity has multiple mentions, and the **M2** subset denotes those entity pairs in which both head and tail entities contain multiple mentions.

The results in the DWIE dataset are shown in Figure 3. It can be observed that as the number of mentions increases, the relation prediction results are more accurate. It indicates that with more mentions included, the information about a particular entity is more comprehensive, which is beneficial for relation classification. Notably, our method has consistently shown improvement over the strong baseline model for all cases, even for those entities that only have a single mention. Experimental results show that KMGRE can more accurately infer the relations between entities from the context than the previous models by directly modeling mention-level relations.

4.7 Case Studies

Figure 4 shows a case study of KMGRE and the previous state-of-the-art baseline ATLOP. We could observe that the head entity *Genc Ruli* and the tail entity *University of Tirana* are mentioned multiple times in the document. And this entity pair expresses multiple relations, i.e., *educated at* and *employer*. These two relations can be inferred from mention pairs in sentences [s3] and [s5], respectively. Also, there are a considerable amount of mention pairs of (*Genc Ruli*, *University of Tirana*) that do not express any relation.

We notice that both KMGRE and ATLOP can successfully identify the *educated at* relation between *Genc Ruli* and *University of Tirana*. However, ATLOP fails to extract the *employer* relation between the same entity pair, while KMGRE deduces it successfully. It indicates that treating all mention pairs equally would introduce unrelated information to mislead the relation extrator.

[s1] Genc Ruli (born April 11, 1958) is an Albanian politician. … [s3] Ruli holds a bachelor's degree in Economics and a bachelor's degree in Law from the University of Tirana. [s4] He holds a PhD in Economics from the Faculty of Economics, University of Tirana. [s5] Ruli is given the title Professor from the Faculty of Economics, University of Tirana. [s6] He has served as a Professor of Finance and Accounting in the Faculty of Economics, at the University of Tirana. [s7] Ruli has an extensive experience as the Minister of Finance and Economy in early 90's and as the Minister of Economy, Trade and Energy during 2005 - 2009. [s8] Ruli resigned from his position as Finance Minister on 9 November 1993, following allegations of corruption. … [s14] Ruli has written several publications in the areas of economics and public policies.

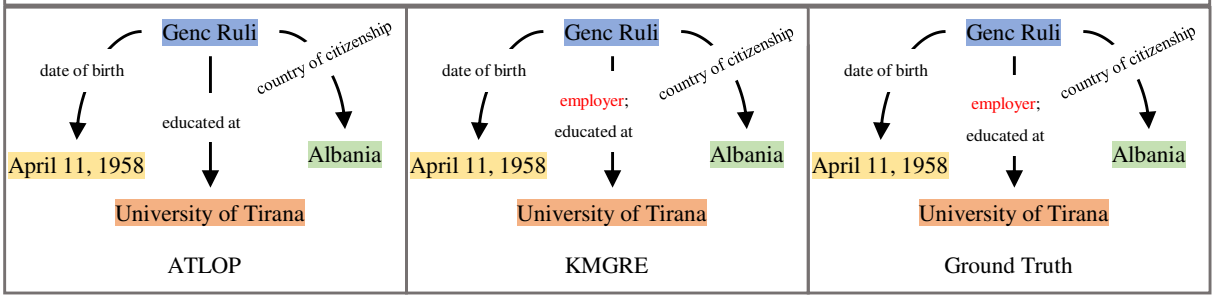


Figure 4: The case study of our proposed KMGRE and the state-of-the-art model, ATLOP (Zhou et al., 2021). The models take the document as input and predict relations among different entities in different colors. We only show a part of entities within the document and the according sentences due to the space limitation.

5 Conclusion

In this paper, we propose a new DocRE model called KMGRE for the multi-mention problem, containing a mention-level relation extractor and a key instance classifier. Our method uses the key instance classifier to identify those key mention pairs responsible for the entity pair relation label. Also, we propose a new optimization method to solve the multi-label problem in optimizing the mention-level relation extractor, as directly assigning the entity-level labels to the key instances can lead to misguidance. Experimental results on two public DocRE datasets show KMGRE outperforms previous state-of-the-art methods. The ablation study also confirms the effectiveness of our new method for optimizing the mention-level relation extractor in multi-label cases.

Acknowledgements

We thank anonymous reviewers for their responsible attitude and helpful comments. This work was supported by National Natural Science Foundation of China (62106013).

References

Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. 2018. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4925–4936, Hong Kong, China. Association for Computational Linguistics.

Aron Culotta and Jeffrey Sorensen. 2004. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 423–429, Barcelona, Spain.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Markus Eberts and Adrian Ulges. 2021. An end-to-end model for entity-level relation extraction using multi-instance learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 3650–3660. Association for Computational Linguistics.

Zhijiang Guo, Yan Zhang, and Wei Lu. 2019. Attention guided graph convolutional networks for relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Lin-*

guistics, pages 241–251, Florence, Italy. Association for Computational Linguistics.

Quzhe Huang, Shibo Hao, Yuan Ye, Shengqi Zhu, Yansong Feng, and Dongyan Zhao. 2022. Does recommend-revise produce reliable annotations? an analysis on missing instances in docred. *CoRR*, abs/2204.07980.

Quzhe Huang, Shengqi Zhu, Yansong Feng, Yuan Ye, Yuxuan Lai, and Dongyan Zhao. 2021. Three sentences are all you need: Local path enhanced document relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 998–1004, Online. Association for Computational Linguistics.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. 2018. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2132–2141. PMLR.

Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. 2021. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, Online. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Zhekun Luo, Devin Guillory, Baifeng Shi, Wei Ke, Fang Wan, Trevor Darrell, and Huijuan Xu. 2020. Weakly-Supervised Action Localization with Expectation-Maximization Multi-Instance Learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIX*, volume 12374 of *Lecture Notes in Computer Science*, pages 729–745. Springer.

Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1546–1557, Online. Association for Computational Linguistics.

Radford M. Neal and Geoffrey E. Hinton. 1998. A view of the em algorithm that justifies incremental, sparse, and other variants. In Michael I. Jordan, editor, *Learning in Graphical Models*, volume 89 of *NATO ASI Series*, pages 355–368. Springer Netherlands.

Dongyu Ru, Changzhi Sun, Jiangtao Feng, Lin Qiu, Hao Zhou, Weinan Zhang, Yong Yu, and Lei Li. 2021. Learning logic rules for document-level relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1239–1250, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniil Sorokin and Iryna Gurevych. 2017. Context-aware representations for knowledge base relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1784–1789, Copenhagen, Denmark. Association for Computational Linguistics.

Qingyu Tan, Lu Xu, Lidong Bing, and Hwee Tou Ng. 2022. Revisiting docred - addressing the overlooked false negative problem in relation extraction. *CoRR*, abs/2205.12696.

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 *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. 2021. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14149–14157. AAAI Press.

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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777,

Florence, Italy. Association for Computational Linguistics.

Jiaxin Yu, Deqing Yang, and Shuyu Tian. 2022. Relation-specific attentions over entity mentions for enhanced document-level relation extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1523–1529. Association for Computational Linguistics.

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. DWIE: an entity-centric dataset for multi-task document-level information extraction. *Inf. Process. Manag.*, 58(4):102563.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762, Lisbon, Portugal. Association for Computational Linguistics.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double graph based reasoning for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1630–1640, Online. Association for Computational Linguistics.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-Level Relation Extraction with Adaptive Thresholding and Localized Context Pooling. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14612–14620. AAAI Press.