Document-level Relation Extraction as Semantic Segmentation

Ningyu Zhang\textsuperscript{1,2}\textsuperscript{*}, Xiang Chen\textsuperscript{1,2}\textsuperscript{*}, Xin Xie\textsuperscript{1,2}, Shumin Deng\textsuperscript{1,2}, Chuanqi Tan\textsuperscript{3}, Mosha Chen\textsuperscript{3}, Fei Huang\textsuperscript{3}, Luo Si\textsuperscript{3}, Huajun Chen\textsuperscript{1,2,}\textsuperscript{1}

\textsuperscript{1} Zhejiang University \& AZFT Joint Lab for Knowledge Engine
\textsuperscript{2} Hangzhou Innovation Center, Zhejiang University
\textsuperscript{3} Alibaba Group

\{zhangningyu, xiang.chen.xx2020, 231sm, huajunsir\}@zju.edu.cn
\{chuanqi.tcq, chenmosha.cms, f.huang, luo.si\}@alibaba-inc.com

Abstract

Document-level relation extraction aims to extract relations among multiple entity pairs from a document. Previously proposed graph-based or transformer-based models utilize the entities independently, regardless of global information among relational triples. This paper approaches the problem by predicting an entity-level relation matrix to capture local and global information, parallel to the semantic segmentation task in computer vision. Herein, we propose a Document U-shaped Network for document-level relation extraction. Specifically, we leverage an encoder module to capture the context information of entities and a U-shaped segmentation module over the image-style feature map to capture global interdependency among triples. Experimental results show that our approach can obtain state-of-the-art performance on three benchmark datasets DocRED, CDR, and GDA\textsuperscript{1}.

1 Introduction

Relation extraction (RE) is an important task in the field of information extraction, which has widespread applications [Zhang et al., 2021b; Zhang et al., 2021a]. Previous works [Zeng et al., 2015; Feng et al., 2018] focused on identifying relations within a single sentence, which failed to recognize relations between entities across sentences. However, many relations are expressed over multiple sentences in real-world applications. According to [Yao et al., 2019], above 40.7% of relations can only be identified at the document level. Therefore, it is crucial for models to be able to extract document-level relations.

Recent studies [Yao et al., 2019; Tang et al., 2020; Zeng et al., 2020; Wang et al., 2020a; Zhou et al., 2021] have extended sentence-level RE to the document level. Compared with sentence-level RE that only contains one entity pair to classify in a sentence, document-level RE requires the model to classify the relations of multiple entity pairs at once. Besides, the subject and object entities involved in a relation may appear in different sentences. Therefore a relation cannot be identified based solely on a single sentence. For example, as shown in Figure 1, it is easy to identify the intra-sentence relations, such as (Maryland, country, U.S.), (Baltimore, located_in, Maryland), and (Eldersburg, located_in, Maryland), owing to the occurrence of entities in the same sentence. However, it is more challenging for a model to recognize inter-sentence relations, such as those between Eldersburg and U.S. and between Baltimore and U.S. because these mentions occur in different sentences and have long-distance dependencies.

To extract relations among these inter-sentence entity pairs, most current studies constructed document-level graph module based on heuristics, structured attention or dependency structures [Peng et al., 2017; Christopoulou et al., 2019; Nan et al., 2020; Zeng et al., 2020; Wang et al., 2020a], followed by reasoning with graph neural models. Meanwhile, considering the transformer architecture can implicitly model long-distance dependencies, some studies [Wang et al., 2019; Tang et al., 2020; Zhou et al., 2021] directly applied pre-trained language models rather than explicit graph reasoning. In general, current approaches obtain entity representation via information passing through nodes on document-level graphs or transformer-based structure learning. However, they mainly focus on token-level syntactic features or contextual information rather than global interactions between entity

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure.png}
\caption{Example document with entity pairs and relations from DocRED. Entity mentions and relations only involved in these relation instances are colored.}
\end{figure}

\textsuperscript{*} Equal contribution and shared co-first authorship.
\textsuperscript{1} Corresponding author.
\textsuperscript{1} The code and datasets are available in https://github.com/zjunlp/DocuNet.
pairs, neglecting the interdependency among the multiple relations in one context. Concretely, the interdependency among multiple triples is advantageous and can provide guidance for relation classification in the case of many entities. For example, if the intra-sentence relation (Maryland, country, U.S.) has been identified, it is implausible for U.S. to be in any other person-social relationship, such as “is the father of...”. Besides, according to the triples that Eldersburg is located in Maryland and Maryland belongs to U.S., we can infer that Eldersburg belongs to U.S.. As described above, each relation triple can provide information to other relation triples in the same text.

To capture the interdependency among the multiple triples, we reformulate the document-level RE task as an entity-level classification problem [Jiang et al., 2019], also known as table filling [Miwa and Sasaki, 2014; Gupta et al., 2016], as shown in Figure 2. It is analogous to semantic segmentation (a well-known computer vision task), whose goal is to label each pixel of the image with the corresponding represented class by convolution network. Inspired by the above, we propose a novel model called Document U-shaped Network (DocuNet), which formulates document-level RE as semantic segmentation. In this manner, given relevant features between entity pairs as an image, the model predicts the relation type for each entity pair as a pixel-level mask. Specifically, we introduce an encoder module to capture the context information of entities and a U-shaped segmentation module over the image-style feature map to capture global interdependency among triples. We further propose a balanced softmax method to handle the imbalance relation distribution. Our contributions can be summarized as follows:

- To the best of our knowledge, this is the first approach that regards document-level RE as a semantic segmentation task.
- We introduce the model DocuNet to capture both local context information and global interdependency among triples for document-level RE.
- Experimental results on three benchmark datasets show that our model DocuNet can achieve state-of-the-art performance compared with baselines.

2 Related Work

Previous relation extraction approaches mainly concentrate on identifying the relation between two entities within a sentence. Many approaches [Zeng et al., 2015; Feng et al., 2018; Zhang et al., 2018; Zhang et al., 2019; Zhang et al., 2020b; Zhang et al., 2020a; Wang et al., 2020b; Ye et al., 2021; Yu et al., 2020; Wang et al., 2020b; Wu et al., 2021; Chen et al., 2021; Zheng et al., 2021] have been proposed to tackle the sentence-level RE task effectively. However, sentence-level RE faces an inevitable restriction in that many real-world relations can only be extracted by reading multiple sentences. For this reason, document-level RE appeals to many researchers [Tang et al., 2020; Nan et al., 2020; Zeng et al., 2020; Wang et al., 2020a; Xiao et al., 2020].

Various approaches for document-level RE mainly include graph-based models and transformer-based models. Graph-based approaches are now widely adopted in RE because of their effectiveness and strength in relational reasoning. Jia et al. [2019] proposed a model that combines representations learned over various text spans throughout the document and across the sub-relation hierarchy. Christopoulou et al. [2019] proposed an edge-oriented graph neural model (EoG) for document-level RE. Li et al. [2020] characterized the complex interaction between sentences and potential relation instances with a graph-enhanced dual attention network (GEDA). Zhang et al. [2020c] proposed a novel graph-based model with a Dual-tier Heterogeneous Graph (DHG), which contains a structure modeling layer followed by a relation reasoning layer. Zhou et al. [2020] proposes a global context-enhanced graph convolutional network (GCGCN), composed of entities as nodes and the contexts of entity pairs as edges between nodes. Wang et al. [2020a] proposed a novel model (GLRE) that encodes the document information in terms of global and local entity representations as well as context relation representations. Nan et al. [2020] proposed a novel model (LSR) that enables relation reasoning across sentences by automatically inducing a latent document-level graph. Zeng et al. [2020] proposed the graph aggregation-and-inference network (GAIN) with double graphs for document-level RE. Xu et al. [2021] proposed an encoder-classifier reconstructor model (HeterGSAN), which manages to reconstruct the ground-truth path dependencies from the graph representation. Explicit graph reasoning can bridge the gap between entities that occur in different sentences, thus mitigating long-distance dependency and achieving promising performance.

In contrast, considering the transformer architecture can implicitly model long-distance dependencies, some researchers directly leverage pre-trained language models without generating document graphs. Wang et al. [2019] proposed a two-step training paradigm on DocRED using BERT as pre-trained word embedding. They observed an imbalance in the distribution of relation and disentangled the relation identification and classification for better inference. Tang et al. [2020] proposed a hierarchical inference network (HIN) to make full use of the abundant information from the entity,
Elias Brown (May 9, 1793– July 7, 1857) was a Representative from near , and is interred in a private cemetery near , . Born near , Brown attended the common schools. He studied law and was admitted to the bar in , , , and is interred in a private cemetery near , . Zhou et al. [2021] proposed a novel transformer-based model (ATLOP) of adaptive thresholding and localized context pooling based on BERT. However, most previous studies focused on the local entity representation, regardless of the high-level global connections between triples, which overlooked the interdependency between multiple relations.

On the one hand, our work is inspired by Jin et al., 2020, which was the first to consider the issue of global interaction between relations, and there have been few studies on RE. On the other hand, as these studies[Nguyen and Grishman, 2015; Shen and Huang, 2016] have done, convolutional neural networks have been long used in the relation extraction area, which enlightens us to pay attention to the role of CNN in extracting information of the image-style feature map. Hence, our work is also related to the study of Liu et al., 2020, who formulated incomplete utterance rewriting as a semantic segmentation task and motivated us to study the RE problem from a computer vision perspective. In this study, we leveraged the U-Net [Ronneberger et al., 2015], which consists of a contracting path to capture context and a symmetric expanding path that enables precise localization. To the best of our knowledge, this is the first approach to formulate RE as a semantic segmentation task.

3 Methodology

3.1 Preliminary

We first introduce the problem definition. With a document \(d\) containing a set of entities \(\{e_i\}_{i=1}^N\), the task is to extract the relations between entity pairs \((e_s, e_o)\). In one document, each entity \(e_i\) may occur multiple times. To model relation extraction between \(e_s\) and \(e_o\), we define a \(N \times N\) matrix \(Y\), where entry \(Y_{s,o}\) indicates the relation type between \(e_s\) and \(e_o\). Then, we obtain the output of matrix \(Y\), analogous to the task of semantic segmentation. Entities in \(Y\) are arranged according to their first appearance in the document. We obtain the feature map via the entity-to-entity relevance estimation and take the feature map as an image. Note that the output entity-level relation matrix \(Y\) is parallel to the pixel-level mask in semantic segmentation, which bridges relation extraction and semantic segmentation. Our approach can also be applied to sentence-level relation extraction. Since the document has relatively more entities, thus, entity-level relation matrix can learn more global information to boost the performance.

3.2 Encoder Module

Given the document \(d = [x_t]_{t=1}^L\), we insert special symbols "<e>" and "</e>" at the start and end of mentions to mark the entity positions. We leverage the pre-trained language model as an encoder to obtain the embedding as follows:

\[
H = [h_1, h_2, ..., h_L] = \text{Encoder}([x_1, x_2, ..., x_L]).
\] (1)

where \(h_i\) is the embedding of the token \(x_i\). Note that some documents are longer than 512, we thus leverage a dynamic window to encode whole documents. We average the embeddings of overlapping tokens of different windows to obtain the final representations. Then, we utilize the embeddings of "<e>" to represent mention following Verga et al., 2018. We leverage a smooth version of max pooling, namely, log-sumexp pooling [Jia et al., 2019] each entity \(e_i\), to obtain the entity embedding \(e_i\):

\[
e_i = \log \sum_{j=1}^{N_{e_i}} \exp (m_j).
\] (2)

This pooling accumulates signals from mentions in the document. Thus, we obtain the entity embedding \(e_i\).
We calculate the entity-level relation matrix based on entity-to-entity relevance. For each entity $e_i$ in the matrix, their relevance is captured by a $D$-dimensional feature vector $F(e_s, e_o)$. We introduce two strategies for computing $F(e_s, e_o)$, namely, similarity-based method and context-based method. Similarity-based method is produced by concatenating operation result of element-wise similarity, cosine similarity and bi-linear similarity between $e_s$ and $e_o$ as:

$$F(e_s, e_o) = [e_s \odot e_o; \cos(e_s, e_o); e_s W_1 e_o]$$ \hspace{1cm} (3)

For the context-based strategy, we leverage entity-aware attention with affine transformation to obtain the feature vector as follows:

$$F(e_s, e_o) = W_2 H a^{(s,o)}$$ \hspace{1cm} (4)

$$a^{(s,o)} = \text{softmax}(\sum_{i=1}^{K} A_i^s \cdot A_i^o)$$ \hspace{1cm} (5)

where $a^{(s,o)}$ is the attention weight for entity-aware attention and $A_i^s$ refers to the tokens’ importance to the $i$-th entity, $H$ is the document embedding, $W_1, W_2$ is the learnable weight matrix, $K$ is the number of head in the transformer.

### 3.3 U-shaped Segmentation Module

Taking the entity-level relation matrix $F \in \mathbb{R}^{N \times N \times D}$ as a $D$-channel image, we formulate the document-level relation prediction as the pixel-level mask in $F$. Where $N$ is the largest number of entities, counted from all the dataset samples. Specifically $N$ is the largest number of entities, counted from all the dataset samples. To this end, we utilize U-Net [Ronneberger et al., 2015], which is a famous semantic segmentation model in computer vision. As can be seen in Figure 3, the module is formed as a U-shaped segmentation structure, which contains two down-sampling blocks and two up-sampling blocks with skip connections. On the one hand, each down-sampling block has two subsequent max pooling and separate convolution modules. Further, the number of channels is doubled in each down-sampling block. As it shows in the Figure 2, the segmentation area in the entity-level relation matrix refers to the co-occurrence of relations between entity pairs. The U-shaped segmentation structure can promote the information exchange between entity pairs in the receptive field analogy to implicit reasoning. Specifically, CNN and down-sampling block can enlarge the receptive field of current entity pair embedding $F(e_s, e_o)$, thus, providing rich global information for representation learning. On the other hand, the model has two up-sampling blocks with a subsequent deconvolution neural network and two separate convolution modules. Different from down-sampling, the number of channels is halved in each up-sampling block, which can distribute the aggregated information to each pixel.

Finally, we incorporate an encoding module and a U-shaped segmentation module to capture both local and global information $Y$ as follows:

$$Y = U(W_2 F)$$ \hspace{1cm} (6)

where $U$ and $Y \in \mathbb{R}^{N \times N \times D'}$ denote the U-shaped segmentation module and entity-level relation matrix respectively. $W_3$ is the learnable weight matrix in order to reduce the dimension of $F$ and $D'$ is much smaller than $D$.

### 3.4 Classification Module

Given the entity pair embedding $e_s$ and $e_o$ with the entity-level relation matrix $Y$, we map them to hidden representations $z$ with a feedforward neural network. Then, we obtain the probability of relation via a bilinear function. Formally, we have:

$$z_s = \tanh(W_s e_s + Y_{s,o}),$$ \hspace{1cm} (7)

$$z_o = \tanh(W_e e_o + Y_{s,o}),$$ \hspace{1cm} (8)

$$P(r|e_s, e_o) = \sigma(z_s W_r z_o + b_r),$$ \hspace{1cm} (9)

where $Y_{s,o}$ is the entity-pair representation of $(s, o)$ in matrix $Y$, $W_r \in \mathbb{R}^{d \times d}$, $b_r \in \mathbb{R}$, $W_s \in \mathbb{R}^{d \times d}$, and $W_o \in \mathbb{R}^{d \times d}$, are learnable parameters.

Since previous work [Wang et al., 2019] observed that there is an imbalance relation distribution for RE (many entity pairs have relation of NA), we introduce a balanced softmax method for training, which is inspired by the circle loss [Sun et al., 2020] from computer vision. Specifically, we introduce an additional category 0, hoping that the scores of the target category are all greater than $\delta_0$ and the scores of the non-target categories are all less than $\delta_0$. Formally, we have:

$$L = \log \left( e^{s_0} + \sum_{i \in \Omega_{neg}} e^{s_i} \right) + \log \left( e^{-s_0} + \sum_{j \in \Omega_{pos}} e^{-s_j} \right).$$ \hspace{1cm} (10)

For simplicity, we set the threshold as zero and have the following:

$$L = \log \left( 1 + \sum_{i \in \Omega_{neg}} e^{s_i} \right) + \log \left( 1 + \sum_{j \in \Omega_{pos}} e^{-s_j} \right).$$ \hspace{1cm} (11)

### 4 Experiments

#### 4.1 Dataset

We evaluated our DocuNet model on three document-level RE datasets. We listed the dataset statistics in Table 1.

<table>
<thead>
<tr>
<th>Statistics / Dataset</th>
<th>DocRED</th>
<th>CDR</th>
<th>GDA</th>
</tr>
</thead>
<tbody>
<tr>
<td># Train</td>
<td>3,053</td>
<td>500</td>
<td>23,353</td>
</tr>
<tr>
<td># Dev</td>
<td>1,000</td>
<td>500</td>
<td>5,839</td>
</tr>
<tr>
<td># Test</td>
<td>1,000</td>
<td>500</td>
<td>1,000</td>
</tr>
<tr>
<td># Relations</td>
<td>97</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Avg. # entities per Doc.</td>
<td>19.5</td>
<td>7.6</td>
<td>5.4</td>
</tr>
<tr>
<td>Avg. # Ment. per Ent.</td>
<td>1.4</td>
<td>2.7</td>
<td>3.3</td>
</tr>
</tbody>
</table>

Table 1: Statistics of the experimental datasets.

- DocRED [Yao et al., 2019] is a large-scale document-level relation extraction dataset by crowdsourcing. DocRED contains 3,053/1,100/1,000 instances for training, validating and test, respectively.
ADAPT model leverages pre-training with external GDA. We optimize our model with AdamW using learning rate \(2e^{-5}\) with a linear warmup for the first 6% of steps. We set the matrix size \(N = 42\). The context-based strategy is utilized by default. We tuned the hyperparameters on the development set. We trained on one NVIDIA V100 16GB GPU and evaluated our model with Ign F1, and F1 following \[Yao\ et\ al.,\ 2019\].

### 4.3 Results on the DocRED Dataset

We compare DocuNet with graph-based models, including GEDA \[Li\ et\ al.,\ 2020\], LSR \[Nan\ et\ al.,\ 2020\], GLRE \[Wang\ et\ al.,\ 2020\] and GAIN \[Zeng\ et\ al.,\ 2020\], HeterGSAN \[Xu\ et\ al.,\ 2021\]; and transformer-based models, including BERT-base \[Wang\ et\ al.,\ 2019\], BERT-TS-base \[Wang\ et\ al.,\ 2019\], HIN-BERT-base \[Tang\ et\ al.,\ 2020\], CorefBERT-base \[Ye\ et\ al.,\ 2020\], and ATLOP-base on the DocRED dataset. From the Table 2, we observed that our approach DocuNet-BERTbase obtains better results than ATLOP-base. Moreover, we found that our DocuNet model obtain a new state-of-the-art result with RoBERTa-large. As of the IJCAI deadline on 20th of January 2021, we held the first position on the CodaLab scoreboard\(^3\) under the alias DocuNet without external data\(^4\).

### 4.4 Results on the Biomedical Datasets

In the biomedical datasets, we compare DocuNet with lots of baselines including: BRAN \[Verga\ et\ al.,\ 2018\], EoG \[Christopoulou\ et\ al.,\ 2019\], LSR \[Nan\ et\ al.,\ 2020\], DHG \[Zhang\ et\ al.,\ 2020c\], GLRE \[Wang\ et\ al.,\ 2020\], and ATLOP [Zhou et al., 2021]. Following ATLOP[Zhou et al., 2021], we utilize the SciBERT [Beltagy et al., 2019] which

### Table 2: Results (%) on the development and test set of DocRED. We run experiments five times with different random seeds and report the mean and standard deviation on the development set. We report the official test score on the CodaLab scoreboard with the best checkpoint on the development set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEDA-BERT_base [Li et al., 2020]</td>
<td>54.52</td>
<td>56.16</td>
</tr>
<tr>
<td>LSR-BERT_base [Nan et al., 2020]</td>
<td>52.43</td>
<td>59.00</td>
</tr>
<tr>
<td>GLRE-BERT_base [Wang et al., 2020a]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GAIN-BERT_base [Zeng et al., 2020]</td>
<td>59.14</td>
<td>61.22</td>
</tr>
<tr>
<td>HeterGSAN-BERT_base [Xu et al., 2021]</td>
<td>58.13</td>
<td>60.18</td>
</tr>
<tr>
<td>BERT_base [Wang et al., 2019]</td>
<td>-</td>
<td>54.16</td>
</tr>
<tr>
<td>HIN-BERT_base [Tang et al., 2020]</td>
<td>54.29</td>
<td>56.31</td>
</tr>
<tr>
<td>CorefBERT_large [Ye et al., 2020]</td>
<td>55.32</td>
<td>57.51</td>
</tr>
<tr>
<td>ATLOP-BERT_base [Zhou et al., 2021]</td>
<td>59.22</td>
<td>61.09</td>
</tr>
<tr>
<td>DocuNet-BERT_base</td>
<td>59.86±0.13</td>
<td>61.83±0.19</td>
</tr>
<tr>
<td>BERT_large [Ye et al., 2020]</td>
<td>56.67</td>
<td>58.83</td>
</tr>
<tr>
<td>CorefBERT_large [Ye et al., 2020]</td>
<td>56.82</td>
<td>59.01</td>
</tr>
<tr>
<td>RoBERTa_large [Ye et al., 2020]</td>
<td>57.14</td>
<td>59.22</td>
</tr>
<tr>
<td>CorefRoBERTa_large [Ye et al., 2020]</td>
<td>57.35</td>
<td>59.43</td>
</tr>
<tr>
<td>ATLOP-RoBERTa_large [Zhou et al., 2021]</td>
<td>61.32</td>
<td>63.18</td>
</tr>
<tr>
<td>DocuNet-RoBERTa_large</td>
<td>62.23±0.12</td>
<td>64.12±0.14</td>
</tr>
</tbody>
</table>

\(^3\)https://competitions.codalab.org/competitions/2017#results

\(^4\)The SSAN_ADAPT model leverages pre-training with external distance supervised data.

[2] *The Eminem Show* includes the commercially successful singles "Without Me", "Cleanin' Out My Closet", "Superman", and "Sing for the Moment".

**Figure 4**: Case study on our proposed DocuNet and baseline model. The specific number in the figure indicates the corresponding label id.

<table>
<thead>
<tr>
<th>Model</th>
<th>Ign $F_1$</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>DocuNet (Context-based)</td>
<td>59.86</td>
<td>61.83</td>
</tr>
<tr>
<td>DocuNet (Similarity-based)</td>
<td>59.04</td>
<td>60.92</td>
</tr>
<tr>
<td>$w/o$ Balanced Softmax</td>
<td>58.56</td>
<td>60.51</td>
</tr>
<tr>
<td>$w/o$ U-shaped Segmentation</td>
<td>57.51</td>
<td>59.65</td>
</tr>
</tbody>
</table>

Table 4: Ablation study of DocuNet on DocRED.

is pre-trained on the scientific publication corpora. From the Table 3, we observe model DocuNet-SciBERT$_{base}$ improved the $F_1$ score by 6.9% and 1.4% on CDR and GDA compared with ATLOP-SciBERT$_{base}$.

**4.5 Ablation Study**

We conducted an ablation study experiment to validate the effectiveness of different components of our approach. **DocuNet (Similarity-based)** means directly using similarity functions strategy to calculate the correlation between two entities as the input matrix, rather than context-based strategy. **$w/o$ U-shaped Segmentation** means that our segmentation module is replaced by a feed-forward neural network. **$w/o$ balanced softmax** refers to the model only with binary cross-entropy loss. From Table 4, we observe that all models have a performance decay without each module, which indicates that both components are beneficial. Besides, we observed that the U-shaped segmentation module and balanced softmax module are most important to model performance and sensitive to $F_1$, leading to a drop of 2.18\% and 1.32\% in dev $F_1$ score respectively when removed from DocuNet. That reveals that global interdependency among triples captured by our model is effective for document-level RE. Moreover, compared with context-based strategy, our approach based on similarity functions strategy drop by 0.84 $F_1$, which illustrates the context-based strategy is advantageous.

**4.6 Case Study**

We follow GAIN [Zeng et al., 2020] to select the same example and conduct a case study to further illustrate the effectiveness of our model DocuNet compared with the baseline. As shown in Figure 4, we notice that both BERT$_{base}$ and DocuNet-BERT$_{base}$ can successfully extract the “part of” relation between “Without Me” and “The Eminem Show”. However, only our model DocuNet-BERT$_{base}$ is able to deduce that the “performer” and “publication date” of “Without Me” are the same as those of “The Eminem Show”, namely, “Eminem” and “May 26, 2002”, respectively.

Intuitively we can observe that relation extraction mentioned above among those entities requires logical inference across sentences. This interesting observation indicates that our U-shaped segmentation structure over the entity-level relation matrix may implicitly conduct relational reasoning among entities.

**4.7 Analysis**

To assess the effectiveness of DocuNet in modeling global information for multiple entities, we evaluated models respec-
tively trained with or without U-shaped segmentation module on different groups of development set in DocRED, which are divided by the number of entities. From Figure 5, we observe that the model w/ U-shaped segmentation module consistently outperforms the model w/o U-shaped segmentation module. We notice that when the number of entities increases, the improvement becomes larger. This indicates that our U-shaped segmentation module can implicitly learn the interdependency among the multiple triples in one context, thus improving the document-level RE performance.

5 Conclusion and Future Work

In this study, we took the first step in formulating document-level RE as a semantic segmentation task and introducing the Document U-shaped Network. Experimental results showed that our model could achieve better performance by capturing local and global information than baselines. We also empirically observe that convolution over entity-entity relation matrix may implicitly conduct relational reasoning among entities. In the future, we plan to apply our approach to other span-level classification tasks, such as aspect-based sentiment analysis and nest named recognition.

Acknowledgments

We want to express gratitude to the anonymous reviewers for their hard work and kind comments. We thank Ning Ding for helpful discussions and feedback on this paper. This work is funded by National Key R&D Program of China (Funding No. 2018YFB1402800), NSFC91846204.

References

[Beltagy et al., 2019] Iz Beltagy, Kyle Lo, and Arman Co- 

[Chen et al., 2021] Xiang Chen, Xin Xie, Ningyu Zhang, Ji- 

[Christopoulou et al., 2019] Fenia Christopoulou, Makoto 

[Feng et al., 2018] Jun Feng, Minlie Huang, Li Zhao, Yang 

[Gupta et al., 2016] Pankaj Gupta, Hinrich Schütze, and 

[Jia et al., 2019] Robin Jia, Cliff Wong, and Hoifung Poon. 

[Jiang et al., 2019] Zhengbao Jiang, Wei Xu, Jun Araki, and 

[Jin et al., 2020] Zhijing Jin, Yongyi Yang, Xipeng Qiu, 

[Li et al., 2016] J. Li, Yueping Sun, Robin J. Johnson, 


[Nan et al., 2020] G. Nan, Zhijiang Guo, Ivan Sekulic, and 

[Nguyen and Grishman, 2015] Thien Huu Nguyen and 

[Peng et al., 2017] Nanyun Peng, Hoifung Poon, Chris 


