Multi-view Inference for Relation Extraction with Uncertain Knowledge

Bo Li[1,2], **Wei Ye**[1*], **Canming Huang**[3], **Shikun Zhang**[1]
[1]National Engineering Research Center for Software Engineering, Peking University
[2]School of Software and Microelectronics, Peking University
[3]Beijing University of Posts and Telecommunications
deepblue.lb@stu.pku.edu.cn, wye@pku.edu.cn,
huangzzk@bupt.edu.cn, zhangsk@pku.edu.cn

Abstract

Knowledge graphs (KGs) are widely used to facilitate relation extraction (RE) tasks. While most previous RE methods focus on leveraging deterministic KGs, uncertain KGs, which assign a confidence score for each relation instance, can provide prior probability distributions of relational facts as valuable external knowledge for RE models. This paper proposes to exploit uncertain knowledge to improve relation extraction. Specifically, we introduce ProBase, an uncertain KG that indicates to what extent a target entity belongs to a concept, into our RE architecture. We then design a novel multi-view inference framework to systematically integrate local context and global knowledge across three views: mention-, entity- and concept-view. The experiment results show that our model achieves competitive performances on both sentence- and document-level relation extraction, which verifies the effectiveness of introducing uncertain knowledge and the multi-view inference framework that we design.

Introduction

The goal of relation extraction (RE) is to classify the semantic relation between entities in a given context. It has plenty of practical applications, such as question answering (Yu et al. 2017) and information retrieval (Kadry and Dietz 2017). Knowledge graphs (KGs) (Bollacker et al. 2008; Ruppenhofer et al. 2006; Speer, Chin, and Havasi 2017; Wu et al. 2012), which contain pre-defined nodes (usually entities or concepts) and their relations, have been widely incorporated into RE tasks to integrate various prior knowledge in recent years.

The interaction between KGs and RE lies in two main aspects. On the one hand, we can use relational facts in exiting KGs to assign relation labels for entity pairs in unlabeled corpora to build datasets for distant supervision RE (Mintz et al. 2009). On the other hand, we can leverage the external knowledge in KGs to boost the performance of RE models, which is the case of this paper. Generally, there are two ways to integrate prior knowledge into RE. One way is to utilize structured information explicitly. For example, Can et al. (2019) retrieved the synonyms of each entity from WordNet[1]; Lei et al. (2018) extracted the n-gram text matching words of each entity from FreeBase (Bollacker et al. 2008); Li et al. (2019a) used relative semantic frames from FrameNet (Ruppenhofer et al. 2006). The other way is to explore the latent semantics of KGs. Researchers may incorporate pre-trained entity and relation embeddings (Wang et al. 2018; Hu et al. 2019), e.g., from TransE (Bordes et al. 2013), or jointly learn the entity embedding and RE models' parameters (Han, Liu, and Sun 2018; Xu and Barbosa 2019).

Despite various ways of leveraging KGs, most previous RE methods focus on deterministic KGs, where a specific relation either exists between two entities or not. However, in real-world scenarios, prior knowledge resources may contain inconsistent, ambiguous, and uncertain information (Wu et al. 2012). It would not be easy to capture highly-related external information if we treat all relations between entities as equally important. Another type of KGs, called Uncertain KGs (Chen et al. 2019) such as ConceptNet (Speer, Chin, and Havasi 2017) and ProBase (Wu et al. 2012), come to the rescue. Given two words (typically two entities or an entity and a concept) and a relation, uncertain KGs provide a confidence score that measures the possibility of the two words having the given relation. This inspires us to exploit the prior probability distribution of relational facts in uncertain KGs to improve relation extraction.

As a representative uncertain KG, ProBase provides *IsA* relation between an entity and a concept, indicating to what extent the target entity belongs to the concept, which is essential information for RE. Meanwhile, the concise structure of ProBase makes it convenient to be coupled with supervised relation extraction datasets. Therefore, we choose to use ProBase as our external uncertain KG and retrieve highly-related concepts of each entity. The relational facts and their confidence scores provide a prior probability distribution of concepts for an entity, which can serve as valuable supplementary knowledge given limited local contextual information.

ProBase brings global knowledge about related concepts for an entity pair, while the target document (or sentence) provides local semantic information about related mentions. We now have three views for the contextual information of a relational fact: concept view, entity view, and mention view.

* Corresponding author.

[1]https://wordnet.princeton.edu/

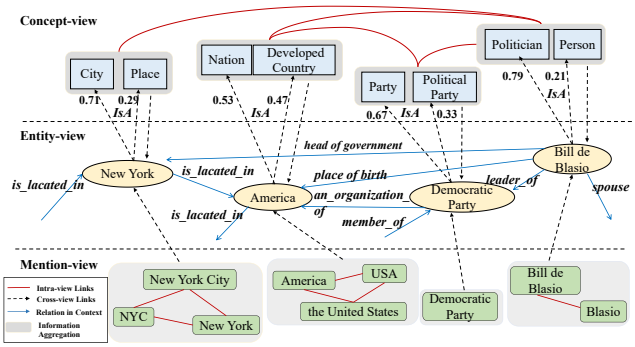

Figure 1: An example of a three-view inference framework for document-level relation extraction. In a document, an entity may have many mentions, usually in different forms. For a given entity, the local interactive representation is aggregated by different mention representations. This is a cross-view interaction between mention- and entity-view. Moreover, by retrieving uncertain KG, the concepts and weighting scores are used to produce the concept representation, which will gather descriptions of entities and concepts to obtain global interactive representations. This is a cross-view interaction between entity- and concept-view. Finally, entity-view is used for contextual information aggregation and relation prediction. Sentence-level relation extraction only involves entity- and concept-view. The figure is better viewed in color.

It is a non-trivial task to aggregate information across different views effectively to obtain discriminative feature vectors. Inspired by Hao et al. (2019), we designed a multi-view inference framework to synthesize contextual semantics from different sources and representation levels systematically. For a given entity, we first retrieve the top K concepts that the target entity most likely belongs to, and then perform cross-view interaction to aggregate the local and global information based on the confidence scores in ProBase. Figure 1 demonstrates the overview of our multi-view inference framework.

Last but not least, since the text descriptions of entities and concepts on Wikipedia also provide rich semantic information for RE, we retrieved the text descriptions from Wikipedia to enrich the representations of entities and concepts, resulting in a corpus that couples with ProBase, which we call ProBase_Desp. The dataset provides high-quality descriptions of more than 15,000,000 entities and concepts, serving as another valuable external knowledge resource for our method. We have released ProBase_Desp [2] to facilitate further research.

In summary, we leverage uncertain KG, ProBase, to improve relation extraction for the first time. To incorporate ProBase, we design a multi-view inference mechanism that integrates local context and global knowledge across mention-, entity-, and concept-view. Experiment results show that our method achieves competitive results on both sentence- and document-level relation extraction tasks. Our contributions are summarized as follows:

- We propose **MIUK**, a **M**ulti-view **I**nference framework for relation extraction with **U**ncertain **K**nowledge. Our work is pioneering research on introducing uncertain KG into relation extraction and investigating interactions among mentions, entities, and concepts.
- We conduct extensive experiments to verify MIUK and achieve competitive results on both sentence- and document-level relation extraction tasks.
- We build and release a corpus with high-quality descriptions of more than 15,000,000 entities and concepts. It can serve as a valuable external knowledge resource for relation extraction and many other natural language processing tasks.

Problem Definition and Input Formalization

Problem Definition

MIUK can handle both sentence- and document-level relation extraction tasks by leveraging the multi-view inference framework. For sentence-level relation extraction, MIUK uses two-view inference framework that exploits the entity- and concept-view representations; for document-level relation extraction, the mention-view representation is added to constitute a three-view inference framework. In this section, we introduce the architecture of MIUK by taking document-level relation extraction as an example, as sentence-level relation extraction is just its simplified case.

For an input document D that contains n sentences ($D = \{s_1, s_2, ..., s_n\}$), and p different entities, there will be $p \cdot (p-1)$ entity pair candidates. The goal of MIUK is to predict the relations of all possible entity pairs in a parallel way. We use m, e, c, s to denote the mention, entity, concept, and sentence respectively, and their corresponding low-dimensional vectors are denoted as **m, e, c, s**. Besides, the weighting score between an entity and a concept is a real number denoted as w.

Input Formalization

This section details the data preparation for our method, which consists of three parts: input document (the context), uncertain KG (ProBase), and descriptions from Wikipedia (ProBase_Desp), as shown in Figure 2. Note that we use Uncased BERT-Base (Devlin et al. 2019) to encode the input document and descriptions.

Input Document Since the input document may contain a lot of entities with various mentions, it is not applicable to add position embeddings (Zeng et al. 2014) directly, especially when using BERT as an encoder (Wu and He 2019). However, the position feature is crucial for relation extraction since it can highlight the target entities. In this paper, we propose to use entity anchors that can mark all the mentions and distinguish different mentions from different entities. Specifically, different entities are marked by different special tokens, while the same token indicates the various

[2] https://github.com/pkuserc/AAAI2021-MIUK-Relation-Extraction

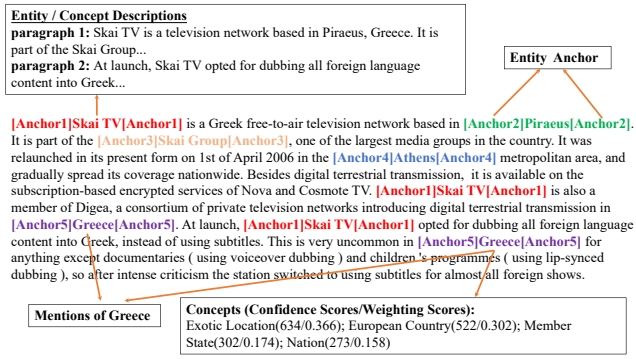

Figure 2: An illustration of entity mention, entity anchor, text descriptions, related concepts, confidence scores, and weighting scores. All the entity mentions are in bold. Different entity mentions are in different colors. There are five different entities in this document, thus 20 potential entity pairs to be predicted.

mentions of the same entity, as shown in Figure 2. These entity anchors make our model pay more attention to the entities. Each word and entity anchor is transformed into a low-dimensional word vectors by BERT.

Uncertain KG For each entity in the document, we retrieve the top-K concepts and their corresponding confidence scores from ProBase, which contain the prior uncertain knowledge. If an entity has less than K concepts, the token $< UNK >$ is used to pad the concept list to the fixed length, with its confidence score set to 0.

Entity and Concept Descriptions For each entity and concept in ProBase, MIUK uses the first two paragraphs from Wikipedia as supplementary descriptions. If the target entity does not exist in ProBase, we then use its entity type instead. Each description is transformed into a low-dimensional vector using BERT and max-pooling operation.

Multi-view Inference

The overall architecture of MIUK is shown in Figure 3. The multi-view inference framework consists of two parts: 1) cross-view links for information extraction, and 2) information aggregation and mixed attention mechanism, which aggregates the various feature vectors generated from cross-view links.

Cross-view Links

Cross-view links aim to extract information from the raw text effectively to generate local and global interactive vectors of entities as well as sentence representations, including the contextual information expressed by the input context and the external knowledge provided by ProBase and ProBase_Desp.

Mention2Entity Links Mention2Entity links (M2E) use the mention embeddings from the input document and the entity description vectors from ProBase_Desp to obtain the local interactive vector $\mathbf{u}_l$. For an entity mention ranging from the a-th to the b-th word, the mention embedding $\mathbf{m}$ is calculated by averaging the embeddings of its anchor token and all the words existing in the mention, where $\mathbf{m} \in \mathbb{R}^{1\times d}$. For a given entity e with t mentions $m_1, m_2, .., m_t$ in the document, unlike previous works that simply use the average of all related mention embeddings as entity representation $\mathbf{e}$, we believe it is best to select the most informative mention embeddings based on the entity description from external knowledge. Therefore, MIUK employs an attention mechanism to generate the local entity representation $\mathbf{e}_l$:

$$\mathbf{e}_l = \sum_{i=1}^{t} \alpha_i \cdot \mathbf{m}_i; \alpha_i = \frac{exp(\mathbf{e}_d \mathbf{m}_i^T)}{\sum_{i=1}^{t} exp(\mathbf{e}_d \mathbf{m}_i^T)}. \quad (1)$$

where the target entity description vector $\mathbf{e}_d \in \mathbb{R}^{1\times d}$ is the query vector, and $\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_t$ are the key vectors.

M2E generates the local representations of the target two entities, denoted as $\mathbf{h}_l$ and $\mathbf{t}_l$. We further add the minimum distance between the two target entities, denoted as d_{ht} and d_{th}, where $d_{ht} = -d_{th}$. They are transformed into low-dimensional vectors $\mathbf{d}_{ht}$ and $\mathbf{d}_{th}$ by a lookup table. Finally, the local interactive vector $\mathbf{u}_l$ is defined as follows:

$$\mathbf{u}_l = f_l([\mathbf{h}_l; \mathbf{d}_{ht}], [\mathbf{t}_l; \mathbf{d}_{th}]), \quad (2)$$

where f_l is the bilinear function, $[\cdot, \cdot]$ denotes concatenation operation, and $\mathbf{u}_l \in \mathbb{R}^{1\times d}$.

Entity2Concept Links Entity2Concept links (E2C) aim to leverage the uncertain knowledge from ProBase to generate a concept vector for the target entity.

For the target entity e, MIUK first retrieves the top K concepts $c_1, c_2, ..., c_k$ and their corresponding confidence scores from ProBase. Then softmax operation is applied to transform the confidence scores into weighting scores $w_1, w_2, ..., w_k$, where $\sum_{i=1}^{k} w_i = 1$, as the confidence scores provided by ProBase are frequency counts. The corresponding concept representations are generated after encoding their descriptions and max-pooling operation, denoted as $\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_k$, where $\mathbf{c}_i \in \mathbb{R}^{1\times d}$.

Here we propose three different techniques to compute the concept vector $\mathbf{c}$:

- Non-weighting Integration (NWI). NWI is a simple technique that averages all the concept representations directly. The intuition behind NWI is that all the related concepts contribute equally to target entity; thus this approach does not need the uncertain knowledge. NWI is defined as follows:

$$\mathbf{c} = \frac{1}{k} \cdot \sum_{i=1}^{k} \mathbf{c}_i. \quad (3)$$

- Attention-based Weighting Integration (AWI). AWI generates the concept vector $\mathbf{c}$ by employing an attention mechanism. The local entity representation $\mathbf{e}_l$ is the query vector, and $\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_k$ are the key vectors. AWI assumes that the local entity representation from the input document is helpful to selecting the most important concept.

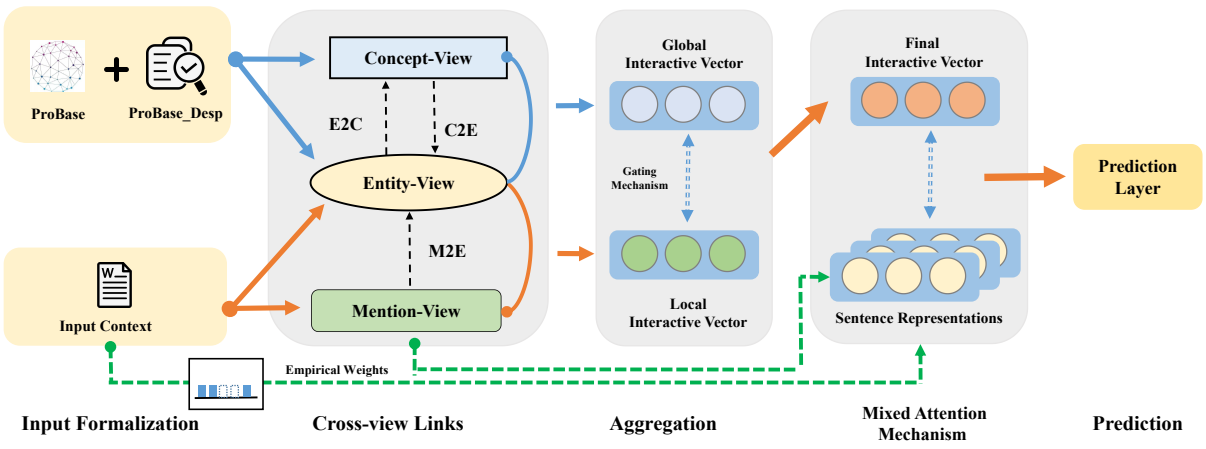

Figure 3: The neural architecture of MIUK.

AWI is defined as follows:

$$\mathbf{c} = \sum_{i=1}^{k} \alpha_i \cdot \mathbf{c}_i; \alpha_i = \frac{exp(\mathbf{e}_l \mathbf{c}_i^T)}{\sum_{i=1}^{k} exp(\mathbf{e}_l \mathbf{c}_i^T)}. \quad (4)$$

- Prior Knowledge-based Weighting Integration (PWI). PWI uses weighting scores $w_1, w_2, ..., w_k$ to aggregate the concept representations. PWI assumes that the weighting scores provide a prior probability distribution of concepts, which is beneficial to the model's performance. PWI is defined as follows:

$$\mathbf{c} = \sum_{i=1}^{k} w_i \mathbf{c}_i. \quad (5)$$

These three techniques are based on different assumptions. We will explore their effects later in Section .

Concept2Entity Links Concept2Entity links (C2E) generate the global interactive vector $\mathbf{u}_g$ by aggregating the entity description vector $\mathbf{e}_d$ and the concept vector $\mathbf{c}$. Specifically, for the target entities h and t, their corresponding description vectors are denoted as $\mathbf{h}_d$ and $\mathbf{t}_d$. The concept vectors can be obtained from Entity2Concept links, written as $\mathbf{c}_h$ and $\mathbf{c}_t$. Then we use another bilinear function f_g to compute the global interactive vector $\mathbf{u}_g$:

$$\mathbf{u}_g = f_g([\mathbf{h}_d; \mathbf{c}_h], [\mathbf{t}_d; \mathbf{c}_t]), \quad (6)$$

where $\mathbf{u}_g \in \mathbb{R}^{1\times d}$.

In addition, cross-view links also output all sentence representations $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n$ from the input document. Given a sentence ranging from the a-th to the b-th word, max-pooling operation is used to generate the sentence representation $\mathbf{s}$, where $\mathbf{s} \in \mathbb{R}^{1\times d}$, which will be used in the next stage.

Information Aggregation and Mixed Attention Mechanism

Information Aggregation The key insight of information aggregation is that the interactive vectors obtained from different sources can learn complementary information for the same entity pair, where $\mathbf{u}_l$ contains contextual information and $\mathbf{u}_g$ contains external knowledge. It is conceivable that aggregating them can achieve optimal performance. MIUK uses a gating mechanism to control the information flow between $\mathbf{u}_l$ and $\mathbf{u}_g$:

$$\mathbf{u} = \mathbf{g} \odot \mathbf{u}_l + (\mathbf{E} - \mathbf{g}) \odot \mathbf{u}_g, \quad (7)$$

where $\odot$ is the element-wise product between two vectors, and $\mathbf{g} \in \mathbb{R}^{1\times d}$ is the gating vector. $\mathbf{E} \in \mathbb{R}^{1\times d}$, and all the elements in $\mathbf{E}$ are 1. MIUK can select the most important information and generates the final interactive vector $\mathbf{u}$.

Mixed Attention Mechanism We use a mixed attention mechanism to generate the document representation $\mathbf{v}$ from sentence representations $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n$. Intuitively, if a target entity mention exists in sentence s, it should be more important than other sentences. Thus we introduce an empirical weight γ for each sentence. For the input document, mixed attention mechanism computes the weight of each sentence by combining both the information-based weight and the empirical weight:

直观地说，如果一个目标实体存在于句子s中，它应该比其他句子更重要

$$\mathbf{v} = \sum_{i=1}^{n} (\frac{\alpha_i + \beta_i}{2} + \gamma_i)\mathbf{s}_i, \quad (8)$$

$$\alpha_i = \frac{exp(\mathbf{u}_l \mathbf{s}_i^T)}{\sum_{i=1}^{n} exp(\mathbf{u}_l \mathbf{s}_i^T)}; \beta_i = \frac{exp(\mathbf{u}_g \mathbf{s}_i^T)}{\sum_{i=1}^{n} exp(\mathbf{u}_g \mathbf{s}_i^T)}, \quad (9)$$

$$\gamma_i = \begin{cases} \frac{1}{z}, s_i \in S, \\ 0, s_i \notin S, \end{cases} \quad (10)$$

where α_i and β_i are the weights based on the local and global interactive vector; $S \in D$ and consists of z different sentences, each of which contains at least one target entity mention.

Prediction

Since the document-level relation extraction is a multi-label problem, i.e., a target entity pair may express more than one relation, we use a single hidden layer and sigmoid activation function for relation extraction:

$$p(r|h,t) = g(\mathbf{W}_r[\mathbf{u};\mathbf{v}]^T + \mathbf{b}_r), \quad (11)$$

where g is the sigmoid activation function, $\mathbf{W}_r \in \mathbb{R}^{l\times 2d}$ and $\mathbf{b}_r \in \mathbb{R}^{l\times 1}$ are trainable parameters, and l is the number of predefined relation types in the dataset.

Experiment Results

Datasets and Hyper-parameter Settings

Datasets and Evaluation Metrics For document-level relation extraction, we use DocRED proposed by Yao et al. (2019). DocRED has 3,053 training documents, 1,000 development documents and 1,000 test documents, with 97 relation types (including "No Relation"). We treat this task as a multi-label classification problem since one or more relation types may be assigned to an entity pair. Following previous works (Yao et al. 2019), we use **F1** and **IgnF1** as the evaluation metrics. IgnF1 is a stricter evaluation metric that is calculated after removing the entity pairs that have appeared in the training set.

For sentence-level relation extraction, we use ACE2005 dataset following Ye et al. (2019). The dataset contains 71,895 total instances with 7 predefined relation types (including "No Relation"), 64,989 of which are "No Relation" instances (negative instances). We use five-fold cross-validation to evaluate the performance, and we report the precision (**P**), recall (**R**) and Micro F1-score (**Micro-F1**) of the positive instances.

Hyper-parameter Settings We use uncased BERT-base (Devlin et al. 2019) to encode the input context and text descriptions, and the size of each word embedding is 768. Note that we do not need to re-train BERT model, the rare words existing in the vocabulary of BERT can be used as the entity anchors. We then use a single layer to project each word embedding into a low-dimensional input vector of size d. Max-pooling operation is further applied to compute the entity description vector or the concept representation. Note that we use the same BERT model to encode both the input context and the entity/concept descriptions. We experiment with the following values of hyper-parameters: 1) the learning rate lr_{BERT} and lr_{Other} for BERT and other parameters $\in \{1\times 10^{-3}, 1\times 10^{-4}, 1\times 10^{-5}\}$; 2) the size of input vector, entity description vector and concept representation $\in \{50, 100, 150, 200\}$; 3) the size of distance embedding $\in \{5, 10, 20, 30\}$; 4) batch size $\in \{4, 8, 12, 16, 20, 24\}$; and 5) dropout ratio $\in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. We tune the hyper-parameters on the development set, and we evaluate the performance on the test set. Table 1 lists the selected hyper-parameter values in our experiments.

Hyper-parameter	Value
lr_{BERT}	1×10^{-5}
lr_{Other}	1×10^{-5}
Input Vector Size d	100
Embedding Size of Entity/Concept	100
Embedding Size of Distance	10
Batch Size	16
Dropout Ratio	0.2

Table 1: Hyper-parameter Settings

Baseline Models

We choose a number of methods as baseline models for sentence- and document-level relation extraction.

Document-level Relation Extraction The following models are used as baseline models. **GCNN** (Sahu et al. 2019) and **EoG** (Christopoulou, Miwa, and Ananiadou 2019a) are graph-based document-level relation extraction models. In addition, the following four methods all use BERT-base as encoder: **BERT-Two-Step** (Wang et al. 2019) classified the relation between two target entities in two steps; **DEMMT** (Han and Wang 2020) proposed an entity mask method, which is similar to entity anchor; **GEDA** (Li et al. 2020) proposed a graph-enhanced dual attention network for document-level relation extraction; **LSR** (Nan et al. 2020) used a multi-hop reasoning framework with external NLP tools, which is the state-of-the-art model for document-level relation extraction.

Sentence-level Relation Extraction The following four approaches are used as baseline models: **SPTree** (Miwa and Bansal 2016) used tree-LSTM for relation extraction; **Walk-based Model** (Christopoulou, Miwa, and Ananiadou 2019b) proposed a graph-based model that considers interactions among various entities; **MLT_Tag** (Ye et al. 2019) exploited entity BIO tag embeddings and multi-task learning for sentence-level relation extraction. Furthermore, we use BERT to replace the encoder in MLT_Tag to form an additional strong baseline named **BERT-MLT_Tag**.

Main Results

Document-level Relation Extraction Table 2 records the performance of the proposed MIUK and other baseline models. We can see that:

- The proposed MIUK (three-view) outperforms other models by a large margin in terms of both F1 and IgnF1. Specifically, compared with the highest scores among baseline models, MIUK achieves 1.11 and 1.08 absolute increase in F1 score on both the development and the test set. Similarly, the absolute increases in IgnF1 score achieved by MIUK are 2.77 and 0.94. The results can verify the effectiveness of our multi-view inference architecture with uncertain knowledge incorporated.
- MIUK and LSR outperform other models (e.g., BERT-Two-Step, GEDA, and DEMMT) by a large margin, even though the latter involve novel training skills or sophisticated architectures (e.g., graph neural networks). We mainly attribute this to the introduction of external knowledge: LSR brings in syntactic dependence information based on external tools, while MIUK incorporates rich information from ProBase and Wikipedia.

Methods	Dev		Test	
	IgnF1%	F1%	IgnF1%	F1%
GCNN§	46.22	51.52	49.59	51.62
EoG§	45.94	52.15	49.48	51.82
BERT-Two-Step	-	54.42	-	53.92
GEDA	54.52	56.16	53.71	55.74
DEMMT	55.50	57.38	54.93	57.13
LSR§	52.43	59.00	56.97	59.05
MIUK (three-view)	**58.27**	**60.11**	**58.05**	**59.99**

Table 2: Document-level relation extraction results on the development set and the test set of DocRED. MIUK (three-view) is designed for document-level relation extraction that contains mention-, entity- and concept-view representations. Results with § are directly cited from (Nan et al. 2020).

- Though LSR achieves impressive results with NLP tools on the test set, the IgnF1 score of LSR on the development set is merely 52.43, far lower than its IgnF1 score on the test set. One possible explanation for this instability could be the usage of external NLP tools (LSR used spaCy[3] to get meta dependency paths of sentences in a document), as external tools may cause error propagation problems. Our method achieves satisfactory results on both the development and the test set simultaneously, showing that leveraging uncertain knowledge can not only boost the model's performance but also improve its generalization ability.

Methods	P%	R%	Micro-F1%
SPTree†	70.1	61.2	65.3
Walk-based Model†	69.7	59.5	64.2
MTL_Tag†	66.5	71.8	68.9
BERT-MTL_Tag*	70.1	74.5	72.0
MIUK (two-view)	74.7	76.9	75.7

Table 3: Comparison between MIUK and the state-of-the-art methods using ACE 2005 dataset for sentence-level relation extraction. The best results are in bold. MIUK (two-view) is used for sentence-level relation extraction that consists of entity- and concept-view representations. Models with $*$ are reproduced based on open implementation. Results with † are directly cited from (Ye et al. 2019).

Sentence-level Relation Extraction For sentence-level relation extraction, MIUK (two-view) outperforms other models by a large margin with an F1 score of 75.7, which clearly sets up a new state-of-the-art. Besides, MIUK also achieves the best **P** (74.7) and **R** (76.9). Compared with BERT-MTL_Tag, a competitive model equipped with BERT, our method still achieves higher F1 score by 3.7, **P** by 4.6, and **R** by 2.4 in absolute value, while balancing **P** and **R** better. These results show that the introduction of external knowledge and our multi-view inference framework that utilizes uncertain knowledge can also benefit sentence-level relation extraction.

[3]https://spacy.io/

Detailed Analyses

Our model mainly involves leveraging uncertain knowledge, multi-view inference, and text descriptions. We will further evaluate the effectiveness of these three components in this section. Due to space limit, we only report the detailed analyses on document-level relation extraction; similar conclusions can be drawn from sentence-level relation extraction as well.

Models	F1%
MIUK	**60.11**
MIUK-NWI	59.37
MIUK-AWI	59.42
MIUK w/o Cross-view Inference	58.03
MIUK w/o Mixed Att	59.70
MIUK w/o Entity Desp	58.21
MIUK w/o Concept Desp	59.02

Table 4: The F1 scores of MIUK and its variants on the development set of DocRED.

Top-K	IgnF1%	F1%
K = 1	55.65	58.22
K = 2	57.03	59.35
K = 3	**58.27**	**60.11**
K = 4	57.22	59.21
K = 5	55.40	58.53

Table 5: The results of using different top K concepts for each entity in DocRED.

Effectiveness of Uncertain Knowledge The main difference between deterministic KGs and uncertain KGs is whether relations of entity pairs are assigned with confidence scores. To explore the effectiveness of uncertain knowledge, we design a model variant named **MIUK-NWI** that uses deterministic knowledge. The only difference between MIUK-NWI and MIUK is that MIUK-NWI uses a KG with all confidence scores removed (or set as the same). As shown in Table 4, we can observe a notable drop in performance (nearly 1.0 in F1 score) comparing MIUK-NWI with MIUK, which shows that the prior probability distribution of concepts for an entity provides valuable global contextual information, and our framework is capable of capturing this information to discriminate relational facts.

We further design **MIUK-AWI** as an enhanced version of MIUK-NWI, which uses a classic attention mechanism to aggregate concept information in a deterministic KG. We find the the performances of MIUK-AWI and MIUK-NWI show no clear difference. We can roughly conclude that compared with the classic attention mechanism, the prior confidence scores can help identify relevant concepts better.

Though ProBase provides a number of concepts for a given entity, using too many concepts may also bring noisy information and thus hinder the model's performance. Therefore, the concept number K for our MIUK is also an important hyper-parameter worth investigating. Table 5 shows the results using different K for concept selection. As expected, using too many concepts (more than three) does not gain better results. If only one or two concepts are selected, we can also observe performance degradation, due to feeding limited external knowledge into the inference framework. In short, we find that $K = 3$ generates the best results in our experiments.

Effectiveness of Multi-view Inference To verify the effectiveness of the multi-view inference framework that we design, we build a model variant named **MIUK w/o Cross-view Inference**, which simply concatenate all the representations from different sources and then feed them into the final classifier. From Table 4 we can see that removing the multi-view inference framework results in significant performance degradation (more than 2 in F1 score). The result shows that our multi-view inference framework provides a better way to integrate and synthesize multi-level information from different sources.

The mixed attention mechanism is another important component of MIUK. We further design a model variant named **MIUK w/o Mixed Att**, which replaces the mixed attention mechanism with the vanilla attention mechanism. Table 4 shows that the mixed attention mechanism improves the F1 score by 0.41. One possible explanation is that when the input document contains too many sentences, the vanilla attention mechanism fail to focus on highly-related sentences. We conclude that the mixed attention mechanism with empirical weights can capture supplementary information that is independent of contextual information.

Effectiveness of Text Descriptions To investigate how the information in text descriptions affects our model, we create two variants of MIUK by removing entity descriptions or concept descriptions, named **MIUK-w/o Entity Desp** and **MIUK-w/o Concept Desp**, respectively. From Table 4 we can see that the F1 scores drop significantly without either entity descriptions (MIUK-w/o Entity Desp) or concept descriptions (MIUK-w/o Concept Desp), which shows that the information from Wikipedia benefits relation extraction and MIUK can capture the rich semantics in these text descriptions well.

Related Work

Uncertain Knowledge Graphs

Uncertain Knowledge Graphs provide a confidence score for each word pair, such as ConceptNet and ProBase. ConceptNet is a multilingual uncertain KG for commonsense knowledge. It gives a list of words with certain relations (such as "located at," "used for," etc.) to the given entity, and provides a confidence score for each word pair. ProBase, otherwise known as Microsoft Concept Graph, is a big uncertain KG with 5,401,933 unique concepts, 12,551,613 unique entities and 87,603,947 *IsA* relations. For a given entity and concept pair that has *IsA* relation, denoted as $(x, y, P_{IsA}(x, y))$, $P_{IsA}(x, y)$ is the confidence score that measures the possibility of the entity x belonging to the concept y.

Since ProBase provides a more concise data structure and is easier to apply to relation extraction, we choose ProBase as the source of uncertain knowledge in this paper.

Relation Extraction with External Knowledge

Most external knowledge-based methods are targeted at distant supervision relation extraction (DSRE). Verga and McCallum (2016) employed FreeBase (Bollacker et al. 2008) and probabilistic model to extract features for DSRE. Weston et al. (2013) first proposed to use both contextual and knowledge representations for DSRE, but they used the two representations independently, and connected them only at the inference stage. Han, Liu, and Sun (2018), and Xu and Barbosa (2019) designed heterogeneous representation methods which jointly learn contextual and knowledge representations. Lei et al. (2018) proposed a novel bi-directional knowledge distillation mechanism with a dynamic ensemble strategy (CORD). For each entity, CORD uses the related words from FreeBase by n-gram text matching, which may bring lots of noise.

Some works leverage external knowledge for supervised relation extraction. Ren et al. (2018) used the descriptions of entities from Wikipedia but did not incorporate KGs for further research. Li et al. (2019b) only focused on Chinese relation extraction with HowNet (Dong and Dong 2003). Li et al. (2019a) incorporated prior knowledge from external lexical resources into a deep neural network. For each relation type, they found all relevant semantic frames from FrameNet and their synonyms from Thesaurus.com. However, they only considered the keywords and synonyms of an entity; therefore, the rich information in the entity description was ignored. MIUK distinguishes itself from previous works by introducing ProBase into relation extraction and systematically investigating the interactions among mentions, entities, and concepts.

Conclusion and Future Work

We have presented MIUK, a **M**ulti-view **I**nference framework for relation extraction with **U**ncertain **K**nowledge. MIUK introduces ProBase, an uncertain KG, into relation extraction pipeline for the first time. To effectively incorporate ProBase, we have designed a multi-view inference mechanism that integrates local context and global knowledge across mention-, entity-, and concept-view. Results of extensive experiments on both sentence- and document-level relation extraction tasks can verify the effectiveness of our method. We have also built a corpus with high-quality descriptions of entities and concepts, serving as a valuable external knowledge resource for relation extraction and many other NLP tasks. We believe it could be a future research direction of relation extraction to further investigate the interactions among mentions, entities, and concepts.

Acknowledgements

This research is supported by the National Key Research And Development Program of China (No. 2019YFB1405802). We would also like to thank Handan Institute of Innovation, Peking University for their support of our work.

References

Bollacker, K. D.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*, 1247–1250. ACM.

Bordes, A.; Usunier, N.; García-Durán, A.; Weston, J.; and Yakhnenko, O. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, 2787–2795.

Can, D.; Le, H.; Ha, Q.; and Collier, N. 2019. A Richer-but-Smarter Shortest Dependency Path with Attentive Augmentation for Relation Extraction. In *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)*, 2902–2912. Association for Computational Linguistics.

Chen, X.; Chen, M.; Shi, W.; Sun, Y.; and Zaniolo, C. 2019. Embedding Uncertain Knowledge Graphs. 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*, 3363–3370.

Christopoulou, F.; Miwa, M.; and Ananiadou, S. 2019a. 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 2019, Hong Kong, China, November 3-7, 2019*, 4924–4935.

Christopoulou, F.; Miwa, M.; and Ananiadou, S. 2019b. A Walk-based Model on Entity Graphs for Relation Extraction. *CoRR* abs/1902.07023.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186.

Dong, Z.; and Dong, Q. 2003. HowNet-a hybrid language and knowledge resource. In *International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003*, 820–824. IEEE.

Han, X.; Liu, Z.; and Sun, M. 2018. Neural Knowledge Acquisition via Mutual Attention Between Knowledge Graph and Text. In *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*, 4832–4839.

Han, X.; and Wang, L. 2020. A Novel Document-Level Relation Extraction Method Based on BERT and Entity Information. *IEEE Access* 8: 96912–96919.

Hao, J.; Chen, M.; Yu, W.; Sun, Y.; and Wang, W. 2019. Universal Representation Learning of Knowledge Bases by Jointly Embedding Instances and Ontological Concepts. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, 1709–1719.

Hu, L.; Zhang, L.; Shi, C.; Nie, L.; Guan, W.; and Yang, C. 2019. Improving Distantly-Supervised Relation Extraction with Joint Label Embedding. 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 2019, Hong Kong, China, November 3-7, 2019*, 3819–3827. Association for Computational Linguistics.

Kadry, A.; and Dietz, L. 2017. Open Relation Extraction for Support Passage Retrieval: Merit and Open Issues. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, 1149–1152.

Lei, K.; Chen, D.; Li, Y.; Du, N.; Yang, M.; Fan, W.; and Shen, Y. 2018. Cooperative Denoising for Distantly Supervised Relation Extraction. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, 426–436.

Li, B.; Ye, W.; Sheng, Z.; Xie, R.; Xi, X.; and Zhang, S. 2020. Graph Enhanced Dual Attention Network for Document-Level Relation Extraction. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, 1551–1560. International Committee on Computational Linguistics.

Li, P.; Mao, K.; Yang, X.; and Li, Q. 2019a. Improving Relation Extraction with Knowledge-attention. 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 2019, Hong Kong, China, November 3-7, 2019*, 229–239.

Li, Z.; Ding, N.; Liu, Z.; Zheng, H.; and Shen, Y. 2019b. Chinese Relation Extraction with Multi-Grained Information and External Linguistic Knowledge. In *Proceedings of the 57th Conference of the Association for Computational*

Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, 4377–4386.

Mintz, M.; Bills, S.; Snow, R.; and Jurafsky, D. 2009. Distant supervision for relation extraction without labeled data. In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore*, 1003–1011.

Miwa, M.; and Bansal, M. 2016. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 1105–1116.

Nan, G.; Guo, Z.; Sekulic, I.; and Lu, W. 2020. Reasoning with Latent Structure Refinement for Document-Level Relation Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 1546–1557.

Ren, F.; Zhou, D.; Liu, Z.; Li, Y.; Zhao, R.; Liu, Y.; and Liang, X. 2018. Neural Relation Classification with Text Descriptions. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, 1167–1177.

Ruppenhofer, J.; Ellsworth, M.; Schwarzer-Petruck, M.; Johnson, C. R.; and Scheffczyk, J. 2006. FrameNet II: Extended theory and practice .

Sahu, S. K.; Christopoulou, F.; Miwa, M.; and Ananiadou, S. 2019. Inter-sentence Relation Extraction with Document-level Graph Convolutional Neural Network. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 4309–4316.

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, 4444–4451.

Verga, P.; and McCallum, A. 2016. Row-less Universal Schema. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction, AKBC@NAACL-HLT 2016, San Diego, CA, USA, June 17, 2016*, 63–68.

Wang, G.; Zhang, W.; Wang, R.; Zhou, Y.; Chen, X.; Zhang, W.; Zhu, H.; and Chen, H. 2018. Label-Free Distant Supervision for Relation Extraction via Knowledge Graph Embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, 2246–2255. Association for Computational Linguistics.

Wang, H.; Focke, C.; Sylvester, R.; Mishra, N.; and Wang, W. 2019. Fine-tune Bert for DocRED with Two-step Process. *CoRR* abs/1909.11898.

Weston, J.; Bordes, A.; Yakhnenko, O.; and Usunier, N. 2013. Connecting Language and Knowledge Bases with Embedding Models for Relation Extraction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, 1366–1371.

Wu, S.; and He, Y. 2019. Enriching Pre-trained Language Model with Entity Information for Relation Classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, 2361–2364.

Wu, W.; Li, H.; Wang, H.; and Zhu, K. Q. 2012. Probase: a probabilistic taxonomy for text understanding. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, Scottsdale, AZ, USA, May 20-24, 2012*, 481–492.

Xu, P.; and Barbosa, D. 2019. Connecting Language and Knowledge with Heterogeneous Representations for Neural Relation Extraction. In *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)*, 3201–3206.

Yao, Y.; Ye, D.; Li, P.; Han, X.; Lin, Y.; Liu, Z.; Liu, Z.; Huang, L.; Zhou, J.; and Sun, M. 2019. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 764–777.

Ye, W.; Li, B.; Xie, R.; Sheng, Z.; Chen, L.; and Zhang, S. 2019. Exploiting Entity BIO Tag Embeddings and Multi-task Learning for Relation Extraction with Imbalanced Data. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 1351–1360.

Yu, M.; Yin, W.; Hasan, K. S.; dos Santos, C. N.; Xiang, B.; and Zhou, B. 2017. Improved Neural Relation Detection for Knowledge Base Question Answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, 571–581.

Zeng, D.; Liu, K.; Lai, S.; Zhou, G.; and Zhao, J. 2014. Relation Classification via Convolutional Deep Neural Network. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, 2335–2344.