From Discourse to Narrative: Knowledge Projection
for Event Relation Extraction

Jialong Tang[1,3], Hongyu Lin[1], Meng Liao[4], Yaojie Lu[1,3],
Xianpei Han[1,2], Le Sun[1,2,*], Weijian Xie[4], Jin Xu[4,*]

[1]Chinese Information Processing Laboratory   [2]State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China
[4]Data Quality Team, WeChat, Tencent Inc., China
{jialong2019,hongyu,xianpei,sunle}@iscas.ac.cn
{maricoliao, vikoxie, jinxxu}@tencent.com

Abstract

Current event-centric knowledge graphs highly rely on explicit connectives to mine relations between events. Unfortunately, due to the sparsity of connectives, these methods severely undermine the coverage of EventKGs. The lack of high-quality labelled corpora further exacerbates that problem. In this paper, we propose a knowledge projection paradigm for event relation extraction: projecting discourse knowledge to narratives by exploiting the commonalities between them. Specifically, we propose **M**ulti-tier **K**nowledge **P**rojection **Net**work (**MKPNet**), which can leverage multi-tier discourse knowledge effectively for event relation extraction.In this way, the labelled data requirement is significantly reduced, and implicit event relations can be effectively extracted. Intrinsic experimental results show that MKPNet achieves the new state-of-the-art performance, and extrinsic experimental results verify the value of the extracted event relations.

1 Introduction

Event-centric knowledge graphs (EventKGs) model the narratives of the world by representing events and identifying relations between them, which are critical for machine understanding and can benefit many downstream tasks, such as question answering (Costa et al., 2020), news reading (Vossen, 2018), commonsense knowledge acquisition (Zhang et al., 2020a) and so on.

Recently, semi-automatically constructing EventKGs have gained much attention (Tandon et al., 2015; Rospocher et al., 2016; Gottschalk and Demidova, 2018; Zhang et al., 2020b). These methods extract event knowledge from massive raw corpora with or without little human intervention, which makes them scalable solutions to build large-scale
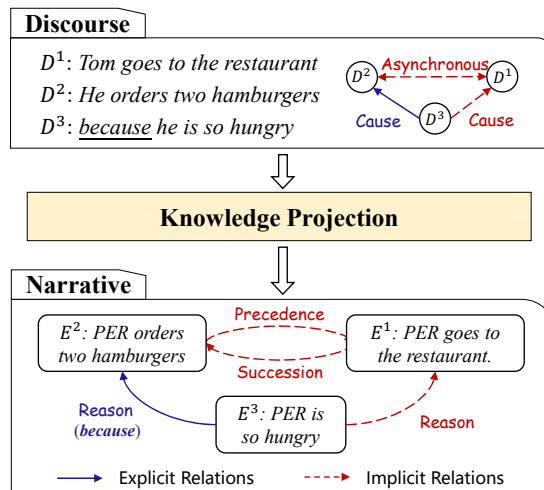
*Corresponding authors.

Figure 1: The knowledge projection paradigm for event relation extraction. The explicit projection directly projects connectives to event relations, e.g., from *"because"* to *Reason*. The implicit projection leverages the discourse knowledge to discover implicit event relations without connectives via MKPNet.

EventKGs. Commonly, each node in EventKGs represents an event, and each edge represents a predefined relation between an event pair[1]. Currently, event relations are majorly extracted based on the explicit connectives between them. For example, in Figure 1, a *Reason* relation is extracted between E^2: *"PER orders two hamburgers"* and E^3: *"PER is so hungry"* using the explicit connective *"because"* between them.

Unfortunately, the connective-based approaches face the critical coverage problem due to the sparsity of connectives. That is, a large proportion of event pairs are not connected with explicit connectives, but with underlying event relations. We denote them as ***implicit event relations***. Further-

[1]Computational and cognitive studies define nodes as eventualities, which include activities, states and events. In this paper, we simplify the definition of each node to "event" due to its popularity.

more, the related events can even not close to each other in a document. For the example in Figure 1, the implicit relation *Reason* between E^1: *"PER goes to the restaurant"* and E^3: *"PER is so hungry"* can not be extracted due to the absence of explicit connective as well as the discontinuity between these two clauses. The common practice in previous connective-based approaches is to ignore all these implicit instances (Zhang et al., 2020b). As a result, the coverage of EventKGs is significantly undermined. Besides, because the scale of the existed event relation corpus (Hong et al., 2016) is limited, it is also impractical to build effective event relation classifiers via supervised learning.

In this paper, we propose a new paradigm for event relation extraction — knowledge projection. Instead of relying on sparse connectives or building classifiers starting from scratch, we project discourse knowledge to event narratives by exploiting the anthropological linguistic connections between them. Enlightened by Livholts and Tamboukou (2015); Altshuler (2016); Reyes and Wortham (2017), discourses and narratives have significant associations, and their knowledge are shared at different levels: 1) token-level knowledge: discourses and narratives share similar lexical and syntactic structures, 2) semantic-level knowledge: the semantics entailed in discourse pairs and event pairs are analogical, e.g., E^3-*Reason*→E^1 and D^3-*Cause*→D^1 in Figure 1., and 3) label-level knowledge: heterogeneous event and discourse relations have the same coarse categories, e.g., both the event relation *Reason* and the discourse relation *Cause* are included in the coarse-grained relation *Contingency*. By exploiting the rich knowledge in manually labelled discourse corpus and projecting them into event relation extraction models, the performance of event relation extraction can be significantly improved, and the data requirement can be dramatically reduced.

Specifically, we design **M**ulti-tier **K**nowledge **P**rojection **Net**work (**MKPNet**), which can leverage multi-tier discourse knowledge effectively for event relation extraction. MKPNet introduces three kinds of adaptors to project knowledge from discourses into narratives: (a) token adaptor for token-level knowledge projection; (b) semantic adaptor for semantic-level knowledge projection; (c) coarse category adaptor for label-level knowledge projection. By sharing the parameters of these three adaptors, the commonalities between discourses and narratives at various levels can be effectively explored. Therefore, we can obtain more general token representations, more accurate semantic representations, and more credible coarse category representations to better predict event relations.

We conduct intrinsic experiments on ASER (Zhang et al., 2020b), one of the representative EventKGs, and extrinsic experiments on Winograd Scheme Challenge (WSC) (Levesque et al., 2012), one of the representative natural language understanding benchmarks. Intrinsic experimental results show that the proposed MKPNet significantly outperforms the state-of-the-art (SoA) baselines, and extrinsic experimental results verify the value of the extracted event relations[2].

The main contributions of this paper are:

- We propose a new knowledge projection paradigm, which can effectively leverage the commonalities between discourses and narratives for event relation extraction.

- We design MKPNet, which can effectively leverage multi-tier discourse knowledge for event relation extraction via token adaptor, semantic adaptor and coarse category adaptor.

- Our method achieves the new SotAevent relation extraction performance, and an enriched EventKG is released by extracting both explicit and implicit event relations. We believe it can benefit many downstream NLP tasks.

2 Background

Event Relation Extraction (ERE). Given an existing EventKG $\mathcal{G} = \{\mathcal{E}, \mathcal{R}\}$, where nodes $\mathcal{E}$ are events and edges $\mathcal{R}$ are their relations. $Y^{ex} \in \mathcal{R}$ are explicit event relations extracted by connective-based methods, and $Y^{im} \notin \mathcal{R}$ are implicit event relations without connectives. Commonly, implicit event relation extraction (IERE) takes two events $E^1 = \{e_1^1, ..., e_{|E^1|}^1\}$, $E^2 = \{e_1^2, ..., e_{|E^2|}^2\}$, $E^1, E^2 \in \mathcal{E}$ as inputs, then uses a neural network to classify their underlying relation.

Discourse Relation Recognition (DRR). DRR aims to recognize the relation of two discourse arguments. Discourse relations can be explicit or implicit, where explicit relations are revealed by

---

[2]Our source codes with corresponding experimental datasets and the enhanced EventKG are openly available at `https://github.com/TangJiaLong/Knowledge-Projection-for-ERE`.
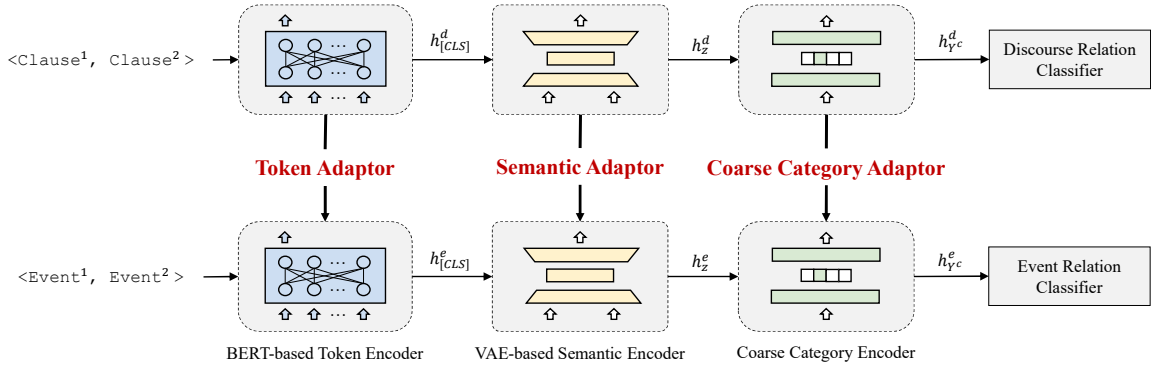
Figure 2: An overview of MKPNet, which projects discourse knowledge for event relation extraction: (a) token adaptor for token-level knowledge projection, (b) semantic adaptor for semantic-level knowledge projection, and (c) coarse category adaptor for label-level knowledge projection.

connectives, while implicit relations lack these surface cues. To resolve the implicit discourse relation recognition (IDRR) task, researchers construct high-quality labelled datasets (Prasad et al., 2008) and design elaborate models (Zhang et al., 2016b; Bai and Zhao, 2018; Kishimoto et al., 2020).

Associations between Discourse and Narrative. Recent NLP studies have proved that discourse and narratives closely interact with each other, and leveraging discourse knowledge benefits narrative analysis significantly, such as subevents detection (Aldawsari and Finlayson, 2019) and main event relevant identification (Choubey et al., 2020). Motivated by the above observation, this paper leverages the knowledge of discourse by a knowledge projection paradigm. Blessed with the associations at token-, semantic- and coarse category-levels, the discourse corpora and knowledge can be effectively exploited for event relation extraction.

3 Multi-tier Knowledge Projection Network for Event Relation Extraction

In this section, we describe how to learn an effective event relation extractor by projecting resource-rich discourse knowledge to the resource-poor narrative task. Specifically, we propose **M**ulti-tier **K**nowledge **P**rojection **Net**work (**MKPNet**) which can effectively leverage multi-tier discourse knowledge for implicit event relation extraction. Figure 2 shows an overview of MKPNet, which uses token adaptor, semantic adaptor and coarse category adaptor to fully exploit discourse knowledge at different levels. In the following, we first describe the neural architecture of MKPNet and then describe the details of three adaptors.

3.1 Neural Architecture of MKPNet

For knowledge projection, we model both event relation extraction (ERE) and discourse relation recognition (DRR) as an instance-pair classification task (Devlin et al., 2019; Kishimoto et al., 2020). For ERE, the input is an event pair such as $<E^1$: *"PER goes to the restaurant"*, E^3: *"PER is so hungry"*> and the output is an event relation such as *Reason*. For DRR, the input is a clause pair such as $<D^1$: *"Tom goes to the restaurant"*, D^3:*"he is so hungry"*> and the output is a discourse relation such as *Cause*.

Specifically, MKPNet extends the SotADRR model — BERT-CLS (Kishimoto et al., 2020) by the VAE-based semantic encoder and the coarse category encoder to model knowledge tier-by-tier (Pan et al., 2016; Guo et al., 2019; Kang et al., 2020; Li et al., 2020b). It 1) first utilizes the BERT-based token encoder to encodes an instance pair as a token representation $h_{[CLS]}$; 2) then obtains the semantic representation h_z via a VAE-based semantic encoder; 3) predicts the coarse-grained label and embeddings it as the coarse category representation h_{Y^c}; 4) finally classifies its relation with the guidance of the aggregate instance-pair representation:

$$Y = Classifier^{Fine}([h_{[CLS]} \oplus h_z \oplus h_{Y^c}]) \quad (1)$$

where $\oplus$ means the concatenation operation. In this way, the parameters of MKPNet can be grouped by $\{\theta^{BERT}, \theta^{Semantic}, \theta^{Coarse}, \theta^{Fine}\}$, where θ^{BERT} for BERT-based token encoder, $\theta^{Semantic}$ for VAE-based semantic encoder, θ^{Coarse} for coarse category encoder and θ^{Fine} for the final relation classifier layer respectively.

3.2 Token Adaptor

Recent studies have shown that similar tasks usually share similar lexical and syntactic structures and therefore lead to similar token representations (Pennington et al., 2014; Peters et al., 2018). The token adaptor tries to improve the token encoding for ERE by sharing the parameters θ^{BERT} of the BERT-based encoders with DRR. In this way, the encoder is more effective due to the more supervision signals and is more general due to the multi-task settings.

Specifically, given an event pair $<E^1, E^2>$, we represent it as a sequence:

$$[CLS], e_1^1, ..., e_{|E^1|}^1, [SEP], e_1^2, ..., e_{|E^2|}^2, [SEP]$$

where[CLS] and [SEP] are special tokens. For each token in the input, its representation is constructed by concatenating the corresponding token, segment and position embeddings. Then, the event pair representation will be inputted into BERT architecture (Devlin et al., 2019) and updated by multi-layer Transformer blocks (Vaswani et al., 2017). Finally, we obtain the hidden state corresponding to the special [CLS] token in the last layer as the token-level event pair representation:

$$h_{[CLS]}^e = BERT(E^1, E^2) \qquad (2)$$

The token-level discourse pair representation $h_{[CLS]}^d$ can be obtained in the same way for DRR.

To project the token-level knowledge, we use the same BERT for event pair and discourse pair encoding. During the optimization process, it is fine-tuned using the supervision signals from both ERE and DRR.

3.3 Semantic Adaptor

Because narrative and discourse analyses need to accurately represent the deeper semantic of the instance pairs, the shallow token-level knowledge captured by the BERT-based token encoder is not enough. However, BERT always induces a non-smooth anisotropic semantic space which is adverse for semantic modelling of large-grained linguistic units (Li et al., 2020a).

To address this issue, we introduce an variational autoencoder-based (VAE-based) semantic encoder to represent the semantics of both events and clauses by transforming the anisotropic semantic distribution to a smooth and isotropic Gaussian distribution (Kingma and Welling, 2014; Rezende
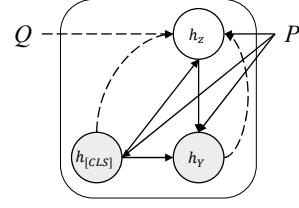

Figure 3: The illustration of the semantic encoder as a directed graph. We use solid lines to denote the generative model $P = p(h_Y|h_{[CLS]}, h_z)p(h_z|h_{[CLS]})$, and dashed lines to denote the variational approximation $Q = q(h_z|h_{[CLS]}, h_Y)$. Both variational parameters and generative parameters are learned jointly.

et al., 2014; Sohn et al., 2015). To better learn the semantic encoder, the semantic adaptor shares the parameters $\theta^{Semantic}$ of it between ERE and DRR and train it using both classification supervision signals and KL divergence.

Specifically, VAE is a directed graphical model with the generative model P and the variational model Q, which learns the semantic representation h_z of the input by an autoencoder framework. Figure 3 illustrates the graphic representation of the semantic encoder. Specifically, we assume that there exists a continuous latent variable $h_z \sim \mathcal{N}(\mu, diag(\sigma^2))$, where μ and σ^2 are mean and variance of the Gaussian distribution respectively. With this assumption, the original conditional probability of the event/discourse relations can be expressed by the following formula:

$$p(h_Y|h_{[CLS]}) = \int_{h_z} p(h_Y|h_{[CLS]}, h_z) \\ p(h_z|h_{[CLS]})dh_z \qquad (3)$$

The posterior approximation is $q(h_z|h_{[CLS]}, h_Y)$, where $h_{[CLS]}$ can be $h_{[CLS]}^e$ or $h_{[CLS]}^d$ and h_Y can be h_Y^e or h_Y^d according to the different tasks. We 1) first obtain the input- and output-side representations via the shared BERT-based token encoder and the individual relation embedding networks, i.e., $h_{[CLS]}$ and h_Y; 2) then perform a non-linear transformation that project them onto the semantic space:

$$h_z' = tanh(W_z[h_{[CLS]}; h_Y] + b_z) \qquad (4)$$

3) obtain the above-mentioned Gaussian parameters μ and $log\sigma^2$ through linear regression:

$$\mu = W_\mu h_z' + b_\mu, \quad log\sigma^2 = W_\sigma h_z' + b_\sigma \qquad (5)$$

where W and b are the parameter matrix and bias term respectively; 4) use a reparameteriza-

tion trick (Kingma and Welling, 2014; Sohn et al., 2015) to get the final semantic representation:

$$h_z = \mu + \sigma \odot \epsilon \qquad (6)$$

where $\epsilon \sim \mathcal{N}(0, I)$ and h_z can be h_z^e or h_z^d.

The neural model for the prior $p(h_z|h_{[CLS]})$ is the same as that for the posterior $q(h_z|h_{[CLS]}, h_Y)$, except for the absence of h_Y. Besides, those two models have parameters independent of each other.

During testing, due to the absence of the output-side representation h_Y, we set h_z to be the mean of $p(h_z|h_{[CLS]})$ (Zhang et al., 2016a), i.e., μ. During training, we minimize the Kullback-Leibler divergence $KL(P||Q)$ between the generation model P and the inference model Q. Intuitively, KL divergence connects these two models:

$$KL(q(h_z|h_{[CLS]}, h_Y)||p(h_z|h_{[CLS]})) \qquad (7)$$

To project the semantic-level knowledge, we use the same VAE for both event pair and discourse pair. Therefore, the commonalities of event semantics and discourse semantics can be captured more accurately.

3.4 Coarse Category Adaptor

The token adaptor and the semantic adaptor commendably cover the knowledge entailed on the input-side. In addition, we found that ERE and DRR share the same coarse-grained categories: *Temporal*, *Contingency*, *Comparison* and *Expansion* (Prasad et al., 2008; Zhang et al., 2020b), although they have different fine-grained categories.

To this end, we design the coarse category adaptor in a coarse-to-fine framework (Petrov, 2009) to bridge the gap between the heterogeneous fine-grained targets. Specifically, we share the parameters θ^{Coarse} of the coarse-grained classifier and the coarse label embedding network to obtain more credible coarse category representations.

Specifically, we first use the token representation $h_{[CLS]}$ and the semantic representation h_z to predict the coarse-grained labels:

$$Y^c = Classifier^{Coarse}(h_{[CLS]}, h_z) \qquad (8)$$

where $Y^c \in \{$*Temporal*, *Contingency*, *Comparison*, *Expansion*$\}$. After that, we use the coarse label embedding network to obtain the corresponding coarse-grained label embedding h_{Y^c}, which is referred as the coarse category representation.

To project that label-level knowledge, we use the same coarse-grained classifier and the same coarse label embedding network. During the optimization process, both event instances and discourse instances can be used to train this coarse category encoder. The more supervision signals make it more effective.

3.5 Full Model Training

In this paper, we utilize multi-task learning (Caruana, 1997) to implement the knowledge projection from discourse to narrative. It expects correlative tasks (ERE and DRR) can help each other to learn better by sharing the parameters of three adaptors. Given ERE and DRR training datasets, an alternate optimization approach (Dong et al., 2015) is used to optimize MKPNet:

$$\begin{aligned} L(\theta) =\ & \alpha(L(\theta; Y) + \lambda KL(P||Q)) \\ & + (1 - \alpha)L(\theta; Y^c) \end{aligned} \qquad (9)$$

where Y can be Y^{im} or Y^d according to the different tasks, λ, α are two hyperparameters, $KL(P||Q))$ is the KL divergence in the semantic encoder, $L(\theta; Y)$ and $L(\theta; Y^c)$ are fine-grained and coarse-grained objectives respectively:

$$L(\theta; Y) = \log p(Y|h_{[CLS]}, h_z, h_{Y^c}) \qquad (10)$$
$$L(\theta; Y^c) = \log p(Y^c|h_{[CLS]}, h_z) \qquad (11)$$

It should be noticed that in MKPNet, $\{\theta^{BERT}, \theta^{Semantic}, \theta^{Coarse}\}$ are the shared parameters of the BERT-based token encoder, the VAE-based semantic encoder and the coarse category encoder between ERE and DRR. And $\{\theta^{Fine}\}$ are separated parameters of the fine-grained ERE and DRR classifiers.

4 Experiments

We conduct intrinsic experiments on ASER (Zhang et al., 2020b) to assess the effectiveness of the proposed MKPNet, and extrinsic experiments on WSC (Levesque et al., 2012) to verify the value of the extracted event relations.

4.1 Intrinsic Experiments

Datasets. For discourse relation recognition (DRR), we use PDTB 2.0 (Prasad et al., 2008) with the same splits of Ji and Eisenstein (2015): sections 2-20/0-1/21-22 respectively for train/dev/test. For event relation extraction (ERE), because there is no labelled training corpus, we construct a new dataset by removing the connectives of the explicit event

relation instances in ASER core version[3] and retaining at most 2200 instances with the highest confidence scores for each category[4]. In this way, we obtain 23,181/1400/1400 train/dev/test instances – we denoted it as implicit event relation extraction (IERE) dataset.

Implementation. We implement our model based on pytorch-transformers (Wolf et al., 2020). We use BERT-base and set all hyper-parameters using the default settings of the SotADRR model (Kishimoto et al., 2020).

Baselines. For ERE, we compare the proposed MKPNet with the following baselines:

- Baselines w/o Discourse Knowledge are only trained on IERE training set. We choose the BERT-CLS as the representative of them due to its SotAperformance.

- Baselines with Discourse Knowledge improve the learning of ERE via transfer learning (Pan and Yang, 2009; Pan et al., 2010) from discourse models, i.e., first pre-train a parameter prior on PDTB 2.0 and then fine-tune it on IERE — we denote it as BERT-Transfer.

For DRR, we compare the proposed MKPNet with the following baselines:

- Bai and Zhao (2018) is a deep neural network model augmented by variable grained text representations like character, sentence and sentence pair levels.

- Kishimoto et al. (2020) is the SotADRR model, BERT-CLS, which incorporating BERT with one additional output layer.

4.1.1 Overall Results

Table 1-3 show the overall ERE/DRR results of baselines and MKPNet. For our approach, we use the full MKPNet and its four ablated settings: MKPNet w/o SA, MKPNet w/o CA, MKPNet w/o SA & CA and MKPNet w/o KP, where SA, CA and KP denote semantic adaptor, coarse category adaptor and knowledge projection correspondingly. We can see that:

1. Based on MKPNet, we enrich the original ASER by abundant implicit event relations. Considering the computational complexity, we classify the event pairs co-occurrence in the same document

---

[4] Higher confidence score means more credible instance.

	Number of Relations
FrameNet (Baker et al., 1998)	1,709
ConceptNet (Speer et al., 2017)	116,097
Event2Mind (Smith et al., 2018)	57,097
ATOMIC (Sap et al., 2019)	877,108
Knowlywood (Tandon et al., 2015)	2,644,415
ASER (Zhang et al., 2020b)	1,287,059
ASER++ (core)	2,034,963
ASER++ (high)	3,530,771
ASER++ (full)	8,766,098

Table 1: Number comparison of event relations in ASER++ and existing event-related resources.

	Acc	F1
Baselines w/o Discourse Knowledge		
BERT-CLS	53.00	52.24
MKPNet w/o KP	53.94	53.52
Baselines with Discourse Knowledge		
BERT-Transfer	54.29	53.44
Multi-tier Knowledge Projection		
MKPNet w/o SA & CA	$54.79_{\triangle 0.85}$	$53.90_{\triangle 0.38}$
MKPNet w/o CA	$55.14_{\triangle 1.20}$	$54.42_{\triangle 0.90}$
MKPNet w/o SA	$55.29_{\triangle 1.35}$	$54.92_{\triangle 1.40}$
MKPNet	$\mathbf{55.86}_{\triangle 1.92}$	$\mathbf{55.36}_{\triangle 1.84}$

Table 2: Experimental results on IERE test set, where $\triangle$ means the improvements when compared with MKPNet w/o KP. All improvements of MKPNet are statistical significance at p<0.01 over the baseline MKPNet w/o KP.

Model	Acc
Bai and Zhao (2018)	48.22
BERT-CLS (Kishimoto et al., 2020)	51.40
BERT-CLS (Ours)	50.91
MKPNet w/o KP	52.86
MKPNet	**54.09**

Table 3: Experimental results on PDTB 2.0 test set. For a fair comparison, the results of baselines are adapted from their original papers.

and filter them by confidence scores. Specifically, we compute the confidence score by multiplying the classification probability and the frequency of the event pair. Integrating with the original explicit event relations, we can obtain the enriched EventKGs ASER++ (core/high/full) with the different threshold confidences (3/2/1). Table 1 shows that when compared with existing event-related resources, ASER++ has an overwhelming advantage in the number of event relations.

2. The proposed MKPNet achieves SotAperformance for ERE. MKPNet can significantly outperform the BERT-Transfer and achieves 55.86 accuracy and 55.36 F1. MKPNet w/o KP obtains considerable performance improvements when com-

pared with BERT-CLS. We believe this is because MKPNet fully explores the knowledge on different tiers, and modelling knowledge tier-by-tier is effective.

3. By projecting knowledge at token-level, semantic level and label level, all three adaptors are useful and are complementary with each other. When compared with the full model MKP-Net, its four variants show declined performance in different degrees. MKPNet outperforms MKPNet w/o CA 0.72 accuracy and 0.94 F1, which indicates that our coarse category adaptor successfully bridges the gap of heterogeneous fine-grained targets. MKPNet outperforms MKPNet w/o SA 0.57 accuracy and 0.44 F1, and therefore we believe that our latent semantic adaptor is helpful for capture the semantic-level commonalities. Finally, there is a significant decline between MKPNet w/o KP and MKPNet w/o SA & CA, which means that token adaptor is indispensable. The insight in those observations is that the commonalities between discourses and narratives under the hierarchical structure, thus projecting them at different levels is effective, and three adaptors can be complementary with each other.

4. The commonalities between discourses and narratives are beneficial for both ERE and DRR. Compared with the baselines w/o discourse knowledge — BERT-CLS and MKPNet w/o KP, both the naive transfer method — BERT-Transfer and our MKPNet achieve significant performance improvements: BERT-Transfer gains 1.29 accuracy and 1.20 F1 when compared to BERT-CLS, and MKP-Net gains 1.92 accuracy and 1.84 F1 when compared to MKPNet w/o KP. Besides, for DRR, our method MKPNet also substantially outperforms the other baselines and its variant MKPNet w/o KP. These results verified the commonalities between discourse knowledge and narrative knowledge.

4.1.2 Detailed Analysis

Effects of Semantic-level Knowledge and Label-level Knowledge. In these experiments, we compare the performance of our models, MKPNet, MKPNet w/o CA and MKPNet w/o SA with or without knowledge projection to find out the effects of semantic-level knowledge and label-level knowledge. From Table 4, we can see that: (1) Compared with their counterparts, MKPNet, MKP-Net w/o CA and MKPNet w/o SA with knowledge projection lead to significant improvements. Thus, it is convincing that the performance improvements

	KP	Fine-grained		Coarse-grained	
		Acc	F1	Acc	F1
BERT-CLS		53.00	52.24	—	—
MKPNet		53.94	53.52	—	—
	✓	**55.86**	**55.36**	—	—
MKPNet w/o CA		53.79	53.39	—	—
	✓	**55.14**	**54.42**	—	—
MKPNet w/o SA		53.21	52.48	66.57	63.04
	✓	**55.29**	**54.92**	**67.93**	**64.76**
MKPNet w/o SA*		70.50	70.32	100.0	100.0

Table 4: Effect of semantic-level knowledge and label-level knowledge on IERE test set, where **KP** stands for knowledge projection and * stands for golden coarse-grained categories.
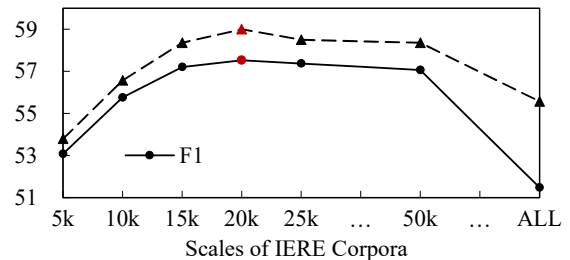

Figure 4: Experimental results of using different sizes of IERE training corpora.

mainly come from the discourse knowledge rather than the neural architecture; (2) Current knowledge projection can be further improved by exploiting more accurate discourse knowledge: MKPNet w/o SA*, which uses golden coarse categories, achieves striking performance (Acc 70.50; F1 70.32).

Tradeoff between Dataset Quality and Size. As described above, the IERE training dataset is constructed using the most confident instances in ASER core version. We can construct a larger but lower quality dataset by incorporating more instances with lower confidence, i.e., the quality-size tradeoff problem. To analyze the tradeoff between the quality and size, we construct a set of datasets with different sizes/qualities, and Figure 4 shows the corresponding results of MKPNet on the development set. We can see that the size is the main factors for performance improvements at the beginning: every 5,000 additional instances can result in a significant improvement (about 2 to 3 F1 gain). When the size is large (more than 20,000 instances in our experiments), more instances will not result in performance improvements, and the low-quality instances will hurt the performance.
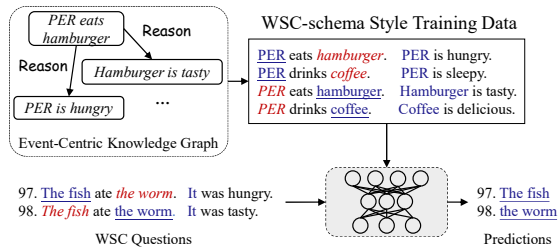
738

Figure 5: An overview of WSC implementation. The blue color means the correct reference while the red color means the wrong one.

4.2 Extrinsic Experiments

The above intrinsic experiments verified the effectiveness of the proposed MKPNet for ERE. In this section, we use the core version of our enriched EventKGs — ASER++, and then conduct extrinsic experiments on Winograd Schema Challenge (WSC) (Levesque et al., 2012) to verify the effect of ASET++.

WSC Implementation. WSC is challenging since its schema is a pair of sentences that differ only in one or two words and that contain a referential ambiguity that is resolved in opposite directions in the two sentences. According to Certu et al. (2019), fine-tuning pre-trained language models on WSC-schema style training sets is a robust method to tackle WSC. Therefore, as Figure 5 shows, we transform ASER++ to WSC-schema style training data in the same way as Zhang et al. (2020b) and fine-tune BERT on it, which we refer to as BERT (ASER++). We compare BERT (ASER++) with these baselines:

- **Pure Knowledge-based Methods** are heuristical rule-based methods, such as Knowledge Hunting (Emami et al., 2018) and String Match (Zhang et al., 2020b).

- **Language Model-based Methods** use language model trained on large-scale corpus and tuned specifically for the WSC task, such as LM (Trinh and Le, 2018).

- **External Knowledge Enhanced Methods** are models based on BERT and trained with the different external knowledge resource, e.g., WscR (Ng, 2012; Certu et al., 2019)

We implement our model based on pytorch-transformers (Wolf et al., 2020). BERT-large is used. All hyper-parameters are default settings as Certu et al. (2019).

Model	WSC
Pure Knowledge-based Methods	
Knowledge Hunting (Emami et al., 2018)	57.3
String Match (Zhang et al., 2020b)	56.6
Language Model-based Methods	
LM (Single) (Trinh and Le, 2018)	54.5
LM (Ensemble) (Trinh and Le, 2018)	61.5
BERT (w/o finetuning) (Devlin et al., 2019)	61.9
External Knowledge Enhanced Methods	
BERT (WscR) Certu et al. (2019)	71.4
BERT (ASER) Zhang et al. (2020b)	64.5
BERT (ASER & WscR) Zhang et al. (2020b)	72.5
BERT (ASER++)	66.2
BERT (ASER++ & WscR)	**74.1**

Table 5: The overall results of extrinsic experiments. The evaluation metric is accuracy.

Extrinsic Results. Table 5 shows the overall results of extrinsic experiments. We can see that: By fine-tuning BERT on our enriched EventKG — ASER++, the WSC performance can be significantly improved. BERT (ASER++) and BERT (ASER++ & WscR) outperform BERT (ASER) and BERT (ASER & WscR) respectively, which verified the effectiveness of ASER++ and implicit event relations are beneficial for downstream NLU tasks.

5 Related Work

Event-centric Knowledge Graphs. Knowledge graphs have come from entity-centric ones (Banko et al., 2007; Suchanek et al., 2007; Bollacker et al., 2008; Wu et al., 2012) to event-centric ones. However, the construction of traditional KGs takes domain experts much effort and time, which are often with limited size and cannot effectively resolve real-world applications, e.g., FrameNet (Baker et al., 1998). Recently, many modern and large-scale KGs have been built semi-automatically, which focus on events (Tandon et al., 2015; Rospocher et al., 2016; Gottschalk and Demidova, 2018; Zhang et al., 2020b) and commonsense (Speer et al., 2017; Smith et al., 2018; Huang et al., 2018; Sap et al., 2019). Specifically, Yu et al. (2020) proposes an approach to extract entailment relations between eventualities, e.g., *"I eat an apple"* entails *"I eat fruit"*, and release an event entailment graph (EEG). Different from EEG, this paper focuses on implicit event relations which are not extracted due to the absences of the connectives and discontinuity.

Knowledge Transfer. Due to the data scarcity problem, many knowledge transfer studies have been proposed, including multi-task learning (Caru-

ana, 1997), transfer learning (Pan and Yang, 2009; Pan et al., 2010), and knowledge distillation (Hinton et al., 2014). Recently, researchers are interested in training/sharing/transferring/distilling models layer by layer to fully excavate the knowledge (Pan et al., 2016; Guo et al., 2019; Kang et al., 2020; Li et al., 2020b). In this paper, we propose a knowledge projection method which can project discourse knowledge to narraties on different tiers.

6 Conclusions

In this paper, we propose a knowledge projection paradigm for event relation extraction and **Multi-tier Knowledge Projection Network** (**MKPNet**) is designed to leverage multi-tier discourse knowledge. By effectively projecting knowledge from discourses to narratives, MKPNet achieves the new state-of-the-art event relation extraction performance, and extrinsic experimental results verify the value of the extracted event relations. For future work, we want to design new data-efficient algorithms to learn effective models using low-quality and heterogeneous knowledge.

Acknowledgments

References

Mohammed Aldawsari and Mark Finlayson. 2019. Detecting subevents using discourse and narrative features. In *Proceedings of ACL 2019*.

Daniel Altshuler. 2016. *Events, States and Times: An Essay on Narrative Discourse in English*. Walter de Gruyter GmbH & Co KG.

Hongxiao Bai and Hai Zhao. 2018. Deep enhanced representation for implicit discourse relation cecognition. In *Proceedings of ICCL 2018*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of COLING-ACL 1998*.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of IJCAI 2007*.

Kurt D. Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of SIGMOD 2008*.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1).

Vid Kocijanand Ana-Maria Certu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. A surprisingly robust trick for the winograd schema challenge. In *Proceedings of ACL 2019*, pages 4837–4842.

Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of ACL 2020*.

Tarcísio Souza Costa, Simon Gottschalk, and Elena Demidova. 2020. Event-qa: A dataset for event-centric question answering over knowledge graphs. *arXiv preprint arXiv:2004.11861*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL 2019*.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of ACL-IJCNLP 2015*.

Ali Emami, Noelia De La Cruz, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2018. A knowledge hunting framework for common sense reasoning. In *Proceedings of EMNLP 2018*, pages 1949–1958.

Simon Gottschalk and Elena Demidova. 2018. Eventkg: A multilingual event-centric temporal knowledge graph. In *Proceedings of ESWC 2018*.

Yunhui Guo, Honghui Shi, Abhishek Kumar, Kristen Grauman, Tajana Rosing, and Rogerio Feris. 2019. Spottune: Transfer learning through adaptive fine-tuning. In *Proceedings of CVPR 2019*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dea. 2014. Distilling the knowledge in a neural network. In *Proceedings of The Workshop on NIPs 2014*.

Yu Hong, Tongtao Zhang, Tim O'Gorman, Sharone Horowit-Hendler, Heng Ji, and Martha Palmer. 2016. Building a cross-document event-event relation corpus. In *Proceedings of ACL 2016*.

Bhavana Dalviand Lifu Huang, Niket Tandon, Wen tau Yih, and Peter Clark. 2018. Tracking state changes in procedural text: A challenge dataset and models for process paragraph comprehension. In *Proceedings of NAACL-HLT 2018*.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. In *Proceedings of TACL 2015*.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. 2020. Decoupling representation and classifier for long-tailed recognition. In *Proceedings of ICLR 2020*.

Kingma and Welling. 2014. Auto-encoding variational bayes. In *Proceedings of ICLR 2014*.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2020. Adapting bert to implicit discourse relation classification with a focus on discourse connectives. In *Proceedings of LREC 2020*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of KR 2012*.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020a. On the sentence embeddings from pre-trained language models. In *Proceedings of EMNLP 2020*.

Jianquan Li, Xiaokang Liu, Honghong Zhao, Ruifeng Xu, Min Yang, and Yaohong Jin. 2020b. Bertemd: Many-to-many layer mapping for bert compression with earth mover's distance. In *Proceedings of EMNLP 2020*.

Mona Livholts and Maria Tamboukou. 2015. *Discourse and Narrative Methods: Theoretical Departures, Analytical Strategies and Situated Writings*. Sage.

Altaf Rahmanand Vincent Ng. 2012. Resolving complex cases of definite pronouns: the winograd schema challenge. In *Proceedings of EMNLP–CoNLL 2012*, pages 777–789.

Jianhan Pan, Xuegang Hu, Peipei Li, Huizong Li, Wei He, Yuhong Zhang, and Yaojin Lin. 2016. Domain adaptation via multi-layer transfer learning. *Neurocomputing*.

Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. 2010. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2).

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *TKDE*, 22(10).

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP 2014*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL 2018*.

Slav Orlinov Petrov. 2009. *Coarse-to-Fine Natural Language Processing*. Ph.D. thesis, EECS Department, University of California, Berkeley.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of LREC 2008*.

Angela Reyes and Stanton Wortham. 2017. *Discourse and Education*, chapter Discourse Analysis Across Events. Springer International Publishing.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of ICML 2014*.

Marco Rospocher, Marieke van Erp, Piek Vossen, Antske Fokkens, Itziar Aldabe, German Rigau, Aitor Soroa, Thomas Ploeger, and Tessel Bogaard. 2016. Building event-centric knowledge graphs from news. *Journal of Web Semantics*.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of AAAI 2019*.

Noah A. Smith, Yejin Choi, Maarten Sap, Hannah Rashkin, and Emily Allaway. 2018. Event2mind: Commonsense inference on events, intents, and reactions. In *Proceedings of ACL 2018*.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Proceedings of NIPS 2015*.

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of AAAI 2017*.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A core of semantic knowledge. In *Proceedings of WWW 2007*.

Niket Tandon, Gerard De Melo, Abir De, and Gerhard Weikum. 2015. Knowlywood: Mining activity knowledge from hollywood narratives. In *Proceedings of CIKM 2015*.

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomezand Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NIPS 2017*.

Piek Vossen. 2018. Newsreader at semeval-2018 task 5: Counting events by reasoning over event-centric-knowledge-graphs. In *Proceedings of The 12th International Workshop on Semantic Evaluation*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP 2020: System Demonstrations*.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q. Zhu. 2012. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of SIGMOD 2012*.

Changlong Yu, Hongming Zhang, Yangqiu Song, Wilfred Ng, and Lifeng Shang. 2020. Enriching large-scale eventuality knowledge graph with entailment relations. In *Proceedings of AKBC 2020*.

Biao Zhang, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. 2016a. Variational neural machine translation. In *Proceedings of EMNLP 2016*.

Biao Zhang, Deyi Xiong, Jinsong Su, Qun Liu, Rongrong Ji, Hong Duan, and Min Zhang. 2016b. Variational neural discourse relation recognizer. In *Proceedings of EMNLP 2016*.

Hongming Zhang, Daniel Khashabi, Yangqiu Song, and Dan Roth. 2020a. Transomcs: From linguistic graphs to commonsense knowledge. In *Proceedings of IJCAI 2020*.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020b. Aser: A large-scale eventuality knowledge graph. In *Proceedings of WWW 2020*.