

ERICA: Improving Entity and Relation Understanding for Pre-trained Language Models via Contrastive Learning

Yujia Qin^{*†‡}, Yankai Lin[†], Ryuichi Takanobu^{*†}, Zhiyuan Liu^{*}, Peng Li[†], Heng Ji[‡],
Minlie Huang^{*}, Maosong Sun^{*}, Jie Zhou[†]

^{*}Department of Computer Science and Technology, Tsinghua University, Beijing, China

[‡]University of Illinois at Urbana-Champaign

[†]Pattern Recognition Center, WeChat AI, Tencent Inc.

yujiaqin16@gmail.com

Abstract

Pre-trained Language Models (PLMs) have shown strong performance in various downstream Natural Language Processing (NLP) tasks. However, PLMs still cannot well capture the factual knowledge in the text, which is crucial for understanding the whole text, especially for document-level language understanding tasks. To address this issue, we propose a novel contrastive learning framework named ERICA in pre-training phase to obtain a deeper understanding of the entities and their relations in text. Specifically, (1) to better understand entities, we propose an entity discrimination task that distinguishes which tail entity can be inferred by the given head entity and relation. (2) Besides, to better understand relations, we employ a relation discrimination task which distinguishes whether two entity pairs are close or not in relational semantics. Experimental results demonstrate that our proposed ERICA framework achieves consistent improvements on several document-level language understanding tasks, including relation extraction and reading comprehension, especially under low resource setting. Meanwhile, ERICA achieves comparable or better performance on sentence-level tasks. We will release the datasets, source codes and pre-trained language models for further research explorations.

1 Introduction

Pre-trained Language Models (PLMs) (Devlin et al., 2018; Yang et al., 2019; Liu et al., 2019) have achieved great success on various Natural Language Processing (NLP) tasks such as text classification (Wang et al., 2018), named entity recognition (Sang and De Meulder, 2003), and question answering (Talmor and Berant, 2019). Through the self-supervised learning objectives such as Masked Language Modeling (MLM), PLMs can effectively

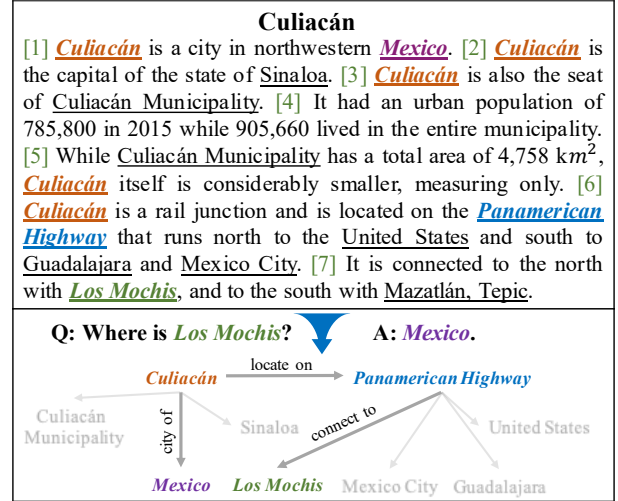


Figure 1: An example for a document “Culiacán”. All entities in the document are underlined. We show entities and their relations as a relational graph, and highlight the important entities and relations to find out “Where is Los Mochis”.

capture the syntax and semantics in the text to generate informative representations for downstream NLP tasks.

Recently, researchers attempt to address the limitation of textual understanding of PLMs by leveraging external knowledge. They propose to enhance PLMs’ ability of understanding entities (Joshi et al., 2020; Xiong et al., 2019; Zhang et al., 2019; Wang et al., 2019; Peters et al., 2019; Soares et al., 2019; Peng et al., 2020) or relations (Soares et al., 2019; Peng et al., 2020) in the text.

However, these knowledge-guided PLMs only focus on understanding individual entities or relations, ignoring the complex interactions between entities and relations in the text. Therefore, these models still cannot well capture the factual knowledge in the text, which is crucial to understand the whole text, especially for documents, since a document may contain multiple entities and their

relations are more complex. As depicted in Figure 1, to understand that “Los Mochis is located in Mexico”, we need to consider the following clues jointly: (i) “Mexico” is the country of “Culiacán” from sentence 1; (ii) “Culiacán” is a rail junction located on “Panamerican Highway” from sentence 6; (iii) “Panamerican Highway” connects to “Los Mochis” from sentence 7. From the example, we can see that there are two main challenges to capture the factual knowledge in natural language text:

1. How to **understand the entities** via considering various kinds of relations among them in the text. In the example, the entity “Culiacán”, occurring in sentence 1, 2, 3, 5, and 6, plays an important role in finding out the answer. To understand “Culiacán”, we should consider all its connected entities and the diverse relations between them.

2. How to **understand the relations** among entities with the complex reasoning patterns in the text. For example, to understand the complex inference chain in the figure, we need to perform both multi-hop (inferring that “Panamerican Highway” is located in “Mexico” through the first two clues) and coreferential reasoning (inferring that the word “It” in sentence 7 refers to “Panamerican Highway”).

In this paper, we propose ERICA, a novel framework to improve the PLMs’ capability of **Entity** and **Relation** understanding by **ContrAstive** learning, aiming to better capture the factual knowledge in the text by considering the complex interactions between entities and relations. Specifically, (1) to better understand entities, we introduce an entity discrimination task, that distinguishes which tail entity can be inferred by the given head entity and relation. It enhances the entity representations via considering its relations to other entities in the text. (2) Besides, to better understand relations, we employ the relation discrimination task which distinguishes whether two entity pairs are close or not in relational semantics. Through constructing entity pairs with document-level distant supervision, this task takes complex relational reasoning chains into consideration in an implicit way and thus improve the relation representations.

We conduct experiments on a suite of entity-centric downstream tasks, including relation extraction, entity typing and reading comprehension. The experimental results show that ERICA effectively improves the performance of both BERT and RoBERTa on all document-level downstream tasks,

especially under low resource setting. Meanwhile, ERICA achieves comparable or better results in both general and entity-centric sentence-level tasks, which verifies the robustness of our knowledge-guided learning objective.

2 Related Work

Pre-trained Language Models (PLMs) have a long history in NLP. Early studies focus on learning distributed and static word representations (Mikolov et al., 2013; Pennington et al., 2014) from unlabeled corpus. Peters et al. (2018) pre-train a BiLSTM language model to obtain contextualized word embeddings (ELMo) for downstream finetuning. With the introduction of the powerful Transformer (Vaswani et al., 2017) architecture, OpenAI GPT (Radford et al., 2018) and BERT (Devlin et al., 2018) further improve the performance and achieve state-of-the-art on various NLP tasks. Since then, several methods about PLMs emerged (Yang et al., 2019; Liu et al., 2019; Lan et al., 2020), such as XLNet (Yang et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020). Although these PLMs have achieved great successes, they still only focus on understanding the individual word tokens in the text, ignoring informative entities and their relations.

To address this issue, Joshi et al. (2020) and Kong et al. (2020) further introduce span-level pre-training objectives to enhance PLMs’ ability in representing continual spans, such as entity mentions. Besides, researcher also try to incorporate external knowledge such as Knowledge Graphs (KGs) into PLMs to enhance the entity representations. Zhang et al. (2019) take the KG embedding obtained by TransE (Bordes et al., 2013) as additional input features for PLMs to fuse both textual and KG’s information. Xiong et al. (2019), on the other hand, employ a zero-shot fact completion task and explicitly forces PLMs to memorize entity knowledge. Yamada et al. (2020) propose a novel pre-training task that predicts masked words and entities in a large entity-annotated corpus. Designed specifically for entity representations, these methods fail to consider the relational information between entities. To this end, Peters et al. (2019) propose a general method to embed multiple KGs into PLMs and enhance their representations with structured, human-curated knowledge. Moreover, Wang et al. (2019) propose to encode entity descriptions with PLMs as entity embedding and learn them in the

same way as conventional Knowledge Embedding (KE) methods. However, although introducing extra relational facts in KGs, the PLMs themselves still cannot fully capture the relation information between entities within the context, leading to a gap between pre-training and downstream finetuning.

In contrast, some work explore obtaining task-agnostic relation representations via contrastive learning. Soares et al. (2019) encourage relation representations in the text to be similar if they share the same entity pair in PLMs. Inspired by distant supervision (Mintz et al., 2009), Peng et al. (2020) further extend the above idea by grouping entity pairs that share the same relation annotated distantly by KGs. However, on one hand, these methods focus on the relation extraction task, and only consider relation representations while ignoring the entity representations; on the other hand, they only explore capturing easier sentence-level relational semantics, which limits the performance in dealing with more complex reasoning patterns of relations in real-world applications.

In this work, we propose to improve PLMs’ understanding of entities and relations jointly in the text at document-level to better capture the factual knowledge for language understanding.

3 Methodology

In this section, we introduce the details of ERICA. ERICA is a general framework that enhances PLM’s capability at capturing factual knowledge, via better modeling both entities and their relations in the text. We first describe how to represent entities and relations in documents and then detail the two contrastive learning-based tasks:

1. **Entity Discrimination (ED)** task improves the PLM’s entity representations via considering the diverse relations among them.

2. **Relation Discrimination (RD)** task improves the PLM’s relation representations via learning from their complex reasoning chains in the text.

3.1 Entity & Relation Representation

Let $\mathcal{D} = \{d_i\}_{i=1}^{|\mathcal{D}|}$ be a batch of documents, $\mathcal{E}_i = \{e_{ij}\}_{j=1}^{|\mathcal{E}_i|}$ be all named entities in d_i , where e_{ij} is the j -th entity in d_i . We first use PLM to encode d_i and get a series of hidden states $\{\mathbf{h}\}$, then we apply *mean pooling* operation over the consecutive tokens that mention e_{ij} to obtain local entity representations:

$$\mathbf{m}_{e_{ij}}^k = \text{MeanPool}(\mathbf{h}_{n_{start}^k}, \dots, \mathbf{h}_{n_{end}^k}), \quad (1)$$

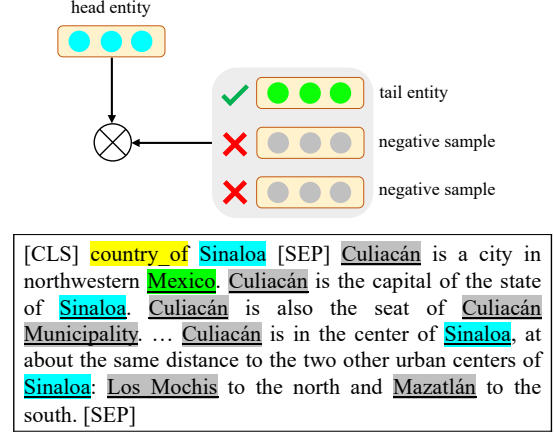


Figure 2: An example of Entity Discrimination (ED) task. For an entity pair with its distantly supervised relation in the text, the ED task requires the ground-truth tail entity to be closer to the head entity than other entities.

where the k -th occurrence of e_{ij} in d_i contains the tokens from index n_{start}^k to n_{end}^k , as e_{ij} may appear multiple times in d_i . To aggregate all information about e_{ij} , we average¹ all representations of each occurrence $\mathbf{m}_{e_{ij}}^k$ as the global entity representation \mathbf{e}_{ij} . Following Soares et al. (2019), we concatenate the final representations of two entities e_{ij_1} and e_{ij_2} as their relation representation, i.e., $\mathbf{r}_{j_1j_2}^i = [\mathbf{e}_{ij_1}; \mathbf{e}_{ij_2}]$.

3.2 Entity Discrimination Task

Entity Discrimination (ED) task aims at finding the tail entity given a head entity and a relation, it trains PLMs to understand an entity via considering its relations with other entities in the text. Formally, let \mathcal{K} be the used external KG with the relation set \mathcal{R} , for each document d_i , we enumerate all entity pairs (e_{ij}, e_{ik}) and link them to their corresponding relation r_{jk}^i in \mathcal{K} (if possible) and obtain a document-triple set $\mathcal{T}_i = \{(d_i; e_{ij}, r_{jk}^i, e_{ik}) | j \neq k\}$. We assign `no_relation` to those entity pairs that cannot find any relation in \mathcal{K} . Then we obtain the overall document-triple set $\mathcal{T} = \{\mathcal{T}_1, \dots, \mathcal{T}_{|\mathcal{D}|}\}$ for this batch.

As shown in Figure 2, in practice, we first sample a triple $(d_i; e_{ij}, r_{jk}^i, e_{ik})$ from \mathcal{T}_i^+ , PLMs are then asked to find e_{ik} given e_{ij} and r_{jk}^i . To inform PLMs of which head entity and relation to be conditioned on, we concatenate the relation name of r_{jk} , the

¹Although weighted summation by attention mechanism is an alternative, the method of entity information aggregation is not our main concern since it keeps the same in both pre-training and fine-tuning phases.

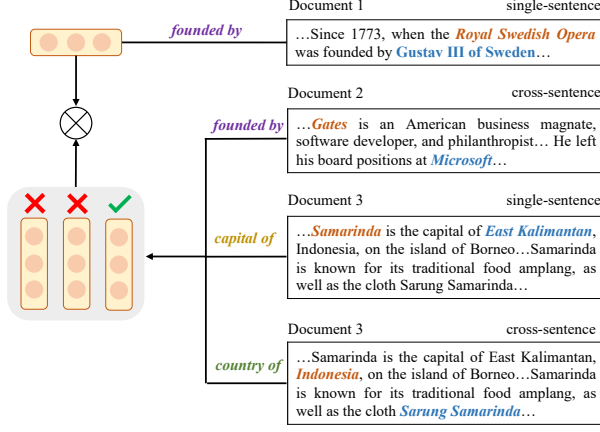


Figure 3: An example of Relation Discrimination (RD) task. For entity pairs belonging to the same relations, the RD task requires their relation representations to be closer.

mention of head entity e_{ij} and a separation token [SEP] in front of d_i , i.e., $d_i^* = \text{"relation_name entity_mention[SEP] } d_i"$. Under this situation, we call e_{ij} and e_{ik} entity-level “neighbors”, while other entities in d_i are “non-neighbors” of e_{ik} . Intuitively, the relation representations between “neighbors” should be closer than those of “non-neighbors”, therefore the entity representations of e_{ij} and e_{ik} should be closer than those of e_{ij} and other entities in the document. We can then leverage contrastive learning to push “neighbors” closer and “non-neighbors” apart. The loss function of ED task can be formulated as:

$$s_{jk}^e = \exp(\cos(\mathbf{e}_{ij}, \mathbf{e}_{ik})/\tau),$$

$$\mathcal{L}_{ED} = -\log \frac{s_{jk}^e}{s_{jk}^e + \sum_{l \notin \{j,k\}} s_{jl}^e}, \quad (2)$$

where $\cos(\cdot, \cdot)$ denotes the cosine similarity between two vectors, τ (temperature) is a hyper-parameter. All entities \mathcal{E}_i recognized in d_i except for e_{ij} and e_{ik} serve as negative samples.

3.3 Relation Discrimination Task

Relation Discrimination (RD) task aims at classifying the relations between two entities in the text. To further make PLMs understand the complex reasoning chains in real-world scenarios, we employ document-level rather than sentence-level distant supervision compared to existing relation-enhanced PLMs. Compared with sentence level, document-level distant supervision includes both intra-sentence (relatively simple cases) and inter-

sentence relations, which involves cross-sentence, multi-hop, or coreferential reasoning. Therefore, it has better coverage and generality of the complex reasoning chains. PLMs are required to perform reasoning in an implicit way to understand those “hard cases”.

As depicted in Figure 3, for two document-triple pairs $A = (d_A; e_{A1}, r_A, e_{A2})$ and $B = (d_B; e_{B1}, r_B, e_{B2})$ in \mathcal{T}^+ , if $r_A = r_B$, we call them relational “neighbors”. Similarly, we define the loss function of RD task as follows:

$$s_{AB}^r = \exp(\cos(\mathbf{r}_A, \mathbf{r}_B)/\tau),$$

$$\mathcal{L}_{RD} = -\log \frac{s_{AB}^r}{s_{AB}^r + \sum_{C \in \mathcal{T}^+ / \{A,B\}} s_{AC}^r}, \quad (3)$$

where N_{neg} (the number of negative samples) is a hyper-parameter. We define the positive document-triple set \mathcal{T}^+ by removing all document-triple pairs with no_relation in \mathcal{T} . In practice, we linearly² sample A and B from \mathcal{T}^+ , and sample triple C from \mathcal{T} instead of \mathcal{T}^+ and require that triple pairs do not share the same head or tail entities. In experiments, we find introducing no_relation entity pairs as negative samples further improves the model performance and the reason is that seeing more diverse entity pairs is beneficial to PLMs. Note A , B and C may come from the same document or different documents, therefore, both inter-document and intra-document interactions are involved in the RD task.

3.4 Overall Objective

Now we present the overall training objective of ERICA. To avoid catastrophic forgetting (McCloskey and Cohen, 1989) of general language understanding ability, we train the masked language modeling task together with ED and RD tasks. Hence, the overall learning objective is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{ED} + \mathcal{L}_{RD} + \mathcal{L}_{MLM}. \quad (4)$$

It’s worth mentioning that we also try to mask entities as suggested by Soares et al. (2019) and Peng et al. (2020), aiming to avoid simply relearning an entity linking system or let PLMs overfit on memorizing entity mentions. However, we do not observe performance gain by such a masking

²In each batch, the sampling rate of each relation is proportional to the total number of it in the current batch.

strategy. We conjecture that in our document-level setting, it’s hard for PLMs to memorize the diverse entity pairs and find “short cuts” on entity mentions due to the better coverage and generality of document-level distantly supervised training data. Besides, masking entities creates a gap between pre-training and fine-tuning, which may be a shortcoming of previous relation-enhanced PLMs.

4 Experiments

In this section, we first describe how we construct the distantly supervised dataset and pre-training details for ERICA. Then we introduce the experiments we did on three entity-centric tasks, including relation extraction (RE), entity typing (ET), and reading comprehension (RC), and eight tasks in GLUE benchmark. We test ERICA on base models of BERT (ERICA_{BERT}) and RoBERTa (ERICA_{RoBERTa}).

4.1 Distantly Supervised Dataset Construction

Following Yao et al. (2019), we construct our pre-training dataset leveraging distant supervision from the English Wikipedia and Wikidata. First, we use spaCy³ to perform *Named Entity Recognition*, and then link these entity mentions as well as Wikipedia’s mentions with hyper-links to Wikidata items. This could obtain the Wikidata ID for each entity. After that, the relations between different entities are annotated distantly by querying Wikidata. We only keep the documents containing more than 128 words, 4 entities and 4 relational triples. In addition, we ignore those entity pairs appearing in the test sets of RE and RC tasks to avoid test set leakage. In the end, we collect 1,000,000 documents (about 1G) in total with more than 4,000 relations annotated distantly. On average, each document contains 186.9 tokens, 12.9 entities and 7.2 relational triples. On average, an entity appears 1.3 times per document.

4.2 Pre-training Details

We initialize ERICA_{BERT} and ERICA_{RoBERTa} with *bert-base-uncased* and *roberta-base* checkpoints released by Google⁴ and Huggingface⁵, respectively. For both models, we keep the original hyper-parameters: we adopt Adam (Kingma and

Ba, 2014) as the optimizer and warm up the learning rate for the first 20% steps and then linearly decay it. We set the learning rate to $3e-5$, weight decay to $1e-5$, batch size to 2,048 and temperature τ to 0.05. For L_{RD} , we randomly select up to 64 negative samples per document. We trained our model with 8 NVIDIA Tesla P40 GPUs for 2,500 steps.

4.3 Relation Extraction

Relation Extraction (RE) aims to extract the relation between two recognized entities from a pre-defined relation set. We conduct experiments on both document-level and sentence-level RE.

Document-level RE. For document-level RE, we choose DocRED (Yao et al., 2019), which requires reading multiple sentences in a document and synthesizing all information to identify the relation between two entities. More than 40% of the relational facts in DocRED should be deduced from multiple sentences. We choose the following baselines: (1) CNN (Zeng et al., 2014), BILSTM (Hochreiter and Schmidhuber, 1997), BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) are widely used as text encoders for relation classification tasks. (2) HINBERT (Tang et al., 2020) propose a hierarchical inference network to leverage the abundant information from different sources. (3) CorefBERT (Ye et al., 2020) propose a pre-training method to help BERT capture the coreferential relations in context. (4) SpanBERT (Joshi et al., 2020) mask and predict contiguous random spans instead of random tokens. (5) ERNIE (Zhang et al., 2019) incorporate KG information into BERT to enhance entity representations. (6) MTB (Soares et al., 2019) and CP (Peng et al., 2020) introduce sentence-level relation contrastive learning for BERT via distant supervision. For fair comparison, we pre-train these two models on our constructed pre-training data based on the implementation released by Peng et al. (2020)⁶.

For training details, all entities are encoded in the same way as in pre-training phase. The relation representations are obtained by adding a bilinear layer on top of two entity representations. We did experiments on three partitions of the original training set (1%, 10% and 100%) and adopt batch size of 10, 32, 32 and training epochs of 400, 400, 200, respectively. We adopt Adam as the optimizer and

³<https://spacy.io/>

⁴<https://github.com/google-research/bert>

⁵<https://github.com/huggingface/transformers>

⁶<https://github.com/thunlp/RE-Context-or-Names>

Size	1%				10%				100%			
	Dev		Test		Dev		Test		Dev		Test	
	F1	IgF1	F1	IgF1	F1	IgF1	F1	IgF1	F1	IgF1	F1	IgF1
CNN			-				-		43.5	41.6	42.3	40.3
BILSTM			-				-		50.9	48.9	51.1	50.3
BERT	30.3	28.8	30.4	28.9	47.4	45.5	47.1	44.9	57.3	55.1	56.8	54.5
HINBERT			-				-		56.3	54.3	55.6	53.7
CorefBERT	32.4	30.9	32.8	31.2	46.4	44.3	46.0	43.7	57.5	55.3	57.0	54.5
SpanBERT	32.3	30.7	32.2	30.4	47.5	45.6	46.4	44.5	56.9	54.8	57.3	55.0
ERNIE	25.3	24.0	26.7	25.5	47.3	45.0	46.7	44.2	57.4	55.2	56.6	54.2
MTB	28.7	27.2	29.0	27.6	46.5	44.6	46.1	44.1	57.0	54.6	56.9	54.3
CP	30.9	29.3	30.3	28.7	45.8	43.8	44.8	42.6	56.0	53.7	55.2	52.7
ERICA _{BERT}	37.4	35.6	37.8	36.0	51.2	49.0	50.8	48.3	58.8	56.7	58.2	55.9
RoBERTa	35.9	34.2	35.3	33.5	48.1	46.1	48.0	45.9	57.8	55.5	58.5	56.1
ERICA _{RoBERTa}	39.9	37.9	40.1	38.0	50.8	48.9	50.3	48.3	58.6	56.3	59.0	56.6

Table 1: Results on DocRED. We report micro F1 (F1) and micro ignore F1 (IgF1) on both dev and test sets. IgF1 metric ignores the relational facts shared by the train and dev/test sets.

Dataset	TACRED			SemEval			Wiki80		
	1%	10%	100%	1%	10%	100%	1%	10%	100%
BERT	36.0	58.5	68.1	43.6	79.3	88.1	60.8	85.0	91.3
MTB	35.7	58.8	68.2	44.2	79.2	88.2	61.8	85.9	91.5
CP	37.1	60.6	68.1	40.3	80.0	88.5	66.3	89.0	92.4
ERICA _{BERT}	36.5	59.7	68.5	47.9	80.1	88.0	72.0	86.7	91.6
RoBERTa	26.3	61.2	69.7	46.0	80.3	88.8	60.8	85.8	91.3
ERICA _{RoBERTa}	40.0	61.9	69.8	46.3	80.4	89.2	67.2	87.3	91.7

Table 2: Results on TACRED, SemEval-2010 Task8 and Wiki80. We report F1 on test sets.

set the learning rate to $4e-5$. We evaluate on dev set every 20/20/5 epochs and then test the best checkpoint on test set.

From the results shown in Table 1, we can see that ERICA outperforms all baselines significantly on each partition, which demonstrates that ERICA is better at aggregating information from multiple sources and understanding relations between entities. We observe that both MTB and CP achieve worse results than BERT, which means sentence-level pre-training hurts models’ performance on document-level tasks to some extent. We also observe that ERICA outperforms baselines by a larger margin with smaller training set, which means ERICA can improve the performance extensively under low-resource settings.

Sentence-level RE. For sentence-level RE, we choose three widely used datasets: TACRED (Zhang et al., 2017), SemEval-2010 Task 8 (Hendrickx et al., 2019) and Wiki80 (Han et al., 2019). We choose BERT, RoBERTa, MTB and CP as baselines. For training details, we insert extra marker tokens to highlight the subject and object in each sentence. The relation representation for

each entity pair is obtained in the same way as in pre-training phase. Other settings are kept the same as Peng et al. (2020). From the results shown in Table 2, we observe that ERICA achieves comparable results at sentence-level with CP, which means document-level pre-training does not hurt PLM’s performance in sentence-level tasks.

4.4 Entity Typing

Metrics	Macro	Micro
BERT	75.50	72.68
MTB	76.37	72.94
CP	76.27	72.48
ERNIE	76.51	73.39
ERICA _{BERT}	77.85	74.71
RoBERTa	79.24	76.38
ERICA _{RoBERTa}	80.77	77.04

Table 3: Results on FIGER. We report macro F1 (macro) and micro F1 (micro) on the test set.

Entity typing aims at classifying entity mentions into pre-defined entity types. We select FIGER (Ling et al., 2015), which is a sentence-level entity typing dataset. BERT, RoBERTa, MTB, CP

Setting	Standard			Masked		
Size	1%	10%	100%	1%	10%	100%
FastQA	-		27.2	-		38.0
BiDAF	-		49.7	-		59.8
BERT	35.8	53.7	69.5	37.9	53.1	73.1
CorefBERT	38.1	54.4	68.8	39.0	53.5	70.7
SpanBERT	33.1	56.4	70.7	34.0	55.4	73.2
MTB	36.6	51.7	68.4	36.2	50.9	71.7
CP	34.6	50.4	67.4	34.1	47.1	69.4
ERICA _{BERT}	46.5	57.8	69.7	40.2	58.1	73.9
RoBERTa	37.3	57.4	70.9	41.2	58.7	75.5
ERICA _{RoBERTa}	47.4	58.8	71.2	46.8	63.4	76.6

Table 4: Results (accuracy) on the dev set of WikiHop. We tested both the standard and masked settings.

and **ERNIE** are chosen as baselines. In finetuning phrase, we encode the entities in the same way as pre-training. We set the learning rate as $3e-5$ and batch size as 256, and finetune the models for three epochs. From the results listed in Table 3, we can see that, ERICA outperforms all baselines, which demonstrate ERICA helps the model better understand and represent entities in the text.

4.5 Reading Comprehension

The task of Reading Comprehension aims at extracting specific text spans as the answer given a query. We choose WikiHop (Welbl et al., 2018), which requires models to answer specific properties of an entity after reading multiple documents. WikiHop has both standard and masked settings where all entities are masked with a random ID in the masked setting. We choose the following baselines: (1) **FastQA** (Weissenborn et al., 2017) and **BiDAF** (Seo et al., 2016) are widely used question answering systems. (2) Similar to RE tasks, we also test **BERT**, **CorefBERT**, **SpanBERT**, **MTB**, **CP** and **RoBERTa**.

For the implementation details, we concatenate the question and documents into one long sequence, encode each candidate entity in the sequence and classify them by a prediction layer. Since the standard setting of WikiHop does not provide the index for each candidate, we then find them by exactly matching them in the documents. We did experiments on three partitions of the original training data (1%, 10% and 100%). We set the batch size to 8 and learning rate to $5e-5$, and train for two epochs.

From the results listed in Table 4, we observe that ERICA outperforms baselines in both standard and masked settings of WikiHop, which indicates

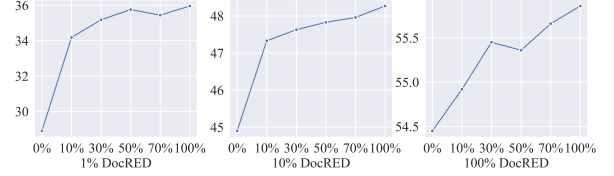


Figure 4: Impacts of pre-training data’s size. X axis denotes different ratios of pre-training data, Y axis denotes test IgF1 on different partitions of DocRED.

that ERICA can better understand the the entities and their relations in the documents. Also, ERICA achieves much more improvements in the masked setting compared to standard setting, indicating that ERICA can better synthesize and analyze information from contexts, instead of relying on entity mention “shortcuts” (Jiang and Bansal, 2019).

4.6 GLUE

The General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018) provides several natural language understanding tasks, which is often used to evaluate PLMs. To test whether L_{ED} and L_{RD} decrease the PLMs’ performance on these tasks, we compare BERT, ERICA_{BERT}, RoBERTa and ERICA_{RoBERTa}. We follow the widely used setting and use the [CLS] token as representation for the whole sentence or sentence pair for classification or regression. Table 5 shows the results on dev set of GLUE Benchmark. It can be observed that, both ERICA_{BERT} and ERICA_{RoBERTa} achieve comparable performance than the original model, which suggests that jointly training L_{ED} , L_{RD} and MLM does not hurt PLMs’ general ability of language understanding.

5 Analysis

In this section, we first explore how \mathcal{L}_{ED} and \mathcal{L}_{RD} impact the performance of ERICA. Then we give a thorough analysis on how size and domain of pre-training data impact the performance of ERICA. We also demonstrate that ERICA has consistent improvements regardless of the methods for entity encoding.

5.1 Ablation Study

To explore how \mathcal{L}_{ED} and \mathcal{L}_{RD} impact the performance of ERICA, we test on DocRED, WikiHop (masked version) and FIGER by using only one of these two losses and compare the results. In addition, we also investigate how single-sentence entity pairs ($\mathcal{L}_{RD}^{single}$) and cross-sentence entity pairs

Dataset	MNLI(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE
BERT	84.0/84.4	88.9	90.6	92.4	57.2	89.7	89.4	70.1
ERICA _{BERT}	84.5/84.7	88.3	90.7	92.8	57.9	89.5	89.5	69.6
RoBERTa	87.5/87.3	91.9	92.8	94.8	63.6	91.2	90.2	78.7
ERICA _{RoBERTa}	87.5/87.5	91.6	92.6	95.0	63.5	90.7	91.5	78.5

Table 5: Results on dev set of GLUE Benchmark. We report matched/mismatched (m/mm) accuracy for MNLI, F1 score for QQP and MRPC, spearman correlation for STS-B and accuracy for other tasks.

Dataset	DocRED	WikiHop	FIGER
BERT	54.5	73.1	72.7
-NSP	54.6	73.3	72.6
-NSP+ \mathcal{L}_{ED}	55.8	74.8	73.8
-NSP+ \mathcal{L}_{RD}^{cross}	54.7	72.8	72.6
-NSP+ $\mathcal{L}_{RD}^{single}$	55.5	72.5	73.5
-NSP+ \mathcal{L}_{RD}^{both}	55.6	72.7	74.0
ERICA _{BERT}	55.9	73.9	74.7

Table 6: Ablation study. We report test IgF1 on DocRED, dev accuracy on the masked setting of WikiHop and test micro F1 on FIGER.

(\mathcal{L}_{RD}^{cross}) contribute to \mathcal{L}_{RD} by sampling only one of them during pre-training. For fair comparison, we also include a baseline by training MLM only. As shown in Table 6, for DocRED and FIGER, either \mathcal{L}_{RD} or \mathcal{L}_{ED} is beneficial, and combining them further improves the performance; For WikiHop, \mathcal{L}_{ED} dominates the improvement while \mathcal{L}_{RD} hurts the performance slightly. This is possibly because reading comprehension resembles the tail entity discrimination process of \mathcal{L}_{ED} more in pre-training phase, while the relation discrimination process of \mathcal{L}_{RD} may have a little conflict. Also, for \mathcal{L}_{RD} , both single-sentence and cross-sentence entity pairs contribute, which demonstrate that incorporating both of them are beneficial for PLMs to understand entities and relations in the text.

5.2 Effects of Pre-training Data’s Size

To explore the effects of pre-training data’s size, we train ERICA on 10%, 30%, 50% and 70% of the original pre-training dataset, respectively. We report the results in Figure 4, from which we observe that with the scale of pre-training data becoming larger, ERICA is performing better.

5.3 Effects of Domain Shifting

We investigate two domain shifting factors: entity distribution and relation distribution, to see how they impact ERICA’s performance.

Entity Distribution Shifting. The entities in supervised datasets of DocRED are recognized by

Size	1%	10%	100%
BERT	28.9	44.9	54.5
ERICA _{BERT}	36.0	48.3	55.9
ERICA _{DocRED} _{BERT}	36.3	48.6	55.9

Table 7: Results (IgF1) on how entity distribution of pre-training data influences ERICA’s performance on DocRED.

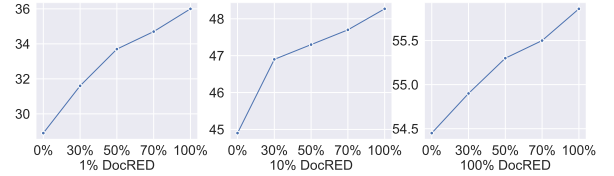


Figure 5: Impacts of relation distribution shifting. X axis denotes different ratios of relations, Y axis denotes test IgF1 on different partitions of DocRED.

human annotators while our pre-training data is processed by spaCy. Hence there may exist an entity distribution gap between pre-training and fine-tuning. To test the effects of entity distribution shifting, we fine-tune a BERT model on training set of DocRED for NER tagging and re-tag entities in our pre-training dataset. Then we pre-train on it (denoted as ERICA_{DocRED}_{BERT}) and compare the performances. As is shown in Table 7, we observe that training on a dataset that shares similar entity distributions with downstream tasks further improves the performance.

Relation Distribution Shifting. Our pre-training data contains over 4,000 Wikidata relations. Ideally, training on a more diverse relation domain should benefit PLMs. We randomly keep only 30%, 50% and 70% relations of the original pre-training data and compare the performances. From the results listed in Figure 5, we observe that increasing the diversity of relation domain in the pre-training phase improves the performance.

5.4 Effects of Methods for Entity Encoding

For all the experiments mentioned above, we encode each occurrence of an entity by mean pooling

Size	1%	10%	100%
Mean Pool			
BERT	28.9	44.9	54.5
ERICA _{BERT}	36.0	48.3	55.9
Entity Marker			
BERT	23.9	44.3	55.6
ERICA _{BERT}	34.8	48.0	57.6

Table 8: Results (IgF1) on how entity encoding strategy influences ERICA’s performance on DocRED.

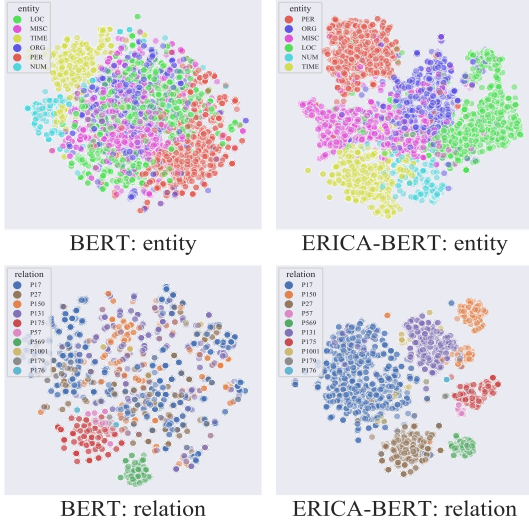


Figure 6: t-SNE plots of learned entity and relation embeddings on DocRED comparing BERT and ERICA_{BERT}.

over all its tokens in both pre-training and downstream tasks. Ideally, ERICA should have consistent improvements on other kinds of methods for entity encoding. To demonstrate this, we try another entity encoding method mentioned by Soares et al. (2019), which inserts a special start token [S] in front of an entity and an end token [E] after it. The representation for this entity is calculated by averaging the representations of all its start tokens in the document. To help PLMs discriminate different entities, we randomly assign different marker pairs ([S1], [E1]; [S2], [E2], ...) for each entity in a document in both pre-training and downstream tasks⁷. All occurrences of one entity in a document share the same marker pair. We show in Table 8 that ERICA achieves consistent performance improvements for both methods (denoted as Mean Pool and Entity Marker) and the improvement is more evident in the latter method.

⁷In practice, we randomly initialize 100 entity marker pairs.

5.5 Embedding Visualization

In Figure 6, we show the learned entity and relation embeddings of BERT and ERICA_{BERT} on DocRED’s test set by t-distributed stochastic neighbor embedding (t-SNE) (Hinton and Roweis, 2002). We label each two-dimensional point with different colors to represent its corresponding category of entities or relations in Wikidata, and we only visualize the most frequent 10 relations here. From the figure, we can see that jointly training MLM with \mathcal{L}_{ED} and \mathcal{L}_{RD} leads to a more compact clustering of both entities and relations that belong to the same category. In contrast, training with MLM only exhibits random distribution. This verifies that ERICA could better represent the entities and relations in the text.

6 Conclusions and Future Work

In this paper, we present ERICA, a general framework for PLMs to improve entity and relation understanding via contrastive learning. We demonstrate the effectiveness of our method on several entity-centric tasks, including relation extraction, reading comprehension and entity typing. The experimental results show that ERICA outperforms baselines by a large margin on document-level tasks while keep comparable performance on sentence-level tasks. In future, we aim to explore explicitly modeling the complex reasoning chains in documents for PLMs.

References

- Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 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, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [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)*, pages 4171–4186.
- Xu Han, Tianyu Gao, Yuan Yao, Demin Ye, Zhiyuan Liu, and Maosong Sun. 2019. [OpenNRE: An open and extensible toolkit for neural relation extraction](#). 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): System Demonstrations*, pages 169–174.
- Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2019. [SemEval-2010 Task 8: Multi-way classification of semantic relations between pairs of nominals](#). In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 94–99.
- Geoffrey E Hinton and Sam Roweis. 2002. [Stochastic neighbor embedding](#). In *Advances in neural information processing systems 15: 16th Annual Conference on Neural Information Processing Systems 2002. Proceedings of a meeting held September 12, 2002, Vancouver, British Columbia, Canada*, volume 15, pages 857–864.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. [Long short-term memory](#). *Neural computation*, 9(8):1735–1780.
- Yichen Jiang and Mohit Bansal. 2019. [Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop qa](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019, July 28, 2019, Florence, Italy*, pages 2726–2736. Association for Computational Linguistics.
- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. [SpanBERT: Improving pre-training by representing and predicting spans](#). *Transactions of the Association for Computational Linguistics*, 8:64–77.
- Diederik P Kingma and Jimmy Ba. 2014. [Adam: A method for stochastic optimization](#). In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7, 2015, Conference Track Proceedings*.
- Lingpeng Kong, Cyprien de Masson d’Autume, Lei Yu, Wang Ling, Zihang Dai, and Dani Yogatama. 2020. [A mutual information maximization perspective of language representation learning](#). In *Proceedings of 8th International Conference on Learning Representations, ICLR 2020, Virtual Conference, April 26, 2020, Conference Track Proceedings*.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. [ALBERT: A lite BERT for self-supervised learning of language representations](#). In *Proceedings of 8th International Conference on Learning Representations, ICLR 2020, Virtual Conference, April 26, 2020, Conference Track Proceedings*.
- Xiao Ling, Sameer Singh, and Daniel S Weld. 2015. [Design challenges for entity linking](#). *Transactions of the Association for Computational Linguistics*, 3:315–328.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692.
- Michael McCloskey and Neal J Cohen. 1989. [Catastrophic interference in connectionist networks: the sequential learning problem](#). In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. [Distributed representations of words and phrases and their compositionality](#). In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.
- Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. [Distant supervision for relation extraction without labeled data](#). In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011. Association for Computational Linguistics.
- Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. [Learning from context or names? an empirical study on neural relation extraction](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3661–3672. Online. Association for Computational Linguistics.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. [Glove: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25, 2014, Doha, Qatar; A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543.
- Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. [Deep contextualized word representations](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1, 2018, Volume 1 (Long Papers)*, page 2227–2237.
- Matthew E Peters, Mark Neumann, Robert L Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. [Knowledge enhanced contextual word representations](#). 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)*. Association for Computational Linguistics.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. [Improving language understanding by generative pre-training](#).
- Erik F Sang and Fien De Meulder. 2003. [Introduction to the conll-2003 shared task: Language-independent named entity recognition](#). In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*.
- Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. [Bidirectional attention flow for machine comprehension](#). In *Proceedings of 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24, 2017, Conference Track Proceedings*.
- Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. [Matching the blanks: Distributional similarity for relation learning](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905. Association for Computational Linguistics.
- Alon Talmor and Jonathan Berant. 2019. [MultiQA: An empirical investigation of generalization and transfer in reading comprehension](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4911–4921. Association for Computational Linguistics.
- Hengzhu Tang, Yanan Cao, Zhenyu Zhang, Jiangxia Cao, Fang Fang, Shi Wang, and Pengfei Yin. 2020. [Hin: Hierarchical inference network for document-level relation extraction](#). In *Advances in Knowledge Discovery and Data Mining-24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11, 2020, Proceedings, Part I, volume 12084 of Lecture Notes in Computer Science*, pages 197–209. Springer.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4 December 2017, Long Beach, CA, USA*, pages 5998–6008.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP1*. Association for Computational Linguistics.
- Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2019. [KEPLER: A unified model for knowledge embedding and pre-trained language representation](#). *arXiv preprint arXiv:1911.06136*.
- Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. [Making neural QA as simple as possible but not simpler](#). In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 271–280. Association for Computational Linguistics.
- Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. [Constructing datasets for multi-hop reading comprehension across documents](#). *Transactions of the Association for Computational Linguistics*, 6:287–302.
- Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. 2019. [Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model](#). In *Proceedings of 8th International Conference on Learning Representations, ICLR 2020, Virtual Conference, April 26, 2020, Conference Track Proceedings*.
- Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. [LUKE: Deep contextualized entity representations with entity-aware self-attention](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. [XLNet: Generalized autoregressive pretraining for language understanding](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*.
- Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. [DocRED: A large-scale document-level relation extraction dataset](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 764–777.
- Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Maosong Sun, and Zhiyuan Liu. 2020. [Coreferential reasoning learning for language representation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7170–7186. Association for Computational Linguistics.
- Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. [Relation classification via convolutional deep neural network](#). In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344. Dublin City University and Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. [Position-aware attention and supervised data improve slot filling](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. [ERNIE: Enhanced language representation with informative entities](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451. Association for Computational Linguistics.