EasyChair Preprint

№ 8086

Open Relation Extraction via Query-Based Span Prediction

Huifan Yang, Da-Wei Li, Zekun Li, Donglin Yang, Jinsheng Qi and Bin Wu

May 24, 2022

Open Relation Extraction via Query-based Span Prediction

Huifan Yang[1*][0000−0002−1874−4863], Da-Wei Li[2], Zekun Li[1], Donglin Yang[1], Jinsheng Qi[1], and Bin Wu[1]

[1] Beijing Key Laboratory of Intelligence Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications
[2] Bing Multimedia Team, Microsoft Software Technology Center Asia
{huifunny,lizekun,iceberg,qijs,wubin}@bupt.edu.cn, daweilee@microsoft.com

Abstract Open relation extraction (ORE) aims to assign semantic relationships between arguments, essential to the automatic construction of knowledge graphs. The previous methods either depend on external NLP tools (e.g., PoS-taggers) and language-specific relation formations, or suffer from inherent problems in sequence representations, thus leading to unsatisfactory extraction in diverse languages and domains. To address the above problems, we propose a **Q**uery-based **O**pen **R**elation **E**xtractor (**QORE**). QORE utilizes a Transformers-based language model to derive a representation of the interaction between arguments and context, and can process multilingual texts effectively. Extensive experiments are conducted on seven datasets covering four languages, showing that QORE models significantly outperform conventional rule-based systems and the state-of-the-art method LOREM [8]. Regarding the practical challenges [1] of *Corpus Heterogeneity* and *Automation*, our evaluations illustrate that QORE models show excellent zero-shot domain transferability and few-shot learning ability.

Keywords: Open relation extraction · Information extraction · Knowledge graph construction · Transfer learning · Few-shot learning

1 Introduction

Relation extraction (RE) from unstructured text is fundamental to a variety of downstream tasks, such as constructing knowledge graphs (KG) and computing sentence similarity. Conventional closed relation extraction considers only a predefined set of relation types on small and homogeneous corpora, which is far less effective when shifting to general-domain text mining that has no limits in relation types or languages. To alleviate the constraints of closed RE, Banko et al. [1] introduce a new paradigm: open relation extraction (ORE), predicting a text span as the semantic connection between arguments from within a context, where a span is a contiguous sub-sequence. This paper proposes a novel

* Corresponding author.

query-based open relation extractor QORE that can process multilingual texts for facilitating large-scale general-domain KG construction.

Open relation extraction identifies an arbitrary phrase to specify a semantic relationship between arguments within a context. (An argument is a text span representing an adverbial, adjectival, nominal phrase, and so on, which is not limited to an entity.) Taking a context "*Researchers develop techniques to acquire information automatically from digital texts.*" and an argument pair ⟨*Researchers, information*⟩, an ORE system would extract the span "*acquire*" from the context to denote the semantic connection between "*Researchers*" and "*information*".

Conventional ORE systems are largely based on syntactic patterns and heuristic rules that depend on external NLP tools (e.g., PoS-taggers) and language-specific relation formations. For example, ReVerb [5], ClausIE [2], OpenIE4 [16] for English and CORE [22], ZORE [18] for Chinese, leverage external tools to obtain part-of-speech tags or dependency features and generate syntactic patterns to extract relational facts. Faruqui et al. [6] present a cross-lingual ORE system that first translates a sentence to English, performs ruled-based ORE in English, and finally projects the relation back to the original sentence. These pattern-based approaches cannot handle the complexity and diversity of languages well, and the extraction is usually far from satisfactory.

To alleviate the burden of designing manual features, multiple neural ORE models have been proposed, typically adopting the methods of either sequence labeling or span selection. MGD-GNN [15] for Chinese ORE constructs a multi-grained dependency graph and utilizes a span selection model to predict based on character features and word boundary knowledge. Compared with our method, MGD-GNN heavily relies on dependency information and cannot deal with various languages. Ro et al. [19] propose sequence-labeling-based Multi^{2}OIE that performs multilingual open information extraction by combining BERT with multi-head attention blocks, whereas Multi^{2}OIE is constrained to extract the predicate of a sentence as the relation. Jia et al. [10] transform English ORE into a sequence labeling process and present a hybrid neural network NST, nonetheless, a dependency on PoS-taggers may introduce error propagation to NST. Improving NST, the current state-of-the-art ORE method LOREM [8] works as a multilingual-embedded sequence-labeling model based on CNN and BiLSTM. Identical to our model, LOREM does not rely on language-specific knowledge or external NLP tools. However, based on our comparison of architectures in Section 4.1, LOREM suffers from inherent problems in learning long-range sequence dependencies [23] that are basic to computing token relevances to gold relations, thus resulting in less satisfactory performances compared with QORE model.

The benchmark ORE datasets in English, Chinese, French, and Russian (denoted as *En, Zh, Fr,* and *Ru*, respectively) include OpenIE4$^{\text{En}}$ [16], LSOIE-wiki$^{\text{En}}$ [20], LSOIE-sci$^{\text{En}}$, COER$^{\text{Zh}}$ [22], SAOKE$^{\text{Zh}}$ [21], WMORC$^{\text{Fr}}$ [6], and WMORC$^{\text{Ru}}$. In the above datasets, contexts are complex or multiple sentences, and the used relation triples are in the form of (*Argument$_1$, Single-span Relation, Argument$_2$*) following the pattern of triplets in common KGs.

Inspired by the broad applications of machine reading comprehension (MRC) and Transformers-based pre-trained language models (LM) like BERT [3] and SpanBERT [11], we design a query-based open relation extraction framework QORE to solve the ORE task effectively and avoid the inherent problems of previous extractors. Given an argument pair and its context, we first create a query template containing the argument information and derive a contextual representation of query and context via a pre-trained language model, which provides a deep understanding of query and context, and models the information interaction between them. Finally, the span extraction module finds an open relation by predicting the start and end indexes of a sub-sequence in the context.

Besides introducing the ORE paradigm, Banko et al. [1] identified major challenges for ORE systems, including *Corpus Heterogeneity* and *Automation*. Thus, we carry out the evaluation on the two challenges from the aspects of **zero-shot domain transferability** and **few-shot learning ability**, which we interpret in the following. (a) *Corpus Heterogeneity*: Heterogeneous datasets form an obstacle for profound linguistic tools such as syntactic or dependency parsers, since they commonly work well when trained and applied to a specific domain, but are prone to produce incorrect results when used in a different genre of text. As QORE models are intended for domain-independent usage, we do not require using any external NLP tool, and we assess the performances in this challenge via zero-shot domain transferring. (b) *Automation*: The manual labor of creating suitable training data or extraction patterns must be reduced to a minimum by requiring only a small set of hand-tagged seed instances or a few manually defined extraction patterns. The QORE framework does not need predefined extraction patterns but trains on amounts of data. We conduct few-shot learning by shrinking the size of training data for the evaluation of this challenge.

To summarize, the main contributions of this work are:

- We propose a novel query-based open relation extractor QORE that utilizes a Transformers-based language model to derive a representation of the interaction between the arguments and context.
- We carry out extensive experiments on seven datasets covering four languages, showing that QORE models significantly outperform conventional rule-based systems and the state-of-the-art method LOREM.
- Considering the practical challenges of ORE, we investigate the zero-shot domain transferability and the few-shot learning ability of QORE. The experimental results illustrate that our models maintain high precisions when transferring or training on fewer data.

2 Approach

An overview of our QORE framework is visualized in Figure 1. Given an argument pair and its context, we first create a query from the arguments based on a template and encode the combination of query and context using a Transformers-

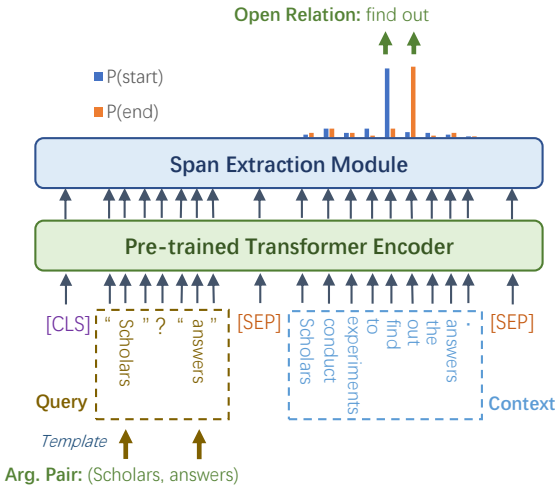

Figure 1. An overview of QORE framework

based language model. Finally, the span extraction module predicts a continuous sub-sequence in the context as an open relation.

2.1 Task Description

Given a context $\boldsymbol{C}$ and an argument pair $\boldsymbol{A} = (A_1, A_2)$ in $\boldsymbol{C}$, an open relation extractor needs to find the semantic relationship between the pair $\boldsymbol{A}$. We denote the context as a word token sequence $\boldsymbol{C} = \{x_i^c\}_{i=1}^{l_c}$ and an argument as a text span $A_k = \{x_i^{a_k}\}_{i=1}^{l_{a_k}}$, where l_c is the context length and l_{a_k} is the k-th argument length. Our goal is to predict a span $R = \{x_i^r\}_{i=1}^{l_r}$ in the context as an open relation, where l_r is the length of a relation span.

2.2 Query Template Creation

Provided an argument pair (A_1, A_2), we adopt a rule-based method to create the query template

$$\boldsymbol{T} = \langle s_1 \rangle\, A_1\, \langle s_2 \rangle\, A_2\, \langle s_3 \rangle \tag{1}$$

having three slots, where $\langle s_i \rangle$ indicates the i-th slot. The tokens filling a slot are separators of the adjacent arguments (e.g., double-quotes, a comma, or words of natural languages) or a placeholder for a relation span (e.g., a question mark or words of natural languages). In this paper, we design two different query templates: (1) the question-mark (QM) style $\boldsymbol{T}_{QM}$, taking the form of a structured argument-relationship triple, and (2) the language-specific natural-language (NL) style $\boldsymbol{T}_{NL}$, where each language has a particular template that is close in meaning. (English: *En*, Chinese: *Zh*, French: *Fr*, Russian: *Ru*.)

$$\boldsymbol{T}_{QM} = \text{“}A_1\text{”?“}A_2\text{”} \tag{2}$$

$$\boldsymbol{T}_{NL^{En}} = \text{What is the relation from“}A_1\text{”to“}A_2\text{”?} \tag{3}$$

$$\boldsymbol{T}_{NL^{Zh}} = \text{“}A_1\text{”和“}A_2\text{”的关系是?} \tag{4}$$

$$\boldsymbol{T}_{NL^{Fr}} = \text{Quelle est la relation entre“}A_1\text{”et“}A_2\text{”?} \tag{5}$$

$$\boldsymbol{T}_{NL^{Ru}} = \text{Какое отношение имеет“}A_1\text{”к“}A_2\text{”?} \tag{6}$$

2.3 Encoder

BERT [3] is a pre-trained encoder of deep bidirectional transformers [23] for monolingual and multilingual representations. Inspired by BERT, Joshi et al. [11] propose SpanBERT to better represent and predict text spans. SpanBERT extends BERT by masking random spans based on geometric distribution and using span boundary objective (SBO) that requires the model to predict masked spans based on span boundaries for structure information integration into pre-training. The two language models both achieve strong performances on the span extraction task. We use BERT and SpanBERT as the encoders of QORE.

Given a context $\boldsymbol{C} = \{x_i^c\}_{i=1}^{l_c}$ with l_c tokens and a query $\boldsymbol{Q} = \{x_j^q\}_{j=1}^{l_q}$ with l_q tokens, we employ a pre-trained language model as the encoder to learn the contextual representation for each token. First, we concatenate the query $\boldsymbol{Q}$ and the context $\boldsymbol{C}$ to derive the input $\boldsymbol{I}$ of encoder:

$$\boldsymbol{I} = \{[CLS], x_1^q, ..., x_{l_q}^q, [SEP], x_1^c, ..., x_{l_c}^c, [SEP]\} \tag{7}$$

where $[CLS]$ and $[SEP]$ denote the beginning token and the segment token, respectively.

Next, we generate the initial embedding $\boldsymbol{e}_i$ for each token by summing its word embedding $\boldsymbol{e}_i^w$, position embedding $\boldsymbol{e}_i^p$, and segment embedding $\boldsymbol{e}_i^s$. The sequence embedding $\boldsymbol{E} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_m\}$ is then fed into the deep Transformer layers to learn a contextual representation with long-range sequence dependencies via the self-attention mechanism [23]. Finally, we obtain the last-layer hidden states $\boldsymbol{H} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_m\}$ as the contextual representation for the input sequence $\boldsymbol{I}$, where $\boldsymbol{h}_i \in \mathbb{R}^{d_h}$ and d_h indicates the dimension of the last hidden layer of encoder. The length of the sequences $\boldsymbol{I}$, $\boldsymbol{E}$, $\boldsymbol{H}$ is denoted as m where $m = l_q + l_c + 3$.

2.4 Span Extraction Module

The span extraction module aims to find a continuous sub-sequence in the context as an open relation. We utilize two learnable parameter matrices (feed-forward networks) $f_{start} \in \mathbb{R}^{d_h}$ and $f_{end} \in \mathbb{R}^{d_h}$ followed by the softmax normalization, then take each contextual token representation $\boldsymbol{h}_i$ in $\boldsymbol{H}$ as the input to produce the probability of each token i being selected as the start/end of relation span:

$$p_i^{start} = softmax(f_{start}(\boldsymbol{h}_1), ..., f_{start}(\boldsymbol{h}_m))_i \tag{8}$$

$$p_i^{end} = softmax(f_{end}(\boldsymbol{h}_1), ..., f_{end}(\boldsymbol{h}_m))_i \tag{9}$$

We denote $\boldsymbol{p}^{start} = \{p_i^{start}\}_{i=1}^m$ and $\boldsymbol{p}^{end} = \{p_i^{end}\}_{i=1}^m$.

2.5 Training and Inference

The training objective is defined as minimizing the cross entropy loss for the start and end selections,

$$p_k = p_{y_k^s}^{start} \times p_{y_k^e}^{end} \tag{10}$$

$$\mathbb{L} = -\frac{1}{N}\sum_{k}^{N} \log p_k \tag{11}$$

where y_k^s and y_k^e are respectively ground-truth start and end positions of example k. N is the number of examples.

In the inference process, an open relation is extracted by finding the indices (s, e):

$$(s, e) = \arg\max_{s \leq e}(p_s^{start} \times p_e^{end}) \tag{12}$$

3 Experimental Setup

We propose the following hypotheses and design a set of experiments to examine the performances of QORE models. We arrange the hypotheses based on the considerations as follows: (1) $\mathbf{H_1}$: By conducting extensive comparisons with the existing ORE systems, we aim to analyze the advantages of QORE framework. (2) $\mathbf{H_2}$ and $\mathbf{H_3}$: As stated in the Introduction, it is significant to evaluate an open relation extractor on the challenges of *Corpus Heterogeneity* and *Automation*. Thus, we investigate the zero-shot domain transferability and the few-shot learning ability of QORE models.

- $\mathbf{H_1}$: For extracting open relations in seven datasets of different languages, QORE models can outperform conventional rule-based extractors and the state-of-the-art neural method LOREM.
- $\mathbf{H_2}$: Considering the zero-shot domain transferability, QORE model is able to perform effectively when transferring to another domain.
- $\mathbf{H_3}$: When the training data size reduces, QORE model shows an excellent few-shot learning ability and maintains high precision.

3.1 Datasets

We evaluate the performances of our proposed QORE framework on seven public datasets covering four languages, i.e., English, Chinese, French, and Russian (denoted as *En, Zh, Fr,* and *Ru*, respectively). In the data preprocessing, we only retain binary-argument triples whose components are spans of the contexts. Table 1 lists the statistics of the used training, development and test sets. (Con.: Contexts; Tri.: Triples.)

- **OpenIE4$^{\mathbf{En}}$**[3] was bootstrapped from extractions of OpenIE4 [16] from Wikipedia and annotated with part-of-speech and dependency information by Zhan and Zhao [25].

[3] https://github.com/zhanjunlang/Span_OIE

Table 1. Statistics of seven datasets over four languages

	OpenIE4$^{\text{En}}$		**LSOIE-wiki$^{\text{En}}$**		**LSOIE-sci$^{\text{En}}$**		**COER$^{\text{Zh}}$**		**SAOKE$^{\text{Zh}}$**		**WMORC$^{\text{Fr}}$**		**WMORC$^{\text{Ru}}$**	
	#Con.	#Tri.	#Con.	#Tri.	#Con.	#Tri.	#Con.	#Tri.	#Con.	#Tri.	#Con.	#Tri.	#Con.	#Tri.
Train	40000	65994	27764	27764	48542	48542	30000	30000	14890	26517	40000	40000	40000	40000
Dev	5000	8186	3164	3164	146	146	5000	5000	1600	2898	4139	4139	5000	5000
Test	5000	8381	3230	3230	10234	10234	5000	5000	2400	4217	525	525	574	574

- **LSOIE-wiki$^{\text{En}}$** and **LSOIE-sci$^{\text{En}}$**[4] [20] were algorithmically re-purposed from the QA-SRL BANK 2.0 dataset [7], covering the domains of Wikipedia and science, respectively.
- **COER$^{\text{Zh}}$**[5] is a high-quality Chinese knowledge base, created by an unsupervised open extractor [22] from heterogeneous web text.
- **SAOKE$^{\text{Zh}}$**[6] [21] is a human-annotated large-scale dataset for Chinese open information extraction.
- **WMORC$^{\text{Fr}}$** and **WMORC$^{\text{Ru}}$**[7] [6] consist of manually annotated open relation data (WMORC$_{\text{human}}$) for French and Russian, and automatically tagged (thus less reliable) relation data (WMORC$_{\text{auto}}$) for the two languages by a cross-lingual projection approach. The sentences are gathered from Wikipedia. We take WMORC$_{\text{auto}}$ for the training and development sets while using WMORC$_{\text{human}}$ as the test data.

3.2 Implementations

Encoders We utilize the *bert-base-cased* or *spanbert-base-cased* language models as the encoders on English datasets (SpanBERT only provides the English version up to now), and *bert-base-chinese* on Chinese datasets. Since there exist few high-quality monolingual LMs for French and Russian, we employ a multilingual LM *bert-base-multilingual-cased* on the datasets of the two languages.

Model Training Adam optimizer [12] is used with a learning rate of 3e-5, dropout rate of 0.1, weight decay of 0.01 and warmup proportion of 0.1. We train on a single NVIDIA Tesla P100 GPU with a batch size of 12 for 15 epochs with an early-stopping patience of 4. Evaluation is performed with our token-level evaluation script.

3.3 Evaluation Metrics

We keep track of the token-level open relation extraction metrics of F1 score, precision, and recall. The F1 score measures the average overlap between a model's

[4] https://github.com/Jacobsolawetz/large-scale-oie
[5] https://github.com/TJUNLP/COER
[6] https://ai.baidu.com/broad/introduction?dataset=saoke
[7] https://www.kaggle.com/shankkumar/multilingualopenrelations15

prediction and the ground-truth relation. Formally, F1 denotes the harmonic mean of precision and recall, where precision is defined as the ratio of correctly predicted tokens to the total number of predicted relation tokens. Recall, meanwhile, is the ratio of correctly predicted tokens to the total number of tokens in the ground-truth relation.

3.4 Baselines

In the experiments, we compare QORE models with a variety of previously proposed methods, some of which were used in the evaluation of the SOTA open relation extractor LOREM [8]. We denote the English (*En*) and Chinese (*Zh*) extractors and the models capable of processing multilingual (*Mul*) texts using the superscripts.

- **OLLIE$^{\mathbf{En}}$** [17] is a pattern-based extraction approach with complex relation schemas and context information of attribution and clausal modifiers.
- **ClausIE$^{\mathbf{En}}$** [2] exploits linguistic knowledge about English grammar to identify clauses as relations and their arguments.
- **Open IE-4.x$^{\mathbf{En}}$** [16] combines a rule-based extraction system and a system analyzing the hierarchical composition between semantic frames to generate relations.
- **MGD-GNN$^{\mathbf{Zh}}$** [15] constructs a multi-grained dependency graph and predicts based on character features and word boundary knowledge.
- **LOREM$^{\mathbf{Mul}}$** [8] is a multilingual-embedded sequence-labeling method based on CNN and BiLSTM, not relying on language-specific knowledge or external NLP tools.
- **Multi$^{\mathbf{2}}$OIE$^{\mathbf{Mul}}$** [19] is a multilingual sequence-labeling-based information extraction system combining BERT with multi-head attention blocks.

4 Experimental Results

4.1 $\mathbf{H_1}$: QORE for Multilingual Open Relation Extraction

In $\mathbf{H_1}$, we evaluate our QORE models on seven datasets of different languages (Tables 2 and 3) to compare with the rule-based and neural baselines. By contrast, QORE models outperform all the baselines on each dataset. We next explain the advantages of QORE framework, compare different query templates, and conduct an ablation study on our model.

For OLLIE, ClausIE, Open IE-4.x, and MGD-GNN, their dissatisfactory results are primarily due to the dependence on intricate language-specific relation formations and error propagation by the used external NLP tools (e.g., MGD-GNN utilizes dependency parser for constructing a multi-grained dependency graph.). If we contrast with the SOTA method LOREM, the neural sequence-labeling-based model outperforms the rule-based systems, but still cannot gain comparable outcomes to our QORE models. In the following, we focus on comparing the architectures of QORE and LOREM.

Table 2. Comparison on English datasets. Bolds indicate the best values per dataset. [Query templates: the question-mark (QM) style and the language-specific natural-language (NL) style.]

Model	OpenIE4^{En}			LSOIE-wiki^{En}			LSOIE-sci^{En}		
	P	R	F1	P	R	F1	P	R	F1
OLLIE	-	-	-	18.02	39.77	23.11	21.96	44.46	27.44
ClausIE	-	-	-	28.78	36.24	31.14	37.18	46.58	40.13
Open IE-4.x	-	-	-	32.06	40.79	34.70	37.73	48.07	40.88
LOREM	83.58	81.56	81.50	71.46	70.58	70.87	76.33	75.13	75.53
QORE$_{\text{BERT+QM}}$	97.89	97.81	97.75	96.74	97.26	96.82	97.35	97.89	97.50
QORE$_{\text{BERT+NL}}$	97.59	97.74	97.51	97.01	97.43	97.11	97.49	98.01	97.63
QORE$_{\text{SpanBERT+QM}}$	**98.85**	**99.10**	**98.76**	97.28	97.72	97.37	**97.60**	98.08	**97.71**
QORE$_{\text{SpanBERT+NL}}$	98.65	98.71	98.50	**97.38**	**97.96**	**97.51**	97.52	**98.15**	97.70

Table 3. Comparison on Chinese, French and Russian datasets

Model	COER^{Zh}			SAOKE^{Zh}			WMORC^{Fr}			WMORC^{Ru}		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
MGD-GNN	77.84	86.06	81.74	53.38	65.83	58.96	-	-	-	-	-	-
LOREM	41.49	42.40	41.42	46.68	52.92	48.70	83.30	83.88	82.75	82.63	87.86	83.80
QORE$_{\text{BERT+QM}}$	**98.11**	98.16	97.88	92.76	**95.00**	92.55	94.94	83.79	85.89	**91.74**	**92.48**	**90.62**
QORE$_{\text{BERT+NL}}$	98.10	98.16	**97.89**	**93.19**	94.78	**92.83**	**95.01**	**84.85**	**86.88**	91.51	92.29	90.24

LOREM encodes an input sequence using pre-trained word embeddings and adds argument tag vectors to the word embeddings. The argument tag vectors are simple one-hot encoded vectors indicating if a word is part of an argument. Then LOREM utilizes CNN and BiLSTM layers to form a representation of each word. The CNN is used to capture the local feature information, as LOREM considers that certain parts of the context might have higher chances of containing relation words than others. Meanwhile, the BiLSTM captures the forward and backward context of each word. Next, a CRF layer tags each word using the NST tagging scheme [10]: `S` (Single-word relation), `B` (Beginning of a relation), `I` (Inside a relation), `E` (Ending of a relation), `O` (Outside a relation).

Advantages of QORE over LOREM. Our QORE framework generates an initial sequence representation with word, position, and segment embeddings. Unlike the simple one-hot argument vectors of LOREM, QORE derives the argument information by creating a query template of arguments. We combine the query with the context to form the input of encoder, and the encoder outputs a contextual representation that we utilize to compute the relevance of each token to a gold relation (Equations 8 and 9). Moreover, by employing the self-attention mechanism of a Transformers-based encoder, QORE has the benefit of learning long-range dependencies easier and deriving a better representation for computing relevances, which we interpret in the following. Learning long-range dependencies is a key challenge in encoding sequences and solving related tasks

[23]. One key factor affecting the ability to learn such dependencies is the length of the paths forward and backward signals have to traverse between any two input and output positions in the network. The shorter these paths between any combination of positions in the input and output sequences, the easier it is to learn long-range dependencies. Vaswani et al. [23] also provide the maximum path length between any two input and output positions in self-attention, recurrent, and convolutional layers, which are $O(1)$, $O(n)$, and $O(\log_k(n))$, respectively. (k is the kernel width of a convolutional layer.) The constant path length of self-attention makes it easier to learn long-range dependencies than CNN and BiLSTM layers. Overall, QORE achieves substantial improvements over LOREM due to the better sequence representations with long-term dependencies, a basis of computing token relevances to gold relations.

In Table 2, if we concentrate on the BERT-encoded and SpanBERT-encoded QORE models, we find that the results from the SpanBERT-encoded models are relatively more significant than the BERT-encoded on all English datasets, which is in line with the advantage of SpanBERT over BERT on the span extraction task [11].

Different Query Templates. We analyze the effects of different query templates by experimenting with the question-mark (QM) style and the language-specific natural-language (NL) style. However, the overall evaluation results of both templates have marginal differences, as observed on all datasets used in this paper. The possible reason for the marginal results is that both the templates consist of the necessary argument information, and the representations learned via a pre-trained Transformer encoder are similar due to the robust expression ability of the encoder.

Ablation Study. To further trace the origin of the significant improvement of QORE, we conduct an ablation study to explore the improvement from two perspectives: the query-based span extraction framework and the pre-trained language model. To insulate the impact from a pre-trained LM, we compare the LSTM sequence-labeling-based model (i.e., the non-query LSTM Tagger, implemented by LOREM) with a query-based model QANet [24] which does not rely on LM. In the LM setting, we compare the BERT sequence-labeling-based model (i.e., the non-query BERT Tagger, performed by Multi^{2}OIE) with our query-based QORE$_{\text{BERT+QM}}$ model. Table 4 illustrates the results on COER$^{\text{Zh}}$. In the non-LM experiments, QANet significantly outperforms LSTM Tagger. Likewise, QORE$_{\text{BERT+QM}}$ model exceeds BERT Tagger, showing the improvement from the query-based extraction framework. Meanwhile, we observe that BERT Tagger surpasses LSTM Tagger, and QORE$_{\text{BERT+QM}}$ gains higher metrics than QANet, indicating that the pre-trained LM also has a considerable impact on promoting the extraction.

Table 4. Ablation study on non-query/query-based and non-LM/LM models

	non-LM			LM		
	LSTM Tagger			**BERT Tagger**		
non-query-based	P	R	F1	P	R	F1
	41.49	42.40	41.42	90.34	89.19	89.76
	QANet			**QORE**		
query-based	P	R	F1	P	R	F1
	95.98	96.52	96.08	**98.11**	**98.16**	**97.88**

4.2 $\mathbf{H_2}$: Zero-shot Domain Transferability of QORE

A model trained on data from the general domain does not necessarily achieve an equal performance when testing on a specific domain such as biology or literature. In $\mathbf{H_2}$, we evaluate the zero-shot domain transferability of QORE by training models on the general-domain LSOIE-wiki$^{\text{En}}$ and testing them on the benchmark of science-domain LSOIE-sci$^{\text{En}}$. We compare our QORE$_{\text{BERT+QM}}$ model with BERT Tagger (non-query-based, performed by Multi^{2}OIE). Table 5 illustrates that when transferring from the general to science domain, QORE$_{\text{BERT+QM}}$ decreases by F1 score (-0.15%) whereas BERT Tagger reduces by F1 (-14.58%). The slighter decline in QORE's performance shows that our model has superior domain transferability.

Table 5. Results of domain transfer

Model	**Train**	**Test**	**F1**
BERT Tagger	LSOIE-sci$^{\text{En}}$	LSOIE-sci$^{\text{En}}$	74.53
QORE$_{\text{BERT+QM}}$	LSOIE-sci$^{\text{En}}$	LSOIE-sci$^{\text{En}}$	97.50
BERT Tagger	LSOIE-wiki$^{\text{En}}$	LSOIE-sci$^{\text{En}}$	59.95 (-14.58)
QORE$_{\text{BERT+QM}}$	LSOIE-wiki$^{\text{En}}$	LSOIE-sci$^{\text{En}}$	97.35 (-0.15)

4.3 $\mathbf{H_3}$: Few-shot Learning Ability of QORE

For few-shot learning ability, we carry out a set of experiments with shrinking training data to compare our QORE model with BERT Tagger (non-query-based, performed by Multi^{2}OIE) on LSOIE-wiki$^{\text{En}}$. Figure 2 indicates that by reducing training samples to 50%, BERT Tagger declines by F1 (-2.75%) while QORE$_{\text{BERT+QM}}$ achieves an even higher F1 (+0.26%). When reducing the training set to 6.25%, QORE$_{\text{BERT+QM}}$ results in a decreased F1 (-1.28%) totally compared with using the whole training set, whereas BERT Tagger reduces by F1 (-7.02%) in total. The comparison results imply that our query-based span extraction framework may bring more enhanced few-shot learning ability to QORE model.

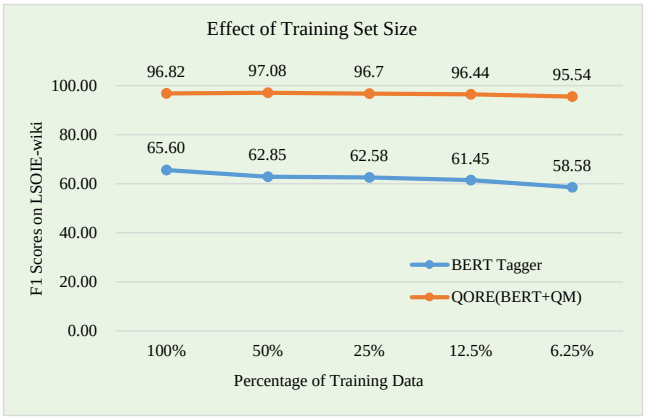

Figure 2. Results of different training set sizes

4.4 Case Study

We conduct case studies on the test set of LSOIE-sci$^{\text{En}}$ to compare the outputs of QORE$_{\text{BERT+QM}}$ and the SOTA extractor LOREM. We choose the English benchmark in this section for the convenience of readers.

Correct extractions: We notice that QORE is better at handling cases where arguments separately locate in main and subordinate clauses than LOREM. For instance, given a context "*The scientific method is a set of steps that help us to answer questions.*" and an argument pair ⟨*The scientific method, answer questions*⟩, LOREM wrongly predicts "*to*" as the relation while QORE extracts "*help*" that is identical with the ground-truth relation. Considering another case where the context is "*Different enzymes that catalyze the same chemical reaction are called isozymes.*" and the argument pair is ⟨*Different enzymes, the same chemical reaction*⟩, our QORE gives out the gold relation "*catalyze*" whereas LOREM predicts "*called*" incorrectly.

False extractions: We analyze the false cases of QORE, and summarize that the majority of errors occur in situations where the semantic relation between arguments concerns modal verbs or auxiliary verbs, such as "*can*" and "*have been*". In these cases, QORE extracts either more or fewer words than gold relations, e.g., provided a context "*Algae had covered moist land areas for millions of years.*" and an argument pair ⟨*Algae, moist land areas*⟩, QORE predicts "*covered*" as the relation that is fewer than the ground truth "*had covered*". Meanwhile, we observe that LOREM also outputs "*covered*". Take another example where the context is "*If the scientist can picture the data the results may be easier to understand.*" and the arguments are ⟨*the scientist, the data*⟩, QORE extracts "*can picture*", more than the gold relation "*picture*".

5 Related Work

We have reviewed the previously proposed methods, datasets, and challenges of ORE in the Introduction. Here, we focus on the related works of our model framework.

Increasing studies have cast various NLP tasks as machine reading comprehension over a context, leading to comparable or superior performances in the following tasks: named entity recognition [13], joint entity-closed-relation extraction [14,26], semantic role labeling [9], and event extraction [4]. Inspired by the MRC framework, we transform ORE into a query-based span prediction task to better construct semantic relations between arguments via well-designed queries.

6 Conclusion

Our work targets open relation extraction using a novel query-based extraction framework QORE. The evaluation results show that our model achieves significant improvements over the SOTA method LOREM. Regarding some practical challenges, we investigate that QORE models show excellent zero-shot domain transferability and few-shot learning ability. In the future, we will explore further demands of the ORE task (e.g., extracting multi-span open relations and detecting non-existent relationships) and present corresponding solutions.

Acknowledgements. This work is supported by the NSFC-General Technology Basic Research Joint Funds under Grant (U1936220), the National Natural Science Foundation of China under Grant (61972047) and the National Key Research and Development Program of China (2018YFC0831500).

References

1. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In: Veloso, M.M. (ed.) IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007. pp. 2670–2676 (2007), http://ijcai.org/Proceedings/07/Papers/429.pdf
2. Corro, L.D., Gemulla, R.: Clausie: clause-based open information extraction. In: Schwabe, D., Almeida, V.A.F., Glaser, H., Baeza-Yates, R., Moon, S.B. (eds.) 22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013. pp. 355–366. International World Wide Web Conferences Steering Committee / ACM (2013). https://doi.org/10.1145/2488388.2488420, https://doi.org/10.1145/2488388.2488420
3. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) 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). pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/n19-1423, https://doi.org/10.18653/v1/n19-1423
4. Du, X., Cardie, C.: Event extraction by answering (almost) natural questions. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020. pp. 671–683. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-main.49, https://doi.org/10.18653/v1/2020.emnlp-main.49
5. Fader, A., Soderland, S., Etzioni, O.: Identifying relations for open information extraction. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL. pp. 1535–1545. ACL (2011), https://aclanthology.org/D11-1142/
6. Faruqui, M., Kumar, S.: Multilingual open relation extraction using cross-lingual projection. In: Mihalcea, R., Chai, J.Y., Sarkar, A. (eds.) NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015. pp. 1351–1356. The Association for Computational Linguistics (2015). https://doi.org/10.3115/v1/n15-1151, https://doi.org/10.3115/v1/n15-1151
7. FitzGerald, N., Michael, J., He, L., Zettlemoyer, L.: Large-scale QA-SRL parsing. In: Gurevych, I., Miyao, Y. (eds.) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers. pp. 2051–2060. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/P18-1191, https://aclanthology.org/P18-1191/
8. Harting, T., Mesbah, S., Lofi, C.: LOREM: language-consistent open relation extraction from unstructured text. In: Huang, Y., King, I., Liu, T., van Steen, M. (eds.) WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020. pp. 1830–1838. ACM / IW3C2 (2020). https://doi.org/10.1145/3366423.3380252, https://doi.org/10.1145/3366423.3380252
9. He, L., Lewis, M., Zettlemoyer, L.: Question-answer driven semantic role labeling: Using natural language to annotate natural language. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 643–653. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015). https://doi.org/10.18653/v1/D15-1076, https://aclanthology.org/D15-1076
10. Jia, S., Xiang, Y., Chen, X.: Supervised neural models revitalize the open relation extraction. CoRR **abs/1809.09408** (2018), http://arxiv.org/abs/1809.09408
11. Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O.: Spanbert: Improving pre-training by representing and predicting spans. Trans. Assoc. Comput. Linguistics **8**, 64–77 (2020), https://transacl.org/ojs/index.php/tacl/article/view/1853
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6980
13. Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., Li, J.: A unified MRC framework for named entity recognition. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) Proceedings of the 58th Annual Meeting of the Association for Compu-

tational Linguistics, ACL 2020, Online, July 5-10, 2020. pp. 5849–5859. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.519, https://doi.org/10.18653/v1/2020.acl-main.519

14. Li, X., Yin, F., Sun, Z., Li, X., Yuan, A., Chai, D., Zhou, M., Li, J.: Entity-relation extraction as multi-turn question answering. In: Korhonen, A., Traum, D.R., Màrquez, L. (eds.) Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. pp. 1340–1350. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/p19-1129, https://doi.org/10.18653/v1/p19-1129
15. Lyu, Z., Shi, K., Li, X., Hou, L., Li, J., Song, B.: Multi-grained dependency graph neural network for chinese open information extraction. In: Karlapalem, K., Cheng, H., Ramakrishnan, N., Agrawal, R.K., Reddy, P.K., Srivastava, J., Chakraborty, T. (eds.) Advances in Knowledge Discovery and Data Mining - 25th Pacific-Asia Conference, PAKDD 2021, Virtual Event, May 11-14, 2021, Proceedings, Part III. Lecture Notes in Computer Science, vol. 12714, pp. 155–167. Springer (2021). https://doi.org/10.1007/978-3-030-75768-7_13, https://doi.org/10.1007/978-3-030-75768-7_13
16. Mausam: Open information extraction systems and downstream applications. In: Kambhampati, S. (ed.) Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016. pp. 4074–4077. IJCAI/AAAI Press (2016), http://www.ijcai.org/Abstract/16/604
17. Mausam, Schmitz, M., Soderland, S., Bart, R., Etzioni, O.: Open language learning for information extraction. In: Tsujii, J., Henderson, J., Pasca, M. (eds.) Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea. pp. 523–534. ACL (2012), https://aclanthology.org/D12-1048/
18. Qiu, L., Zhang, Y.: ZORE: A syntax-based system for chinese open relation extraction. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL. pp. 1870–1880. ACL (2014). https://doi.org/10.3115/v1/d14-1201, https://doi.org/10.3115/v1/d14-1201
19. Ro, Y., Lee, Y., Kang, P.: Multi^{2}oie: Multilingual open information extraction based on multi-head attention with BERT. In: Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020. Findings of ACL, vol. EMNLP 2020, pp. 1107–1117. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.99, https://doi.org/10.18653/v1/2020.findings-emnlp.99
20. Solawetz, J., Larson, S.: LSOIE: A large-scale dataset for supervised open information extraction. In: Merlo, P., Tiedemann, J., Tsarfaty, R. (eds.) Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021. pp. 2595–2600. Association for Computational Linguistics (2021), https://aclanthology.org/2021.eacl-main.222/
21. Sun, M., Li, X., Wang, X., Fan, M., Feng, Y., Li, P.: Logician: A unified end-to-end neural approach for open-domain information extraction. In: Chang, Y., Zhai, C., Liu, Y., Maarek, Y. (eds.) Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining,

WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018. pp. 556–564. ACM (2018). https://doi.org/10.1145/3159652.3159712, https://doi.org/10.1145/3159652.3159712

22. Tseng, Y., Lee, L., Lin, S., Liao, B., Liu, M., Chen, H., Etzioni, O., Fader, A.: Chinese open relation extraction for knowledge acquisition. In: Bouma, G., Parmentier, Y. (eds.) Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden. pp. 12–16. The Association for Computer Linguistics (2014). https://doi.org/10.3115/v1/e14-4003, https://doi.org/10.3115/v1/e14-4003
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 5998–6008 (2017), https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
24. Yu, A.W., Dohan, D., Luong, M., Zhao, R., Chen, K., Norouzi, M., Le, Q.V.: Qanet: Combining local convolution with global self-attention for reading comprehension. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018), https://openreview.net/forum?id=B14TlG-RW
25. Zhan, J., Zhao, H.: Span model for open information extraction on accurate corpus. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. pp. 9523–9530. AAAI Press (2020), https://aaai.org/ojs/index.php/AAAI/article/view/6497
26. Zhao, T., Yan, Z., Cao, Y., Li, Z.: Asking effective and diverse questions: A machine reading comprehension based framework for joint entity-relation extraction. In: Bessiere, C. (ed.) Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020. pp. 3948–3954. ijcai.org (2020). https://doi.org/10.24963/ijcai.2020/546, https://doi.org/10.24963/ijcai.2020/546