Large-Scale Multi-granular Concept Extraction Based on Machine Reading Comprehension

Siyu Yuan[1], Deqing Yang[1(✉)], Jiaqing Liang[2], Jilun Sun[2],
Jingyue Huang[1], Kaiyan Cao[1], Yanghua Xiao[2,3(✉)], and Rui Xie[4]

[1] School of Data Science, Fudan University, Shanghai, China
`{yuansy17,yangdeqing,jingyuehuang18,kycao20}@fudan.edu.cn`
[2] School of Computer Science, Fudan University, Shanghai, China
`{jlsun18,shawyh}@fudan.edu.cn`
[3] Fudan-Aishu Cognitive Intelligence Joint Research Center, Shanghai, China
[4] Meituan, Beijing, China
`rui.xie@meituan.com`

Abstract. The concepts in knowledge graphs (KGs) enable machines to understand natural language, and thus play an indispensable role in many applications. However, existing KGs have the poor coverage of concepts, especially fine-grained concepts. In order to supply existing KGs with more fine-grained and new concepts, we propose a novel concept extraction framework, namely *MRC-CE*, to extract large-scale multi-granular concepts from the descriptive texts of entities. Specifically, MRC-CE is built with a machine reading comprehension model based on BERT, which can extract more fine-grained concepts with a pointer network. Furthermore, a random forest and rule-based pruning are also adopted to enhance MRC-CE's precision and recall simultaneously. Our experiments evaluated upon multilingual KGs, i.e., English *Probase* and Chinese *CN-DBpedia*, justify MRC-CE's superiority over the state-of-the-art extraction models in KG completion. Particularly, after running MRC-CE for each entity in CN-DBpedia, more than 7,053,900 new concepts (instanceOf relations) are supplied into the KG. The code and datasets have been released at https://github.com/fcihraeipnusnacwh/MRC-CE.

Keywords: Concept extraction · Knowledge graph · Machine reading comprehension · Multi-granular concept · Concept overlap

1   Introduction

The concepts in knowledge graphs (KGs) [3,36,41] enable machines to better understand natural languages and thus play an increasingly significant role in many applications, such as question answering [7], personalized recommendation [32], commonsense reasoning [25], etc. Particularly, fine-grained concepts greatly promote the downstream applications. For example, if entity *Google* has 'technology company' and 'search engine company' as its concepts, a job recommender system would recommend Google rather than Wal-Mart to a graduate from computer science department based on such fine-grained concepts.

Although there have been a great number of efforts on constructing KGs in recent years, the concepts in existing KGs are still far from being complete, especially for fine-grained concepts. In the widely used Chinese KG *CN-DBpedia* [41], there are nearly 17 million entities but only 0.27 million concepts in total, and more than 20% entities even have no concepts. In general, fine-grained concepts contain more than one modifier, and thus have longer characters (words). However, the average character length of CN-DBpedia concepts is only 3.62, implying that most of them are coarse-grained.

The poor coverage of concepts, especially fine-grained concepts, in the existing KGs is due to their approaches' drawbacks. Most of the existing concept acquisition approaches are based on *generation* or *extraction* from texts. Generation methods often generate coarse-grained concepts from free texts since they are inclined to generate high-frequency words [2,34]. Extraction methods mainly have three types of models. Pattern-matching models [3,24,36,45] focus on extracting concepts from texts based on handcrafted patterns or rules, but the recall of concept extraction (CE) is low due to their limited generalization ability. Classification models [9,15,18,40] identify concepts through classifying a given entity into a predefined concept set based on text information, but can not find new concepts. Sequence labeling models [22,38] treat the CE problem as a sequence labeling task, but can not handle the problem of *concept overlap*. The concept overlap refers to the phenomenon that a concept term is the subsequence of another concept term. For example, in Fig. 1, once a sequence labeling model labels 'company' as Concept 1, it would not label 'multinational technology company' as a fine-grained concept, since it can not mark a token with different labels.

Machine reading comprehension (MRC) model can extract the answer from the contextual texts for a proposed question. Inspired by MRC model's excellent extraction capability, we handle CE problem as an MRC task in this paper, and propose a novel CE framework named *MRC-CE*, to implement large-scale multi-granular CE from the descriptive texts of given entities. MRC-CE can extract more multi-granular concepts than previous concept extraction models due to the following reasons. Firstly, MRC-CE is built with an MRC model based on BERT [8] (BERT-MRC), which can find abundant new concepts and handle the problem of *concept overlap* well with a pointer network [37]. Secondly, a random forest [30] and rule-based pruning are adopted to filter the candidate concepts output by BERT-MRC, which enhances MRC-CE's precision and recall
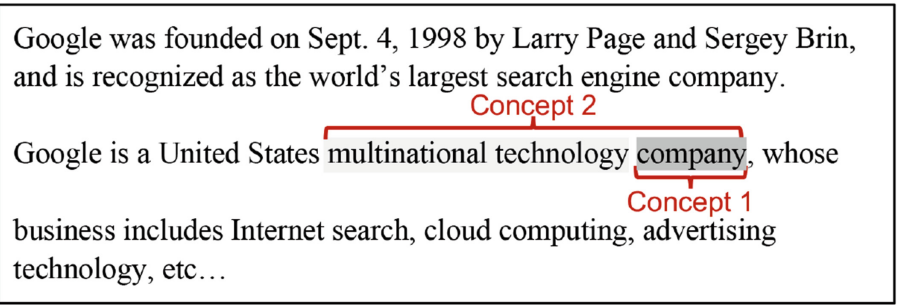
Fig. 1. The abstract text of entity *Google*. The problem of concept overlap in the text is challenging for traditional extraction models to extract multiple concepts from the same span.

simultaneously. Furthermore, MRC-CE has been proven capable of acquiring large-scale multi-granular concepts through our extensive experiments.

The major contributions of this paper can be summarized as follows:

1. To the best of our knowledge, this is the first work to employ MRC model in text-based CE for large-scale KG completion.
2. Through our experiments, more than 7,053,900 new concepts (instanceOf relations) were extracted by our MRC-CE for completing the large-scale Chinese KG *CN-DBpedia*. Furthermore, the online service of CE based on MRC-CE is released on http://kw.fudan.edu.cn.

The rest of this paper is organized as follows. Section 2 is the review of related works and Sect. 3 is the detailed introduction of our framework, respectively. We display our experiment results in Sect. 4 to justify the effectiveness and rationality of MRC-CE. We also display the CE results of applying our framework on CN-DBpedia and Meituan[1] in Sect. 5, and at last conclude our work in Sect. 6.

2 Related Work

In this section, we review some works related to concept acquisition.

2.1 Ontology Extension

Ontology [5] extension focuses more on identifying hypernym-hyponym relationships [23, 46, 47]. Unlike it, our goal is to acquire more new concepts, especially fine-grained concepts, to complete the existing KGs. Thus, we did not compare MRC-CE with these methods in our experiments.

---

[1] https://www.meituan.com.

2.2   IsA Relation and Entity Typing

IsA relation extraction [12,29] is extracting the subsumption (subClassOf) relation between two classes. Entity type aims to classify an entity into a predefined set of types (concepts), such as person, location and organization without new concept recognition [17,40,42]. However, in our setting in this paper, the entity is given at first and its candidate concepts are not pre-defined. Thus, our work is different to isA relation extraction and entity typing, and thus the methods of these two tasks were not compared in our experiments.

2.3   Text-Based Concept Extraction

In this paper, we only focus on extracting the concepts already existing in the texts rather than concept generation. Hence we pay attention to extraction models rather than generative models [2,34]. The existing CE methods can be divided into three categories.

Pattern-Matching Method. Pattern-matching methods [3,4,16,24,36,39,41,45] try to extract concepts from free texts based on handcrafted patterns or rules. They can acquire accurate concepts, but only have low recall due to the poor generalization ability. Comparatively, our MRC-CE achieves CE task based on MRC model beyond the limitation of handcrafted patterns, and thus acquires more concepts.

Learning-Based Method. Classification models transform CE into classification [9,17,18,40] to determine which concept in a predefined set meets hypernym-hyponym relation with the given entity, but they can not acquire new concepts. Sequence labeling models have been proven effective on extraction tasks [1,21,22,38]. Given the extraction feature, sequence labeling models can also extract concepts from the texts describing entities as our MRC-CE. Recently, pre-trained contextual embeddings have been widely used in sequence labeling models [6,35,43,44] to gain good performance, but can not handle the problem of *concept overlap*.

Knowledge-Based Method. These methods [4,26,27] mainly use the external information from KGs to achieve extraction tasks, resulting in good CE. However, these models have poor generalization ability and only focus on a specific field.

2.4   Machine Reading Comprehension

MRC [19] is a task to seek the answer from contextual texts for a proposed question, which can be categorized into four classes according to answer format, i.e., cloze test, multiple choice, span extraction and free answering. The span extraction [31,33] is most related and similar to our task. The state-of-the-art pre-trained language models [8,20] are often applied in MRC tasks. More

recently, many researchers have employed MRC model in accomplishing other NLP tasks, including nested NER [13] and RE [14]. Comparatively, our MRC-CE is the first attempt of employing MRC model in text-based CE.

3 Methodology

In this section, we introduce our CE framework in detail.

3.1 Task Definition

Our CE task can be formulated as follows. Given an entity e and its relevant descriptive text, denoted as a sequence of words $X = \{x_1, x_2, ..., x_n\}$ where $x_i (1 \leq i \leq n)$ is a word token, the CE task aims to find one or multiple spans from X as e's concept(s).

3.2 Data Construction

The descriptive text of high quality plays an important role in text-based CE. Since our task scenario is span extraction, the input text should contain the concept(s) of the given entity. We consider the given entity's abstract in encyclopedia since it is well structured and explicitly mentions the concept(s) of the given entity. In the following introduction, the input text is always referred to the abstract of a given entity. The construction details of our English and Chinese datasets will be introduced in Subsect. 4.1.

3.3 Summary

To employ MRC model for CE, we need to construct appropriate questions towards which the spans in the text are extracted. Since our target is to extract concepts from the abstract of a given entity, we set the question Q as 'What is the concept for [entity]?'. The pipeline of our MRC-CE is displayed in Fig. 2, which can be divided into three modules, i.e., the concept extractor, concept selector and concept pruner. The first module is BERT-MRC built with a pointer network [37], which receives the input text and extracts candidate multi-granular concepts from the text. The second module adopts a random forest [30] to select the desirable concepts from the candidates output by BERT-MRC. The third module prunes away the unsatisfactory concepts from the second module's outputs based on some pruning rules.

3.4 MRC-Based Concept Extractor

The BERT-MRC in our framework is built with a BERT [8] encoder and a pointer network to generate candidate spans along with their corresponding confidence scores of being concepts. We took each entity's concepts from the KG which also exist in the entity's abstract text as the real labels of one training sample.
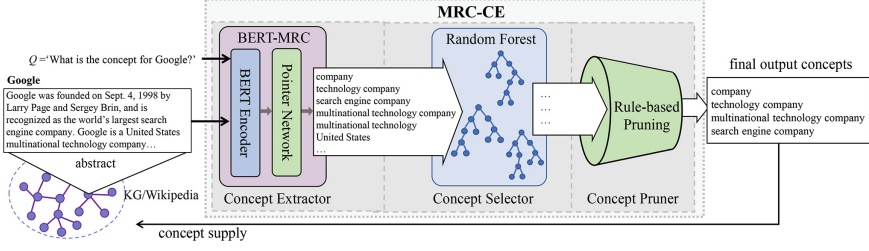
Fig. 2. The pipeline of our MRC-CE framework.

BERT Encoding. In this module, BERT is used as an encoder layer to generate an embedding for each input token, based on which the subsequent pointer network predicts the confidence score for each candidate extracted span. At first, the tokens of input X and the tokens of input X and the question Q are concatenated together as follows, to constitute the input of BERT encoder.

$$\{[CLS]q_1, q_2, ..., q_m[SEP]x_1, x_2, ..., x_n\}$$

where [CLS] and [SEP] are special tokens, and $q_i (1 \leq i \leq m)$ is a token of Q. Then, BERT encoder outputs the text token embedding matrix $\mathbf{E} \in \mathbb{R}^{n \times d}$, where d is the embedding dimension.

Generating Candidate Concepts by Pointer Network. With the token embeddings in $\mathbf{E}$, a pointer network is built to predict the probability of a token being the start position and end position of the answer, through a fully connected layer after the BERT encoder. Then, the confidence score of each span can be calculated as the sum of the probability of its start token and end token. One span can be output repeatedly as the same subsequence of multiple extracted concepts through an appropriate selection threshold. This strategy enables extracting multi-granular concepts. For example, in Fig. 3, 'company' is extracted for multiple times corresponding to three concepts of different granularity, if the confidence score threshold is set to 0.85.

Specifically, we use $\mathbf{p}^{start}, \mathbf{p}^{end} \in \mathbb{R}^n$ to denote the vectors storing each token's probability of being the start position and end position, respectively. They are calculated based on $\mathbf{E}$ as

$$[\mathbf{p}^{start}; \mathbf{p}^{end}] = softmax(\mathbf{EW} + \mathbf{B}) \qquad (1)$$

where $\mathbf{W}, \mathbf{B} \in \mathbb{R}^{n \times 2}$ are both trainable parameters. Given a span with x_i and x_j as the start token and the end token, respectively, its confidence score $cs_{ij} \in \mathbb{R}$ can be calculated as

$$cs_{ij} = p_i^{start} + p_j^{end} \qquad (2)$$

Then, BERT-MRC generates a ranking list of candidate concepts (spans) along with confidence scores, and outputs the concepts with the confidence scores larger than the selection threshold.
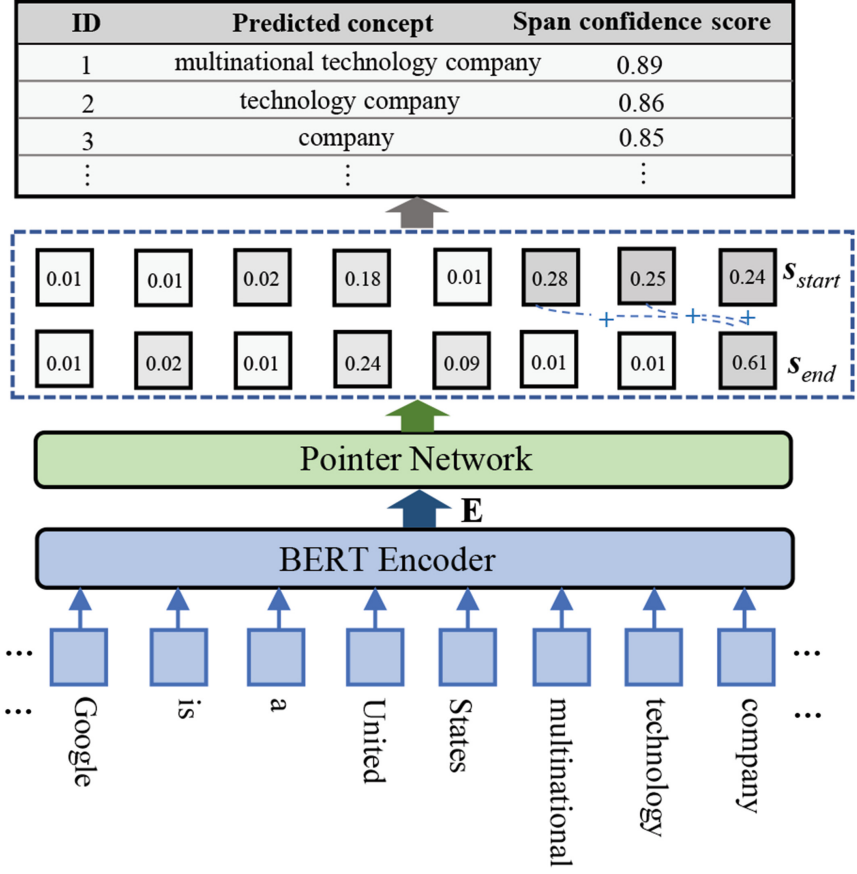
ID	Predicted concept	Span confidence score
1	multinational technology company	0.89
2	technology company	0.86
3	company	0.85
⋮	⋮	⋮

Fig. 3. The structure of BERT-MRC.

BERT-MRC Loss. We adopt cross entropy function $CrossEntropy(\cdot)$ as the loss function of BERT-MRC. Specifically, suppose the set $Y_{start} \in \mathbb{R}^n$ (or $Y_{end} \in \mathbb{R}^n$) contains the real label of an input token $x_i (1 \leq i \leq n)$ being the start (or end) position of a concept. And $Y_{span} \in \mathbb{R}^{C_n^2}$ contains the real label of a span $x_i x_j$ where x_i and x_j are the start position and end position of a concept, respectively. Then, we have the following three losses for the predictions of the three situations:

$$\mathcal{L}_{start} = CrossEntropy(\mathbf{p}^{start}, Y_{start})$$
$$\mathcal{L}_{end} = CrossEntropy(\mathbf{p}^{end}, Y_{end}) \tag{3}$$
$$\mathcal{L}_{span} = CrossEntropy(\mathbf{cs}, Y_{span})$$

Then, the comprehensive loss of training BERT-MRC is

$$\mathcal{L} = \alpha \mathcal{L}_{start} + \beta \mathcal{L}_{end} + (1 - \alpha - \beta) \mathcal{L}_{span} \tag{4}$$

where $\alpha, \beta \in (0,1)$ are control parameters. We use Adam [10] to optimize the loss.

3.5   Selecting Concepts by Random Forest

The second module (concept selector) is built to select desirable concepts from the candidate spans extracted by BERT-MRC. We argue that it is unsatisfactory to output the concepts by choosing a specific threshold to directly truncate BERT-MRC's output ranking list. If we adopt a relatively big threshold, we can get more accurate concepts but may lose some correct concepts. If recall is preferred, precision might be hurt. To achieve a better trade-off between precision and recall, we adopt a classifier to predict whether a candidate extracted by BERT-MRC deserves being reserved. Such concept selector filters out the desirable concepts in terms of the significant features rather than the span confidence score, and thus improves the performance of concept filtering.

Specifically, we feed the classifer with the following features. At first, we adopt the span confidence score (feature A), start position probability (feature B) and end position probability (feature C) of each candidate span, which are obtained from BERT-MRC. Furthermore, we use another two features. Feature D is whether the current candidate span is an existing concept in the KG. In addition, the most important objective of MRC-CE is to handle the problem of concept overlap for fine-grained CE. Therefore, we consider another feature E whether the current span contains other candidate span(s).

Since there are only five simple features, we do not need to choose deep models. For the traditional classification model, random forest [30] has better accuracy than many other models, since it adopts ensemble learning. Therefore, we adopt random forest as the classifier.

We did not directly use the concepts existing in the KG as the training samples' real labels of random forest given that many correct concepts not existing in the KG can be extracted by BERT-MRC. Thus, we invited some volunteers to label the training samples manually. Specifically, we randomly sampled 1,000 results from BERT-MRC's outputs and requested the volunteers to label whether the result is a correct concept of the corresponding entity. Then, we saved the features and the labels of the results to constitute the training samples. The significance values of the five features obtained through training the random forest towards the Chinese corpus (CN-DBpedia) are listed in Table 1. From the table, we find that except for feature A, feature D is the most significant feature to select desirable concepts. It implies that referring to the existing concepts in KGs is very crucial to extract new concepts.

Table 1. The significance values of the five features in MRC-CE's random forest for Chinese KG *CN-DBpedia*.

Feature	A	B	C	D	E
Significance	0.34	0.20	0.16	0.25	0.05

3.6   Rule-Based Concept Pruning

There are still some wrong concepts being reserved after above concept selection. The errors can be roughly categorized into three classes as below.

1. The extracted concepts are semantically exclusive with each other. For example, both 'president' and 'vice president' could be extracted by BERT-MRC, but they are mutually exclusive in terms of conceptualization. Obviously, a president cannot be a vice president simultaneously.
2. The modifier of a concept is mistakenly extracted as another concept. For example, 'railway' and 'station' are both extracted as concepts from the span 'railway station'.
3. A correct concept is mistakenly mixed with a functional word. For example, both 'is ancient costume drama' and 'in high school' are wrong concepts, where 'is' and 'in' are the redundant tokens.

In this step, some pruning strategies are executed based on explicitly rules, since above errors can be easily recognized according to some pre-defined patterns. For example, for the class that the extracted concepts are mutually exclusive in semantics, we set our pruning rule case by case rather than setting a general rule.

4   Experiments

In this section, in order to justify our framework's CE capability for KG completion, we evaluate its performance upon a Chinese KG *CN-DBpedia* and an English KG *Probase*.

4.1   Datasets

Chinese Dataset. We obtained the latest data of CN-DBpedia from its open websites[2] and randomly selected 100,000 entities along with their concepts and abstract texts, as the sample pool. Then, we randomly selected 500 samples from the sample pool as the test set, and the rest samples were divided as the training and validation set according to 9:1, to learn the models. For BERT-MRC, we took each entity's concepts from the KG which also exist in the entity's abstract text as the real labels of one training sample. For training the random forest, we invited some volunteers to label whether the candidate concepts from BERT-MRC are the given entity' concepts since many correct concepts not existing in the KG can be extracted by BERT-MRC. When evaluating the models with the test samples, we invited some volunteers to assess whether the concepts extracted by the models are correct, given that many new concepts not existing in the KG may be extracted.

---

² http://kw.fudan.edu.cn/cndbpedia.

English Dataset. Probase[3] has the hypernym-hyponym relations between concepts and entities, but no abstract texts of entities. Hence, we first fetched the entities along with their concepts from Probase, and then crawled entities' abstract texts from Wikipedia.[4] Particularly, we totally sampled 50,000 entities with their concepts and abstract texts. Then, we constructed the training, validation and test set the same as the Chinese dataset.

4.2   Baselines

We compared our MRC-CE with the following five state-of-the-art models to justify MRC-CE's performance. Please note that, the models we selected are NER and Open IE models since NER and Open IE are extraction tasks which are mostly meet our task scenario. Besides, We also compare the pattern matching approaches.

We did not compare the methods for ontology extension and the generation models, since both of them do not meet our task scenario, i.e., text-based CE. The entity typing models and classification models were also not included because they can not meet the objective to complete existing KGs with new concepts. The knowledge-based methods were also excluded since most of them are only applicable to specific fields.

Hearst Patterns [28]. Hearst patterns are the basic and popular rules for extracting concepts from free texts. We totally designed 5 Hearst patterns to achieve concept extraction, which are listed in Table 2.[5] We allowed leading and following noun phrases to match limited modifiers to extract fine-grained concepts.

Table 2. The full set of Hearst patterns for *CN-DBpedia* and *Probase*.

CN-DBpedia	X is Y
	X is a type/a of Y...
	X is one of Y
	X belongs to Y
	Y is located/founded/ in...
Probase	X is a Y that/which/who
	X is one of Y
	X is a member/part/form... of Y
	X refers to Y
	As Y, X is ...

XLNet-NER [44]. With the capability of modeling bi-directional contexts, XLNet demonstrates an excellent sequence labeling model in many NLP tasks, such as NER and Open IE, which is also applicable to our task.

---

[3] https://www.microsoft.com/en-us/research/project/probase/.
[4] https://www.wikipedia.org/.
[5] We translate Chinese patterns for CN-DBpedia into English.

FLAIR [1]. FLAIR is a novel NLP framework combining different word and document embeddings to achieve excellent results. FLAIR can be employed in our CE task since it can extract spans from the text.

KVMN [22]. As a sequence labeling model, KVMN was proposed to handle NER by leveraging different types of syntactic information through attentive ensemble.

XLM-R [6]. It is a transformer-based multilingual masked language model incorporating XLM [11] and RoBERTa [20]. It can achieve CE task since it is also a sequential labeling model.

BERT-BiLSTM-CRF [35, 43]. It is an advanced version of BERT built with BiLSTM and CRF. This language model is used to obtain optimal token embeddings, based on which the downstream tasks such as NER and Open IE, CE are well achieved.

4.3   Experiment Settings

In MRC-CE, we adopted a BERT-base with 12 layers, 12 self-attention heads and H = 768 as BERT-MRC's encoder. The training settings are: batch size = 4, learning rate = 3e−5, dropout rate = 0.1 and training epoch = 2. In addition, we set $\alpha = 0.3, \beta = 0.25$ in Eq. 4 based on parameter tuning. Besides, we chose 50 as the tree number of the random forest.

4.4   Experiment Results

CE Performance Comparison. We refer to an extracted concept of a given entity that already exists in the KG as an *EC* (existing concept), and refer to an extracted correct concept of a given entity that does not exist in the KG as a *NC* (new concept). NCs are more significant than ECs given that the primary objective of our work is KG completion, i.e., supplying new instanceOf relations. Please note that a NC in fact corresponds to a new instanceOf relation to a given entity rather than being a unique new concept, because the concept linked by this new instanceOf relation may already exist in the KG as another entity's concept.

All models' CE results of the Chinese dataset and the English dataset are listed in Table 3, where EC # and NC # are the number of existing concepts and new concepts (instanceOf relations) of the entities, respectively. From the table we find that, compared with other models, MRC-CE extracts more NCs. As we claimed before, our framework is capable of extracting fine-grained or long-tail concepts from the texts. To justify it, we counted the average number of characters (for Chinese) or words (for English) constituting ECs (EC length) and NCs (NC length). The larger NC length is, the more possible the NC is to be fine-grained/long-tail. Although Hearst Patterns's NC length is larger than MRC-CE's in the Chinese dataset, it ignores some coarse-grained concepts and thus it can not achieve multi-granular concept extraction. For example, as shown in Table 6, Hearst Patterns only regards 'the railway station of JR East Japan' as a whole concept, whereas MRC-CE can extract 'station', 'railway station' and 'the railway station of JR East Japan' simultaneously from the same span.

Table 3. CE performance comparisons of 500 test samples.

Dataset	Model	EC #	NC #	EC length	NC length	OC ratio	Prec.	R-Recall	R-F1
CN-DBpedia	Hearst Patterns	158	222	2.63	**7.1**	NA	95.24%	27.24%	42.36%
	XLNet-NER	391	46		2.61	NA	89.92%	5.64%	10.62%
	FLAIR	405	63		3.11	NA	**95.51%**	7.73%	14.30%
	KVMN	247	253		4.03	NA	64.27%	31.04%	41.86%
	XLM-R	89	250		5.35	NA	76.35%	30.67%	43.77%
	BERT-BiLSTM-CRF	411	25		4.32	NA	88.26%	3.07%	5.93%
	MRC-CE	**519**	**323**		4.91	**29.35%**	92.22%	**39.63%**	**55.44%**
Probase	Hearst Patterns	308	402	1.25	2.04	NA	**95.56%**	20.18%	33.32%
	XLNet-NER	534	398		1.42	NA	92.00%	19.98%	32.83%
	FLAIR	307	141		1.67	NA	84.69%	7.08%	13.06%
	KVMN	186	404		1.96	NA	47.50%	20.28%	28.43%
	XLM-R	**672**	322		1.48	NA	81.74%	16.16%	26.99%
	BERT-BiLSTM-CRF	191	154		1.68	NA	81.18%	7.73%	14.12%
	MRC-CE	636	**626**		**2.31**	**36.82%**	90.08%	**31.38%**	**46.54%**

To prove MRC-CE's capability of extracting overlapped concepts, we further recorded the ratio of overlapped concepts (one is another one's subsequence) to all extracted concepts, denoted as OC ratio in Table 3. The precision (Prec.) is the ratio of the correct concepts assessed by the volunteers to all concepts extracted by the model. FLAIR obtains the highest Prec. mainly due to its high precision of ECs, but it is howbeit meaningless to KG completion. The denominator of recall is the number of all new (correct) concepts in the input text. Since it is difficult to know all new concepts in the input text except for the costly human assessment, we report the relative recall (R-Recall) to measure the new concept extraction ability of the models. Specifically, the new concepts extracted by all models are regarded as the overall NCs. Then, the relative recall is calculated as NC # divided by the number of overall NCs. Accordingly, the relative F1 (R-F1) can also be calculated with Prec. and R-Recall. The results of Table 3 show that MRC-CE can gain satisfactory precision and recall simultaneously.

Ablation Study. We further display the results of ablation study on the Chinese dataset, to investigate the effectiveness of each module in MRC-CE. At first, we took the BERT-MRC with fixed threshold truncation (FTT) as one ablated variant of our framework [13], denoted as BERT-MRC+FTT. In this variant, we simply chose the spans output by BERT-MRC that have the confidence score higher than 0.8, as the extracted concepts. Furthermore, we respectively appended the rest two modules to BERT-MRC, to propose another two ablated variants, denoted as BERT-MRC+RF and BERT-MRC+RULE.

The comparison results of all variants and MRC-CE (BERT-MRC+RF+RULE) are listed in Table 4. The results show that Prec. drops 1.39% and 0.64% without RULE (BERT-MRC+RF vs. BERT-MRC+RF+RULE), as well as R-Recall drops 3.23% and 17.58% without RF (BERT-MRC+RULE vs. BERT-MRC+RF+RULE) in CN-DBpedia and Probase, respectively. It proves that

Table 4. CE results of ablation study.

Dataset	Model	EC #	NC #	Prec.	R-Recall	R-F1
CN-DBpedia	BERT-MRC+FTT	**522**	311	87.96%	83.83%	85.84%
	BERT-MRC+RF	519	**323**	90.83%	**87.06%**	88.91%
	BERT-MRC+RULE	**522**	311	91.84%	83.83%	87.65%
	BERT-MRC+RF+RULE	519	**323**	**92.22%**	**87.06%**	**89.57%**
Probase	BERT-MRC+FTT	**646**	502	79.56%	71.21%	75.15%
	BERT-MRC+RF	636	**626**	89.44%	**88.79%**	89.12%
	BERT-MRC+RULE	**646**	502	84.72%	71.21%	77.38%
	BERT-MRC+RF+RULE	636	**626**	**90.08%**	**88.79%**	**89.43%**

Table 5. Comparisons between RoBERTa and BERT in CN-DBpedia.

Encoder	Model	EC#	NC#	EC length	NC length	OC ratio	Prec.	R-Recall.	R-F1.
RoBERTa	+FTT	403	105	2.63	4.97	4.75%	92.87%	25.80%	40.38%
	+RF	404	119		4.92	6.69%	89.71%	29.24%	44.10%
	+RULE	403	105		4.91	4.73%	**95.31%**	25.80%	40.61%
	+RF+RULE	404	119		**5.01**	6.93%	94.92%	29.24%	44.71%
BERT	+FTT	**522**	311	2.63	4.72	27.74%	87.96%	76.41%	81.78%
	+RF	519	**323**		4.75	27.26%	90.83%	**79.36%**	84.71%
	+RULE	**522**	311		4.74	27.83%	91.84%	76.41%	83.42%
	+RF+RULE	519	**323**		4.91	**29.35%**	92.22%	**79.36%**	**85.31%**

MRC-CE can obtain better performance with RF and RULE since they are complementary to each other.

Meanwhile, we investigate different BERT-based encoders' influence on CE performance. To this end, we replaced BERT with RoBERTa [20] in Concept Extractor module. Then, RF and RULE were also adopted to filter out the candidate spans output by RoBERTa-MRC. The comparison results are shown in Table 5, showing that although the variants with RoBERTa gain higher Prec., they are inferior to the indicators with BERT on other metrics. It is possibly due to that, RoBERTa tends to extract the spans the same as the concepts existing in the KG from the texts, and thus it is easy to ignore the new concepts.

Case Study. We further delve into the concepts extracted by MRC-CE and baselines through some case studies. Table 6 lists the correct concepts extracted by the models along with the existing concepts of two entities.[6] It shows that MRC-CE extracts more fine-grained and overlapped concepts from the texts than the baselines.

---

[6] *Prince Station*'s abstract text and CE results were translated from Chinese.

Table 6. The CE results of two specific entities along with their existing concepts in the KG.

Dataset	Entity	Abstract text	Model	Correct extracted concept
CN- DBpedia	PrinceStation	Prince station is the railway station of JR East Japan and Tokyo Metro	Existing concepts	location, station
			XLNet-NER	station
			FLAIR	station
			KVMN	railway station
			XLM-R	railway station
			BERT-BiLSTM-CRF	station
			Hearst Patterns	the railway station of JR East Japan
			MRC-CE	station, railway station, the railway station of JR East Japan
Probase	Franklin Delano Roosevelt	Franklin Delano Roosevelt was an American politician who served as the 32nd president. He became a central figure in world events during the first half of the 20th century. As a dominant leader of his party. He built New Deal Coalition	Existing concepts	figure, president, leader, politician
			XLNet-NER	president,leader figure, politician
			FLAIR	politician
			KVMN	politician
			XLM-R	figure, president, leader, politician
			BERT-BiLSTM-CRF	politician
			Hearst Patterns	American politician
			MRC-CE	figure, politician, leader, president, American politician

Table 7. The statistics of extracted concepts for all entities in CN-DBpedia.

Data type	instanceOf relation #	Unique concept #	NC	Avg. NC # per entity	Avg. concept # per entity	Avg. character # per concept
Original	11,494,627	270,025	–	–	2.04	3.62
Extracted	9,021,805	894,689	7,053,900	3.16	5.20	4.92

5   Applications

We further demonstrate MRC-CE's advantages on the KG completion of some real applications.

5.1   KG Completion

After training MRC-CE, we ran it for all entities in CN-DBpedia to supply substantial new instanceOf relations including unique new concepts. Please note that the instanceOf relations only focus on the relations between the entity and concept. The related statistics of CE results are listed in Table 7. Please note that the extracted concepts counted in the table include some wrong concepts. According to the table, MRC-CE extracts more than 7,053,900 new instanceOf relations (NC #). In addition, the extracted concepts have more characters than

the original concepts existing in CN-DBpeida. The results justify MRC-CE's capability of extracting large-scale multi-granular concepts for KG completion.

5.2   Domain Concept Acquisition

Our MRC-CE have achieved excellent results concept acquisition on the abstract texts of given entities. In order to verify MRC-CE's capability of acquiring concepts for a certain domain, we also employed MRC-CE in the domain of Meituan which is a famous Chinese e-business platform. Specifically, we collected 117,489 Food & Delight entities in Meituan along with their descriptive texts from CN-DBpedia. After running MRC-CE, we got 458,497 new instanceOf relations from the texts, and the CE precision is 78.0% based on human assessments on 100 samples, justifying that MRC-CE can be successfully applied to domain concept acquisition.

6   Conclusion

In this paper, we propose a concept extraction framework MRC-CE to achieve large-scale multi-granular concept completion for existing KGs. MRC-CE is capable of extracting massive multi-granular concepts from entities' descriptive texts. In our framework, a BERT-based MRC model with a pointer network is built to handle the problems of concept overlap. Meanwhile, a random forest and rule-based pruning are also employed to obtain satisfactory concept extraction (CE) precision and recall simultaneously. Our extensive experiments have justified that our MRC-CE not only has excellent CE performance, but also is competent to acquire large-scale concepts for multilingual KGs. Furthermore, MRC-CE makes a great contribution to supply sufficient concepts for the Chinese KG *CN-Dbpedia*.

References

1. Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R.: FLAIR: an easy-to-use framework for state-of-the-art NLP. In: Proceedings of NAACL, pp. 54–59 (2019)
2. Alomari, S., Abdullah, S.: Improving an AI-based algorithm to automatically generate concept maps. Comput. Inf. Sci. **12**(4), 72 (2019)
3. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: a nucleus for a web of open data. In: Aberer, K., et al. (eds.) ASWC/ISWC -2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76298-0_52
4. Bai, H., Xing, F.Z., Cambria, E., Huang, W.B.: Business taxonomy construction using concept-level hierarchical clustering. arXiv preprint arXiv:1906.09694 (2019)
5. Budin, G.: Ontology-driven translation management. In: Knowledge Systems and Translation (2005)
6. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116 (2019)

7. Cui, W., Xiao, Y., Wang, W.: KBQA: an online template based question answering system over freebase. In: Proceedings of IJCAI (2016)

8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

9. Ji, J., Chen, B., Jiang, H.: Fully-connected LSTM–CRF on medical concept extraction. Int. J. Mach. Learn. Cybern. **11**(9), 1971–1979 (2020). https://doi.org/10.1007/s13042-020-01087-6

10. Kingma, J., Ba, J.: Adam: a method for stochastic optimization. In: Proceedings of ICLR (2015)

11. Lample, G., Conneau, A.: Crosslingual language model pretraining. In: Proceedings of NeurIPS (2019)

12. Li, N., Tian, M., Lv, S.: Extracting hierarchical relations between the back-of-the-book index terms. In: Hong, J.-F., Zhang, Y., Liu, P. (eds.) CLSW 2019. LNCS (LNAI), vol. 11831, pp. 433–443. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-38189-9_45

13. Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., Li, J.: A unified MRC framework for named entity recognition. In: Proceedings of ACL (2020)

14. Li, X., et al.: Entity-relation extraction as multi-turn question answering. arXiv preprint arXiv:1905.05529 (2019)

15. Liang, J., Xiao, Y., Wang, H., Zhang, Y., Wang, W.: Probase+: inferring missing links in conceptual taxonomies. IEEE Trans. Knowl. Data Eng. **29**(6), 1281–1295 (2017)

16. Liang, J., Xiao, Y., Wang, H., Zhang, Y., Wang, W.: Probase+: inferring missing links in conceptual taxonomies. IEEE TKDE **29**(6), 1281–1295 (2017)

17. Liang, J., Zhang, Y., Xiao, Y., Wang, H., Wang, W., Zhu, P.: On the transitivity of hypernym-hyponym relations in data-driven lexical taxonomies. In: Proceedings of AAAI, vol. 31 (2017)

18. Liao, J., Sun, F., Gu, J.: Combining concept graph with improved neural networks for Chinese short text classification. In: Wang, X., Lisi, F.A., Xiao, G., Botoeva, E. (eds.) JIST 2019. CCIS, vol. 1157, pp. 205–212. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-3412-6_20

19. Liu, S., Zhang, X., Zhang, S., Wang, H., Zhang, W.: Neural machine reading comprehension: methods and trends. Appl. Sci. **9**(18), 3698 (2019)

20. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

21. Nguyen, A.D., Nguyen, K.H., Ngo, V.V.: Neural sequence labeling for Vietnamese POS tagging and NER. In: Proceedings of IEEE-RIVF, pp. 1–5. IEEE (2019)

22. Nie, Y., Tian, Y., Song, Y., Ao, X., Wan, X.: Improving named entity recognition with attentive ensemble of syntactic information. arXiv preprint arXiv:2010.15466 (2020)

23. Petrucci, G., Rospocher, M., Ghidini, C.: Expressive ontology learning as neural machine translation. JWS **52**, 66–82 (2018)

24. Ponzetto, S.P., Strube, M.: WikiTaxonomy: a large scale knowledge resource. In: Proceedings of ECAI, vol. 178, pp. 751–752. Citeseer (2008)

25. Poria, S., Hussain, A., Cambria, E.: EmoSenticSpace: dense concept-based affective features with common-sense knowledge. In: Multimodal Sentiment Analysis. SC, vol. 8, pp. 85–116. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-95020-4_5

26. Preum, S.M., Shu, S., Alemzadeh, H., Stankovic, J.A.: EMSContExt: EMS protocol-driven concept extraction for cognitive assistance in emergency response. In: Proceedings of AAAI, pp. 13350–13355 (2020)
27. Qiu, J., Chai, Y., Tian, Z., Du, X., Guizani, M.: Automatic concept extraction based on semantic graphs from big data in smart city. IEEE Trans. Comput. Soc. Syst. **7**(1), 225–233 (2019)
28. Roller, S., Kiela, D., Nickel, M.: Hearst patterns revisited: automatic hypernym detection from large text corpora. In: Proceedings of ACL (2018)
29. Ruan, D.R., He, X.Y., Li, D.Y., Gao, K.: Modeling and extracting hyponymy relationships on Chinese electric power field content. In: 2016 8th International Conference on Modelling, Identification and Control (ICMIC), pp. 439–443. IEEE (2016)
30. Sammut, C., Webb, G.I. (eds.): Encyclopedia of Machine Learning and Data Mining. Springer, Boston (2017). https://doi.org/10.1007/978-1-4899-7687-1
31. Seo, M., Kembhavi, A., Farhadi, A., Hajishirzi, H.: Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603 (2016)
32. Sharma, R., Gopalani, D., Meena, Y.: Concept-based approach for research paper recommendation. In: Shankar, B.U., Ghosh, K., Mandal, D.P., Ray, S.S., Zhang, D., Pal, S.K. (eds.) PReMI 2017. LNCS, vol. 10597, pp. 687–692. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69900-4_87
33. Shen, Y., Huang, P.S., Gao, J., Chen, W.: ReasoNet: learning to stop reading in machine comprehension. In: Proceedings of ACM SIGKDD, pp. 1047–1055 (2017)
34. Shvets, A., Wanner, L.: Concept extraction using pointer-generator networks. arXiv preprint arXiv:2008.11295 (2020)
35. Song, Y., Tian, S., Yu, L.: A method for identifying local drug names in Xinjiang based on BERT-BiLSTM-CRF. Autom. Control Comput. Sci. **54**(3), 179–190 (2020). https://doi.org/10.3103/S0146411620030098
36. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: a core of semantic knowledge. In: Proceedings of WWW (2007)
37. Vinyals, O., Fortunato, M., Jaitly, N.: Pointer networks. In: Proceeedings of NIPS, pp. 2692–2700 (2015)
38. Wei, Z., Su, J., Wang, Y., Tian, Y., Chang, Y.: A novel hierarchical binary tagging framework for joint extraction of entities and relations. arXiv preprint arXiv:1909.03227 (2019)
39. Wu, W., Li, H., Wang, H., Zhu, K.Q.: Probase: a probabilistic taxonomy for text understanding. In: Proceedings of ACM SIGMOD, pp. 481–492 (2012)
40. Xu, B., et al.: METIC: multi-instance entity typing from corpus. In: Proceedings of CIKM, pp. 903–912 (2018)
41. Xu, B., et al.: CN-DBpedia: a never-ending Chinese knowledge extraction system. In: Benferhat, S., Tabia, K., Ali, M. (eds.) IEA/AIE 2017. LNCS (LNAI), vol. 10351, pp. 428–438. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-60045-1_44
42. Xu, B., Zhang, Y., Liang, J., Xiao, Y., Hwang, S., Wang, W.: Cross-lingual type inference. In: Navathe, S.B., Wu, W., Shekhar, S., Du, X., Wang, X.S., Xiong, H. (eds.) DASFAA 2016. LNCS, vol. 9642, pp. 447–462. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-32025-0_28
43. Yang, Y., Shen, X., Wang, Y.: BERT-BiLSTM-CRF for Chinese sensitive vocabulary recognition. In: Li, K., Li, W., Wang, H., Liu, Y. (eds.) ISICA 2019. CCIS, vol. 1205, pp. 257–268. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-5577-0_19

44. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: XLNet: generalized autoregressive pretraining for language understanding. In: Proceedings of NIPS, pp. 5753–5763 (2019)
45. Yao, J., Cui, B., Cong, G., Huang, Y.: Evolutionary taxonomy construction from dynamic tag space. WWW **15**(5–6), 581–602 (2012). https://doi.org/10.1007/s11280-011-0150-4
46. Yilahun, H., Abdurahman, K., Imam, S., Hamdulla, A.: Automatic extraction of Uyghur domain concepts based on multi-feature for ontology extension. IET Netw. **9**(4), 200–205 (2020)
47. Zhao, G., Zhang, X.: Domain-specific ontology concept extraction and hierarchy extension. In: Proceedings of NLPIR, pp. 60–64 (2018)