SCIENCE CHINA
Technological Sciences

Special Topic: Natural Language Processing Technology
• Review •

October 2020 Vol. 63 No. 10: 1971–1989
https://doi.org/10.1007/s11431-020-1673-6

A survey on neural relation extraction

LIU Kang[1,2*]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China;
[2] University of Chinese Academy of Sciences, Beijing 100049, China

Received March 9, 2020; accepted June 9, 2020; published online September 15, 2020

Relation extraction is a key task for knowledge graph construction and natural language processing, which aims to extract meaningful relational information between entities from plain texts. With the development of deep learning, many neural relation extraction models were proposed recently. This paper introduces a survey on the task of neural relation extraction, including task description, widely used evaluation datasets, metrics, typical methods, challenges and recent research progresses. We mainly focus on four recent research problems: (1) how to learn the semantic representations from the given sentences for the target relation, (2) how to train a neural relation extraction model based on insufficient labeled instances, (3) how to extract relations across sentences or in a document and (4) how to jointly extract relations and corresponding entities? Finally, we give out our conclusion and future research issues.

knowledge graph, relation extraction, event extraction and information extraction

Citation: Liu K. A survey on neural relation extraction. Sci China Tech Sci, 2020, 63: 1971–1989, https://doi.org/10.1007/s11431-020-1673-6

1 Introduction

In a past decade, along with the prosperous progress of deep learning, knowledge engineering, as another important branch and fundamental infrastructure of artificial intelligence, has played more and more significant roles in many applications, such as natural language processing, information retrieval and recommendation. Through knowledge engineering, researchers or developers try to distill valuable information from explosive data and make them understandable for machine. In this way, artificial intelligence related applications could not be restricted on a shallow or superficial level and could have deep understanding of the data.

Knowledge graph (KG), a recent well known form of knowledge, also called as knowledge base (KB), could provide structured semantic information about the complicated real world and has attracted a widespread attentions. Actually, KG could be regarded as a kind of structured and linked data, which stores the relational facts in a graph structure. In current KGs, the nodes usually denote entities (including entities, concepts, classes, properties etc.) and the edges/links between any two nodes represent their corresponding semantical relations. The basic factual unit in a knowledge graph is a triplet which is composed of two entities (nodes) and their semantic relation (edge). As shown in Figure 1, Chicago and United States are two entities, named as head entity and tail entity, respectively. "belongs_to_Country" is their relation. Based on this graph structure, the explicit or implicit semantics behind the data could be expressed or inferred.

So far, most existing KGs are built in three ways. The first is annotated through experts, like CYC[1)], Wordnet[2)] and Hownet[3)]. And the second kind of KGs like Freebase [1], ConceptNet[4)][2] and Wikidata[5)] are crowd-sourcing annotated resources. The third is extracted from semi-structured

*Corresponding author (email: kliu@nlpr.ia.ac.cn)

1) http://opencyc.org

2) http://wordnet.princeton.edu

3) http://keenage.com

4) http://conceptnet.io

5) https://www.wikidata.org

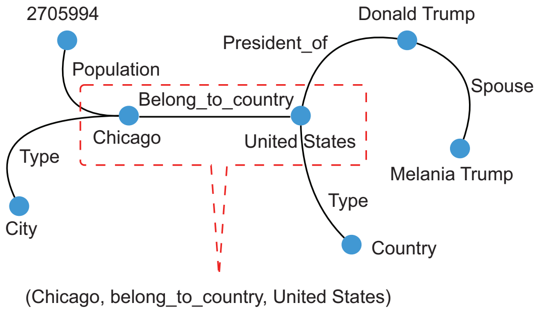


Figure 1 (Color online) An example of factual triplet in knowledge graph.

web-pages, like DBpedia [3], BabelNet[6], Yago [4] and Knowledge Vault [5]. Although the most of these knowledge graphs are huge and contain billions of facts, they still are sparseness from the view of the real applications. Actually, most knowledge scatter widely in the plain texts. Therefore, many researchers start and continue to investigate extracting knowledge from texts, such as NELL [6], KnowItAll [7], TextRunner [8] and Reverb [9]. In the extraction process, several must-do tasks are investigated, like entity recognition, entity disambiguation, relation extraction, event extraction, and relation prediction. This paper mainly investigates extracting relations from texts and the recent research progresses.

However, because of the diversity of the textual expressions, extracting relations from texts is not a easy task, which heavily depends on the ability of the textual understanding. Classical approaches mainly focused on feature engineering which aims to extract effective features for indicating target relations. However, their main deficiency is the problem of "semantic gaps". It means that all features are based on symbolic representation (characters, words, phrases, etc.) which have polysemy and ambiguity problems. With the development of deep learning, deep neural networks have expressed no-doubt advantages in many research fields. Their basic advantage is the usage of distributed representation rather than symbols, where the exact matching between symbols could be replaced by the operations among distributed vectors. In this way, the semantic gap problem could be alleviated in a great extent. Meanwhile, deep learning based models could automatically learn feature representations from the inputing raw data via non-linear activation function in an end2end manner, even including complex and intricate features. As a result, traditional error accumulation and propagation problems in the feature engineering process could be avoided. Such paradigm has been widely adapted for extracting relation and the corresponding researches are proposed explosively. Therefore, the typical neural models for relation extraction are the main target in this paper.

The remains of this paper are organized as follows. The Sect. 2 presents the task description of relation extraction, current used evaluation datasets and metrics. Then we show the main challenging research problems and technical branches. Then the following Sects. 4–7 mainly present the main challenges and corresponding recent research progresses. At last, the conclusion is summarized and the future research issues are presented.

2 Preliminaries of relation extraction

2.1 Task description

Relation extraction from texts is to identify or extract the semantic relations between two given entities in the sentence. According to the input, current relation extraction task could be divided into two categories, including sentence-level, document-level and corpus-level.

Sentence-level relation extraction focuses on identifying entities' relations in a sentence. For example, in Figure 2(a), we need to know that there is a belongs_to_Country relation between Chicago and United States.

Document-level relation extraction or **relation extraction across sentence** focuses on identifying entities' relations in a document or multiple sentences. It means the relational information about an entity pair is expressed in multiple continue or separated sentences, but we do not know where are the aimed sentences.

Corpus-level relation extraction is to identify the relations by given two entities without considering the contexts of the given two entities. As shown in Figure 2(c), system needs to say there is a belongs_to_Country relation between Chicago and United States. It does not consider such relation is expressed in which sentence. Actually, existing researches usually employ the given two entities (e_1 and e_2) to retrieve some evidences, that the texts contain e_1 and e_2. Then their relation will be judged based on those evidences. Here, an important assumption: at-least-one-assumption is given. That is, if there are only one sentence containing the given two entities say there are a relation r between them, we can say that there have r relation between e_1 and e_2 in the knowledge base.

Moreover, according to the definition of the relations, current task could be divided into other two categories, including pre-defined relation extraction and open relation extraction. As the name suggests, in pre-defined relation extraction, all extracted relations are pre-defined or already known. Thus, it could be regarded as a classification problem. In open relation extraction, the relations are unknown. This task is to

[6] http://babelnet.org/stats

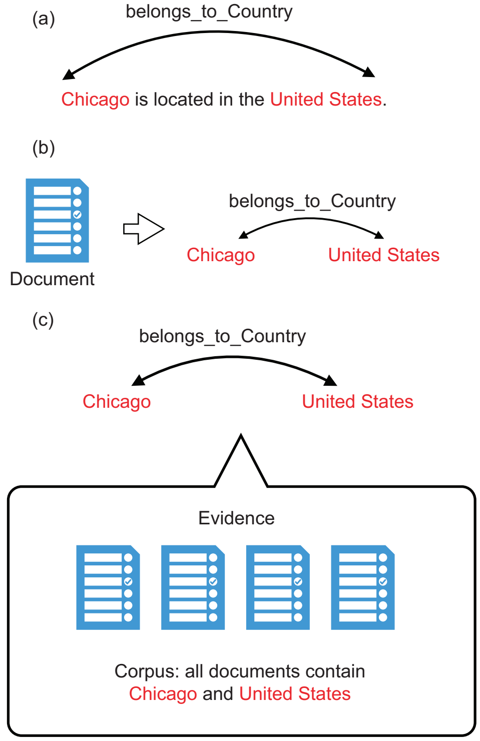


Figure 2 (Color online) Sentence-level relation extraction vs. Corpus-level relation extraction. (a) Sentence-level relation extraction; (b) document-level relation extraction; (c) corpus-level relation extraction.

extract some textual expressions which reflect or indicate some kinds of relations. This paper mainly focuses on the former, i.e., pre-defined relation extraction.

2.2 Evaluation datasets and metrics

Fore relation extraction, there are several open accessed datasets for evaluation, including ACE, SemEval-2010, TAC Relation Extraction, New York Times, and Google-IISc Distant Supervision. We briefly introduce them as follows.

2.2.1 ACE dataset

The objective of the ACE was to develop technology to automatically infer the mentioned entities and their relations. There are five general types of relations in ACE corpus[7] in total, some of which are further sub-divided, yielding a total of 24 types/subtypes of relations, like Founder, Client, Affiliate-Partner, and Citizen-Of. The whole dataset contains the labeled data of English, Arabic and Chinese for evaluation. Recently, the ACE 2004 and ACE 2005 corpora are widely used for document-level relation extraction, that provide entity and relation labels for a collection of documents (348 in ACE 2004 and 511 in ACE 2005).

2.2.2 SemEval 2010 dataset

SemEval-2010 Taks 8 dataset [10] was widely used for sentence-level relation classification. There are nine relation types, including Cause-Effect, Component-Whole, Content-Container, Entity-Destination, Entity-Origin, Instrument-Agency, Member-Collection, Message-Topic and Product-Producer and additional “Other” type. Totally, the whole dataset contains 10717 annotated examples, including 8000 training instances and 2717 test instances. The distribution for training and testing examples are listed in Table 1.

2.2.3 TAC relation extraction dataset (TACRED)

TACRED [11] is a large-scale relation extraction dataset, which consists of 119474 instances with 41 relation types and additional “no relation” label. It is built over newswire and web text from the corpus used in the yearly TAC Knowledge Base Population (TAC KBP) challenges. The dataset uses instances from TAC KBP 2009 to TAC KBP 2012 as the training set, TAC KBP 2013 as the development set, and TAC KBP 2014 as the testing set. Totally, there are 75050 instances for training, 25764 for development and 18660 for testing.

2.2.4 CoNLL 2004 dataset

The CoNLL04 dataset [12] contains sentences from TREC data (news articles) with annotated named entities and relations. It includes 5336 entities with four types (Location, Organization, People, Other) and 19048 relation labels with five types (Work-For, Kill, Organization-Based-In, Live-In, Located-In).

2.2.5 New York Times (NYT) dataset

NYT dataset [13] was widely used for corpus level relation

Table 1 The distributions of training data and testing data in SemEval-2010 dataset

Relation type	Training data	Testing data
Other	17.63%	16.71%
Cause-Effect	12.54%	12.07%
Component-Whole	11.76%	11.48%
Entity-Destination	10.56%	10.75%
Product-Producer	8.96%	9.61%
Entity-Origin	8.95%	9.50%
Member-Collection	8.63%	8.58%
Message-Topic	7.92%	8.50%
Content-Container	6.75%	7.07%
Instrument-Agency	6.30%	5.74%

[7] https://catalog.ldc.upenn.edu/LDC2006T06

extraction. Previous methods mostly employed weakly/distant supervision for this task, where freebase is employed as knowledge graph for distant supervision. In NYT, over 1.8 million articles written and published between January 1, 1987 and June 19, 2007 are used, where 3.2 million relation instances with 430 relation types and 1.8 million entities are included. To align the NYT dataset with freebase entities, Standford named entity recognizer[8)] is used to tag the texts, and the sentences containing the entity pairs in the freebase will be figured out. In details, the NYT subset in the years 2005–2006 is selected as the training set. The subset in the year 2007 is used as the testing set.

2.2.6 Google-IISc distant supervision (GIDS) dataset

In distantly supervised relation extraction, the dataset automatically generated through distant supervision usually contains many noisy labels that means that not all returned sentences correctly express the target relation. To address this problem, Jat et al. [14] build a Google-IISc Distant Supervision (GIDS) dataset, which guarantees that at least one sentence in a bag expresses the target relation. The whole dataset consists of five relation types, including perGraduatedFromInstitution, perHasDegree, perPlaceOfBirth, perPlaceOfDeath and NA. In total, it contains 10832 instances (i.e., entity-pair bags), where 6498 bags are used as the training set, 1082 and 3247 bags are used as the developing and testing set, respectively.

2.2.7 WebNLG dataset

WebNLG dataset was originally created for natural language generation (NLG) task. This dataset totally contains 246 valid relations. In this dataset, a instance including a group of triplets and several standard sentences (written by human). For relation extraction, where the original triplets are used as the annotations for the generated sentences. As a result, it could be used for multiple factual triplets (relations and entities) extraction task [15–17]. The used WebNLG dataset totally contains 5019 sentences as the training set, 500 sentences as the developing set and 703 sentences as the testing set, respectively.

2.2.8 Few-shot relation classification (FewRel) dataset

FewRel dataset was proposed by Han et al. [18]. This dataset is specifically designed for extracting long-tail relations in a knowledge graph, called as few-shot relation classification. It contains 70000 sentences on 100 relations selected from Wikipedia. The dataset was annotated through crowdsourcing. In details, the relation information are firstly labeled through distant supervision, and then are checked by crowd-workers.

2.2.9 Evaluation metrics

Most methods used Precision (P), Recall (R) and F1 score (macro-averaged F1-scores) as the evaluation metrics, which compare the system outputs with the manual labeled golden standard. Note that, for automatic evaluation in corpus-level relation extraction task, existing methods usually used Held-out evaluation which compares the predicted relation of the entity pair with those in existing knowledge graphs, such as Freebase. However, such evaluation has a problem that the extracted correct relations not occurring in the testing set will be regarded as the errors, named as false negatives. To remedy this problem, manual evaluation is often used. With the help of human annotators, existing methods usually calculate the precision of the top N extracted relation instances. Table 2 shows so far state-of-the-art F1 scores for all aforementioned evaluation datasets in relation extraction.

3 Recent research challenges and taxonomy of approaches

This paper mainly investigates the typical approaches on relation extraction, especially for pre-defined relation extraction that could be modeled as a classification problem. Similar with other natural language processing tasks, the statistical learning models were widely employed. The recent investigated research issues are as follows.

(1) How to learn the semantic representations from the given texts for the target relation? For relation extraction, early approaches designed multiple different features based on natural language processing results, such as lexical, syntactic and kernel-based features [26–30]. Recently, with the

Table 2 The state-of-the-art F1 for each dataset (some statics come from https://paperswithcode.com/task/relation-extraction. F1 score in ACE 2004 dataset is for relation and entity extraction. Others are F1 scores for relation extraction)

Dataset	F1 (%)	Methods
ACE 2004	59.7	DYGIE [19]
ACE 2005	63.2	DYGIE [19]
SemEval 2010	90.2	EPGNN [20]
TACRED	71.5	BERTEM+MTB [21]
CoNLL 2004	71.47	SpERT [22]
NYT	89.5	HBT [23]
GIDS	81.8	RELE [24]
WebNLG	88.8	HBT [23]
FewRel	88.32	ERNIE [25]

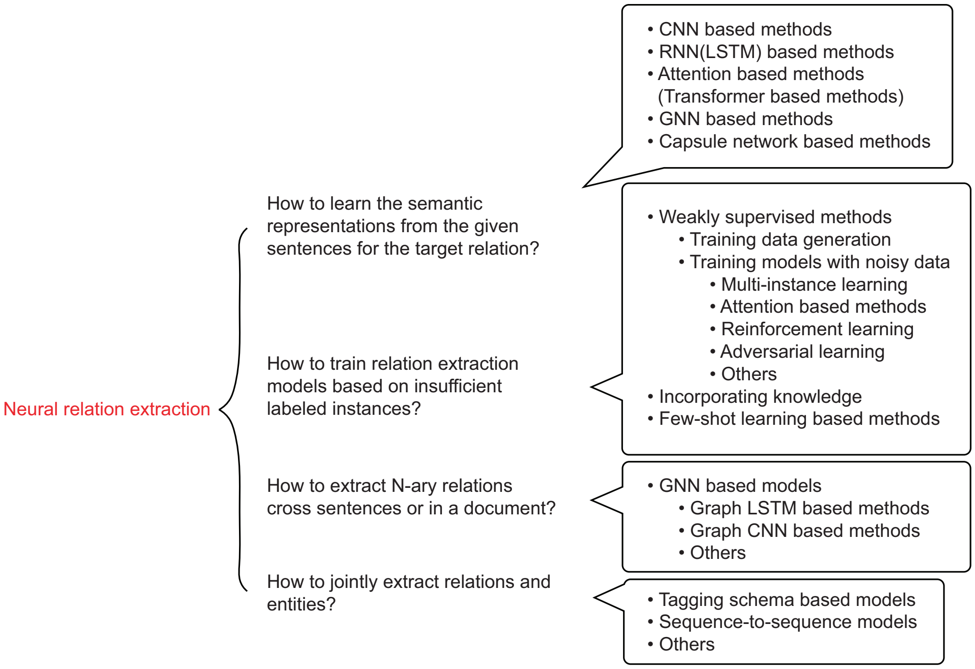
[8)] https://nlp.stanford.edu/software/CRF-NER.shtml development of deep learning, researchers have employed various neural networks to learn the better textual semantic representations. The typical neural networks are convolutional neural network (CNN), recurrent neural networks, recursive neural network, etc. Moreover, attention mechanism has been proven to be effective for semantic embeddings beyond the words. Then various attention based variations are proposed. Furthermore, to learn more precise semantic representation for the target relations, recent work start to employ the syntactic structures and external knowledge. Consequently, graph based neural network and various strategies are employed.

(2) How to train relation extraction models based on insufficient labeled instances? The other challenge for machine learning based approaches is the scale of the training data. Because of the big scales of the relation types in current knowledge base (for example, DBpedia has more than 48000 relations), human annotation for all relations is very cost and not practical. Therefore, researchers started to focus on training data acquisition through weak/distant supervision and effective learning models based on small data. In weakly supervised relation extraction, only a knowledge base with some triplets are provided. Therefore, most approaches employ those provided triplets and some heuristic rules to automatically re-label training instances in the texts. Such process is also named as distant supervision. Of course, the generated training set is noisy, which is the reason that such methods are weakly supervised. Thus, more efforts have been proposed for avoiding noisy data in the training process. In the scenario of learning model based on small data, the current work try the solutions by incorporating external knowledge and regarding this problem as a few-shot learning problem, where several networks were proposed.

(3) How to extract N-ary relations cross sentences or in a document? Previous approaches usually considered extracting relational information from a given sentence, whatever the task is sentence level or corpus level extraction. Their basic setting is that the given entities occur in a sentence and the target relation belong to a binary type. Actually, there is n-ary relation type and the corresponding entities usually scatter across different sentences in a document. Therefore, how to extract such N-ary relations between entities in document level is a challenging problem, which needs to consider the textual expressions for the target relation in a wider contexts, even for considering sentences' relations.

(4) How to jointly extract relations and entities? Relation extraction is a subtask of knowledge extraction that still needs to extract entities in the given texts besides the relations. Previous methods often regarded them as two separated tasks and may adopt a pipeline extraction strategy which may suffer from error propagation. Could extracting entities and their relations boost each other and are extracted jointly? In general, these two extractions are regraded as a multi-task learning problem, where several joint models have been proposed to fulfill this aim.

To give a clearer description of current relation extraction categories, Figure 3 shows the taxonomy of introduced methods/challenges in this paper for neural relation extraction. In



Neural relation extraction
How to learn the semantic representations from the given sentences for the target relation?
How to train relation extraction models based on insufficient labeled instances?
How to extract N-ary relations cross sentences or in a document?
How to jointly extract relations and entities?
• CNN based methods
• RNN(LSTM) based methods
• Attention based methods (Transformer based methods)
• GNN based methods
• Capsule network based methods
• Weakly supervised methods
• Training data generation
• Training models with noisy data
• Multi-instance learning
• Attention based methods
• Reinforcement learning
• Adversarial learning
• Others
• Incorporating knowledge
• Few-shot learning based methods
• GNN based models
• Graph LSTM based methods
• Graph CNN based methods
• Others
• Tagging schema based models
• Sequence-to-sequence models
• Others

Figure 3 (Color online) Challenging research problems and taxonomy of approaches for neural relation extraction.

the remain parts, we introduce them in details.

4 Neural networks for relation extraction

Existing common neural networks based models for relation extraction include convolutional neural networks (CNN) and long short term memory networks (LSTM). Recently, to learn the semantic representations more precisely, some advanced neural models have been introduced, like attention based methods, and graph neural networks (GNN).

4.1 Convolutional neural network based methods

Zeng et al. [31] proposed an earlier work that applied a CNN for relation extraction. Given a sentence, they employ a CNN to learn a sentence-level semantic representation. The framework of CNN part for learning the sentence-level semantic representation is shown in Figure 4.

The first step is word semantic representation, where each input word token is transformed into a dense vector with a pre-defined dimension. The word vector contains two parts: word embedding (WF) and position embedding (PF). Word embeddings are low-dimensional vectors of tokens and every token corresponds to a word embedding, where word embeddings could be pre-trained from the large unlabeled texts. Commonly used word embedding models include Google Word2Vec [32], Stanford Glove [33], and Facebook fastText [34]. In addition, position embeddings are low-dimensional vectors of relative positions of the words in a sentence. The relative position is the distance between the token and the given entities. As shown in Figure 5, the relative position from token "co-founder" to entity "Steve Jobs" and "Apple" is 3 and -4, respectively. Every relative position also corresponds a dense vector v_p which is initialized randomly in the start of the training process. The vector representation v is concatenated by word embedding v_w and position embedding v_p as shown in Figure 4. That is, $d_v = d_w + 2d_p$, where d_v, d_w and d_p are the dimension of $\mathbf{v}$, $\mathbf{v}_w$ and $\mathbf{v}_p$, respectively.

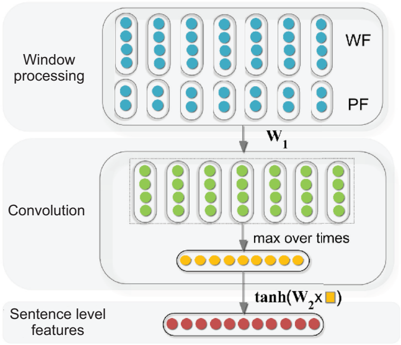


Figure 4 (Color online) The framework used for extracting sentence level features in ref. [31].

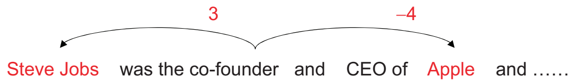


Figure 5 (Color online) Relative positions from "co-founder" to "Steve Jobs" and "Apple".

The second step is a convolution operation. The convolution of $\mathbf{A} \in \mathbf{R}^{m\times n}$ and $\mathbf{B} \in \mathbf{R}^{m\times n}$ is defined as

$$\mathbf{A} \otimes \mathbf{B} = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} b_{ij}, \tag{1}$$

where a_{ij} and b_{ij} are the elements of $\mathbf{A}$ and $\mathbf{B}$ in i-th row and j-th column, respectively. Let v_j^i denotes the vector of j-th token in sentence s_i and let $\mathbf{S}_i$ represents the matrix concatenated by $[v_1^i; v_2^i; \ldots; v_{|s_i|}^i]$, where $|s_i|$ is the number of tokens contained by s_i. Given $\mathbf{S}_i$, we use a filter $\mathbf{W}_q \in \mathbf{R}^{w\times d_v}$ to extract local features from s_i. By sliding $\mathbf{W}_q$ along the sentence s_i, we could obtain a feature map $\mathbf{c}_{iq} \in \mathbf{R}^{|s_i|-w+1}$:

$$c_j^{iq} = f([\mathbf{v}_j^i; \mathbf{v}_{j+1}^i; \ldots; \mathbf{v}_{j+w-1}^i] \otimes \mathbf{W}_q + b), \tag{2}$$

where $b \in \mathbf{R}$ is a bias and $f(\cdot)$ is non-linear function such as Tanh and ReLU. For sentence s_i, we apply three filters $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3$ and get three feature maps $\mathbf{c}_{i1}, \mathbf{c}_{i2}$ and $\mathbf{c}_{i3}$.

The third step is a pooling operation. Zeng et al. [31] selected a max-pooling operation, which is to find the most useful feature in each feature map. The max-pooling of vector $\mathbf{c}_{iq} = [c_1^{iq}, c_2^{iq}, \ldots, c_n^{iq}]$ is defined as: $\max(c_1^{iq}, c_2^{iq}, \ldots, c_n^{iq})$.

The last step is a classify/output layer. Assume that there are m relations, after we get the representation $\mathbf{rep}_i$ of sentence s_i, a single layer fully connected network is used to output the confidence vector $\mathbf{O}_i \in R^m$. Then the conditional probability of j-th relation is

$$p(\mathrm{rel}_j|\theta, s_i) = \frac{\exp(o_j^i)}{\sum_{k=1}^{m} \exp(o_k^i)}, \tag{3}$$

where o_k^i is the k-th element of $\mathbf{O}_i$.

Moreover, there are several CNN based variations for relation extraction task. Zeng et al. [35] proposed a solution that proposed a piece-wise max pooling operation on the original CNN for relation extraction. They believed that max-pooling is too coarse to capture fine-grained features between two entities [35]. Thus, they applied a piece-wise max pooling for relation extraction. In specific, for each feature map c_i, they split the sentence into three segments $\{c_1,c_2,c_3\}$ based on the position of two given entities and do max-pooling for each segment as follows:

$$p_{ij} = \max(C_{ij}), j = 1, 2, 3. \tag{4}$$

All pooling vectors are concatenated as $P_{i:n}$ and a nonlinear function is applied to obtain the output vector of piecewise max pooling: $g = \tanh(p_{1:n})$. Then, g is feed into a softmax classifier to obtain the probability distribution of each relation class.

dos Santos et al. [36] proposed a CNN ranking based model (CR-CNN) that also exploited a CNN to extract sentence-level features. A new pairwise ranking loss function is proposed to reduce the impact of artificial classes:

$$L = \log(1+\exp(\gamma(m^{+}-s_{\theta}(x)_{y^{+}})))+\log(1+\exp(\gamma(m^{-}-s_{\theta}(x)_{c^{-}}))), \tag{5}$$

where m^+ and m^- are margins and γ is a scaling factor that magnifies the difference between the score and the margin, and helps to penalize more on the prediction errors. The first term in the right side of eq. (5) decreases when the score $s_\theta(x)_{y^+}$ increases, where $s_\theta(x)_{y^+}$ means the score of example x for the relation class label y^+. The second term in the right side decreases when the score $s_\theta(x)_{c^-}$ decreases, where $s_\theta(x)_{c^-}$ means the score of example x for other relation class label rather than y^+.

Wang et al. [37] incorporated an attention mechanism into CNN at two levels for relation extraction. They employed the attention to determine which parts of the sentence are most influential with respect to the two entities of interest. The first level attention is calculated as follows:

$$\begin{aligned}
\alpha_i^j &= \frac{\exp(A^j_{i,i})}{\sum_{i'=1}^{n}\exp(A^j_{i',i'})},\\
\mathbf{r}_i &= \mathbf{z}_i\frac{\alpha_i^1+\alpha_i^2}{2},
\end{aligned} \tag{6}$$

where A^j is a diagonal attention matrix with values $A^j_{i,i} = f(e_j, w_i)$ to characterize the strength of contextual correlations and connections between entity mention e_j and word w_i. α_i^j quantifies the relative degree of relevance of the i-th word with respect to the j-th entity ($j \in \{1, 2\}$). $\mathbf{z}_i$ is the contextual information for the i-th word. After that, Wang et al. [37] applied a convolutional max-pooling with another secondary attention model to extract more abstract features at the phrase level.

Moreover, to effectively employ the syntactic information in the sentence, He et al. [38] proposed a syntax-aware CNN model to learn syntax-aware entity representations for neural relation extraction.

4.2 Recurrent neural network based methods

Similar to convolutional neural network, recurrent neural network (RNN) is also used to learn textual semantic representation for relations. Currently, long short-term memory network (LSTM) is a widely used model, which is designed to address the gradient vanishing and exploding problems in RNN via introducing gate mechanism and memory cell. For a given sentence $x = w_1, w_2, , w_n$, each word w_i has its embedding x_t by getting one column of the pre-trained embedding matrix. The size of the word embedding is d^w. The computation process in a LSTM is

$$\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i),\\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f),\\
g_t &= \tanh(W_{xc}x_t + W_{hc}h_{t-1} + W_{cc}c_{t-1} + b_c),\\
c_t &= i_t g_t + f_t c_{t-1},\\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o),\\
h_t &= o_t\tanh(c_t),
\end{aligned} \tag{7}$$

where i, f and o are the input gate, forget gate and output gate, respectively. c is the cell memory. W and b are trainable weight matrix and bias, respectively. $\sigma(\cdot)$ is a sigmoid function.

Because that standard LSTM learns the semantic representations of the given sequence only in a single direction (forward or backward), which may have the input data flexibility problem. Thus, similar to other NLP tasks, bidirectional long short-term memory networks (BLSTM), that performs forward LSTM and backward LSTM simultaneously, are widely used for relation extraction. Zheng et al. [39] employed a BLSTM to extract sentence level features for relation extraction. They concatenated the outputs of a forward LSTM and a backward LSTM together at each time to form new hidden vectors. Hidden vectors at all times are combined (average, adding, etc.) as the final feature representation, which are then fed into a softmax classifier to predict the semantic relation labels.

Moreover, Xu et al. [40] also applied a LSTM on the syntactic tree of the given sentence to perform feature representation. Their model leveraged the shortest dependency path between two entities, where Standford parser[9] was used to gain the dependency tree of the given sentence. Then, four channels of information (words, part-of-speech tags, grammatical relations and WordNet hypernyms) along the shortest dependency path will be modeled by a distinct LSTM. After that, a max-pooling layer is applied after the LSTM. The pooling results of each channel will be concatenated. Similarly, a softmax classifier will be used to obtain the classify results.

Sorokin et al. [41] proposed a context-aware LSTM model for relation extraction. Their model utilizes a LSTM model to jointly learn representations for all relations in a single

[9] https://nlp.stanford.edu/software/lex-parser.shtml

sentence. In this way, the employed contextual relations are helpful for predicting the target relation.

Yang et al. [42] proposed an ensemble LSTM model for neural relation extraction. They employed an adaptive boosting algorithm to combine multiple LSTMs. With such ensemble learning neural framework, more precise and robust semantic representations could be learned.

4.3 Attention based methods

Attention mechanism is well-known because of its capability of learning the “importance” distribution over the inputs and has been proven to be successful in a wide range, such as machine translation and question answering. Similarly, for relation extraction, Shen et al. [43] and Zhou et al. [44] applied attention mechanism to learn crucial information. In specific, Shen et al. [43] believed that not all words in a sentence contributed equally to the representation of the semantic relation. For example,

> *The **women** that caused the **accident** was in the cell phone and ran through the intersection without pausing on the median.*

In this sentence, the type of relation is Cause-Effect (accident, women). “caused” is very important to predict the relation Cause-Effect, but phone is less useful. Therefore, attention mechanism is exploited to extract such crucial and primary information for relation prediction.

Zhou et al. [44] also proposed an attention based BLSTM for relation extraction. After extracting sentence level features through a BLSTM, the model leveraged the attention mechanism to learn important semantical information as follows:

$$\begin{aligned}
&M = \tanh(H),\\
&\alpha = \mathrm{softmax}(w^T M),\\
&r = H\alpha^{\mathrm{T}},\\
&h^* = \tanh(r),
\end{aligned} \tag{8}$$

where H is the output of BiLSTM. After obtaining the final representation, the model exploits a softmax classifier to predict the label.

Du et al. [45] proposed a multi-level structured self-attention mechanism for distantly supervised relation extraction. The model exploited a 2-D matrix based word level attention, which contained multiple vectors, each focusing on different aspects of the sentence for better context representation learning.

Most of methods handle each relation in isolation, regardless of rich semantic correlations located in relation hierarchies. Han et al. [46] proposed a hierarchical attention scheme to incorporate the hierarchical information of relations for distantly supervised relation extraction. The multiple layers of the hierarchical attention scheme provided coarse-to-fine granularity to better identify valid instances, which was especially effective for extracting those long-tail relations.

Deep transformer based methods. Recently, Transformer, proposed by Vaswani et al. [47], dispensed with recurrence and convolutions entirely. Transformer utilized stacked self-attention and point-wise, fully connected layers to build basic blocks. Based on transformer, Radford et al. [48] proposed generative pre-trained transformer (GPT) for language understanding. Unlike GPT (a left-to-right architecture), Bidirectional Encoder Representations from Transformers (BERT) [49] was recently proposed to pre-train deep bidirectional Transformer by jointly conditioning on both left and right context in all layers. BERT has been proven to be the state-of-the-art for several NLP tasks.

Alt et al. [50] utilized a pre-trained language model, OpenAI generative pre-trained transformer (GPT), to learn the semantic representations for the given sentence. The pre-trained language model can capture semantic and syntactic features, and also a notable amount of “common-sense” knowledge, which is very beneficial to recognizing a more diverse set of relations, especially for those long-tail relations instead of high-frequent relations only. Moreover, Wang et al. [51] used a BERT based model to complete the multiple entity-relation extraction task with only one-pass encoding on the input corpus.

4.4 Graph neural network based methods

Existing CNNs and LSTMs could only model the sequential data. However, there are tree-like or graph-like structures in the sentences, such as dependency tree and syntactic tree. Those syntactic trees could convey more rich structural information that are useful for extracting relations among entities in texts. Actually, whatever dependency trees or syntactic trees are, they could be regarded as a kind of graph. Thus, graph neural networks (GNN) are naturally exploited to model this kind of complex data structure. In specific, graph convolutional network (GCN) [52] is most representative model of GNN, which defines convolution and readout operations on irregular graph structures to capture common local and global structural patterns. A multi-layer GCN works with the following layer-wise propagation rule:

$$H^{l+1} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^l W^l), \tag{9}$$

where $\tilde{A} = A + I$. A is the adjacency matrix of graph and I is the identity matrix. $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ and W^l is a layer-specific trainable weight matrix. $\sigma(\cdot)$ denotes an activation function.

Zhang et al. [53] proposed an extension of graph convolutional networks for relation extraction. Their model could pool information over arbitrary dependency structures efficiently in parallel. To incorporate relevant information while maximally removing irrelevant contents, they further applied a pruning strategy to the input trees by keeping words immediately around the shortest path between the two entities.

Guo et al. [54] proposed an attention guided graph neural network for relation extraction. Based on the dependency tree, they built a graph and exploited an attention based GNN to selectively attend to the relevant and useful sub-structures for indicating relational semantics.

Zhu et al. [55] proposed a graph neural network with generated parameters for relation extraction. The model took natural language sentences as inputs to generate parameters, which aimed to adapt to the natural language relational reasoning problem.

Song et al. [56] utilized dependency forests for deep semantic features in the medical domain. In their method, a graph neural network was used to represent the forests and automatically distinguished the useful syntactic information from parsing noises.

Christopoulou et al. [57] proposed an edge-oriented graph neural model for relation extraction. The model utilized different types of nodes and edges to create a graph. An inference mechanism on the graph edges enabled to consider different relation paths between two entities.

Moreover, there are more work employing GNN for relation extraction, such as N-ary relation extraction across sentences [58], jointly extracting entities and relation [59], incorporating external knowledge [60]. We will present them in the following parts.

4.5 Capsule network based methods

Capsule network [61] is designed for better representations than CNNs and RNNs, where it uses vector-output capsules instead of feature detectors in CNNs and introduce an iterative routing process to decide the credit attribution between nodes from lower and higher layers.

Zhang et al. [62] proposed a novel neural approach based on capsule networks with attention mechanisms for relation extraction. Zhang et al. [63] also proposed a capsule network for multi-labeled relation extraction. To better cluster the used features and improve the extraction performance, they proposed an attention-based routing algorithm and a sliding-margin loss function.

5 Relation extraction with insufficient training data

As mentioned in the Sect. 2, in many knowledge extraction scenarios, the scale of the target relation types is huge, where many long-tail relation types are included. As a result, the training data are usually insufficient for training precise extraction models. Generally, existing approaches addressed this problem in several ways. (1) Generating training data automatically through distant supervision that also are named as weakly supervised relation extraction. (2) Employing external knowledge as the supplement. (3) Exploiting effective learning models, like few-shot or adversarial learning, which could train an effective model by leveraging a small dataset.

5.1 Weakly supervised learning based methods

This kind of methods try to generate labeled data automatically as the training set. In the setting of weakly supervised relation extraction, only a knowledge base with some triplets are provided. Therefore, most approaches employ those provided triplets and some heuristic rules to automatically relabel training instances in the texts. Such process is also named as distant supervision. Of course, the generated training set contains noises, which is the reason that such methods are weakly supervised. A sentence that mentions two entities does not necessarily express their relation in a knowledge base. It is possible that these two entities may simply share the same topic. How to alleviate the influence of such noisy data have draw much attentions.

5.1.1 Training data generation

Mintz et al. [64] proposed distant supervision paradigm to automatically align knowledge bases with free texts. They assume that if two entities have a relation in a knowledge base, then all sentences that mention these two entities will express such relation. Specifically, for a triplet $r(e_1, e_2)$ in a knowledge graph, all sentences that mention both entities e_1 and e_2 are aligned with relation r. We call the set containing an entity pair with sentences mentioned them as a bag. We show this process in Figure 6. In this example, Steve Jobs and Apple are two related entities in a knowledge graph. All sentences that contain these two entities are selected as training instances.

5.1.2 Training based on noisy data

Actually, the labeled sentences through distant supervision contain noises, i.e., the selected sentences do not express the target relation, also named as false positive examples. For

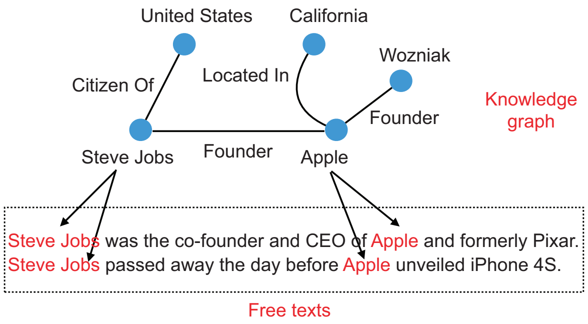

United States
California
Wozniak
Citizen Of
Located In
Founder
Knowledge graph
Steve Jobs
Founder
Apple
Steve Jobs was the co-founder and CEO of Apple and formerly Pixar.
Steve Jobs passed away the day before Apple unveiled iPhone 4S.
Free texts

Figure 6 (Color online) Distant Supervision: aligning relational triplets in knowledge base with free texts to automatically generate labeled training dataset.

example, in Figure 6, the second sentence "Steve Jobs passed away the day before Apple unveiled iPhone 4S" does not express Founder relation although Steve Jobs and Apple cooccur. As a result, such labeled sentences are regarded as a noise for training extraction models. To alleviate the noises from distant supervision in the training process, existing methods usually select the following strategies, include multi-instance learning, attention strategy, reinforcement learning, adversarial learning, etc.

Multi-instance learning All multi-instance learning based relation extraction methods follow an important assumption: at-least-one-assumption. It assumes that at least one sentence in a bag expresses the relation. Zeng et al. [35] proposed an extended Piece-wise CNN (PCNN) approach for distantly supervised relation extraction. They employ a PCNN model to learn the semantic representations of the given sentences. To avoid noises in the training process, similar to ref. [64], they [35] employed multi-instance learning for PCNNs. In the training process, PCNN model selects the most probable valid sentence to represent an entity pair bag for training, while the remaining sentences in the bag were ignored. In this way, the noisy data will be filtered out. In specific, suppose we have T bags and m_i^j represents the j-th instance of i-th bag. The object function is defined as

$$\begin{aligned} J(\theta) &= \sum_{i=1}^{T} \log p(y_i | m_i^{j^*}; \theta), \\ j^* &= \arg\max_j p(y_i | m_i^j; \theta). \end{aligned} \tag{10}$$

Attention based methods Only selecting the most possible sentence (only one instance in each bag) for each entity pair in training will lose a large amount of information. To consider this problem, Lin et al. [65] proposed to use attention mechanism to automatically learn weights for different instances in a bag. They believed that the true positive examples would obtain more weights. By the contrast, the noises would obtain less weights.

In details, there are two main parts in their model: Sentence Encoder and Selective Attention over Instances. Sentence Encoder is used to obtain the distributed representation of a sentence which is a CNN/PCNN as similar to ref. [35]. The bag is denoted as $S = \{x_1, x_2, ..., x_n\}$, i-th sentence representations as $\mathbf{x}_i$ and bag representation as $\mathbf{s}$. Then the representation of the bag $\mathbf{s}$ can be calculated as the weighted sum of all its sentences:

$$\mathbf{s} = \sum_i \alpha_i \mathbf{x}_i, \tag{11}$$

where α_i is the weight of each sentence. Then, in Selective Attention over Instances, Lin et al. [65] applied attention mechanism to calculate those weights automatically:

$$\begin{aligned} \alpha_i &= \frac{\exp(e_i)}{\sum_k \exp(e_k)}, \\ e_i &= \mathbf{x}_i \mathbf{A} \mathbf{r}, \end{aligned} \tag{12}$$

where $\mathbf{A}$ is a weighted diagonal matrix and $\mathbf{r}$ is the query vector associated with relation r.

Finally, the conditional probability $p(r|S, \theta)$ can be defined as

$$\begin{aligned} p(r|S, \theta) &= \frac{\exp(o_r)}{\sum_{k=1}^{n_r} \exp(o_k)}, \\ \mathbf{o} &= \mathbf{M}\mathbf{s} + \mathbf{d}, \end{aligned} \tag{13}$$

where $\mathbf{d}$ is bias vector and $\mathbf{M}$ is the representation matrix of relations.

Ji et al. [66] also proposed a sentence-level attention based convolutional neural network model to exploit all information of sentences in bag for weakly supervised relation extraction. Motivated by TransE [67], the model exploits the $E(e_2) - E(e_1)$ to denote the embedding of the relation $E(r)$, which is used to compute the weight of each sentence in a bag, where $E(e_2)$ and $E(e_1$ are embeddings of the head entity and tail entity, respectively. In addition, existing knowledge graphs contain entity descriptions. They extracted entity descriptions from Freebase and Wikipedia to learn the entity representations in the attention module. Moreover, Feng et al. [68] proposed two memory based models, including word-level and relation-level memory network, to improve the attention weights computation in the training process.

Reinforcement learning Here, researchers used a hard-selection strategy to filter out the noises in the training set. They believed that it is not a good choice to deal wrong labeled sentences with soft attention weights and those sentences should be treated with a hard decision. Thus, they usually employed reinforcement learning to make selection.

Feng et al. [69] proposed a reinforcement learning based model to this task. Their model contained two key modules: instance selector and relation classifier. These two modules will interacts with each other during training process as shown in Figure 7. Given a bag, the instance selector selected valid sentences from the bag. Then, the selected sentences were feed to the relation classifier to obtain the reward. Finally, both the selector and classifier were trained with policy gradient algorithm.

The instance selector selects sentences one by one through reinforcement learning with action "select". The state s_i of the reinforcement learning process includes the current sentence, the selected sentences and the entity pair. Sentence representation is obtained through a CNN. The selected sentences are represented by the average of their representation. The vector representation of s_i is denoted as $\mathbf{F}(s_i)$.

The reinforcement learning agent have two actions $\{0, 1\}$ for "select", which indicates whether to select the current sentence or not. The action a_i in state s_i is sampled from the policy function $\pi_\theta(s_i, a_i)$, where θ is the parameters of the policy function. A logistic function is used as the policy function and $\theta = \{\mathbf{W}, \mathbf{b}\}$:

$$\pi_\theta(s_i, a_i) = a_i\sigma(\mathbf{W}\times\mathbf{F}(s_i)+b)+(1-a_i)(1-\sigma(\mathbf{W}\times\mathbf{F}(s_i)+b)). \tag{14}$$

The reward is the director of training, which indicates the utility of the chosen sentences. Only when the selection is finished, the terminal reward could be determined. Therefore, the reward of states before finishing all the selection are set to 0. The terminal reward is calculate as

$$r(s_i|B) = \frac{1}{\hat{B}}\sum_{x_j\in B}\log p(r|x_j), \tag{15}$$

where $\hat{B}$ is the set of selected sentences and B is the original bag, r is the label of bag B. $p(r|x_j)$ is obtained by the CNN relation classifier.

The CNN model used in relation classifier is similar to ref. [35]. The vector representation of each word including its word embedding and position embedding. Then, a convolution layer and a pooling layer are applied to obtain the sentence representation.

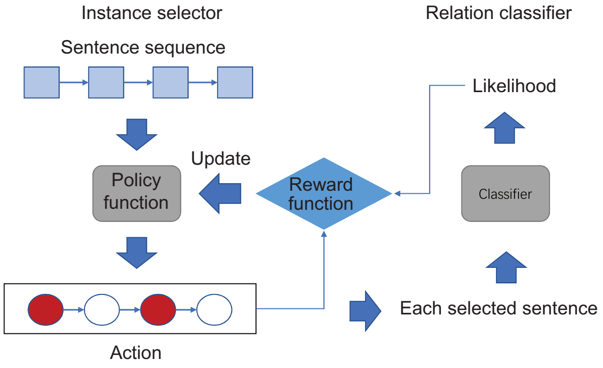


Figure 7 (Color online) Instance selector and relation classifier.

Qin et al. [70] also proposed a deep reinforcement learning model for distantly supervised relation extraction. The work is very similar to Feng et al. [69], they both adopt reinforcement learning to learn an instance selector. The reward of ref. [69] is calculated via the prediction probabilities. By contrast, the reward of Qin et al. [70] is intuitively reflected by the performance change of the relation classifier.

Moreover, most of aforementioned methods mainly focused on solving the false positive examples but overlooked false negative ones. However, a lot of false negative instances express similar semantic information with positive data and provide evidence for the target relation. To this problem, Yang et al. [71] firstly exploited a discriminator to split the noisy data into correctly labeled data and incorrectly labeled data with reinforcement learning. Then the model regarded the incorrectly labeled data as unlabeled data and adopted a semi-supervised learning way to leverage the data.

Different from removing the wrong labeled instances, Zeng et al. [72] applied reinforcement learning to predict the bag relation for distantly supervised relation extraction. The model integrated the predicted relations of sentences to predict the relation of the bag via the following rules. When predicting the relation of a bag, the bag is NA relation iff all sentences in bag represents NA relation. Otherwise, the bag is the real relation represent by its sentences. The model compared the predicted bag relation with the gold bag relation to determine the long term reward.

Adversarial learning Besides reinforcement learning, adversarial training has also been applied to distantly supervised relation extraction. Qin et al. [73] introduced an adversarial learning based method to filter the wrong labeled instances. The model firstly employed a sentence-level true-positive generator to generate positive instances, and regarded the generated positive samples as the negative samples to train a discriminator. When the performance of the discriminator has greatest decline, the generator will be optimized. Then the model leverages the generator to remove the wrong labeled data.

Other weakly supervised learning based methods Most of previous models for distantly supervised relation extraction only focuses on sentence level denoising problem and used the hard-label (i.e., the label of the instance is one-hot and is not changed during training) to train the model. Liu et al. [74] believed that such hard-label methods cannot fully address the wrong labeling problem. They assumed that the correctly labeled instances are dominant in distantly supervised relation extraction. Based on this assumption, they proposed a soft-label method at entity-pair level for distantly supervised relation extraction. The model exploited semantic information from correctly labeled entity pairs to correct wrong labels dynamically during training. They proposed a joint score function to obtain the soft label for a bag by combining the relational scores based on the entity-pair representation and the confidence of the hard label. During training, the obtained soft label served as the target for model training.

For distantly supervised relation extraction, most methods only make use of sentences that contain two target entities, while ignoring those sentences containing only one of the entities. However, these sentences can provide useful information. For example, if we know that "h is the father of e" and "e is the father of t", we can infer that h is the grandfather of t. Based on the observation, Zeng et al. [75] proposed a path-based neural relation extraction model with relation paths. The model firstly exploited a CNN to encode the sentence, and then proposed a relation path encoder to model the inference chains. Finally, the information of direct sentences and inference chains was combined to predict the label.

Similar to ref. [75], Deng et al. [76] also explored additional training data for weakly supervised relation extraction. They referred to the conventional distant supervision strategy as 1-hop distant supervision. Then, they proposed a new strategy to exploit such data and called the strategy as 2-hop distant supervision. They combined the generated data via 1-hop and 2-hop distant supervision to improve the extraction performance.

Beltagy et al. [77] proposed a neural network method to combine direct supervision data for improving the performance of weakly supervised relation extraction. Instead of concatenating instances from direct and distant supervision into one large dataset, they proposed a multi-task learning framework to jointly exploit the two types of supervision. The direct supervision data was employed as supervision for attention weights, which can improve the models ability to identify noisy sentences.

Huang et al. [78] firstly proposed a collaborative curriculum learning framework for weakly supervised relation extraction. In order to learn a better sentence representation, the model firstly exploited a self-attention mechanism which was performed after extracting features via a convolutional layer. To leverage the advantages of maximum probability sentence model [35] and selective sentence attention model [79], the curriculum learning was then leveraged to combine these two models. The two relation extractors were expect to collaboratively alleviate noisy effects and boost each other.

Moreover, noisy bag in distant supervision is an important problem, which means that all sentences in one bag are incorrectly labeled. Ye et al. [80] proposed intra-bag and inter-bag attentions to deal with the noise at sentence-level and bag-level, respectively. For the bag-level noise problem, they proposed a new concept, named as "group", which contains the bags sharing the same relation label. The representation of the group is calculated by weighting bag representations using a similaritybased inter-bag attention module. The bag group is utilized as a training sample to train the proposed model.

Jia et al. [81] proposed an attention regularization based noise reduction framework for distantly supervision relation extraction. The model firstly exploited attention regularization on the neural model to focus on relation patterns. Then, if the learned model can discover patterns for candidate instances, these candidates are selected as correct labeled data for the further training step.

Lin et al. [82] leveraged a semi-supervised learning method to exploit the unlabeled data for relation extraction. They believed that retrieving sentences expressing a relation is a dual task of predicting relation label for a given sentence. The model contained prediction and retrieval modules. The retrieval module retrieved unlabeled sentences from the unlabeled corpus and prediction module annotated the sentences. The two modules are complementary to each other and can be optimized jointly for mutual enhancement.

5.2 Incorporating knowledge into the networks

To resolve the problem of lacking labeled training data and avoid noises in the training set, several work tried to incorporate external knowledge into the network as the remedies. In addition, with the help of knowledge, the semantic representation of the sentences could be learned more precisely,

For the wrong labeled problem in weakly supervised relation extraction, Wang et al. [83] proposed a label-free distant supervision method, which made no use of the relation labels. The model only used the prior knowledge derived from the knowledge graph to supervise the learning process directly and softly. Specifically, the TransE encodes entities and relations of a KG into embeddings with the translation law $h + r \approx t$, where h, r and t are the head entity, relation and tail entity, respectively. The proposed model used the $t - h$ as the supervision signal and make the sentence embedding close to $t - h$. In addition, the model also made full use of the entity type information in the knowledge graph.

Vashishth et al. [60] utilized additional supervision from a knowledge graph through its neural network based architecture. In specific, the model made use of entity type and relation alias information from knowledge graph, to impose soft constraints when predicting the relation. They also employ GCNs to encode syntactic information and improve neural relation extraction.

Long-tail relations are important and cannot be ignored. Zhang et al. [84] took advantage of the knowledge from data-rich classes at the head of the distribution to boost the performance of the data-poor classes at the tail. Firstly, they proposed to leverage implicit relational knowledge among class labels from knowledge graph embeddings and learned explicit relational knowledge using GCNs. Secondly, they integrated such relational knowledge into a neural relation extraction model by coarse-to-fine knowledge-aware attention mechanism.

To exploit the information in existing knowledge graphs, Han et al. [85] proposed joint representation learning framework for knowledge representation and relation extraction. The representations of knowledges graphs and texts are embedded in a sharing semantic space. To achieve better fusion, they proposed a mutual attention between the knowledge graph and texts.

Li et al. [86] proposed a knowledge-attention encoder which incorporated prior knowledge from external lexical resources into neural networks. They also selected three integration strategies to take the advantages of both knowledge-attention and self-attention.

Hu et al. [24] proposed a multi-layer attentionbased model to improve relation extraction with joint label embedding. The model made full use of both structural information from knowledge graphs and textual information from entity descriptions to learn label embeddings. They tried to avoid the imposed noises with attention mechanism. Then the learned label embeddings were used as another attention over the instances (whose embeddings are also enhanced with the entity descriptions) for improving relation extraction.

For Chinese relation extraction, most existing methods typically suffer from segmentation errors and ambiguity of polysemy. To address the issues, Li et al. [87] proposed a multi-grained lattice framework (MG lattice) for Chinese relation extraction to take advantage of multi-grained language information and external linguistic knowledge.

5.3 Few-shot learning for relation extraction

Han et al. [18] formulated relation classification, especially for long-tail relations, as a few-shot learning task. They presented a Few-Shot Relation Classification Dataset (FewRel), consisting of 70,000 sentences on 100 relations derived from Wikipedia and annotated by crowd-workers.

To deal with the noisy label in few-shot relation classification task, Gao et al. [88] proposed hybrid attentionbased prototypical networks. The model embedded all instances in a support set and computed a feature vector (prototype) for each relation, and adopt a hybrid attention consisting of an instance-level attention and a feature-level attention to select more informative instances and highlight important dimensions in the feature space, respectively.

Ye et al. [89, 90] both proposed a multi-level matching and aggregation network for few-shot relation classification. Their models were also based on prototypical network. To model the dependencies between the prototypes of query instance and support set, the model encoded the query instance and each support set in an interactive way by considering their matching information at both local and instance levels.

To address new relations with few-shot instances, Gao et al. [91] proposed a bootstrapping approach to learn new relations by transferring semantic knowledge about existing relations. Given a new relation with few instances, the model found reliable instances from unlabeled datasets. Then these instances were used to train a relation classifier, which can further identify new facts for new relations. Above processes were repeated iteratively.

Soares et al. [21] proposed to learn relation representations directly from texts. They applied the method to few-shot relation classification task and achieved state-of-the-art performance.

6 Jointly extracting relations and entities

Aforementioned approaches mainly focused on extracting relations from the texts when the entities are pre-given. However, from the view of knowledge graph construction, we still need to identify or extract entities from the texts. Most of earlier methods, such as refs. [92, 93], usually adopt a pipeline strategy which first extract entities and their relations step by step. However, such strategy may suffer from error propagation. Therefore, several recent work start to focus on the joint extraction model which assume that the entity extraction and relation extraction are related and could boost each other [94–99]. Refs. [94–96] relied on NLP tools to do feature engineering for jointly extraction. Refs. [97–99] applied neural networks to jointly extract entities and relations. They converted the relation extraction task into a table filling task. However these methods could not handle the multiple triplet extraction problem, i.e., there are more than one triplet in the sentence.

Fu et al. [59] proposed a graph convolutional network to jointly learn named entities and relations, which considered the interaction between named entities and relations via a relation-weighted GCN. Zheng et al. [17] proposed a tagging schema to cover two tasks jointly. Under this schema, jointly extracting relations and entities is regarded as a tagging problem. Figure 8 shows an example of factual triplet extraction based on this tagging schema. Here, “CP” means the President_of_Country relation, and “CF” means the

Founder_of_Country relation. “B”,“E”,“S” denote the begin of an entity, the end of an entity and single words for an entity, respectively. “O” means unrelated words. Moreover, “1” and “2” denote the head and tail entity in a factual triplet, respectively. Based on these tags, we could group them and output the extracted triplets. Finally, we could obtain two triplets, i.e., (United States, President_of_Country, Trump) and (Apple Inc., Founder_of_Company, Steven Jobs). The total number of the defined tags is $N = 2\times4\times|R|+1$, where $|R|$ denotes the number of defined relations.

Then, Zheng et al. [17] employed a Recurrent Neural Networks (RNN) based sequence labeling model to jointly extract entities and relations. In specific, an encoder part of a RNN model is employ to learn the embeddings of the input sentences and an output part plus a CRFs are employed to decode the final tags.

However, a deficiency of aforementioned tagging based method is that it could not deal with the triplet overlapping problem, i.e. two different factual triplets may share one/two same entity/entities. To resolve this issue, Zeng et al. [15] proposed a sequence-to-sequence model with copy mechanism to directly generated the factual triplets from the sentences. In specific, the main components include two parts: encoder and decoder. As shown in Figure 9, an encoder employs a LSTM based model to convert a natural language sentence (the input sentence) into a fixed length semantic vector. Then, a decoder reads this vector and generates triplets directly. To generate a triplet, the decoder generates the relation firstly. For example, leaderName is firstly generated. Secondly, the decoder uses a copy mechanism to identify the corresponding entities in the sentence, such as Aarhus airport and Aarhus are copied in turn. After the first triplet is output (first three units in the decoder), the next three units are used to generate the next triplet, like (Aarhus., cityServed, Jacob Bundsgaard) in Figure 9. Because one entity is allowed to be copied several times when it needs to participate in different triplets, their model could handle the triplet overlap issue.

Moreover, since extracting entities and relations in a single end2end neural network, this model could extract entities and relations jointly. Zeng et al. [16] also extend this model. They employ a reinforcement learning framework for training. As a result, the triplets extraction order could be determined optimally.

7 Document-level relation extraction

All of the aforementioned methods focused on the relation extraction in a sentence, which means that the relation between any two entities is expressed in a singleton sentence. However, in several texts, such relation could be inferred across multiple sentences. According to the statistic from ref. [100], there are at least 40.7% relational facts can only be extracted from multiple sentences. As an example shown in ref. [100], we could extract an factual triplet, i.e., (Riddarhuset, Belong_to_Country, Sweden). This triplet is inferred from two sentences, including the first and last one.

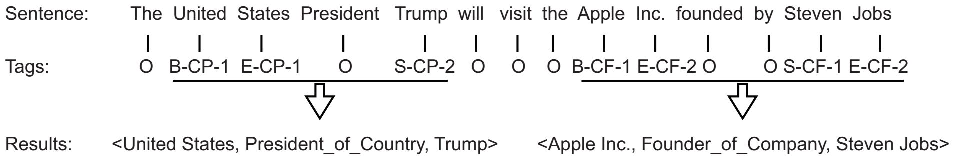

Sentence:
The United States President Trump will visit the Apple Inc. founded by Steven Jobs
Tags:
O B-CP-1 E-CP-1 O S-CP-2 O O O B-CF-1 E-CF-2 O O S-CF-1 E-CF-2
Results:
<United States, President_of_Country, Trump>
<Apple Inc., Founder_of_Company, Steven Jobs>

Figure 8 (Color online) An example of triplet extraction based on tagging schema.

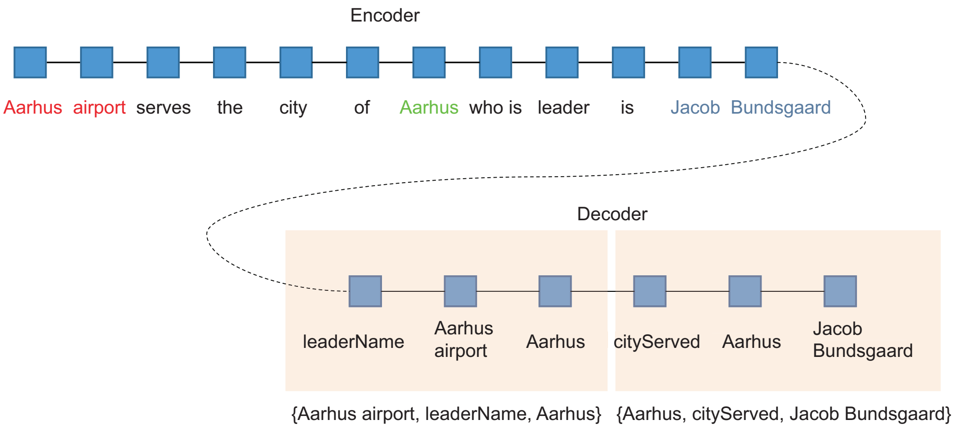

Encoder
Aarhus airport serves the city of Aarhus who is leader is Jacob Bundsgaard
Decoder
leaderName
Aarhus airport
Aarhus
cityServed
Aarhus
Jacob Bundsgaard
{Aarhus airport, leaderName, Aarhus}
{Aarhus, cityServed, Jacob Bundsgaard}

Figure 9 (Color online) An example of triplet extraction based on an end2end neural model.

> *Kungliga Hovkapellet (The Royal Court Orchestra) is a Swedish orchestra, originally part of the Royal Court in **Sweden**'s capital **Stockholm**. ...... From 1731, public concerts were performed at **Riddarhuset** in **Stockholm**.*

Thus, how to extract such relational information is a challenging problem. To this end, Yao et al. [100] proposed a so far biggest large-scale dataset based on Wikipedia and Wikidata. This corpus contains five thousands human annotated documents, and also provides 101,873 distant-supervised documents. They also show that the current state-of-the-arts sentence-level relation extraction neural models could obtain around 50% F1 value, which is a big gap compared with the human's score (88% F1 value). Quirk et al. [101] adopt a document-level graph representation that augments conventional intra-sentential dependencies with new dependencies introduced for adjacent sentences and discourse relations. Their results showed that feature extraction along multiple paths leads to more robust extraction.

Furthermore, besides binary relations, several researchers started to focusing on extracting N-ary relations which are common in some specific domains, such as biomedical domain. In the following example from ref. [58], there is a "Response" relation among three entities, including a mutation **858E**, a gene **EGFR** and a drug **gefitinib**.

> *The deletion mutation on exon-19 of **EGFR** gene was present in 16 patients, while the **858E** point mutation on exon-21 was noted in 10. All patients were treated with **gefitinib** and showed a partial response.*

As mentioned in ref. [58], the key difference between N-ary relation and binary relation extraction is to capture the relational expression in the given texts. Moreover, such N-ary relation is usually expressed across sentences. Thus, ref. [58] proposed a graph LSTM based model to learn a continuous representation for words and entities. Here, graph LSTM was employed to capture the multiple syntactic paths among different entities.

Song et al. [102] extended Peng's method. They proposed a graph state LSTM model for learning semantic representation. Compared with graph LSTM, the used graph state LSTM could use a parallel state to model each word and keep the original graph structure that may be lost in the syntactic graph construction process. Moreover, the computation could be sped up by allowing more parallelization.

Sahu et al. [103] presented a novel inter-sentence relation extraction model that built a labeled edge graph convolutional neural network model on a document-level graph. Inter-sentence relation extraction dealt with a number of complex semantic relationships in documents, which required local, non-local, syntactic and semantic dependencies. The graph was constructed using various inter-sentence and intra-sentence dependencies to capture local and non-local dependency information.

However, similar to training binary relation extraction models, the lack of sufficient positive instances is still a big obstacle for the case of cross-sentence extraction. Akimoto et al. [104] proposed a universal schema model training framework that could employ more dense lower-arity (unary and binary) facts that result from decomposing higher-arity facts instead of only employing original sparse N-ary facts. In this way, the sparsity problem could be alleviated.

8 Conclusion and future work

This paper introduces a survey on the task of neural relation extraction. We mainly focuses on recent four hot research topics and corresponding methods, including recent neural relation extraction models, relation extraction models with insufficient training data, joint models for relation extraction, document-level relation extraction models. We also give the widely used evaluation datasets, metrics, and current performances on each dataset. Nevertheless, not all important researches are included in this paper, such as unknown relation discovery, multiple relation extraction, open information extraction and event extraction. Moreover, the performance of current relation extraction methods still do not satisfy the demand in real applications, especially when the number of the relation types is huge, which make relation extraction is a still challenging task.

More important, with the arise and fast development of pre-trained language models, like BRET [49], ELMo [105], GPT-2 [106], which could conduct state-of-the-arts for many NLP tasks. They usually automatically learn some knowledge from free texts through a unsupervised learning paradigm or naturally labeled data. Thus, we should ask ours an important question:

> *Could pre-trained language models learn relational knowledge from texts and even replace current symbolic knowledge graph?*

Petroni et al. [107] compared BERT and knowledge graphs built by relation extraction on an open domain QA task. They found that BERT captures relational knowledge comparable to that of a symbolic knowledge base extracted by relation extractors, and factual knowledge can be recovered well by pre-trained language models except some relation types (N-to-M relation types). Thus, are BERT and similar language models become a "viable alternative to traditional knowledge bases extracted from text"? By contrast, Pöerner et al. [108] argued that the impressive performance of BERT is partly due to reasoning about (the surface form of) entity names rather than factual knowledge on QA task. When filtering those easy-to-guess facts, the performance drops dramatically. They stated that pre-trained language model could not replace current symbolic KGs. Thus, is there overlapping between pre-trained language modeling and symbolic knowledge bases? Could we really do not need to extract structured knowledge from texts? It is a valuable problem needs more deep researches.

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61922085 and 61533018), the Natural Key R&D Program of China (Grant No. 2018YFC0830101), the Key Research Program of the Chinese Academy of Sciences (Grant No. ZDBS-SSW-JSC006), Beijing Academy of Artificial Intelligence (BAAI2019QN0301), the Open Project of Beijing Key Laboratory of Mental Disorders (2019JSJB06), and the independent research project of National Laboratory of Pattern Recognition.

1 Bollacker K, Evans C, Paritosh P, et al. Freebase: A collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. New York, 2008

2 Speer R, Havasi C. Representing general relational knowledge in ConceptNet 5. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). Istanbul: European Language Resources Association (ELRA), 2012. 3679–3686

3 Bizer C, Lehmann J, Kobilarov G, et al. DBpedia—A crystallization point for the Web of Data. J Web Semantics, 2009, 7: 154–165

4 Suchanek F M, Kasneci G, Weikum G. YAGO: A large ontology from Wikipedia and WordNet. J Web Semantics, 2008, 6: 203–217

5 Dong X, Gabrilovich E, Heitz G, et al. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. London, 2014. 601–610

6 Mitchell T, Cohen W, Hruschka E, et al. Never-ending learning. Commun ACM, 2018, 61: 103–115

7 Etzioni O, Cafarella M J, Downey D, et al. Web-scale information extraction in knowitall. In: Proceedings of the 13th International Conference on World Wide Web. New York, 2004. 100–110

8 Banko M, Cafarella M J, Soderland S, et al. Open information extraction from the web. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence. Hyderabad, 2007. 2670–2676

9 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. Edinburgh, 2011. 1535–1545

10 Hendrickx I, Kim S N, Kozareva Z, et al. 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. Boulder, 2009. 94–99

11 Zhang Y, Zhong V, Chen D, et al. Position-aware attention and supervised data improve slot filling. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017. 35–45

12 Roth D, Yih W t. A linear programming formulation for global inference in natural language tasks. In: Proceedings of the Eighth Conference on Computational Natural Language Learning. Boston, 2004

13 Riedel S, Yao L, McCallum A. Modeling relations and their mentions without labeled text. In: Proceedings of the Learning and Knowledge Discovery in Databases, European Conference. Barcelona, 2010. 148–163

14 Jat S, Khandelwal S, Talukdar P P. Improving distantly supervised relation extraction using word and entity based attention. In: Proceedings of the 6th Workshop on Automated Knowledge Base Construction. Long Beach, 2017

15 Zeng X, Zeng D, He S, et al. Extracting relational facts by an end-to-end neural model with copy mechanism. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018. 506–514

16 Zeng X, He S, Zeng D, et al. Learning the extraction order of multiple relational facts in a sentence with reinforcement learning. 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). Hong Kong: Association for Computational Linguistics, 2019. 367–377

17 Zheng S, Wang F, Bao H, et al. Joint extraction of entities and relations based on a novel tagging scheme. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: Association for Computational Linguistics, 2017. 1227–1236

18 Han X, Zhu H, Yu P, et al. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 4803–4809

19 Luan Y, Wadden D, He L, et al. A general framework for information extraction using dynamic span graphs. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019. 3036–3046

20 Zhao Y, Wan H, Gao J, et al. Improving relation classification by entity pair graph. In: Proceedings of The Eleventh Asian Conference on Machine Learning. Nagoya, 2019. 1156–1171

21 Soares L B, FitzGerald N, Ling J, et al. Matching the blanks: Distributional similarity for relation learning. In: Proceedings of the 57th Conference of the Association for Computational Linguistics. Florence, 2019. 2895–2905

22 Eberts M, Ulges A. Span-based joint entity and relation extraction with transformer pre-training. CoRR, 2019

23 Wei Z, Su J, Wang Y, et al. A novel hierarchical binary tagging framework for joint extraction of entities and relations. CoRR, 2019

24 Hu L, Zhang L, Shi C, et al. Improving distantly-supervised relation extraction with joint label embedding. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 3821–3829

25 Zhang Z, Han X, Liu Z, et al. ERNIE: Enhanced language representation with informative entities. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 1441–1451

26 Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. In: Proceedings of the ACL Interactive Poster and Demonstration Sessions. Barcelona: Association for Computational Linguistics, 2004. 178–181

27 Zhou G, Su J, Zhang J, et al. Exploring various knowledge in relation extraction. In: Proceedings of the Conference on the 43rd Annual Meeting of the Association for Computational Linguistics. University of Michigan, The Association for Computer Linguistics, 2005. 427–434

28 Lodhi H, Saunders C, Shawe-Taylor J, et al. Text classification using string kernels. J Mach Learn Res, 2002. 2: 419–444

29 Collins M, Duffy N. Convolution kernels for natural language. In: Advances in Neural Information Processing Systems 14. Vancouver: MIT Press, 2001. 625–632

30 Bunescu R C, Mooney R J. A shortest path dependency kernel for relation extraction. In: Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. Vancouver: The Association for Computational Linguistics, 2005. 724–731

31 Zeng D, Liu K, Lai S, et al. Relation classification via convolutional deep neural network. In: Proceedings of the Conference on 25th International Conference on Computational Linguistics. Technical Papers. Dublin, 2014. 2335–2344

32 Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, 2013. 3111–3119

33 Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha: Association for Computational Linguistics, 2014. 1532–1543

34 Joulin A, Grave E, Bojanowski P, et al. Fasttext.zip: Compressing text classification models. CoRR, 2016

35 Zeng D, Liu K, Chen Y, et al. Distant supervision for relation extraction via piecewise convolutional neural networks. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: The Association for Computational Linguistics, 2015. 1753–1762

36 dos Santos C N, Xiang B, Zhou B. Classifying relations by ranking with convolutional neural networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. Beijing: The Association for Computer Linguistics, 2015. 626–634

37 Wang L, Cao Z, de Melo G, et al. Relation classification via multi-level attention CNNs. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016. 1298–1307

38 He Z, Chen W, Li Z, et al. Syntax-aware entity representations for neural relation extraction. Artificial Intelligence, 2019, 275: 602–617

39 Zhang S, Zheng D, Hu X, et al. Bidirectional long short-term memory networks for relation classification. In: Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation. Shanghai, 2015

40 Xu Y, Mou L, Li G, et al. Classifying relations via long short term memory networks along shortest dependency paths. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: The Association for Computational Linguistics, 2015. 1785–1794

41 Sorokin D, Gurevych I. Context-aware representations for knowledge base relation extraction. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017. 1784–1789

42 Yang D, Wang S, Li Z. Ensemble neural relation extraction with adaptive boosting. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. Stockholm, 2018. 4532–4538

43 Shen Y, Huang X. Attention-based convolutional neural network for semantic relation extraction. In: Proceedings of the 26th International Conference on Computational Linguistics. Technical Papers. Osaka, 2016. 2526–2536

44 Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: The Association for Computer Linguistics, 2016

45 Du J, Han J, Way A, et al. Multi-level structured self-attentions for distantly supervised relation extraction. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 2216–2225

46 Han X, Yu P, Liu Z, et al. Hierarchical relation extraction with coarse-to-fine grained attention. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 2236–2245

47 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Advances in Neural Information Processing Systems 30. Curran Associates, Inc., 2017. 5998–6008

48 Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. Technical Report. Computer Sciences, OpenAI.com, 2018

49 Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019. 4171–4186

50 Alt C, Hübner M, Hennig L. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 1388–1398

51 Wang H, Tan M, Yu M, et al. Extracting multiple-relations in one-pass with pre-trained transformers. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019 1371–1377

52 Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. arXiv:1609.02907

53 Zhang Y, Qi P, Manning C D. Graph convolution over pruned dependency trees improves relation extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 2205–2215

54 Guo Z, Zhang Y, Lu W. Attention guided graph convolutional networks for relation extraction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 241–251

55 Zhu H, Lin Y, Liu Z, et al. Graph neural networks with generated parameters for relation extraction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 1331–1339

56 Song L, Zhang Y, Gildea D, et al. Leveraging dependency forest for neural medical relation extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 208–218

57 Christopoulou F, Miwa M, Ananiadou S. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 4925–4936

58 Peng N, Poon H, Quirk C, et al. . Trans Association Comput Linguistics, 2017, 5: 101–115

59 Fu T J, Li P H, Ma W Y. GraphRel: Modeling text as relational graphs for joint entity and relation extraction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 1409–1418

60 Vashishth S, Joshi R, Prayaga S S, et al. RESIDE: Improving distantly-supervised neural relation extraction using side information. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 1257–1266

61 Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules.

In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems. Long Beach, 2017. 3856–3866

62 Zhang N, Deng S, Sun Z, et al. Attention-based capsule networks with dynamic routing for relation extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 986–992

63 Zhang X, Li P, Jia W, et al. Multi-labeled relation extraction with attentive capsule network. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, the Thirty-First Innovative Applications of Artificial Intelligence Conference, the Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, Honolulu : AAAI Press, 2019. 7484–7491

64 Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data. In: Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Singapore: The Association for Computer Linguistics, 2009. 1003–1011

65 Lin Y, Liu Z, Sun M. Neural relation extraction with multi-lingual attention. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: Association for Computational Linguistics, 2017. 34–43

66 Ji G, Liu K, He S, et al. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. San Francisco: AAAI Press, 2017. 3060–3066

67 Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data. In: Advances in Neural Information Processing Systems 26. Curran Associates, Inc., 2013. 2787–2795

68 Feng X, Guo J, Qin B, et al. Effective deep memory networks for distant supervised relation extraction. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. Melbourne, 2017. 4002–4008

69 Feng J, Huang M, Zhao L, et al. Reinforcement learning for relation classification from noisy data. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18). New Orleans: AAAI Press, 2018. 5779–5786

70 Qin P, Xu W, Wang W Y. Robust distant supervision relation extraction via deep reinforcement learning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018. 2137–2147

71 Yang K, He L, Dai X Y, et al. Exploiting noisy data in distant supervision relation classification. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019. 3216–3225

72 Zeng X, He S, Liu K, et al. Large scaled relation extraction with reinforcement learning. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18). New Orleans: AAAI Press, 2018. 5658–5665

73 Qin P, Xu W, Wang W Y. DSGAN: Generative adversarial training for distant supervision relation extraction. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018. 496–505

74 Liu T, Wang K, Chang B, et al. A soft-label method for noise-tolerant distantly supervised relation extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017. 1790–1795

75 Zeng W, Lin Y, Liu Z, et al. Incorporating relation paths in neural relation extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017. 1768–1777

76 Deng X, Sun H. Leveraging 2-hop distant supervision from table entity pairs for relation extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 410–420

77 Beltagy I, Lo K, Ammar W. Combining distant and direct supervision for neural relation extraction. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019. 1858–1867

78 Huang Y, Du J. Self-attention enhanced CNNs and collaborative curriculum learning for distantly supervised relation extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 389–398

79 Lin Y, Shen S, Liu Z, et al. Neural relation extraction with selective attention over instances. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016. 2124–2133

80 Ye Z X, Ling Z H. Distant supervision relation extraction with intra-bag and inter-bag attentions. In: Proceedings of theConference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019. 2810–2819

81 Jia W, Dai D, Xiao X, et al. ARNOR: Attention regularization based noise reduction for distant supervision relation classification. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 1399–1408

82 Lin H, Yan J, Qu M, et al. Learning dual retrieval module for semi-supervised relation extraction. In: The World Wide Web Conference. San Francisco, 2019. 1073–1083

83 Wang G, Zhang W, Wang R, et al. Label-free distant supervision for relation extraction via knowledge graph embedding. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 2246–2255

84 Zhang N, Deng S, Sun Z, et al. Long-tail relation extraction via knowledge graph embeddings and graph convolution networks. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019. 3016–3025

85 Han X, Liu Z, Sun M. Neural knowledge acquisition via mutual attention between knowledge graph and text. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18). New Orleans: AAAI Press, 2018. 4832–4839

86 Li P, Mao K, Yang X, et al. Improving relation extraction with knowledge-attention. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 229–239

87 Li Z, Ding N, Liu Z, et al. Chinese relation extraction with multi-grained information and external linguistic knowledge. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 4377–4386

88 Gao T, Han X, Liu Z, et al. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, the Thirty-First Innovative Applications of Artificial Intelligence Conference, the Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. Honolulu: AAAI Press, 2019. 6407–6414

89 Ye Z X, Ling Z H. Multi-level matching and aggregation network for few-shot relation classification. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 2872–2881

90 Ye Q, Liu L, Zhang M, et al. Looking beyond label noise: Shifted label distribution matters in distantly supervised relation extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019. 3839–3848

91 Gao T, Han X, Xie R, et al. Neural snowball for few-shot relation learning. CoRR, 2019

92 Zelenko D, Aone C, Richardella A. Kernel methods for relation extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2002. 71–78

93 Chan Y S, Roth D. Exploiting syntactico-semantic structures for relation extraction. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland: Association for Computational Linguistics, 2011. 551–560

94 Yu X, Lam W. Jointly identifying entities and extracting relations in encyclopedia text via a graphical model approach. In: Coling 2010: Posters. Beijing: Coling 2010 Organizing Committee, 2010. 1399–1407

95 Li Q, Ji H. Incremental joint extraction of entity mentions and relations. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore: Association for Computational Linguistics, 2014. 402–412

96 Miwa M, Bansal M. End-to-end relation extraction using LSTMs on sequences and tree structures. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016. 1105–1116

97 Miwa M, Sasaki Y. Modeling joint entity and relation extraction with table representation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Doha: Association for Computational Linguistics, 2014. 1858–1869

98 Gupta P, Schütze H, Andrassy B. Table filling multi-task recurrent neural network for joint entity and relation extraction. In: Proceedings of the 26th International Conference on Computational Linguistics. Technical Papers. Osaka: The COLING 2016 Organizing Committee, 2016. 2537–2547

99 Zhang M, Zhang Y, Fu G. End-to-end neural relation extraction with global optimization. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017. 1730–1740

100 Yao Y, Ye D, Li P, et al. DocRED: A large-scale document-level relation extraction dataset. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 764–777

101 Quirk C, Poon H. Distant supervision for relation extraction beyond the sentence boundary. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia: Association for Computational Linguistics, 2017. 1171–1182

102 Song L, Zhang Y, Wang Z, et al. N-ary relation extraction using graph-state lstm. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 2226–2235

103 Sahu S K, Christopoulou F, Miwa M, et al. Inter-sentence relation extraction with document-level graph convolutional neural network. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 4309–4316

104 Akimoto K, Hiraoka T, Sadamasa K, et al. Cross-sentence n-ary relation extraction using lower-arity universal schemas. 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). Hong Kong: Association for Computational Linguistics, 2019. 6225–6231

105 Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans: Association for Computational Linguistics, 2018. 2227–2237

106 Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. OpenAI Blog, 2018

107 Petroni F, Rocktäschel T, Riedel S, et al. Language models as knowledge bases? In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 2463–2473

108 Pörner N, Waltinger U, Schütze H. BERT is not a knowledge base (yet): Factual knowledge vs. name-based reasoning in unsupervised QA. CoRR, 2019