Distantly Supervised Relation Extraction using Multi-Layer Revision Network and Confidence-based Multi-Instance Learning

Xiangyu Lin[1], **Tianyi Liu**[2], **Weijia Jia**[2,3*], and **Zhiguo Gong**[1*]
[1]**SKL-IOTSC and Department of Computer and Information Science, University of Macau†**
[2]**Department of Computer Science and Engineering, Shanghai Jiao Tong University**
[3]**BNU-UIC Institute of AI and Future Networks, Beijing Normal University (Zhuhai)**
yc07403@um.edu.mo, liutianyi@sjtu.edu.cn, jiawj@sjtu.edu.cn, fstzgg@um.edu.mo

November 28, 2023

Abstract

Distantly supervised relation extraction is widely used in the construction of knowledge bases due to its high efficiency. However, the automatically obtained instances are of low quality with numerous irrelevant words. In addition, the strong assumption of distant supervision leads to the existence of noisy sentences in the sentence bags. In this paper, we propose a novel Multi-Layer Revision Network (MLRN) which alleviates the effects of wordlevel noise by emphasizing inner-sentence correlations before extracting relevant information within sentences. Then, we devise a balanced and noise-resistant Confidence-based Multi-Instance Learning (CMIL) method to filter out noisy sentences as well as assign proper weights to relevant ones. Extensive experiments on two New York Times (NYT) datasets demonstrate that our approach achieves significant improvements over the baselines.

1 Introduction

Relation Extraction (RE), which aims to classify the relations between a pair of entities in a sentence, is crucial to various applications like questionanswering and construction of knowledge bases. However, supervised relation extraction requires large amounts of manually labeled training data, which is hard to obtain. Therefore, Mintz et al. (2009) proposed Distantly Supervised Relation Extraction (DSRE) to automatically generate training data by aligning the knowledge base with text corpus. However, DSRE is based on the strong assumption that for an entity pair participating in a relation in the knowledge base, all sentences mentioning this entity pair in the corpus express the same relation. This brings a large number of noisy

Entity1: dreamworks   Entity2: steven_spielberg
Relation: /business/company/founders

S1	with mr. eastwood as director and **steven_spielberg** as a producer , ... and ferocious backing from paramount and **dreamworks** .
S2	it is hard to say ... the fund-raiser : the **dreamworks** <u>co-founders</u> david geffen , jeffrey katzenberg and **steven_spielberg** .
S3	the outsize robot adventure movie was born ... with **dreamworks** , paramount and another longtime associate , **steven_spielberg** , among others .

Figure 1: An instance from NYT corpus along with its corresponding entity pair and relation type. Relevant words are <u>underlined</u>.

2 Related Work

Distant supervision (DS) for relation extraction paperone is proposed for efficient knowledge base construction. However, DS brings about the wrong labeling problem as well. Riedel et al. Bollacker et al. 2008 proposes multi-instance learning for DSRE to address this issue. Most of the current work uses two types of MIL strategies: to remove noisy sentences or to apply soft weights. Following the at-least-one assumption, Zeng et al. Bollacker et al. 2008 selects the instance with the highest probability within the bag. Qin et al. Bollacker et al. 2008 and Feng et al. Bollacker et al. 2008 employ reinforcement learning for instance selection. For better information utilization, Lin et al.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 165–174,
November 7–11, 2021. c 2021 Association for Computational Linguistics

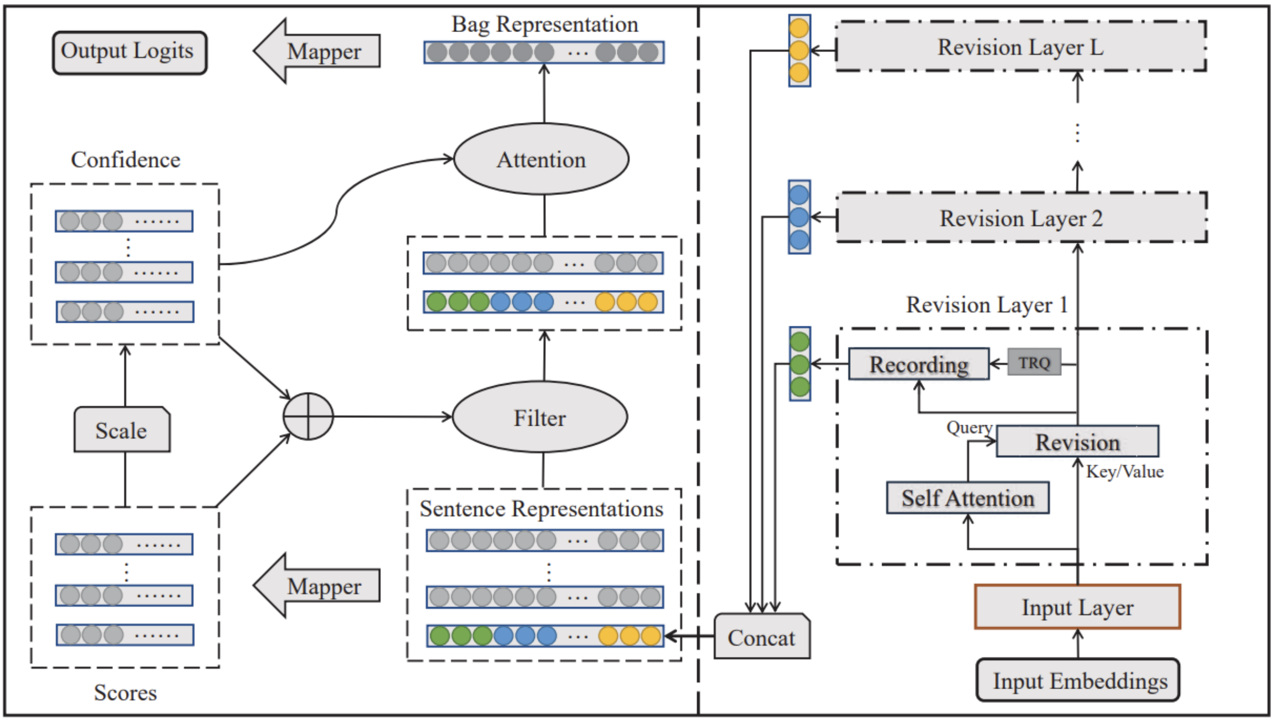

Figure 2: An overview of our MLRN+CMIL model. The revision network used for generating sentence representations is on the right, while confidence-based multi-instance learning is on the left.

3 Methodology

The overall structure of our model is shown in Figure 2. Our model can be divided into three parts: embedding layer, revision network and multiinstance learning layer. In this section, we introduce them respectively. The overall structure of our model is shown in Figure 2. Our model can be divided into three parts: embedding layer, revision network and multiinstance learning layer. In this section, we introduce them respectively

3.1 Embedding Layer

Before being fed into the revision network, the input instances are transformed into distributed representations. The representation of each word token consists of two parts: word embedding and position embeddings.

Word Embeddings are distributed representations for word tokens. Formally, we define the jth word token in the ith sentence as w_{ij}, which is mapped to a d_w-dimensional word vector $v_{ij} \in \mathbb{R}^{d_w}$. Similar to previous studies, we adopt the Skip-Gram method to obtain the pre-trained word embedding matrix. Position Embeddings are distributed representations for the relative distances from each word to the two entities, which are represented as low-dimensional vectors. $pe1_{ij}$, $pe2_{ij} \in \mathbb{R}^{dp}$

Finally, the input embedding x_{ij} is generated by concatenating word embedding v_{ij}, position embeddings $pe1_{ij}$, and $pe2_{ij}$, which is formulated as below:

$$x_{ij} = [v_{ij}; pe1_{ij}; pe2_{ij}] \tag{1}$$

where the dimension of x_{ij} is $d_h = d_w + 2d_p$.

3.2 Revision Network

Formally, the revision network takes a sequence of word representations $X_i = \{x_{i1}, x_{i2}, x_{i3}, ..., x_{il}\}$ with length l as the input and outputs a d-dimensional representation $\text{US}_i \in \mathbb{R}^d$ for the sentence. The revision layer for word-level noise reduction includes two types of attention sub-layers: self-attention layer and query-attention layer. By applying self-attention on the input, the correlations between each pair of tokens are calculated. In order to emphasize the correlations, the attention weights are revised in a query-attention layer before updating the representations. Afterwards, we apply a Translation Query (TRQ) inspired by TransE **bordes2013translating** to extract relevant information as the record for each layer. Finally, these records are concatenated to form the sentence representation used for multi-instance learning. The compositions of revision network will be discussed in detail in this section.

3.2.1 Input Layer

The input layer serves as an encoding layer which calculates feature representations from input embeddings. The input is the embeddings of ith instance, denoted as Xi. For convenience, the subscript i is omitted in the equations of this part. Instead of using CNN or RNN input layers as in most of the previous work, we apply an attention layer to model the long-distance dependencies in the sentence. The attention mechanism used can be formulated as follows:

$$\text{Att}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (2)$$

where Q is the query, K is the key, and V is the value as described in Vaswani et al. (2017). d_k is the dimension of the key and serves as a scaling factor.

In order to explore various semantic spaces of the sentence, we use Multi-Head Self-Attention (MHSA) in the input layer, which is shown as follows:

3.2.2 Revision Layer

3.3 Confidence-based Multi-Instance Learning

3.4 Optimization

4 Experiments

4.1 Datasets and Evaluation Metrics

In order to evaluate the performance of our model, we conduct experiments on widely used NYT-10 dataset (Riedel et al., 2010) and complex NYT-18 dataset (Zhang et al., 2020). NYT-10 is a standard dataset constructed by aligning relation facts in Freebase (Bollacker et al., 2008) with the New York Times corpus, where sentences from 2005 to 2006 are used for training and sentences from 2007 are used as the test set. NYT-18 is a larger dataset containing NYT documents from 2008 to 2017. Both datasets are labeled with Freebase and Stanford Named Entity Recognizer (Finkel et al., 2005). All the sentences are divided into five parts with the same relation distribution for five-fold crossvalidation. The details of the datasets are shown in Table 1

Datasets	Rel.	Training (k)		Testing	
		Ent.	Sen.	Ent.	Sen.
NYT-10	1234	1234	1234	1234	1234
NYT-10	1234	1234	1234	1234	1234

Table 1: The details of datasets. **Rel.**, **Ent.** and **Sen.** indicate numbers of relations, entity pairs and sentences respectively.

4.2 Experimental Settings

4.3 Evaluation on NYT-10

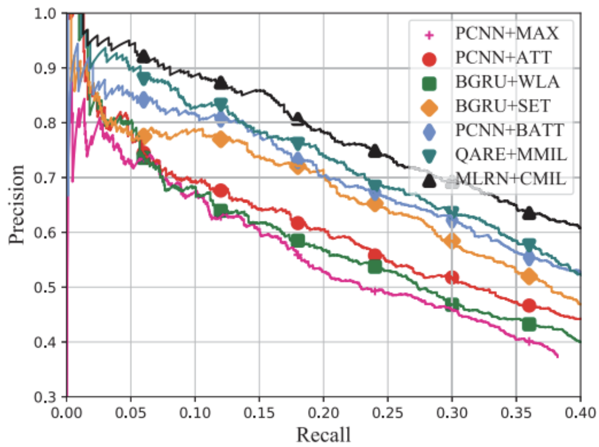

Figure 3: Precision-recall curves of models on NYT10.

4.4 Ablation Study

5 Case Study

6 Conclusion and Future Work

Acknowledgment

This work was supported by National Key D&R Program of China (2019YFB1600704), The Science and Technology Development Fund, Macau SAR (0068/2020/AGJ, 0045/2019/A1, 0007/2018/A1, SKL-IOTSC-2021-2023), GSTIC (201907010013, EF005/FST-GZG/2019/GSTIC), University of Macau (MYRG2018-00129-FST) and GDST (2020B1212030003).

References

Bollacker, Kurt et al. (2008). "Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge". In: *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*. SIGMOD '08. Vancouver, Canada: Association for Computing Machinery, pp. 1247–1250. ISBN: 9781605581026. DOI: 10.1145/1376616.1376746. URL: https://doi.org/10.1145/1376616.1376746.