Improving Distantly Supervised Relation Extraction via Textual Representation of Multi-hop Inference over Text

Qin Dai[1], **Naoya Inoue**[1,2], **and Kentaro Inui**[1,2]
[1]Tohoku University, Japan
[2]RIKEN Center for Advanced Intelligence Project, Japan
{daiqin, naoya-i, preisert, ryo.t, inui}@ecei.tohoku.ac.jp

Abstract

Distantly Supervised Relation Extraction (DS-RE) is a widely applied approach to collect relational facts from unstructured text, but often accompanies with the problem of fuzzy data. To alleviate this issue, existing approach uses multi-hop reasoning over text to facilitate the relation classification. However, the method only represents the textual relation (e.g., the middle context between an entity pair) as a unit, rather than a textual sequence of words and so cannot utilize the semantic information (e.g., pretrained word embedding) of the textual relation for DS-RE. Therefore, in this work, we propose a new DS-RE model that represents the textual relation as a sequence of words and encodes the sequence with a world-level attention mechanism. Experimental results prove the effectiveness or the proposed model for DS-RE, because the proposed model achieves significant and consistent improvement as compared with baselines, and obtains a competitive results with the state-of-the-art models on the NYT10 dataset.

1 Introduction

Knowledge Base (KB) provides large collections of relations between entities, typically stored as (e_h, r, e_t) triplets, e.g., (*Tokyo*, *capital_of*, *Japan*). KBs, such as Freebase (Bollacker et al., 2008) and DBpedia (Lehmann et al., 2015), are extremely crucial for many Natural Language Processing (NLP) tasks. However, KBs are often highly incomplete (Min et al., 2013) and this would impede their usefulness in real-world applications. To enrich the KBs, it is important to turn unstructured text into an well organized KB, and it belongs to the task of Relation Extraction (RE).

One obstacle that is encountered when building a RE system is the generation of training instances. For coping with this difficulty, (Mintz et al., 2009) proposes distant supervision to automatically generate training samples via aligning KBs with text. They assume that if two entities are connected by a relation in a KB, then all sentences that contain those entity pairs will express the relation. For instance, (*David Yassky*, */people/person/place_lived*, *Brooklyn*), (*Washington Irving*, */people/person/place_lived*, *New York*) and (*Huber Humphrey*, */people/person/place_lived*, *Minneapolis*) are fact triplets in Freebase. Distant supervision will automatically label all sentences, such as Example 1 below, which are taken from NYT10 dataset, as positive instances for the relation */people/person/place_lived*. Although distant supervision could provide a large amount of training data at low cost, it always suffers from fuzzy data. For instance, Example 1 should not be seen as the explicit evidences to support the */people/person/place_lived* relationship between corresponding entity pairs.

(1) *Once predictably democratic in national politics , the anchor of upper midwest liberal populism from the 1920s through* ***Hubert Humphrey*** *and Walter Mondale, Minnesota is now considered a battleground, with the republicans scheduled to hold their national convention in 2008 in* ***Minneapolis*** *and St. Paul around that declaration.*

To identify relation from these fuzzy evidences, (Das et al., 2016) uses Multi-hop Inferences (MI) over text as the evidence to extract relation. However, they only represent the textual relation as a unit relation, rather than a sequence of words. For instance, they represent the textual relation in T1 in Fig 1 as $\{e_1, left\_his\_at\_the, e_2\}$, rather than the textual representation $\{e_1, left, his, homestate, to, attend, college, at, the, e_2\}$. This, thus, cannot utilize the semantic meaning of each words in the textual relation and a world-level attention mechanism.

Contributions. We proposed a DS-RE model

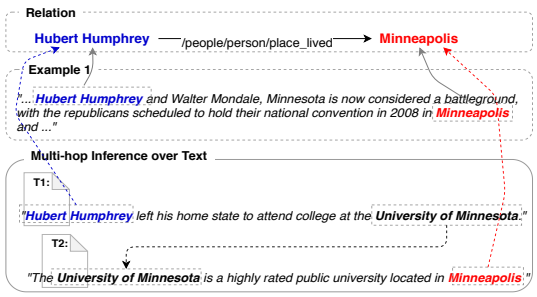

Figure 1: An example of multi-hop inference over Text.

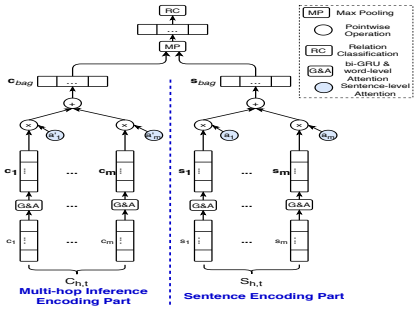

Figure 2: Overview of the proposed model.

that represents MI over text as a sequence of words and conduct evaluation on NYT10 dataset. Our method achieves a competitive result with the state-of-the-art models. The experimental results prove the effectiveness of the proposed model for DS-RE.

2 Related Work

RE is a fundamental task in the NLP community. However, annotation of training data for RE is expensive and time-consuming. To address this issue, distant supervision is proposed by (Mintz et al., 2009). Recent methods utilize external Background Knowledge (BK) for DS-RE. (Ji et al., 2017) applies entity descriptions as BK, (Lin et al., 2017) utilizes multilingual text as BK and (Vashishth et al., 2018) uses entity types and relation alias information as BK for DS-RE and (Alt et al., 2019) utilizes pre-trained language model as BK for DS-RE. (Das et al., 2016) uses MI over text as the source of relation extraction. However, none of these existing approaches mentioned above applies textual representation of MI over text as BK for DS-RE.

3 Proposed Model

3.1 Generation of Multi-hop Inference

Let (e_h, e_t) be an entity pair of interest. The text with entity mention annotations is represented as a graph, where nodes indicates the entities. Since all parts of the sentence (not just the middle context) are useful for relation classification (Vu et al., 2016; Xiao and Liu, 2016), the edges of the graph are represented by the sentence that contains the connected entity pair. To reduce the semantic redundancy and computational cost, if there are multiple sentences containing a same entity pair, we select the sentence with the least words as the edge of the entity pair. To simulate the situation of reasoning across documents, we remove the sentence (i.e., edge) that containing the entity pair of interest from the graph. We apply the random walk procedure [1] to extract MI over text between a given entity pair. Each MI over text is represented as $c = \{e_h, s_{1w_1}, s_{1w_2}, ...s_{1w_n}, s_{2w_1}, s_{2w_2}, ...e_t\} \in C_{h,t}$, where, s_{1w_n} indicates the last word of the first hop of the MI.

3.2 Architecture

The proposed model consists of two parts: Sentence Encoding Part and Multi-hop Inference Encoding Part , as shown in Figure 2.

Sentence Encoding Part (SEP). SEP is the Bi-GRU based DS-RE model with word level and sentence level attention proposed by (Jat et al., 2018). Specifically, given a sentence s with n words $s_i = \{w_1, ..., w_n\}$ from the bag $S_{h,t} = \{s_1, ..., s_m\}$ in which each sentence contains the entity pair (e_h, e_t), vector representation $\mathbf{v}_t$ for each word w_t is calculated as $\mathbf{v}_t = [\mathbf{v}_t^w; \mathbf{v}_t^{wp1}; \mathbf{v}_t^{wp2}]$, where $\mathbf{v}_t^w$ is k-dimensional GloVe embedding (Pennington et al., 2014), $\mathbf{v}_t^{wp1}$ and $\mathbf{v}_t^{wp1}$ are the word position embedding (Zeng et al., 2014). Then, the hidden representation for each word $\mathbf{h}_i$ is obtained by using Bi-GRU (Cho et al., 2014) over the sentence. The vector representation of the sentence $\mathbf{s}_i$ is calculated via the Equation 1, where $\mathbf{r}_w$ is a random query vector for calculating the word level attention a_i^w.

$$\mathbf{s}_i = \sum_{i=1}^{m} a_i^w \mathbf{h}_i, \qquad a_i^w = \frac{\exp(\langle \mathbf{h}_i, \mathbf{r}_w \rangle)}{\sum_{k=1}^{n} \exp(\langle \mathbf{h}_k, \mathbf{r}_w \rangle)} \tag{1}$$

The vector representation of the entire bag $\mathbf{s}_{bag}$ is calculated via Equation 2, where $\mathbf{r}_s$ is a random

[1] Considering computational time and semantic drift, the cut-off criteria of random walk is manually set as 3.

System	P@100	P@200	P@300	P@500	Mean
Mintz†	52.3	50.2	45.0	39.7	46.8
PCNN+ATT†	73.0	68.0	67.3	63.6	68.0
RESIDE†	81.8	**75.4**	**74.3**	**69.7**	**75.3**
DISTRE†	68.0	67.0	65.3	65.0	66.3
SEP+MIEP(Unit)	81.0	73.0	70.0	64.2	72.1
SEP+MIEP(Textual)	**83.0**	75.0	72.3	68.6	74.7

Table 1: Precision@N from previous state-of-the-art DS-RE models and our proposed model, where †represents that these results are quoted from (Alt et al., 2019), where "(Unit)" indicates the method of MI representation proposed by (Das et al., 2016), which just takes the last hidden state of Bi-GRU as the vector representation of MI.

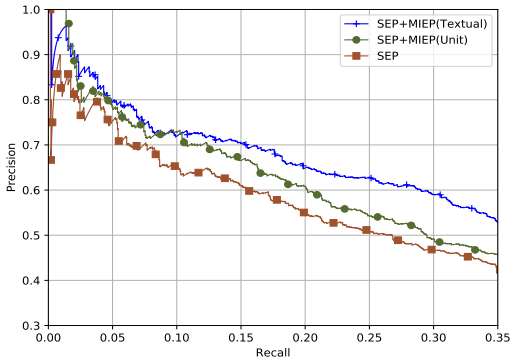

Figure 3: Precision-Recall curves of SEP, SEP+MIEP(Unit) and SEP+MIEP(Textual).

query vector for calculating the sentence level attention a_i^s.

$$\mathbf{s}_{bag} = \sum_{i=1}^{m} a_i^s \mathbf{s}_i, \qquad a_i^s = \frac{\exp(\langle \mathbf{s}_i, \mathbf{r}_s \rangle)}{\sum_{k=1}^{m} \exp(\langle \mathbf{s}_k, \mathbf{r}_s \rangle)} \tag{2}$$

Multi-hop Inference Encoding Part (MIEP). Since a MI could be seen as a sequence of words, or a special sentence, we apply the similar Bi-GRU based model used in the SEP to encode a bag [2] of MI $C_{h,t} = \{c_1, ..., c_m\}$ into vector representation $\mathbf{c}_{bag}$. Finally, we use a max pooling layer over $\mathbf{s}_{bag}$ and $\mathbf{c}_{bag}$ to generate final feature vector and calculate the probability distribution over the target relations via the Equation 3.

$$p(y) = \mathbf{W}\text{max-pool}(\mathbf{s}_{bag}, \mathbf{c}_{bag}) + \mathbf{b} \tag{3}$$

4 Experiments

Dataset. We evaluate our model on NYT10 dataset (Riedel et al., 2010), which is created by aligning Freebase relational facts with the New York Times Corpus. Sentences from the year 2005-2006 are used for training and the sentences from 2007 are used for testing. We use ClueWeb12 with Freebase entity mention annotations (Gabrilovich et al., 2013) to extract MI over text.

Evaluation Metrics. We follow (Mintz et al., 2009) and conduct the held-out evaluation, in which a DS-RE model is evaluated by comparing the fact triplets identified from textual data (i.e., the set of sentences containing the target entity pairs) with those in KB. We report Precision-Recall curves, Precision@N and AUC score, which measures of the area under the Precision-Recall curve.

Results. The Precision-Recall (PR) curves are shown in Figure 4 and Figure 3. Precision@N are listed in Table 1. The results show that, SEP-MIEP(Textual) significantly outperforms the most of of previous models over almost entire range of recall, and achieve a competitive results with the state-of-the-art models. This proves that our proposed model is effective for DS-RE.

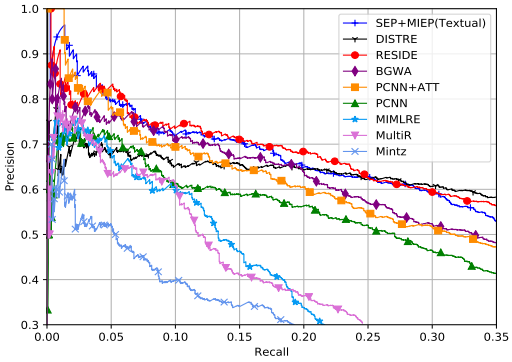

Figure 4: Precision-Recall curves of previous state-of-the-art method and our proposed model SEP+MIEP(Textual).

5 Conclusion and Future Work

In this work, we hypothesize that textual representation of MI over text is useful for DS-RE. Experimental results validates our hypothesis, and moreover, the proposed model obtains a competitive results with the state-of-the-art models on NYT10 dataset. Although the MI over text gathered via the random walk algorithm are proved to be useful for DS-RE, the randomness of the MI would hinder its effectiveness. Therefore, in the future, We will apply more sophisticated strategy such as reinforcement learning (Xiong et al., 2017; Das et al., 2017; Lin et al., 2018) to search MI for DS-RE.

[2] Considering computational cost, the maximum size of bag is manually set as 10.

References

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. *arXiv preprint arXiv:1906.08646*.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. AcM.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, Luke Vilnis, Ishan Durugkar, Akshay Krishnamurthy, Alex Smola, and Andrew McCallum. 2017. Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning. *arXiv preprint arXiv:1711.05851*.

Rajarshi Das, Arvind Neelakantan, David Belanger, and Andrew McCallum. 2016. Chains of reasoning over entities, relations, and text using recurrent neural networks. *arXiv preprint arXiv:1607.01426*.

Evgeniy Gabrilovich, Michael Ringgaard, and Amarnag Subramanya. 2013. Facc1: Freebase annotation of clueweb corpora, version 1 (release date 2013-06-26, format version 1, correction level 0).

Sharmistha Jat, Siddhesh Khandelwal, and Partha Talukdar. 2018. Improving distantly supervised relation extraction using word and entity based attention. *arXiv preprint arXiv:1804.06987*.

Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Xi Victoria Lin, Richard Socher, and Caiming Xiong. 2018. Multi-hop knowledge graph reasoning with reward shaping. *arXiv preprint arXiv:1808.10568*.

Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. Neural relation extraction with multi-lingual attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 34–43.

Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. 2013. Distant supervision for relation extraction with an incomplete knowledge base. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 777–782.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.

Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. 2018. Reside: Improving distantly-supervised neural relation extraction using side information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1257–1266.

Ngoc Thang Vu, Heike Adel, Pankaj Gupta, and Hinrich Schütze. 2016. Combining recurrent and convolutional neural networks for relation classification. *arXiv preprint arXiv:1605.07333*.

Minguang Xiao and Cong Liu. 2016. Semantic relation classification via hierarchical recurrent neural network with attention. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1254–1263.

Wenhan Xiong, Thien Hoang, and William Yang Wang. 2017. Deeppath: A reinforcement learning method for knowledge graph reasoning. *arXiv preprint arXiv:1707.06690*.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, et al. 2014. Relation classification via convolutional deep neural network. In *COLING*, pages 2335–2344.