Improving Medical Relation Extraction with Distantly Supervised Pre-training

Zhen Wan Fei Cheng Zhuoyuan Mao Qianying Liu Haiyue Song Sadao Kurohashi
Kyoto University
{ZhenWan, zhuoyuanmao, ying, song}nlp.ist.i.kyoto-u.ac.jp
{feicheng, kuro}i.kyoto-u.ac.jp

Abstract

Relation extraction (RE) is used to populate knowledge bases that are important to many applications. Traditional RE task largely relies on the sufficiency of labeled training data, but for medical domain relation extraction, it is costly and time-consuming to construct large labeled training data. Meanwhile, there are many available unlabeled medical corpora. To utilize both the abundance of raw corpora and the accuracy of annotated datasets, we propose a two-stage framework to pre-train models on an intermediate task for improving the target RE task performance. In the first stage, we also introduce a distant supervision based method to construct the training data for the intermediate task. The empirical results suggest the proposal significantly improve the target RE task.

1 Introduction

Relation extraction is the task of extracting semantic relationships from a text. Such a relationship occurs between one or more entities of a certain type (eg: person, organization) and belongs to a particular semantic category (eg: date of birth, employed by). Consider the sentence “Joe Biden in the president of America” in figure 1. Here, the relation “president of” connects the subject entity “Joe Biden” to the object entity “America”. Relation extraction has many applications in information extraction, creating or extending knowledge bases, automatically annotating structured information found in text and recently, in evaluating the factual consistency of abstractive text summarization. With the recent advance of deep learning, neural relation extraction (NRE) models (Baldini Soares et al [1];Zeng et al [2];Zhang et al [3]; Chen et al [4]) have achieved the latest state-of-the-art results and some of them are even comparable with human performance on several public RE benchmarks.

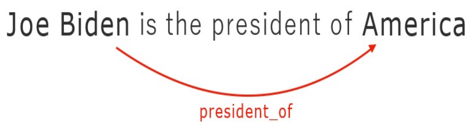

Figure 1 An example for relation extraction to identify relationship between entites mentioned in the text.

However, The success of NRE models on current RE benchmarks largely relies on the sufficiency of training data, but manual annotation is costly and time-consuming in specific domain, such as medical relation extraction. In this case, the insufficient amount of training data may be the reason that hinders the good results of relation extraction tasks. Meanwhile, although The labeling of medical datasets is a complex problem, there are many available unlabeled medical texts. The existence of such rich raw corpora makes us wonder: Can we utilize large raw corpora to improve RE on annotated datasets ?

In this article, we propose a two-stage fine-tuning framework for solving this question. In the first stage, we pre-train models on an intermediate task with weak supervision extracted from large raw corpora, and then in the second stage, we can further fine-tune the output trained models on the original annotated dataset, which is our target relation extraction task. Through such a framework, we can both take advantage of the abundant data brought by distant supervision (DS) and maintain the accuracy brought by manual annotation data.

The idea of a pre-training and fine-tuning framework has been a new trend in the relation extraction field (Peng et al [5], Robert Ormandi et al [6]). The recent work Contrastive Pre-training (CP) by Peng et al [5] has confirmed the effectiveness of the pre-training stage to improve the final target relation extraction task. In their work, they first generate a dataset from Wikipedia data by distant supervision and use this constructed dataset as a pre-training step. Given a triplet in the knowledge base (KB) containing a particular entity pair and their corresponding relation type, any sentence from the raw text sharing the same two entity pairs at the same time will be extracted and labeled the corresponding relation type as shown in figure 2. As for the second fine-tuning stage, Peng et al [5] further fine-tune the model on various target datasets to figure out whether the pre-training step can improve the final performance of the relation extraction task on each dataset.

There are two limitations of their framework. 1) The intermediate pre-training data extracted from wikipedia has a different domain from multiple target datasets. This inconsistency between two stages may lead to unreliable influence on the target task. 2) KBs are not always available in a specific domain, such as the medical domain. Instead of relying on an existing KB, we propose to induce a triplet set from the target manual annotated dataset so that we can generate a distantly supervised dataset for pre-training. In this way, we can ensure consistency between the two steps and avoid the lack of available KBs in the medical domain. In the section 3, we will introduce the details to construct the distantly supervised dataset, and to use two state-of-the-art (SOTA) distant supervision approaches in the pre-training stage. Then in section 4, we will implement experiments to figure out whether the pre-training stage can improve the final target relation extraction task.

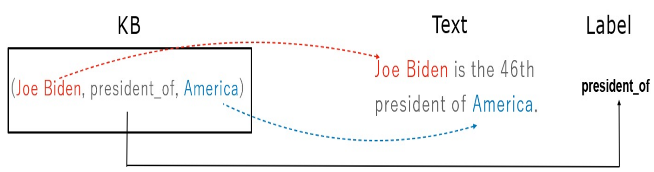

Figure 2 A distant supervision example, from the triplet in the KB, we can align entity pairs with relation to the text to construct a labeled instance.

Generally, we summarize our contributions as follows:

- We introduce a two-stage framework by leveraging the intermediate task in the first stage to improve the target relation extraction task in the second stage.
- We construct intermediate medical RE dataset distantly supervised by the triplets derived from the manual target data, which can serve various future applications.

2 Related Work

With awareness of the existing DS noise, Surdeanu et al [7]introduces the multi-instance learning (MIL) framework to distantly supervised relation extraction (DSRE) by dividing training instances into several bags and using bags as new data units. Regarding the strategy for selecting instances inside the bag, the soft attention mechanism proposed by Lin et al [8] is widely used for its better performance than the hard selection method. The ability to form accurate representations from noisy data makes the MIL framework soon become a paradigm of following-up works.

More recently, Chen et al [9] argues that the long-standing MIL framework can not effectively utilize abundant instances inside MIL bags, they propose a novel contrastive instance learning (CIL) method to boost the distantly supervised relation extraction (DSRE) model performances under the MIL framework. In detail, they regard the initial MIL framework as the bag encoder, which provides relatively accurate representations for different relational triples. Then they develop contrastive instance learning (CIL) to utilize each instance in an unsupervised manner: In short, the goal of their CIL is that the instances sharing the same relational triples (i.e.positive pairs) ought to be close in the semantic space, while the representations of instances with different relational triples (i.e.negative pairs) should be far away. They achieve dramatic improvement on various benchmarks, but as it is a MIL-based method, the reliability of MIL strategy and the noise in DS data still constrain the RE task performance.

In order to reduce the influence of noise labeling in the DS part, another recent work CP [5] has removed the MIL loss in their pre-training stage on DS data, the objective function only focuses on contrastive learning loss to avoid the noise labeling. Their final results on target fine-tuning datasets confirm the effectiveness of the contrastive learning loss in pre-training step.

3 Proposed Method

In this section, we first present an overview of our proposed approach in Section 3.1 and then detail our approach in Section 3.2, 3.3.

Dataset	# Rel	# Train	# Dev	# Test
i2b2 2010VA	6	3,120	11	6,147

Table 1 **Statistics of i2b2 2010VA**, # Rel denotes the number of relation types. # Train, # Dev, # Test denote the number of instances in train, dev and test.

3.1 Overview

The traditional supervised relation extraction task is based on a human-annotated dataset. However, with the large corpora as external knowledge, the task now starts from an annotated dataset, then extracts proper sentences from raw corpora by distant supervision, and uses this generated dataset as a pre-training step to improve the relation extraction task on the original annotated dataset. We show the overview of our proposal in the figure 3.

3.2 External Datasets Construction

i2b2 2010VA shared task collection consists of 170 documents for training and 256 documents for testing, which is the subset of the original dataset [10]. The dataset was collected from three different hospitals and was annotated by medical practitioners for eight types of relations between problems and treatments.

This is our final target dataset, and to construct a knowledge base for distant supervision, normally, we can randomly select two entities from the i2b2 2010VA dataset and combine them with their labeled relation type to generate a triplet. However, this random strategy may involve too many entity pairs, including cross sentence pairs whose relation types are hard to confirm. To make the task more realistic, we will focus on each sentence. First, extract all entities in the particular sentence, and if any two of them are labeled a relation type, they will generate a triplet with a particular relation. Otherwise, they will still generate a triplet but labeled NA (no relation)

Normally, we can randomly select two entities from the annotated dataset and combine them with their labeled relation type to generate a triplet. However, this random strategy may involve too many entity pairs, including cross sentence pairs whose relation types are hard to confirm. To make the task more realistic, we will focus on each sentence. First, extract all entities in the particular sentence, and if any two of them are labeled a relation type, they will generate a triplet with a particular relation. Otherwise, they will still generate a triplet but labeled NA (no relation). For example,

Dataset	# Triplets in KB (NA)	# Instances (NA)
DS from MIMIC-III	2,777 (35,737)	36,084 (76,079)

Table 2 **Statistics of DS dataset from MIMIC-III**, # Triplets denotes the number of triplets in the KB generated from i2b2 2010VA, and NA denoted no-relation triplets.

After constructing the knowledge base, we can extract valuable sentences from raw text based on each triplet in the KB. The strategy here is a standard distant supervision method as in Introduction. To balance the number of sentences extracted by each triplet, we also add an upper bound to the number of extracted sentences. As i2b2 2010VA is a medical domain dataset, for the purpose of consistency, we then choose MIMIC-III ('Medical Information Mart for Intensive Care') as the raw text to extract sentences.

MIMIC-III is a large, single-center database comprising information relating to patients admitted to critical care units at a large tertiary care hospital. Data includes vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, and more. The database supports applications including academic and industrial research, quality improvement initiatives, and higher education coursework.

3.3 Two-Stage Framework

Pre-training Stage The goal of the pre-training step is both to utilize extracted sentences from large raw corpora and to avoid noise in the distant supervision method. We leverage two introduced SOTA contrastive learning methods (CIL, CP) to solve the first-stage distant supervision pre-training. CIL use bag-level information to construct postive and negative pairs for contrastive learning, and CP instead focuses on relation-level information to construct positive and negative pairs.

Fine-tuning Stage After the pre-training step, the output model will be regarded as the input model in the final fine-tuning (FT) step, here we treated the relation extraction task as a sentence classification by replacing two named entities in the sentence with predefined tags (e.g., @GENE$, @DRUG$) (Lee et al [11]). For example, we used "@CHEMICAL$ protected against the RTI-76-induced inhibition of @GENE$ binding." to replace

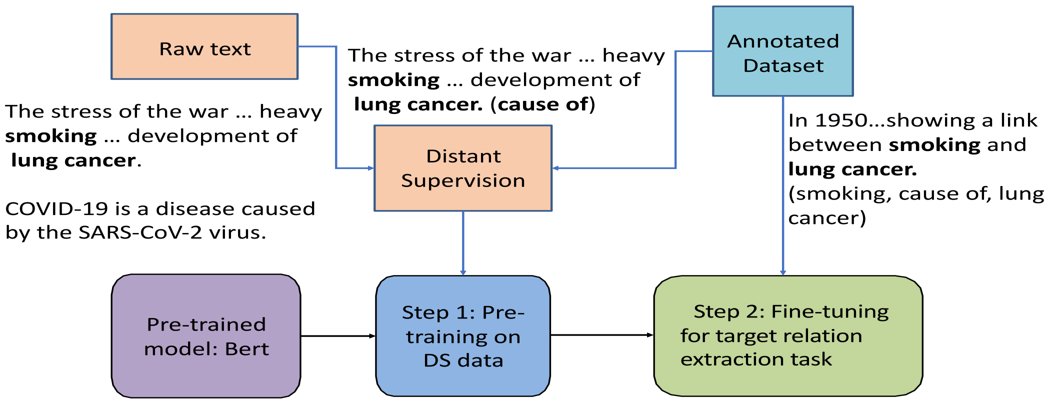

Figure 3 Overview of our proposed approach.

the original sentence "Citalopram protected against the RTI-76-induced inhibition of SERT binding." in which "citalopram" and "SERT" has a chemicalgene relation. The only difference is that the input model will be a pre-trained language model for traditional relation extraction such as BERT, and the input model is the output model from the pre-training step for this task.

4 Experiment

4.1 Baselines

Our task is to improve the RE on i2b2 2010VA dataset, and the fundamental baseline is to directly fine-tune (FT) language models on i2b2 2010VA without the pre-training step. At the same time, we choose two introduced SOTA methods, CIL [9] + FT and CP [5] + FT to utilize the pre-training step best. As for the experiment settings of CIL + FT and CP + FT, we follow the default hyper-parameters in their papers. The implementation of CP can be found on their website[1)], and we also receive the codes for CIL from its author via email.

We use both Bert-base-uncased [12] and the SOTA medical domain BlueBert [13] as the pre-trained language model to better evaluate the performance.

4.2 Results

We summarize the model performances of directly fine-tuning and two-step models in the Table 3. From the results, we can observe that: (1) For both bert-base-uncased and the SOTA BlueBert, with the external pre-training step, both CIL and CP improve the final performance. (2) The

1) https://github.com/thunlp/RE-Context-or-Names

Models	Precision	Recall	Micro-F1
Bert-Base-Uncased			
FT	**75.63**	67.77	71.45
CIL + FT	72.81	**72.51**	72.66
CP + FT	75.19	70.75	**72.90**
BlueBert			
FT	**76.98**	73.54	75.22
CIL + FT	75.13	75.67	75.39
CP + FT	74.28	**76.62**	**75.43**

Table 3 **Overall results.**

SOTA BlueBert can obviously improve the medical relation extraction task. (3) Compared with the CIL method, CP achieves a better result on final evaluation, we assume that as CIL is a MIL based framework, its MIL loss may include more noise generated from distant supervision, which somehow leads to a worse result in the final fine-tuning.

5 Conclusion

We introduce a two-stage framework to improve the traditional medical relation extraction. Experiment results show that through this framework, we can benefit both from the abundance of training data by distant supervision and the accuracy of the human-annotated dataset. We also propose a method to generate a distantly supervised dataset from raw corpora based on the annotated dataset without relying on an exisiting KB, which we can use for future purposes.

References

[1] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In **Proceedings of**

the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2895–2905, Florence, Italy, July 2019. Association for Computational Linguistics.

[2] Daojian Zeng, Haoran Zhang, and Qianying Liu. Copymtl: Copy mechanism for joint extraction of entities and relations with multi-task learning. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 34, No. 05, pp. 9507–9514, Apr. 2020.

[3] Ranran Haoran Zhang, Qianying Liu, Aysa Xuemo Fan, Heng Ji, Daojian Zeng, Fei Cheng, Daisuke Kawahara, and Sadao Kurohashi. Minimize exposure bias of seq2seq models in joint entity and relation extraction, 2020.

[4] Zexuan Zhong and Danqi Chen. A frustratingly easy approach for entity and relation extraction. In **North American Association for Computational Linguistics (NAACL)**, 2021.

[5] Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. Learning from context or names? an empirical study on neural relation extraction, 2020.

[6] Róbert Ormándi, Mohammad Saleh, Erin Winter, and Vinay Rao. Webred: Effective pretraining and finetuning for relation extraction on the web. **CoRR**, Vol. abs/2102.09681, , 2021.

[7] Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. Multi-instance multi-label learning for relation extraction. In **Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning**, pp. 455–465, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[8] Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. Neural relation extraction with selective attention over instances. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2124–2133, Berlin, Germany, August 2016. Association for Computational Linguistics.

[9] Tao Chen, Haizhou Shi, Siliang Tang, Zhigang Chen, Fei Wu, and Yueting Zhuang. CIL: Contrastive instance learning framework for distantly supervised relation extraction. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 6191–6200, Online, August 2021. Association for Computational Linguistics.

[10] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. **Journal of the American Medical Informatics Association**, Vol. 18, No. 5, pp. 552–556, 06 2011.

[11] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining, 2019.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[13] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets, 2019.