Model-Agnostic Meta-Learning for Relation Classification with Limited Supervision

Abiola Obamuyide
Department of Computer Science
University of Sheffield
avobamuyide1@sheffield.ac.uk

Andreas Vlachos
Dept. of Computer Science and Technology
University of Cambridge
andreas.vlachos@cst.cam.ac.uk

Abstract

In this paper we frame the task of supervised relation classification as an instance of meta-learning. We propose a model-agnostic meta-learning protocol for training relation classifiers to achieve enhanced predictive performance in limited supervision settings. During training, we aim to not only learn good parameters for classifying relations with sufficient supervision, but also learn model parameters that can be fine-tuned to enhance predictive performance for relations with limited supervision. In experiments conducted on two relation classification datasets, we demonstrate that the proposed meta-learning approach improves the predictive performance of two state-of-the-art supervised relation classification models.

1 Introduction

Relation classification, the task of determining the relationship that exists between two entities, is a long-standing challenge in artificial intelligence with many downstream applications, including question answering, knowledge base population and web search. A variety of supervised methods have been proposed in the literature for this task (Zelenko et al., 2003; Bunescu and Mooney, 2005; Mintz et al., 2009; Surdeanu et al., 2012; Riedel et al., 2013). Current approaches are predominantly supervised models based on neural networks, for instance recursive neural networks (Socher et al., 2012; Hashimoto et al., 2013), convolutional neural networks (Zeng et al., 2014; Nguyen and Grishman, 2015), recurrent neural networks (Zhang and Wang, 2015; Xu et al., 2015; Zhang et al., 2017) or a combination of recurrent and convolutional neural networks (Vu et al., 2016). The performance of these approaches relies mostly on the quantity of their training data. However, labelled training data can be expensive to obtain and available only in limited quantities. It is therefore pertinent to develop methods that reduce their reliance on large quantities of labelled training data.

In this work we propose a model-agnostic protocol for training supervised relation classification systems to achieve higher predictive performance in limited supervision settings, motivated by the observation that meta-learning leads to learning a better parameter initialization for new tasks than ad hoc multi-task learning across all tasks (Finn et al., 2017). We show that relation classification can be approached from a meta-learning perspective, and propose a model-agnostic meta-learning protocol for training relation classification models that explicitly learns a model parameter initialization for enhanced predictive performance across all relations with limited supervision. During training, our algorithm considers all relations and their instances as coming from a joint distribution, and seeks to learn model parameters that can be quickly adapted using each relation's training instances to enhance predictive performance on its test set.

In experiments on two relation classification datasets, we apply the proposed approach to two relation classification models, the position-aware relation classification model proposed in Zhang et al. (2017) (*TACRED-PA*) and the contextual graph convolution networks proposed in Zhang et al. (2018) (*C-GCN*), with varying amounts of supervision available at training time. We find that our approach improves the accuracy of both relation classification models on the two datasets. For instance our approach improves the F1 performance of *TACRED-PA* from 3.13% to 21.05% with just 1% of the training data on the *SemEval* dataset, and from 2.98% to 34.59% with just 0.5% of the training data on the *TACRED* dataset.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5873–5879
Florence, Italy, July 28 - August 2, 2019.

2 Background

Meta-learning, sometimes referred to as *learning to learn* (Thrun and Pratt, 1998), aims to develop models and algorithms which are able to exploit background knowledge to adaptively improve their learning process with experience. A number of meta-learning approaches have been proposed, and broadly fall into the following lines of work: learning how to update model parameters from background knowledge (for instance, Andrychowicz et al. 2016; Ravi and Larochelle 2017), specific model architectures for learning with limited supervision (for instance, Vinyals et al. 2016; Snell et al. 2017), and model-agnostic methods for learning a good parameter initialization for learning with limited supervision (for instance, Finn et al. 2017; Nichol et al. 2018).

We next give a brief overview of the model-agnostic methods for meta-learning, which learn a good parameter initialization for target tasks from a set of source tasks, as proposed in Finn et al. (2017) and Nichol et al. (2018). These algorithms work by training a meta-model on the set of source tasks, such that the meta-model provides a good parameter initialization for target tasks which are taken from the same distribution as the source tasks. At test time, such an initialization can be fine-tuned with a limited number of gradient steps using a limited amount of training examples from the target tasks, in order to achieve good performance on the target tasks.

In formal terms, let $p(\mathcal{T})$ be the distribution over tasks and f_θ be the function learned by a neural model parametrized by θ. During adaptation to each task $\mathcal{T}_i$ sampled from $p(\mathcal{T})$, the model parameters θ are updated to task-specific parameters θ'_i. For a single gradient step, for instance, this update can be carried out as:

$$\theta'_i = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta) \tag{1}$$

where $\mathcal{L}_{\mathcal{T}_i}$ is the loss on task $\mathcal{T}_i$ and α is the step size hyperparameter.

The model parameters θ are trained to optimize the performance of $f_{\theta'_i}$, after taking a number of gradient steps with limited example instances from tasks sampled from $p(\mathcal{T})$. This is can be achieved by utilizing the meta-objective:

$$\min_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta'_i}) = \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)}) \tag{2}$$

The optimization of the meta-objective is performed across tasks using *SGD*, by making updates to θ:

$$\theta \leftarrow \theta - \epsilon \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta'_i}) \tag{3}$$

where ϵ is the meta step size parameter.

Intuitively, the meta-objective explicitly encourages the model to learn model parameters that can be quickly adapted to achieve optimum predictive performance across all tasks with as few gradient descent steps as possible.

A number of approaches have been proposed for extracting relations with zero or few supervision instances. For the problem of zero-shot extraction of relations, Rocktäschel et al. (2015); Demeester et al. (2016) proposed the use of logic rules, Levy et al. (2017) proposed to address the problem by formulating it as a reading comprehension challenge, while Obamuyide and Vlachos (2018) proposed to address it as a textual entailment challenge.

In this work we address the case where a limited number of supervision instances is available for all relations. In previous work, Obamuyide and Vlachos (2017) explored the use of a Factorization Machine (Rendle, 2010) framework for extracting relations with limited supervision instances. Here we instead propose an approach which is generally applicable to gradient-optimized relation extraction models. Han et al. (2018) proposed a dataset and evaluation setup for few-shot relation classification which assumes access to full supervision for training relations (specifically 700 instances per relation). In contrast, we address a different setting in which only limited supervision is available for all relations. In addition, the setup in Han et al. (2018) requires a model architecture *specific* to few-shot learning based on distance metric learning. On the other hand, our approach has the advantage that it applies to any gradient-optimized relation classification model.

3 Model-Agnostic Meta-Learning for Relation Classification

If we consider each relation $\mathcal{R}_i$ as a task, then one approach to supervised relation classification with limited supervision is to train a multi-class classifier for all relations in a multi-task fashion. For all relations $\mathcal{R}_i$ from a distribution $p(\mathcal{R})$, this approach directly optimizes for the following objec-

tive:

$$\theta^* = \min_{\theta} \sum_{\mathcal{R}_i \sim p(\mathcal{R})} \mathcal{L}_{\mathcal{R}_i}(f_\theta) \quad (4)$$

where $\mathcal{L}_{\mathcal{R}_i}$ is the loss on relation $\mathcal{R}_i$. This assumes that joint training on all relations would naturally result in the optimal model parameters θ^* with good predictive performance for all relations. This is however not necessarily the case, especially for relations with limited training instances from which the model can learn to generalize.

We propose to instead utilize meta-learning to explicitly encourage the model to learn a good joint parameter initialization for all relations, which can then be fine-tuned with limited supervision from each relation's training instances to achieve good performance on its test set. Such parameters would be especially beneficial for enhancing performance on relations with limited training instances.

Observe though that directly optimizing Equation 2 requires computing second order derivatives over the parameters, which can be computationally expensive. Thus, we follow Nichol et al. (2018) by approximating the meta-objective in Equation 2 with the training Algorithm in 1.

Algorithm 1 Meta-Learning Relation Classification (*MLRC*)

Require: distribution over relations $p(\mathcal{R})$
Require: relation classification function f_θ
Require: gradient-based optimization algorithm (e.g. *SGD*)
Require: step size ϵ, learning rate α

1: randomly initialize θ
2: **while** not done **do**
3: Sample batch of $\mathcal{B}$ relations $\mathcal{R}_i \sim p(\mathcal{R})$
4: **for all** $\mathcal{R}_i$ **do**
5: Sample train instances $\mathcal{D} = \{x^{(j)}, y^{(j)}\}$ from $\mathcal{R}_i$
6: Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{R}_i}(f_\theta)$ using $\mathcal{D}$
7: Compute adapted parameters: $\theta_i' = SGD(\theta_i, \nabla_\theta \mathcal{L}_{\mathcal{R}_i}(f_\theta), \alpha)$
8: **end for**
9: Compute update of meta-parameters: $\theta = \theta - \epsilon \frac{1}{\mathcal{B}} \sum_{i=1}^{i=\mathcal{B}} (\theta_i' - \theta)$
10: **end while**
11: Fine-tune f_θ with standard supervised learning.

Subsequently we refer to our overall training procedure as summarized in Algorithm 1 as Meta-learning Relation Classification (*MLRC*). We assume access to f_θ (learner model), which is a relation classification model parameterized by θ and a distribution over relations $p(R)$. The algorithm consists of the meta-learning phase (lines 1-10), followed by the supervised learning phase (line 11) which fine-tunes the meta-learned parameters, both carried out on a relation classification model using the same data for both stages.

In the first phase of learning, each iteration in our approach starts by sampling a batch of relations from $p(R)$ (line 3). Then for each relation we sample a batch of supervision instances $\mathcal{D}$ from its training set (line 5). We then obtain the adapted model parameters θ_i' on this relation by first computing the gradient of the training loss on the sampled relation instances (line 6) and backpropagating the gradients with a gradient-based optimization algorithm such as *SGD* or *Adagrad* (Duchi et al., 2011) (line 7). At the end of the learning iteration, the adapted parameters on each sampled relation in the batch are averaged, and an update is made on the model parameters θ (line 9).

In the second phase of learning, we first initialize the model parameters with that learned during meta-training. We then proceed to fine-tune the model parameters with standard supervised learning by taking a number of gradient descent steps using the same randomly sampled batches of supervision instances from the relations' training set as was used during meta-learning (line 11).

4 Experiments

4.1 Relation Classification Models

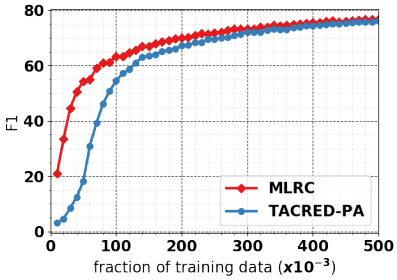
We adopt as the learner model (f_θ) two recent supervised relation classification models, the position-aware model of Zhang et al. (2017) (*TACRED-PA*) and the contextual graph convolution networks proposed in Zhang et al. (2018) (*C-GCN*), both of which are multi-class models with parameters optimized via stochastic gradient descent.

4.2 Setup

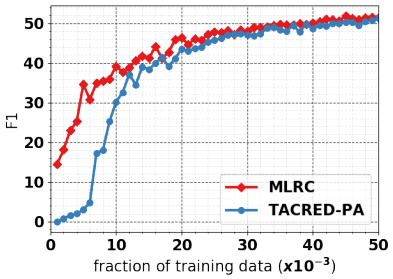
We conduct experiments in a limited supervision setting, where we provide all models with the same fraction of randomly sampled supervision instances during training. Further, for each experiment the supervision instances within each fraction is exactly the same across all models. We report results for each experiment by taking the average over ten (10) different runs.

4.3 Datasets

We evaluate our approach on the SemEval-2010 Task 8 relation classification dataset (Hendrickx et al., 2009) (*SemEval*), and on the recent, more

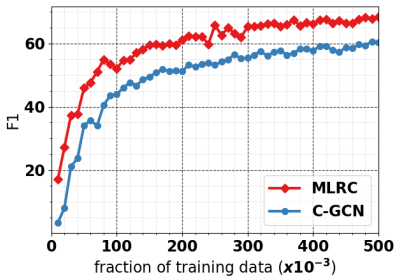


(a)

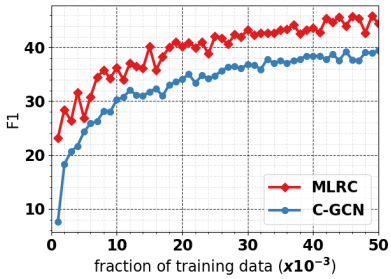


(b)

Figure 1: Results obtained using *TACRED-PA* as the learner model on (a) *SemEval*, and (b) *TACRED* datasets



(a)



(b)

Figure 2: Results obtained using *C-CGN* as the learner model on (a) *SemEval*, and (b) *TACRED* datasets

challenging TACRED dataset (Zhang et al., 2017) (*TACRED*). The *SemEval* dataset has a total of 8000 training and 2717 testing instances respectively. For experiments the training set is split into two, and we use 7500 instances for training and 500 instances for development. For *TACRED*, we use the standard training, development and testing splits as provided by Zhang et al. (2017).

4.4 Experimental Details and Hyperparameters

We initialize word embeddings with Glove vectors (Pennington et al., 2014) and did not fine-tune them during training. Model training and parameter tuning are carried out on the training and development splits of each dataset, and final results reported on the test set.

We ensure all models have access to the same data. For model *MLRC*, for each fraction, we train for 150 meta-learning iterations on *TACRED* dataset and 1000 meta-iterations on the *SemEval* dataset using that fraction of data. We then fine-tune with standard supervised learning using exactly the same data as was used during meta-learning.

For both relation classification models, that is *TACRED-PA* and *C-GCN*, we use the same hyper-parameters as in Zhang et al. (2017) and Zhang et al. (2018) respectively.

Relation	#	F1(%)	
		TC-PA	MLRC
Instrument-Agency	3	0	8.44
Content-Container	4	0.93	30.9
Member-Collection	5	3.04	24.19
Entity-Destination	7	14.33	35.36
Entity-Origin	7	2.85	24.62
Message-Topic	7	0.8	12.32
Component-Whole	8	2.68	14.87
Product-Producer	9	0.68	10.29
Cause-Effect	11	2.93	28.52
Average		3.13	21.05

Table 1: Results with 1% training data on *SemEval*. The # column is the number of instances of each relation during training, and TC-PA denotes the *TACRED-PA* model (trained without meta-learning), while *MLRC* denotes the same model trained with our approach.

4.5 Evaluation Metrics

For the *TACRED* dataset, we follow Zhang et al. (2017) and report micro-averaged F1 scores[1]. For the *SemEval* dataset, we report the official measure, which is the F1 score macro-averaged across relations.[2]

4.6 Results and Discussion

The results obtained on the *SemEval* and *TACRED* datasets using *TACRED-PA* as the learner model (f_θ) are shown in Figures 1(a) and 1(b) respectively. We find that on both datasets, our approach improves performance as more supervision becomes available, with the largest gains obtained at the early stage when very limited supervision is available. For instance on *SemEval*, given just 1% of the training set (first datapoint in Figure 1(a)), our approach improves the F1 performance of *TACRED-PA* from 3.13% to 21.05%, representing an absolute increase of 17.92%. Table 1 gives a further breakdown of the F1 scores of individual relations when both approaches are given access to 1% of the training set. We observe that *MLRC* considerably improves the performance of *TACRED-PA* on relations with the least number of training instances, likely by leveraging background knowledge from relations with more training instances. On the *TACRED* dataset, *MLRC* improves the performance of *TACRED-PA* from 2.98% to 34.59% with just 0.5% of the training data (fifth datapoint in Figure 1(b)), which is an absolute increase of 31.61%.

A similar trend is observed using *C-GCN* as the learner model on both datasets, as presented in Figures 2(a) and 2(b). For instance on *SemEval*, we improve the F1 performance of *C-GCN* from 3.38% to 17.14% using just 1% of the training data (first datapoint in Figure 2(a)). Similarly on *TACRED*, the performance of *C-GCN* is improved from 7.59% to 23.18% (first datapoint in Figure 2(b)) by using 0.1% of its training set.

Further, we find that the proposed approach does not adversely affect performance when full supervision is available during training. For instance, when given full supervision on the *TACRED* dataset, while *TACRED-PA* obtains an F1 score of 65.1%, its performance is improved to 65.2% by using our approach, demonstrating that the proposed approach does not adversely affect performance when provided full supervision during training.

[1]We use the same evaluation script as Zhang et al. (2017).

[2]We compute these measures using the official evaluation script that comes with the dataset.

5 Conclusion and Future Work

We show that the performance of supervised relation classification models can be improved, even with limited supervision at training time, by framing relation classification as an instance of meta-learning, and proposed a model-agnostic learning protocol for training relation classifiers with enhanced predictive performance in limited supervision settings. In future work, we want to extend this approach to other natural language processing tasks.

Acknowledgements

The authors acknowledge support from the EU H2020 SUMMA project (grant agreement number 688139). We are grateful to Yuhao Zhang for sharing his data with us.

References

Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. 2016. Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems*, pages 3981–3989.

Razvan Bunescu and Raymond J Mooney. 2005. A shortest path dependency kernel for relation extraction. *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, B.C.*

Thomas Demeester, Tim Rocktäschel, and Sebastian Riedel. 2016. Lifted rule injection for relation embeddings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1389–1399, Austin, Texas. Association for Computational Linguistics.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135, International Convention Centre, Sydney, Australia.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. 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*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Kazuma Hashimoto, Makoto Miwa, Yoshimasa Tsuruoka, and Takashi Chikayama. 2013. Simple Customization of Recursive Neural Networks for Semantic Relation Classification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1372–1376. Association for Computational Linguistics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. 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*, pages 94–99. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pages 1003–1011.

Thien Huu Nguyen and Ralph Grishman. 2015. Relation Extraction: Perspective from Convolutional Neural Networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 39–48, Denver, Colorado. Association for Computational Linguistics.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *CoRR, abs/1803.02999*.

Abiola Obamuyide and Andreas Vlachos. 2017. Contextual pattern embeddings for one-shot relation extraction. In *Proceedings of the NeurIPS 2017 Workshop on Automated Knowledge Base Construction (AKBC)*.

Abiola Obamuyide and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the EMNLP 2018 Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78, Brussels, Belgium. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Sachin Ravi and Hugo Larochelle. 2017. Optimization As a Model for Few-Shot Learning. In *International Conference on Learning Representations 2017*.

Steffen Rendle. 2010. Factorization machines. *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 995–1000.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. 2013. Relation Extraction with Matrix Factorization and Universal Schemas. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–84.

Tim Rocktäschel, Sameer Singh, and Sebastian Riedel. 2015. Injecting Logical Background Knowledge into Embeddings for Relation Extraction. *North American Association for Computational Linguistics*, pages 1119–1129.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic Compositionality through Recursive Matrix-Vector Spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211. Association for Computational Linguistics.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. 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*, pages 455–465. Association for Computational Linguistics.

Sebastian Thrun and Lorien Pratt. 1998. Learning to Learn: Introduction and Overview. In *Learning to Learn*, pages 3–17. Springer US, Boston, MA.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638.

Ngoc Thang Vu, Heike Adel, Pankaj Gupta, and Hinrich Schütze. 2016. Combining Recurrent and Convolutional Neural Networks for Relation Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 534–539. Association for Computational Linguistics.

Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015. 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*, pages 1785–1794, Lisbon, Portugal. Association for Computational Linguistics.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *The Journal of Machine Learning Research*, 3:1083–1106.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation Classification via Convolutional Deep Neural Network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Dongxu Zhang and Dong Wang. 2015. Relation classification via recurrent neural network. *arXiv preprint arXiv:1508.01006*.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45.