Extracting Multiple-Relations in One-Pass with Pre-Trained Transformers

Haoyu Wang[*†] **Ming Tan**[*†] **Mo Yu**[*‡] **Shiyu Chang**[‡]
Dakuo Wang[‡] **Kun Xu**[§] **Xiaoxiao Guo**[‡] **Saloni Potdar**[†]
[†]IBM Watson [‡]IBM Research [§]Tencent AI Lab

Abstract

The state-of-the-art solutions for extracting multiple entity-relations from an input paragraph always require a multiple-pass encoding on the input. This paper proposes a new solution that can complete the multiple entity-relations extraction task with only **one-pass encoding** on the input corpus, and achieve a **new state-of-the-art accuracy performance**, as demonstrated in the ACE 2005 benchmark. Our solution is built on top of the pre-trained self-attentive models (Transformer). Since our method uses a single-pass to compute all relations at once, it scales to larger datasets easily; which makes it more usable in real-world applications. [1]

1 Introduction

Relation extraction (RE) aims to find the semantic relation between a pair of entity mentions from an input paragraph. A solution to this task is essential for many downstream NLP applications such as automatic knowledge-base completion (Surdeanu et al., 2012; Riedel et al., 2013; Verga et al., 2016), knowledge base question answering (Yih et al., 2015; Xu et al., 2016; Yu et al., 2017), and symbolic approaches for visual question answering (Mao et al., 2019; Hu et al., 2019), *etc*.

One particular type of the RE task is *multiple-relations extraction (MRE)* that aims to recognize relations of multiple pairs of entity mentions from an input paragraph. Because in real-world applications, whose input paragraphs dominantly contain multiple pairs of entities, an efficient and effective solution for MRE has more important and more practical implications. However, nearly all existing approaches for MRE tasks (Qu et al., 2014; Gormley et al., 2015; Nguyen and Grishman, 2015) adopt some variations of the single-relation extraction (SRE) approach, which treats each pair of entity mentions as an independent instance, and requires multiple passes of encoding for the multiple pairs of entities. The drawback of this approach is obvious – it is computationally expensive and this issue becomes more severe when the input paragraph is large, making this solution impossible to implement when the encoding step involves deep models.

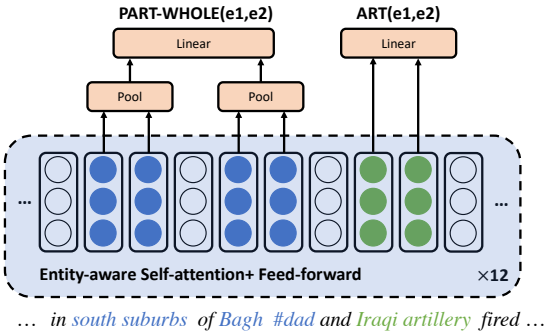

Figure 1: Model Architecture. Different pairs of entities, e.g., (*Iraqi* and *artillery*), (*southern suburbs*, *Baghdad*) are predicted simultaneously.

This work presents a solution that can resolve the inefficient multiple-passes issue of existing solutions for MRE by encoding the input only once, which significantly increases the efficiency and scalability. Specifically, the proposed solution is built on top of the existing transformer-based, pre-trained general-purposed language encoders. In this paper we use *Bidirectional Encoder Representations from Transformers (BERT)* (Devlin et al., 2018) as the transformer-based encoder, but this solution is not limited to using BERT alone. The two novel modifications to the original BERT architecture are: (1) we introduce a structured prediction layer for predicting multiple relations for different entity pairs; and (2) we make the self-attention layers aware of the positions of all en-

[*] Equal contributions from the corresponding authors: {wanghaoy,mingtan,yum}@us.ibm.com. Part of work was done when Kun was at IBM.

[1] https://github.com/helloeve/mre-in-one-pass.

tities in the input paragraph. To the best of our knowledge, this work is the first promising solution that can solve MRE tasks with such high efficiency (encoding the input in one-pass) and effectiveness (achieve a new state-of-the-art performance), as proved on the ACE 2005 benchmark.

2 Background

MRE is an important task as it is an essential prior step for many downstream tasks such as automatic knowledge-base completion and question-answering. Popular MRE benchmarks include ACE (Walker et al., 2006) and ERE (Linguistic Data Consortium, 2013). In MRE, given as a text paragraph $\mathbf{x} = \{x_1, \ldots, x_N\}$ and M mentions $\mathbf{e} = \{e_1, \ldots, e_M\}$ as input, the goal is to predict the relation r_{ij} for each mention pair (e_i, e_j) either belongs to one class of a list of pre-defined relations $\mathcal{R}$ or falls into a special class *NA* indicating no relation. This paper uses "entity mention", "mention" and "entity" interchangeably.

Existing MRE approaches are based on either feature and model architecture selection techniques (Xu et al., 2015; Gormley et al., 2015; Nguyen and Grishman, 2015; F. Petroni and Gemulla, 2015; Sorokin and Gurevych, 2017; Song et al., 2018b), or domain adaptations approaches (Fu et al., 2017; Shi et al., 2018). But these approaches require multiple passes of encoding over the paragraph, as they treat a MRE task as multiple passes of a SRE task.

3 Proposed Approach

This section describes the proposed one-pass encoding MRE solution. The solution is built upon BERT with a structured prediction layer to enable BERT to predict multiple relations with one-pass encoding, and an entity-aware self-attention mechanism to infuse the relational information with regard to multiple entities at each layer of hidden states. The framework is illustrated in Figure 1. It is worth mentioning that our solution can easily use other transformer-based encoders besides BERT, *e.g.* (Radford et al., 2018).

3.1 Structured Prediction with BERT for MRE

The BERT model has been successfully applied to various NLP tasks. However, the final prediction layers used in the original model is not applicable to MRE tasks. The MRE task essentially requires to perform edge predictions over a graph with entities as nodes. Inspired by (Dozat and Manning, 2018; Ahmad et al., 2018), we propose that we can first encode the input paragraph using BERT. Thus, the representation for a pair of entity mentions (e_i, e_j) can be denoted as $\mathbf{o}_i$ and $\mathbf{o}_j$ respectively. In the case of a mention e_i consist of multiple hidden states (due to the byte pair encoding), $\mathbf{o}_i$ is aggregated via average-pooling over the hidden states of the corresponding tokens in the last BERT layer. We then concatenate $\mathbf{o}_i$ and $\mathbf{o}_j$ denoted as $[\mathbf{o}_i : \mathbf{o}_j]$, and pass it to a linear classifier[2] to predict the relation

$$P(r_{ij}|\mathbf{x}, e_i, e_j) = \text{softmax}(\mathrm{W}^L[\mathbf{o}_i : \mathbf{o}_j] + \mathbf{b}), \quad (1)$$

where $\mathrm{W}^L \in \mathbb{R}^{2d_z \times l}$. d_z is the dimension of BERT embedding at each token position, and l is the number of relation labels.

3.2 Entity-Aware Self-Attention based on Relative Distance

This section describes how we encode multiple-relations information into the model. The key concept is to use the relative distances between words and entities to encode the positional information for each entity. This information is propagated through different layers via attention computations. Following (Shaw et al., 2018), for each pair of word tokens (x_i, x_j) with the input representations from the previous layer as $\mathbf{h}_i$ and $\mathbf{h}_j$, we extend the computation of self-attention $\mathbf{z}_i$ as:

$$\mathbf{z}_i = \sum_{j=1}^{N} \frac{\exp e_{ij}}{\sum_{k=1}^{N} \exp e_{ik}} (\mathbf{h}_j \mathrm{W}^V + \mathbf{a}_{ij}^V), \quad (2)$$

$$\text{where} \quad e_{ij} = \mathbf{h}_i \mathrm{W}^Q (\mathbf{h}_j \mathrm{W}^K + \mathbf{a}_{ij}^K)/\sqrt{d_z}. \quad (3)$$

$\mathrm{W}^Q, \mathrm{W}^K, \mathrm{W}^V \in \mathbb{R}^{d_z \times d_z}$ are the parameters of the model, and d_z is the dimension of the output from the self-attention layer.

Compared to standard BERT's self-attention, $\mathbf{a}_{ij}^V, \mathbf{a}_{ij}^K \in \mathbb{R}^{d_z}$ are extra, which could be viewed as the edge representation between the input element x_i and x_j . Specifically, we devise $\mathbf{a}_{ij}^V$ and $\mathbf{a}_{ij}^K$ to encourage each token to be aware of the relative distance to different entity mentions, and vice versa.

[2] We also tried to use MLP and Biaff instead of the linear layer for the classification, which do not show better performance compared to the linear classier, as shown in the experiment section. We hypothesize that this is because the embeddings learned from BERT are powerful enough for linear classifiers. Further experiments is needed to verify this.

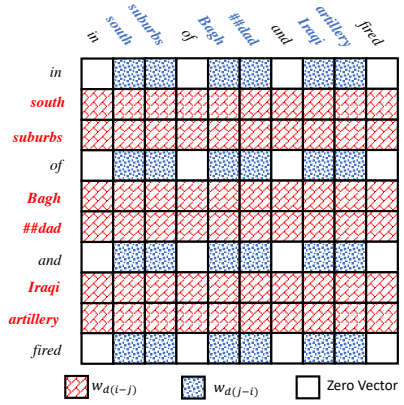

Figure 2: Illustration of the tensor $\{\mathbf{a}_{ij}^K\}$ introduced in self-attention computation. Each red cell embedding is defined by $w_{d(i-j)}$, as the distance from entity x_i to token x_j. Each blue cell embedding is defined by $w_{d(j-i)}$, as the distance from the entity x_j to token x_i . White cells are zero embeddings since neither x_i nor x_j is entity. The $\{\mathbf{a}_{ij}^V\}$ follows the same pattern with independent parameters.

Adapted from (Shaw et al., 2018), we argue that the relative distance information will not help if the distance is beyond a certain threshold. Hence we first define the distance function as:

$$d(i,j) = min(max(-k,(i-j)),k). \quad (4)$$

This distance definition clips all distances to a region $[-k, k]$. k is a hyper-parameter to be tuned on the development set. We can now define $\mathbf{a}_{ij}^V$ and $\mathbf{a}_{ij}^K$ formally as:

$$\mathbf{a}_{ij}^V, \mathbf{a}_{ij}^K = \begin{cases} w_{d(i,j)}^V, w_{d(i,j)}^K, & \text{if } x_i \in e \\ w_{d(j,i)}^V, w_{d(j,i)}^K, & \text{if } x_j \in e \\ \mathbf{0}, & \text{else.} \end{cases} \quad (5)$$

As defined above, if either token x_i or x_j belongs to an entity, we will introduce a relative positional representation according to their distance. The distance is defined in an entity-centric way as we always compute the distance from the entity mention to the other token. If neither x_i nor x_j are entity mentions, we explicitly assign a zero vector to $\mathbf{a}_{ij}^K$ and $\mathbf{a}_{ij}^V$. When both x_i and x_j are inside entity mentions, we take the distance as $d(i,j)$ to make row-wise attention computation coherent as depicted in Figure 2.

During the model fine-tuning, the newly introduced parameters $\{w_{-k}^K, ..., w_k^K\}$ and $\{w_{-k}^V, ..., w_k^V\}$ are trained from scratch.

4 Experiments

We demonstrate the advantage of our method on a popular MRE benchmark, ACE 2005 (Walker et al., 2006), and a more recent MRE benchmark, SemEval 2018 Task 7 (Gábor et al., 2018). We also evaluate on a commonly used SRE benchmark SemEval 2010 task 8 (Hendrickx et al., 2009), and achieve state-of-the-art performance.

4.1 Settings

Data For **ACE 2005**, we adopt the multi-domain setting and split the data following (Gormley et al., 2015): we train on the union of news domain (nw and bn), tune hyperparameters on half of the broadcast conversation (bc) domain, and evaluate on the remainder of broadcast conversation (bc), the telephone speech (cts), usenet newsgroups (un), and weblogs (wl) domains. For **SemEval 2018 Task 7**, we evaluate on its sub-task 1.1. We use the same data split in the shared task. The passages in this task is usually much longer compared to ACE. Therefore we adopt the following pre-processing step – for the entity pair in each relation, we assume the tokens related to their relation labeling are always within a range from the fifth token ahead of the pair to the fifth token after it. Therefore, the tokens in the original passage that are not covered by the range of ANY input relations, will be removed from the input.

Methods We compare our solution with previous works that predict a single relation per pass (Gormley et al., 2015; Nguyen and Grishman, 2015; Fu et al., 2017; Shi et al., 2018), our model that predicts single relation per pass for MRE, and with the following naive modifications of BERT that could achieve MRE in one-pass.

• **BERT$_{\text{SP}}$**: BERT with structured prediction only, which includes proposed improvement in 3.1.

• **Entity-Aware BERT$_{\text{SP}}$**: our full model, which includes both improvements in §3.1 and §3.2.

• **BERT$_{\text{SP}}$ with position embedding on the final attention layer**. This is a more straightforward way to achieve MRE in one-pass derived from previous works using position embeddings (Nguyen and Grishman, 2015; Fu et al., 2017; Shi et al., 2018). In this method, the BERT model encode the paragraph to the last attention-layer. Then, for each entity pair, it takes the hidden states, adds the relative position embeddings corresponding to the target entities, and finally makes the relation prediction for this pair.

• **BERT$_{\text{SP}}$ with entity indicators on input layer**: it replaces our structured attention layer, and adds indicators of entities (transformed to embeddings)

Method	dev	bc	cts	wl	avg
Baselines w/o Domain Adaptation (Single-Relation per Pass)					
Hybrid FCM (Gormley et al., 2015)	-	63.48	56.12	55.17	58.26
Best published results w/o DA (from Fu et al.)	-	64.44	54.58	57.02	58.68
BERT fine-tuning out-of-box	3.66	5.56	5.53	1.67	4.25
Baselines w/ Domain Adaptation (Single-Relation per Pass)					
Domain Adversarial Network (Fu et al., 2017)	-	65.16	55.55	57.19	59.30
Genre Separation Network (Shi et al., 2018)	-	66.38	57.92	56.84	60.38
Multi-Relation per Pass					
BERT_{SP} (our model in §3.1)	64.42	67.09	53.20	52.73	57.67
Entity-Aware BERT_{SP} (our full model)	**67.46**	**69.25**	**61.70**	**58.48**	**63.14**
BERT_{SP} w/ entity-indicator on input-layer	65.32	66.86	57.65	53.56	59.36
BERT_{SP} w/ pos-emb on final att-layer	67.23	69.13	58.68	55.04	60.95
Single-Relation per Pass					
BERT_{SP} (our model in §3.1)	65.13	66.95	55.43	54.39	58.92
Entity-Aware BERT_{SP} (our full model)	68.90	68.52	63.71	57.20	63.14
BERT_{SP} w/ entity-indicator on input-layer	67.12	69.76	58.05	56.27	61.36

Table 1: Main Results on ACE 2005.

directly to each token's word embedding[3]. This method is an extension of (Verga et al., 2018) to the MRE scenario.

Hyperparameters For our experiments, most model hyperparameters are the same as in pre-training. We tune the training epochs and the new hyperparameter k (in Eq. 4) on the development set of ACE 2005. Since the SemEval task has no development set, we use the best hyperparameters selected on ACE. For the number of training epochs, we make the model pass similar number of training instances as in ACE 2005.

4.2 Results on ACE 2005

Main Results Table 1 gives the overall results on ACE 2005. The first observation is that our model architecture achieves much better results compared to the previous state-of-the-art methods. Note that our method was not designed for domain adaptation, it still outperforms those methods with domain adaptation. This result further demonstrates its effectiveness.

Among all the BERT-based approaches, fine-tuning the off-the-shelf BERT does not give a satisfying result, because the sentence embeddings cannot distinguish different entity pairs. The simpler version of our approach, BERT_{SP}, can successfully adapt the pre-trained BERT to the MRE task, and achieves comparable performance at the prior state-of-the-art level of the methods without domain adaptation.

Our full model, with the structured fine-tuning of attention layers, brings further improvement of about 5.5%, in the MRE one-pass setting, and achieves a new state-of-the-art performance when compared to the methods with domain adaptation. It also beats the other two methods on BERT in Multi-Relation per Pass.

Performance Gap between MRE in One-Pass and Multi-Pass The MRE-in-one-pass models can also be used to train and test with one entity pair per pass (*Single-Relation per Pass* results in Table 1). Therefore, we compare the same methods when applied to the multi-relation and single-relation settings. For BERT_{SP} with entity indicators on inputs, it is expected to perform slightly better in the single-relation setting, because of the mixture of information from multiple pairs. A 2% gap is observed as expected. By comparison, our full model has a much smaller performance gap between two different settings (and no consistent performance drop over different domains).

The BERT_{SP} is not expected to have a gap as shown in the table.For BERT_{SP} with position embeddings on the final attention layer, we train the model in the single-relation setting and test with two different settings, so the results are the same.

Training and Inference Time Through our experiment,[4] we verify that the full model with MRE is significantly faster compared to all other methods for both training and inference. The training

[3]Note the usage of relative position embeddings does not work for one-pass MRE, since each word corresponds to a varying number of position embedding vectors. Summing up the vectors confuses this information. It works for the single-relation per pass setting, but the performance lags behind using only indicators of the two target entities.

[4]All evaluations were done on a single Tesla K80 GPU.

Method	dev	bc	cts	wl	avg
Linear	**67.46**	**69.25**	**61.70**	**58.48**	**63.14**
MLP	67.16	68.52	61.16	54.72	61.47
Biaff	67.06	68.22	60.39	55.60	61.40

Table 2: Our model with different prediction modules.

time for full model with MRE is 3.5x faster than it with SRE. As for inference speed, the former could reach 126 relation per second compared the later at 23 relation per second. It is also much faster when compared to the second best performing approach, *BERT$_{SP}$ w/ pos-emb on final att-layer*, which is at 76 relation per second, as it runs the last layer for every entity pair.

Prediction Module Selection Table 2 evaluates the usage of different prediction layers, including replacing our linear layer in Eq.(1) with MLP or Biaff. Results show that the usage of the linear predictor gives better results. This is consistent with the motivation of the pre-trained encoders: by unsupervised pre-training the encoders are expected to be sufficiently powerful thus adding more complex layers on top does not improve the capacity but leads to more free parameters and higher risk of over-fitting.

4.3 Results on SemEval 2018 Task 7

The results on SemEval 2018 Task 7 are shown in Table 3. Our Entity-Aware BERT$_{SP}$ gives comparable results to the top-ranked system (Rotsztejn et al., 2018) in the shared task, with slightly lower Macro-F1, which is the official metric of the task, and slightly higher Micro-F1. When predicting multiple relations in one-pass, we have 0.9% drop on Macro-F1, but a further 0.8% improvement on Micro-F1. Note that the system (Rotsztejn et al., 2018) integrates many techniques like feature-engineering, model combination, pre-training embeddings on in-domain data, and artificial data generation, while our model is almost a direct adaption from the ACE architecture.

On the other hand, compared to the top single-model result (Luan et al., 2018), which makes use of additional word and entity embeddings pre-trained on in-domain data, our methods demonstrate clear advantage as a single model.

4.4 Additional SRE Results

We conduct additional experiments on the relation classification task, SemEval 2010 Task 8, to compare with models developed on this benchmark. From the results in Table 4, our proposed techniques also outperforms the state-of-the-art on this single-relation benchmark.

Method	Averaged F1 Macro	Averaged F1 Micro
Top 3 in the Shared Task		
(Rotsztejn et al., 2018)	**81.7**	82.8
(Luan et al., 2018)	78.9	-
(Nooralahzadeh et al., 2018)	76.7	-
Ours (single-per-pass)	81.4	83.1
Ours (multiple-per-pass)	80.5	**83.9**

Table 3: Results on SemEval 2018 Task 7, Sub-Task 1.1.

Method	Macro-F1
Best published result (Wang et al., 2016)	88.0
BERT out-of-box	80.9
Entity-Aware BERT	88.8
BERT$_{SP}$	88.8
Entity-Aware BERT$_{SP}$	**89.0**

Table 4: Additional Results on SemEval 2010 Task 8.

On this single relation task, the out-of-box BERT achieves a reasonable result after fine-tuning. Adding the entity-aware attention gives about 8% improvement, due to the availability of the entity information during encoding. Adding structured prediction layer to BERT (i.e., BERT$_{SP}$) also leads to a similar amount of improvement. However, the gap between BERT$_{SP}$ method with and without entity-aware attention is small. This is likely because of the bias of data distribution: the assumption that only two target entities exist, makes the two techniques have similar effects.

5 Conclusion

In summary, we propose a first-of-its-kind solution that can simultaneously extract multiple relations with one-pass encoding of an input paragraph for MRE tasks. With the proposed structured prediction and entity-aware self-attention layers on top of BERT, we achieve a new state-of-the-art results with high efficiency on the ACE 2005 benchmark. Our idea of encoding a passage regarding multiple entities has potentially broader applications beyond relation extraction, e.g., entity-centric passage encoding in question answering (Song et al., 2018a). In the future work, we will explore the usage of this method with other applications.

References

Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2018. Near or far, wide range zero-shot cross-lingual dependency parsing. *arXiv preprint arXiv:1811.00570*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Timothy Dozat and Christopher D Manning. 2018. Simpler but more accurate semantic dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 484–490.

L. Del Corro F. Petroni and R. Gemulla. 2015. Core: Context-aware open relation extraction with factorization machines. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Lisheng Fu, Thien Huu Nguyen, Bonan Min, and Ralph Grishman. 2017. Domain adaptation for relation extraction with domain adversarial neural network. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 425–429.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haifa Zargayouna, and Thierry Charnois. 2018. Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 679–688.

Matthew R Gormley, Mo Yu, and Mark Dredze. 2015. Improved relation extraction with feature-rich compositional embedding models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1774–1784.

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.

Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. 2019. Language-conditioned graph networks for relational reasoning. *arXiv preprint arXiv:1905.04405*.

Linguistic Data Consortium. 2013. Deft ere annotation guidelines: Relations v1.1. 05.17.2013.

Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. The UWNLP system at SemEval-2018 task 7: Neural relation extraction model with selectively incorporated concept embeddings. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 788–792, New Orleans, Louisiana. Association for Computational Linguistics.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*.

Thien Huu Nguyen and Ralph Grishman. 2015. Combining neural networks and log-linear models to improve relation extraction. *arXiv preprint arXiv:1511.05926*.

Farhad Nooralahzadeh, Lilja Øvrelid, and Jan Tore Lønning. 2018. SIRIUS-LTG-UiO at SemEval-2018 task 7: Convolutional neural networks with shortest dependency paths for semantic relation extraction and classification in scientific papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, New Orleans, Louisiana. Association for Computational Linguistics.

Lizhen Qu, Yi Zhang, Rui Wang, Lili Jiang, Rainer Gemulla, and Gerhard Weikum. 2014. Senti-lssvm: Sentiment-oriented multi-relation extraction with latent structural svm. *Transactions of the Association for Computational Linguistics*, 2:155–168.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–84.

Jonathan Rotsztejn, Nora Hollenstein, and Ce Zhang. 2018. ETH-DS3Lab at SemEval-2018 task 7: Effectively combining recurrent and convolutional neural networks for relation classification and extraction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, New Orleans, Louisiana. Association for Computational Linguistics.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *NAACL-HLT*, page 464468.

Ge Shi, Chong Feng, Lifu Huang, Boliang Zhang, Heng Ji, Lejian Liao, and Heyan Huang. 2018. Genre separation network with adversarial training for cross-genre relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1018–1023.

Linfeng Song, Zhiguo Wang, Mo Yu, Yue Zhang, Radu Florian, and Daniel Gildea. 2018a. Exploring graph-structured passage representation for multi-hop reading comprehension with graph neural networks. *arXiv preprint arXiv:1809.02040*.

Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018b. N-ary relation extraction using graph-state lstm. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2226–2235.

Daniil Sorokin and Iryna Gurevych. 2017. Context-Aware Representations for Knowledge Base Relation Extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1784–1789. 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.

Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth, and Andrew McCallum. 2016. Multilingual relation extraction using compositional universal schema. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 886–896.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *NAACL 2018*, pages 872–884.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57.

Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention cnns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1298–1307.

Kun Xu, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2015. Semantic relation classification via convolutional neural networks with simple negative sampling. *arXiv preprint arXiv:1506.07650*.

Kun Xu, Siva Reddy, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2016. Question answering on freebase via relation extraction and textual evidence. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2326–2336.

Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic parsing via staged query graph generation: Question answering with knowledge base. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1321–1331.

Mo Yu, Wenpeng Yin, Kazi Saidul Hasan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2017. Improved neural relation detection for knowledge base question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 571–581.