Three Sentences Are All You Need: Local Path Enhanced Document Relation Extraction

Quzhe Huang, Shengqi Zhu, Yansong Feng[*], Yuan Ye, Yuxuan Lai, Dongyan Zhao
Wangxuan Institute of Computer Technology, Peking University, China
The MOE Key Laboratory of Computational Linguistics, Peking University, China
{huangquzhe, zhusq, fengyansong, pkuyeyuan, erutan, zhaody}
@pku.edu.cn

Abstract

Document-level Relation Extraction (RE) is a more challenging task than sentence RE as it often requires reasoning over multiple sentences. Yet, human annotators usually use a small number of sentences to identify the relationship between a given entity pair. In this paper, we present an embarrassingly simple but effective method to heuristically select evidence sentences for document-level RE, which can be easily combined with BiLSTM to achieve good performance on benchmark datasets, even better than fancy graph neural network based methods. We have released our code at https://github.com/AndrewZhe/Three-Sentences-Are-All-You-Need.

1 Introduction

The task of relation extraction (RE) focuses on extracting relations between entity pairs in texts, and has played an important role in information extraction. While earlier works focus on extracting relations within a sentence (Lin et al., 2016; Zhang et al., 2018), recent studies begin to explore RE at document level (Peng et al., 2017; Zeng et al., 2020a; Nan et al., 2020a), which is more challenging as it often requires reasoning across multiple sentences.

Espoo Cathedral
[1] *The Espoo Cathedral* is a medieval stone church in *Espoo*, *Finland* and the seat of the Diocese of Espoo of the Evangelical Lutheran Church of Finland. **[2]** The cathedral is located in the district of *Espoon keskus*, near the Espoonjoki river. ... **[6]** In addition to being the seat of the Diocese of Espoo, *the Espoo Cathedral* serves as the church for *the EC Parish* and hosts ...

Subject:	*Espoo*	*The Espoo Cathedral*	*the EC Parish*
Object:	*Finland*	*Espoon keskus*	*Finland*
Relation:	**country**	**location**	**country**
Evidence:	**[1]**	**[1], [2]**	**[1], [6]**

Figure 1: A case extracted from the DocRED dataset. While the document has 6 sentences, only 1 or 2 sentences form the evidence for each relation instance.

Compared with sentence level extraction, documents are significantly longer with useful information scattered in a larger scale. However, given a pair of entities, one may only need a few sentences, not the entire document, to infer their relationship; reading the whole document may not be necessary, since it may introduce unrelated information inevitably. As we can see in Figure 1, $S[1]$ is sufficient to recognize *Finland* as the country of *Espoo*, and recognizing the rest two instances requires just 2 sentences as supporting evidence as well. Although the document contains 6 sentences and evidence may span from $S[1] \sim S[6]$, identifying *each* relation instance can be achieved by just reading through 1 or 2 related sentences. This naturally leads us to consider a question: *given an entity pair, how many sentences are required to identify a relationship between them?* We perform a pilot study across 3 widely-used document RE datasets, DocRED (Yao et al., 2019), CDR (Li et al., 2016) and GDA (Wu et al., 2019). As shown in Table 1, we find that more than 95% instances require no more than 3 sentences as supporting evidence, and 87% even requires only 2 or less.

Our preliminary finding suggests that, instead of taking the entire document as context, a case-specific selection may be more useful to help a model focus on the most relevant and informative evidence. Previous studies apply graph neural networks (GNNs) for this filtering process (Christopoulou et al., 2019; Zeng et al., 2020b). Here, GNNs are used to collect relevant information from the entire context through an aggregation scheme (Nan et al., 2020a) and achieve great performance, but the selection of crucial evidence from documents is still implicit and lacks interpretability. If, as indicated by our pilot study, most entity relationships can be decided with just $1 \sim 3$ evidence sentences, is there a simpler method that can filter the document explicitly while maintaining the

[*] Corresponding author.

	0	1	2	3	>=4	# Sent
DocRED	3.7%	49.7%	34.3%	8.4%	3.8%	8.0
CDR	0.0%	68.0%	30.0%	0.0%	2.0%	9.7
GDA	0.0%	66.0%	19.0%	3.0%	5.0%	10.2

Table 1: The proportion of instances with different supporting evidence sizes. # Sent shows the average number of sentences in a document.

crucial information?

We take a closer look at how entity pairs are contextually related in the annotated supporting evidence, and find that annotators tend to select sentences that can connect the two entities. We therefore design three heuristic rules to extract a small set of *paths* from the document, which can be seen as an approximation of the supporting evidence. Specifically, the *Consecutive Paths* consider the scenario where the head and tail entities are close in the context: if they are within 3 consecutive sentences, we regard these sentences as one path. The *Multi-Hop Paths* correspond to the entity pairs in distant sentences, which can be bridged via other entities that co-occur with the head entity and tail entity in different sentences. As the third relation in Figure 1 shows, *Finland* co-occurs with *The Espoo Cathedral* in S[1] and with *the EC Parish* in S[6], which makes it a bridge to connect *The Espoo Cathedral* and *the EC Parish*. In this case, S[1] and S[6] compose a multi-hop path. When neither of the above rules applies, we collect all the pairs of sentences where one contains head entity and the other contains tail entity as *Default Paths*.

By comparing our path set with human-annotated supporting evidence, we find that up to 87.5% of the supporting evidence can be fully covered by our heuristically selected paths. In other words, our straightforward and interpretable rules serve as an effective proxy to select supporting evidence from documents. We further feed our selected paths to a simple neural network model and obtain surprisingly good performance on DocRED, showing that our selected evidence can retain sufficient information from the entire document to support document-level relation extraction.

2 Do we need the entire document?

For document RE, the major challenge is that the subject and object involved in a relationship may appear in different sentences. Thus, more than one sentence is required to capture the relations. Nonetheless, how many sentences from the entire document are required to identify the relationship between an entity pair? To address this question, we analyze the supporting evidence presented in DocRED. The supporting evidence for a relation instance refers to all the sentences that can be used to decide whether this relation holds between the entity pair, labeled by human annotators (Yao et al., 2019). Table 1 shows the proportions of entity relation instances with different number of supporting sentences. As can be seen,

more than 96% of the DocRED instances are associated with at most 3 supporting evidence. These only take up 37.5% of a document, since the average document length is 8 sentences. This means that reading a small part of a document is adequate for one to identify an entity relation instance.

We further extend our study to two widely used document RE datasets, CDR (Li et al., 2016) and GDA (Wu et al., 2019), where CDR is manually constructed and GDA is distantly supervised. In order to find the minimal number of sentences required, we ask annotators to label a minimal set of sentences that are exactly sufficient to identify an entity relation instance, instead of including all relation-associated sentences as the original DocRED pattern. We randomly select 100 instances respectively from CDR and GDA for this further annotation, and the results are shown at the bottom of Table 1[1]. Although the average length of documents in GDA and CDR are longer than DocRED, it turns out that one can still use no more than 3 supporting sentences to identify over 95% of the entity relation instances. The results on CDR and GDA confirm our previous finding that, a very small number of sentences (or more exactly, no more than 3 sentences) would make it sufficient for human annotators to recognize almost all entity relation instances in a document in widely-used benchmark datasets.

3 Which sentences are decisive?

Now our question is how to select the supporting sentences that are sufficient to identify an entity relation instance. Intuitively, the supporting evidence should be the sentences that build up the *connection* between a pair of entities. Thus, we aim to extract sentence *paths* from the head entity to the tail entity to describe how they are connected. As for the simplest case, if there exists one sentence that contains

[1] As GDA is a distantly supervised dataset, 7 instances that are found wrongly labeled are discarded.

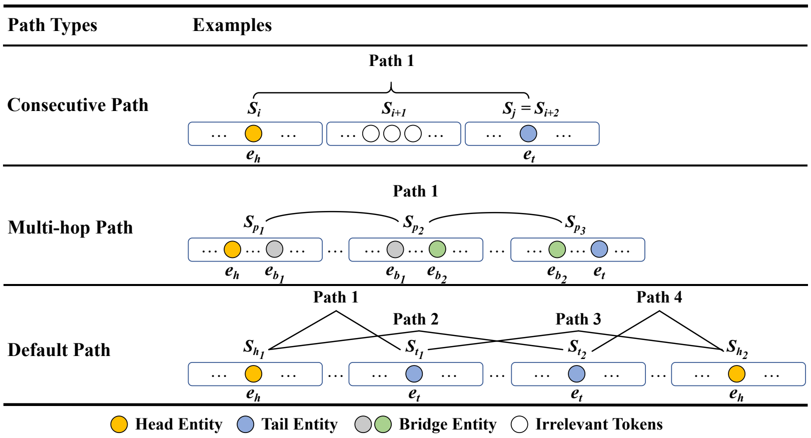

Figure 2: Types of paths connecting head and tail entities. The rounded rectangles represent sentences and the circles are mentions of involved entities or other irrelevant tokens. e_h and e_t stands for a mention of head and tail entities respectively, and S_* represents a sentence.

both the head and tail entities, the sentence itself can be seen as a path (the intra-sentence case). For more complex situations where the head and tail entities do not co-occur in one sentence, we define the following 3 types of paths which indicate how the head and tail entities can be possibly related in the context. Figure 2 provides a visualization of the three types of paths.

Consecutive Paths Previous studies have shown that the majority of inter-sentence relations are often in nearby text (Swampillai and Stevenson, 2010; Quirk and Poon, 2017). We thus select the consecutive sentences to form a path when the head and tail entities are in nearby sentences. Formally, if one mention of the head entity appears in sentence S_i and one mention of the tail entity is in sentence S_j, these two sentences along with the sentence in between, i.e., sentence S_{i+1}, …, S_{j-1} (or S_{j+1}, …, S_{i-1} when $i \geq j$) forms a possible path that connects the two entities. Given that no more than 3 sentences would suffice for inference, we limit the length of these *Consecutive Paths* to be at most 3, which means $|j-i| \leq 2$. Note that this definition can be naturally extended to the intra-sentence case where $j = i$. We thus consider the intra-sentence case as a type of the Consecutive Path. A pair of entities can correspond to multiple consecutive paths since they can be mentioned more than once.

Multi-Hop Paths Another typical case for inter-sentence relation instances is the multi-hop relation (Yao et al., 2019; Zeng et al., 2020a). In such cases, the head and tail entities are far from each other in the document but can be connected through *bridge entities*, just like the entity *The Espoo Cathedral* in Figure 1 bridges *the EC Parish* and *Finland* in sentence 1 and 6.

For these cases, we start from the head entity, go through all the bridge entities, arrive at the tail entity, and select all the corresponding sentences in this route as a path. Formally, for the head entity e_h and the tail entity e_t, the multi-hop relation indicates that there exist a list of bridge entities e_{b_1}, …, e_{b_k} such that $(e_h, e_{b_1}), (e_{b_1}, e_{b_2}), \ldots, (e_{b_k}, e_t)$ form $k+1$ intra-sentence relations respectively in sentence $S_{p_1}, \ldots, S_{p_{k+1}}$. Following this route, we choose these $k+1$ sentences as the *Multi-Hop Path*. Given the discovery in §2 that most instances only needs 3 sentences, we restrict k to be at most 2, i.e., with only 1 or 2 bridge entities. It is possible to have several multi-hop paths for a certain pair with different lists of bridge entities.

Default Paths If neither of the aforementioned rules applies, we consider a rough estimate for the evidence with the most relevant sentences. We collect all pairs of sentences where one contains the head entity and the other contains the tail entity as *Default Paths*. Formally, let $\{S_{h_1}, \ldots, S_{h_p}\}$ and $\{S_{t_1}, \ldots, S_{t_q}\}$ denote the sets of sentences that contain the head entity e_h and the tail entity e_t, respectively. For this entity pair, we will have $p \times q$ Default Paths $\{S_{h_1}, S_{t_1}\}, \ldots, \{S_{h_p}, S_{t_q}\}$. Note that this type of paths is extracted only when no paths are found with the previous two patterns.

4 Comparing with Annotated Evidence

To demonstrate the effectiveness of our heuristic rules, we check the size of our path set on DocRED and their consistency with the gold supporting ev-

	Path Recall	#Sent	#Path
C	71.7%	2.31	1.71
M	31.5%	3.14	2.35
C+M	80.5%	2.73	2.37
C+M+D	87.5%	2.69	2.27
document	-	8.00	-

Table 2: C, M and D stand for Consecutive Paths, Multi-hop Paths, and Default Paths, respectively. #Path and #Sent are the average path numbers and average sentence numbers in the union of all paths.

idence. As mentioned in §2, the gold annotation acts as a collection of all related evidence, while each of our extracted paths represents one possible and minimal sentence set. Ideally, if the path set is sufficient, all connecting sentences between the entity pair should be successfully captured. In other words, they would be presented via various paths in our path set. Therefore, the union of paths is expected to be a superset of the supporting evidence. We use the **Coverage** of the supporting evidence to measure the *sufficiency* of our path set, which stands for the percentage of instances whose supporting evidence is fully covered by the union of our paths. Meanwhile, the total number of paths ($\#Path$) and union size of the paths ($\#Sent$) should also remain at a low standard, so as to avoid *redundancy*.

Table 2 shows the statistics of the path sets extracted via our rules. The Consecutive Paths form a strong baseline that covers 71.7% of instances. Combining the three types, up to 87.5% of instances from the supporting evidence are fully covered by our path sets. The main reason that C+M+D can not cover all the instances is that the supporting evidence annotated in DocRED includes all associated sentences, while C+M+D only find a sufficient set to identify the relation.

Meanwhile, notice that the union of the three types contains only 2.69 different sentences on average, which means that our methods can filter out up to 2/3 of the original text. Also, our method is computationally efficient since only 2.27 paths need to be modeled on average. This demonstrates that our methods form a sufficient and non-redundant estimate for the gold supporting evidence, drastically alleviating the negative impact of irrelevant information.

Model	Dev			Test
	Intra-F1	Inter-F1	F1	F1
CNN	51.87	37.58	43.45	42.26
BiLSTM	57.05	43.49	50.94	51.06
HIN-Glove	60.83	48.35	52.95	53.30
GAT	58.14	43.94	51.44	49.51
GCNN	57.78	44.11	51.52	51.62
EoG	58.90	44.60	52.15	51.82
AGGCN	58.76	45.45	52.47	51.45
LSR-Glove	60.83	48.35	55.17	54.18
GAIN-Glove	61.67	48.77	55.29	55.08
Paths+BiLSTM	**62.73**	**49.11**	**56.54**	**56.23**

Table 3: Model performance on DocRED.

5 Experiments

To further validate the sufficiency of our selected paths, we perform evaluation on DocRED by feeding the paths to an RE model. While previous works take entire documents as input, we replace the document with our selected paths regarding a given entity pair. Intuitively, if the paths can cover all crucial information in the document, we would expect comparable or better performance with identical model architecture, as our paths contain little irrelevant information and may help focus on a few key sentences.

Setup Given a pair of entities, all paths are first extracted as described in §3. Since each path corresponds to one possible connection of the head and tail entities, we predict the relations with each path independently and aggregate the results afterwards.

For every single path c, we concatenate all sentences in it as one segment $[\mathbf{w_1^c}, ..., \mathbf{w_m^c}]$, where the order of sentences is the same as in the original document. The segment is fed to a BiLSTM to obtain the contextual embeddings $[\mathbf{h_1^c}, ..., \mathbf{h_m^c}]$. The representation of an entity mention, which spans from the s-th word to the t-th word, is defined as $\mathbf{m_k^c} = \frac{1}{t-s+1}\sum_{j=s}^{t}\mathbf{h_j^c}$. The representation of an entity e_i^c with K mentions is computed as the average of the representations of its mentions: $\mathbf{e_i^c} = \frac{1}{K}\sum_k \mathbf{m_k^c}$. Then, we use a two-layer perceptron to calculate the probability of each relation r based on the current path c: $P_{ij}^c(r) = \sigma(F([e_i^c; e_j^c; |e_i^c - e_j^c|; e_i^c * e_j^c]))$, where $\sigma(\cdot)$ is the Sigmoid function and $F(\cdot)$ stands for the two-layer perceptron.

After obtaining the prediction of every path between a given entity pair, we aggregate the predicted results by selecting the most likely predictions: $P_{ij}(r) = \max_c P_{ij}^c(r)$.

We use the Glove-100 (Pennington et al., 2014) embedding for the BiLSTM encoder with hidden size 256. Following previous works (Nan et al., 2020b), we report the F1 for intra- and inter-sentence entity pairs along with the overall F1 score as evaluation metrics.

Results We compare our methods with previous sequence-based models and graph-based models. All these models take the entire document as input. As shown in Table 3, our selected path with BiLSTM achieves 56.23% F1 on the test set, which outperforms the sequence-based models. Compared with the baseline BiLSTM, our model brings 5.68% and 5.62% improvement on intra- and inter-sentence entity pairs on the dev set, respectively.

Surprisingly, our simple method achieves a higher performance compared with graph-based models, which are more complex and also possess the ability to filter out irrelevant information. Combined with our path-selection scheme, a BiLSTM can perform 1.25% and 1.15% better on the dev and test set, respectively, compared to the SOTA graph-based model in the same situation. This may indicate that, while graph-based models have shown excellent abilities to focus on important information in a self-adaptive manner, it is more helpful to explicitly select from the document than to fully rely on graph-based models. With a simple filtering scheme inspired by human annotations, we can better explore the potentials of existing models and produce better results.

6 Discussion

So far we have shown from experiments the limited number of sentences required to deduce a relation instance. While the interesting results seem unconventional for Document RE, which features complex inter-sentence relations, it is worth mentioning that possible explanations exist in current works in related fields. The interdisciplinary outlooks may provide helpful insights for community members to understand the causes of the *three-sentences* phenomenon and revisit the problem of Document-level Relation Extraction.

Linguistic Perspective One likely cause of the discussed phenomenon is that the seemingly distant relations are not so difficult given their linguistic form. Stevenson (2006) mentions that a majority of inter-sentence relation instances are in fact due to *co-references* (anaphoric expressions or alternative descriptions). In these cases, relations could be considered to be described entirely within one sentence but with head or tail entities being referred to indirectly. Considering anaphoric expressions are likely to appear in surrounding sentences for the candidate mentions (Chowdhury and Zweigenbaum, 2013), these findings are directly in line with our observation that consecutive paths could support more than 70% relation instances, and provide evidence for *three-sentences* phenomenon.

Cognitive Perspective Another possible explanation is that the RE task is naturally defined within a limited amount of entities and context, given the nature of the human brain. It is widely believed that *Working Memory* (WM) (Baddeley, 1992) plays a vital role to store and manipulate information in inference tasks (Barreyro et al., 2012), but the capacity of separate information chunks in WM are often limited to 4 (Cowan, 2001). As we need to memorize all the separate entities in the inference chain along with their relations, it is natural that we tend to describe a relation within a limited number of sentences, since rendering a relationship with more sentences may cause our WM to exceed its capacity. Daneman and Carpenter (1980) show that the success rate of completing a reading task drastically drops if too much information, exceeding the subject's WM capacity, is required for the task. Therefore, as the datasets are constructed from natural language, the *three-sentences* phenomenon in the data may be a common pattern that we (unconsciously) follow for mutual understanding.

7 Conclusion

In this paper, we perform an analysis over 3 document RE benchmark datasets, and find that human annotators often use a small number of sentences to extract entity relations in document level. This motivates us to think over which sentences are critical for document RE. We carefully design heuristic rules to select informative *path* sets from entire documents, which can be further combined with a simple BiLSTM to achieve competitive performance on a benchmark dataset, even better than complex graph-based methods.

Acknowledgments

We thank the anonymous reviewers for the helpful comments and suggestions. This work is supported in part by the National Hi-Tech R&D Program of China (2018YFC0831900) and the NSFC Grants (No.61672057, 61672058).

References

Alan Baddeley. 1992. Working memory. *Science*, 255(5044):556–559.

Juan Pablo Barreyro, Jazmín Cevasco, Débora Burín, and Carlos Molinari Marotto. 2012. Working memory capacity and individual differences in the making of reinstatement and elaborative inferences. *The Spanish journal of psychology*, 15(2):471.

Md Faisal Mahbub Chowdhury and Pierre Zweigenbaum. 2013. A controlled greedy supervised approach for co-reference resolution on clinical text. *Journal of biomedical informatics*, 46(3):506–515.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4925–4936, Hong Kong, China. Association for Computational Linguistics.

Nelson Cowan. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, 24(1):87–114.

Meredyth Daneman and Patricia A Carpenter. 1980. Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior*, 19(4):450–466.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative V CDR task corpus: A resource for chemical disease relation extraction. *Database: the journal of biological databases and curation*, 2016:baw068.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. 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)*, pages 2124–2133, Berlin, Germany. Association for Computational Linguistics.

Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020a. Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1546–1557, Online. Association for Computational Linguistics.

Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020b. Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1546–1557, Online. Association for Computational Linguistics.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph LSTMs. *Transactions of the Association for Computational Linguistics*, 5:101–115.

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, Doha, Qatar. Association for Computational Linguistics.

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1171–1182, Valencia, Spain. Association for Computational Linguistics.

Mark Stevenson. 2006. Fact distribution in information extraction. *Language resources and evaluation*, 40(2):183–201.

Kumutha Swampillai and Mark Stevenson. 2010. Inter-sentential relations in information extraction corpora. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Ye Wu, Ruibang Luo, Henry CM Leung, Hing-Fung Ting, and Tak-Wah Lam. 2019. Renet: A deep learning approach for extracting gene-disease associations from literature. In *International Conference on Research in Computational Molecular Biology*, pages 272–284. Springer.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020a. Double graph based reasoning for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1630–1640, Online. Association for Computational Linguistics.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020b. Double graph based reasoning for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1630–1640, Online. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.