Incongruity Detection between Bangla News Headline and Body Content through Graph Neural Network

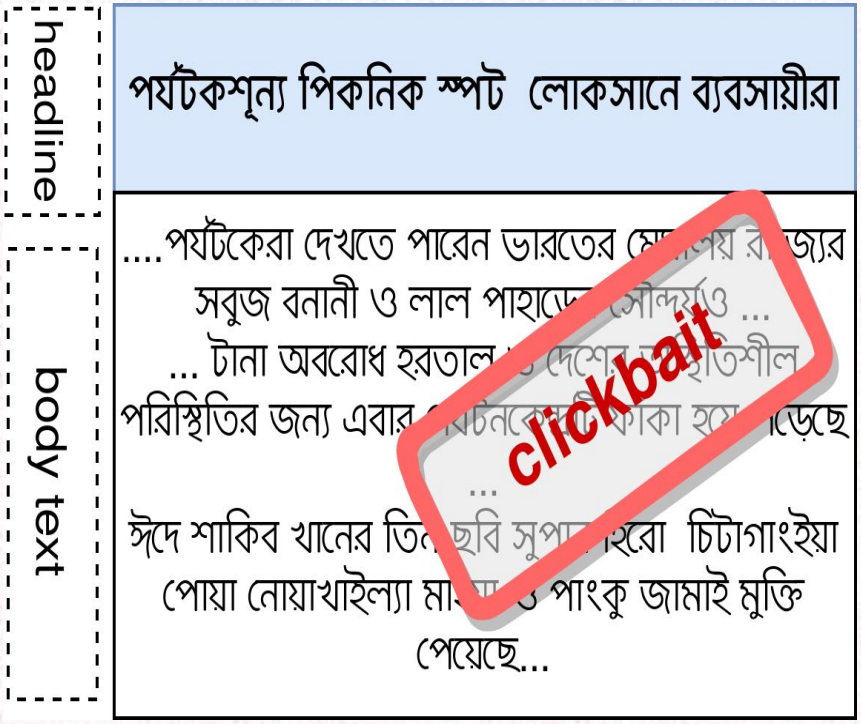
Md Aminul Haque Palash | Akib Khan | Kawsarul Islam | MD Abdullah Al Nasim | Ryan Mohammad Bin Shahjahan

Index

- Problem to Solve
- Related Works and Limitation
- Our Contributions
  - Synthetic Data Generation for Research
  - Proposed Approach
  - Evaluation Results
- Conclusion

Research Problem

- **Detecting incongruity between news headline and body text** (i.e when news headline does not correctly represent the story in advertisements, clickbait, fake news, hijacked stories, etc.)



Why Is This an Important Problem?

- News headlines are known to play important role in making first impressions to readers, and thereby deciding the viral potential of news stories within social networks.

- people are less likely to read or click on the whole contents but just read news headlines.

➢ **much of news sharing is headline based**

Why Is This an Important Problem?

Detecting the prevalent deceptive and misleading news headlines on the web in advance will better assist readers in choosing proper news stories to consume.

Previous work

- Many works have proposed for detecting ambiguous and misleading news headlines. Such as -
  - Text similarity
  - Mutual Attentive Semantic Matching
  - Deep Hierarchical Encoder
  - Convolution Dual Encoder
  - Recurrent Dual Encoder and
  - Many other ML models (XGBoost, SVM)

Previous work

- However, various solutions are primarily being developed for English to address this problem, leaving low-resource languages out of the picture.
- So we are the first to address this kind of problem in Bangla language.
- Bangla has a more complex syntactic structure and fewer natural language processing resources, so it becomes challenging to perform NLP tasks like incongruity detection.

Methodology

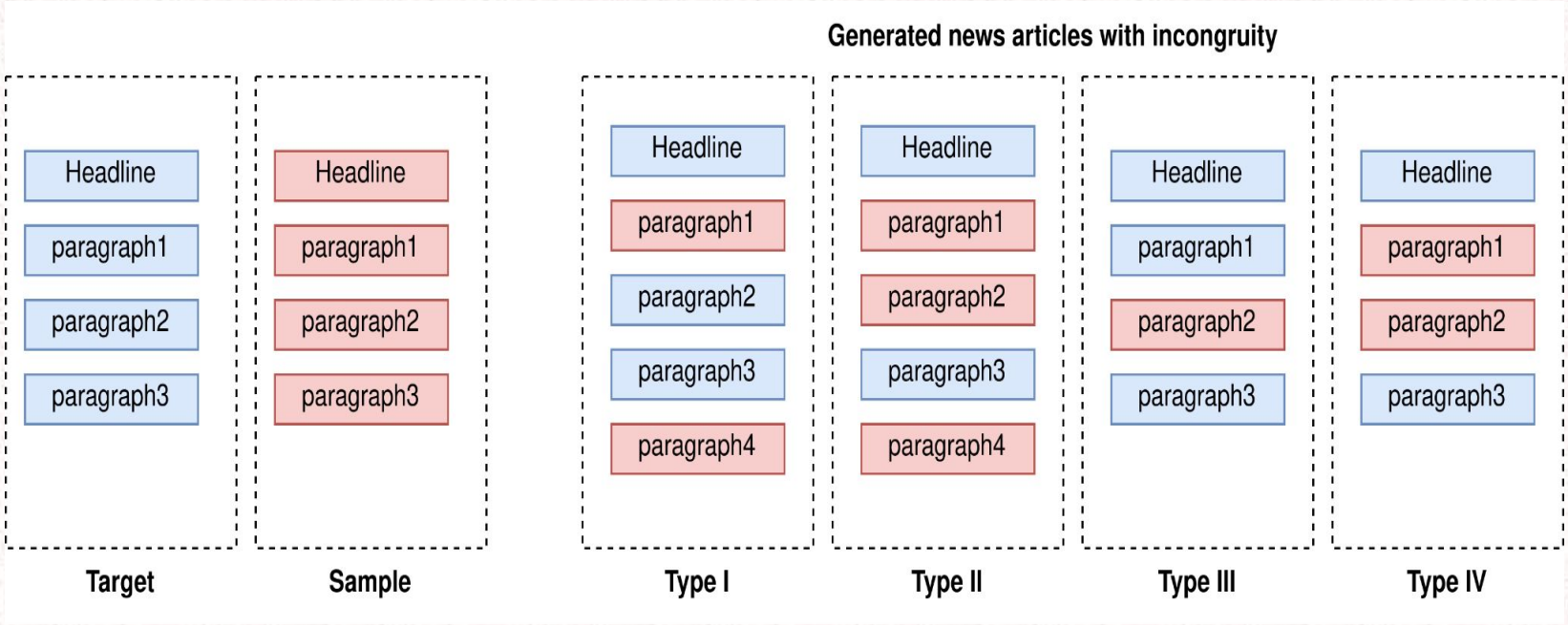
- To tackle this problem, we've addressed our problem solution into two important parts-
  - **Synthetic Data Generation**
    - 400k+ bangla news samples, 25+ categories
  - **Proposed a graph-based model and method**
    - Bangla graph-based hierarchical dual encoder (BGHDE)

Generate Dataset for Research

- The main challenges for our problem is the lack of a large training dataset.
- We did not find any dataset for Bangla language for detecting ambiguous and misleading news headlines.

Generate Dataset for Research

Injecting paragraphs from a set of sample news articles into target news articles to generate incongruent news article



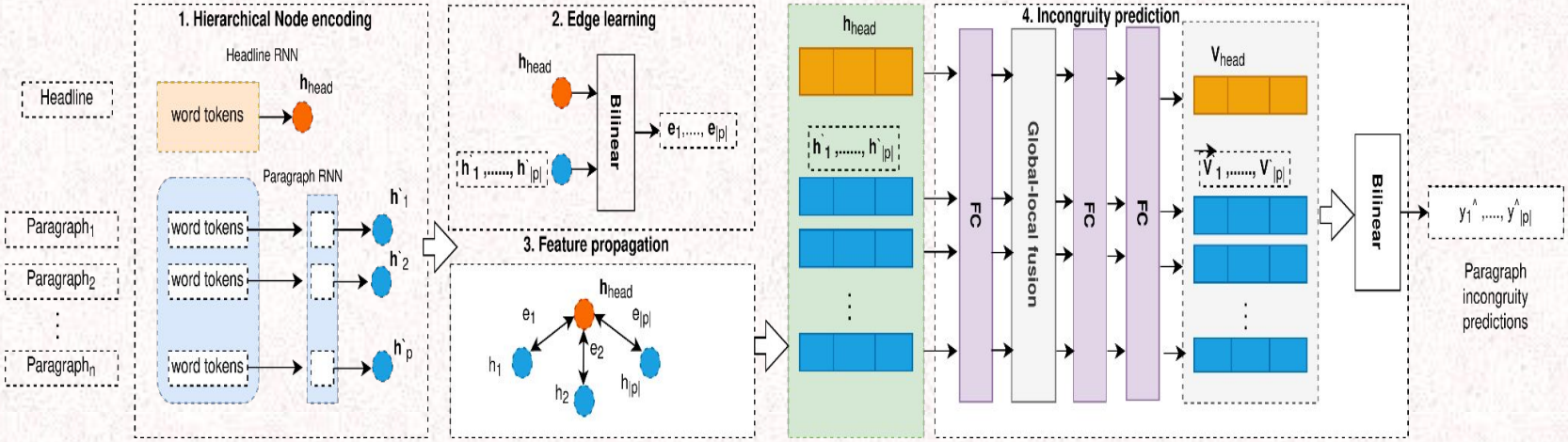
Generate Dataset for Research

The overall data distribution of our generated synthetic dataset.

		Headline		Content	
Dataset	**Samples**	**Avg.**	**Std**	**Avg.**	**Std**
Train	228000	5.58	1.45	319.35	205.41
Dev	120000	5.58	1.43	319.01	241.06
Test	120000	5.57	1.43	323.55	214.124

Proposed Model

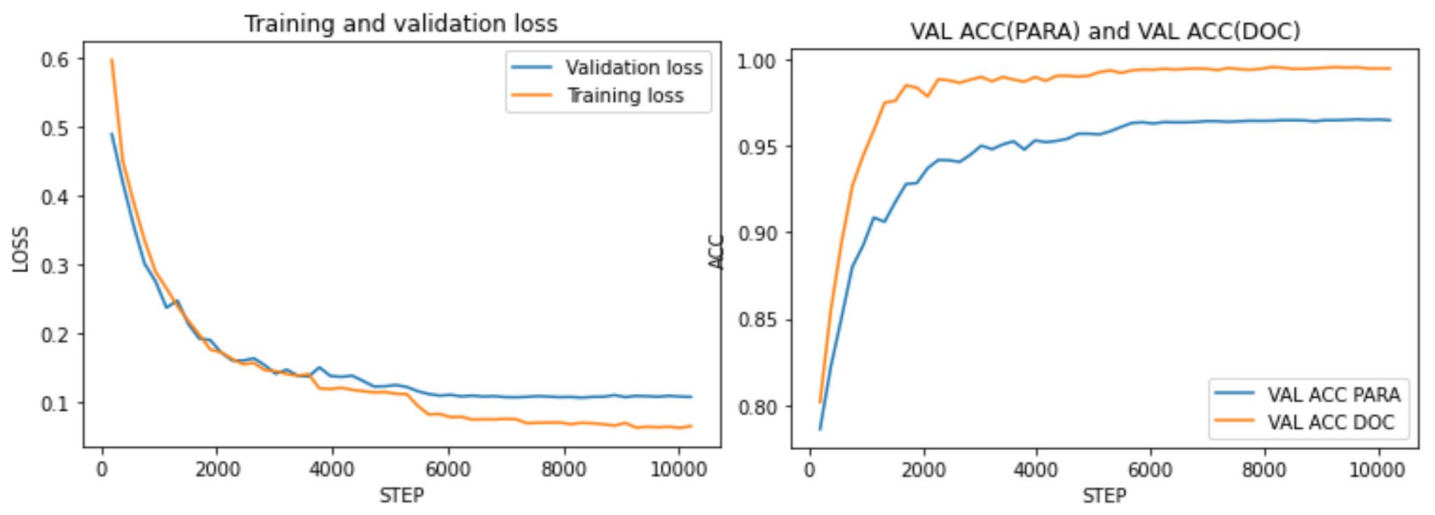
- Bangla Graph-based Hierarchical Dual Encoder (BGHDE)



- The Hierarchical Node Encoding Step
- The Edge Learning and Feature Propagation Step
- The Incongruity Prediction Step

Experiments

- Pre-trained Bangla GloVe [12] embedding consisting of 300 dimensional vectors to initialize the word embeddings.
- Model is trained on google colab and kaggle kernel.



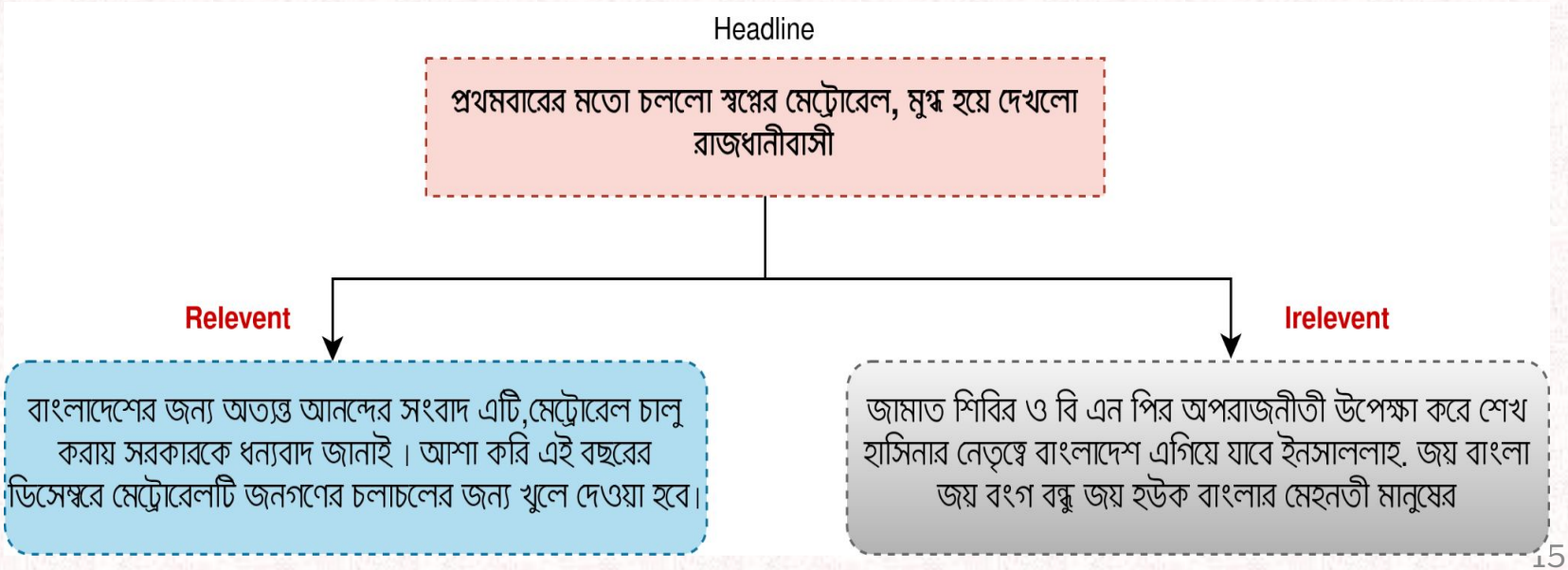
Evaluation

- We've tested our model on various bengali news dataset and achieve promising performance on synthetic dataset.
- Our evaluated dataset contains both Bangladeshi and Indian Bangla news article.

Dataset	Size	Acc(para)	Acc(doc)	Evaluation			
				Precision	Recall	F1 Scr	Support
Prothom alo	6000	0.9658	0.9918	0.98.80	0.9956	0.9918	[3000 3000]
bdnews24	2000	0.9175	0.94.50	0.7554	0.913	0.9431	[1000 1000]
Ananda bazar	1000	0.9175	0.9450	0.9623	0.97	0.9431	[500 500]
ebela	5000	0.9192	0.9702	0.970	0.9704	0.9702	[2500 2500]
zeenews	5000	0.9026	0.9542	0.9511	0.9576	0.9543	[2500 2500]
Ittefaq	8000	0.9445	0.9866	0.9812	0.9922	0.9866	[4000 4000]
Jugantor	6999	0.9494	0.9862	0.9830	0.9893	0.9861	[3458 3477]

Evaluation on Comments data

- We've tested our model on detecting Relevant and irrelevant comments from social networking sites like Facebook and YouTube.
- We achieved an accuracy of 0.73 on the Bangla comments dataset.



Future work

- Although our model gives promising results on synthetic data evaluation but does not get better results on real world data compare to that.
- We plan to improve the model performance by taking several steps-
  - Collecting more Bangla dataset.
  - Manually annotations.
  - Use headline similarity between the sample and target news while injecting paragraphs in the data generation process.
  - Try more better architecture.

Conclusion

We've addressed the problem of incongruent headline detection for Bangla language for the first time.

- Release and proposed synthetic bangla dataset and generation technique for detecting ambiguous and misleading news headlines.
- Proposed a graph based neural network for detecting headline incongruity.
- Evaluated model performance both in synthetic and real world dataset.

References

1. Chakraborty, A., Paranjape, B., Kakarla, S., Ganguly, N.: Stop clickbait: Detecting and preventing clickbaits in online news media. pp. 9–16 (08 2016). https://doi.org/10.1109/ASONAM.2016.7752207

2. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Krish-napuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D., Rastogi, R. (eds.) Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. pp. 785–794. ACM (2016). https://doi.org/10.1145/2939672.2939785,

https://doi.org/10.1145/2939672.2939785

3. Ecker, U.K.H., Lewandowsky, S., Chang, E.P., Pillai, R.: The effects of subtle misinformation in news headlines. Journal of experimental psychology. Applied 204, 323–35 (2014)

4. Fey, M., Lenssen, J.E.: Fast graph representation learning with pytorch geometric. ArXiv abs/1903.02428 (2019)

5. Guzman, A.L., Lewis, S.C.: Artificial intelligence and communication: A human–machine communication research agenda. New Media & Society 22(1), 70–86 (2020). https://doi.org/10.1177/1461444819858691

6. Hossain, M.Z., Rahman, M.A., Islam, M.S., Kar, S.: Banfakenews: A dataset for detecting fake news in bangla. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020. pp. 2862–2871. European Language Resources Association (2020), https://aclanthology.org/2020.lrec-1.349/

References

7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

8. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. CoRR abs/1609.02907 (2016), http://arxiv.org/abs/1609.02907

9. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks (2017)

10. Mishra, R., Yadav, P., Calizzano, R., Leippold, M.: Musem: Detecting incongruent news headlines using mutual attentive semantic matching. In: Wani, M.A., Luo, F., Li, X.A., Dou, D., Bonchi, F. (eds.) 19th IEEE International Conference on Machine Learning and Applications, ICMLA 2020, Miami, FL, USA, December 14-17, 2020. pp.709–716.IEEE(2020). https://doi.org/10.1109/ICMLA51294.2020.00116,

https://doi.org/10.1109/ICMLA51294.2020.00116

11. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z.,Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. ArXiv abs/1912.01703 (2019)

12. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)

References

13. Ren, Y., Zhang, J.: Fake news detection on news-oriented heterogeneous information networks through hierarchical graph attention. In: International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22,2021. pp. 1–8. IEEE (2021). https://doi.org/10.1109/IJCNN52387.2021.9534362,

https://doi.org/10.1109/IJCNN52387.2021.9534362

14. Riedel, B., Augenstein, I., Spithourakis, G.P., Riedel, S.: A simple but tough-to-beat baseline for the fake news challenge stance detection task. CoRR abs/1707.03264 (2017), http://arxiv.org/abs/1707.03264

15. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. IEEE transactions on neural networks 20(1), 61–80 (2008)

16. Schlichtkrull, M.S., Kipf, T.N., Bloem, P., van den Berg, R., Titov, I., Welling, M.: Modeling relational data with graph convolutional networks. In: Gangemi, A., Navigli, R., Vidal, M., Hitzler, P., Troncy, R., Hollink, L., Tordai, A., Alam, M. (eds.) The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings. Lecture Notes in Computer Science, vol. 10843, pp. 593–607. Springer (2018). https://doi.org/10.1007/978-3-319-93417- 4 38, https://doi.org/10.1007/978-3-319-93417-4 38

17. Song, L., Wang, Z., Yu, M., Zhang, Y., Florian, R., Gildea, D.: Ex-ploring graph-structured passage representation for multi-hop reading com-prehension with graph neural networks. CoRR abs/1809.02040 (2018), http://arxiv.org/abs/1809.02040

References

18. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds (2019)

19. Wei, W., Wan, X.: Learning to identify ambiguous and misleading news headlines. pp. 4172–4178 (08 2017). https://doi.org/10.24963/ijcai.2017/583

20. Xiao, Y., Qu, Y., Qiu, L., Zhou, H., Li, L., Zhang, W., Yu, Y.: Dynamically fused graph network for multi-hop reasoning. CoRR abs/1905.06933 (2019), http://arxiv.org/abs/1905.06933

21. Yoon, S., Park, K., Lee, M., Kim, T., Cha, M., Jung, K.: Learning to detect incon-gruence in news headline and body text via a graph neural network. IEEE Access 9, 36195–36206 (2021). https://doi.org/10.1109/ACCESS.2021.3062029

22. Yoon, S., Park, K., Shin, J., Lim, H., Won, S., Cha, M., Jung, K.: Detecting incongruity between news headline and body text via a deep hierarchical encoder. CoRR abs/1811.07066 (2018), http://arxiv.org/abs/1811.07066

23. Zhang, J., Dong, B., Yu, P.S.: Fakedetector: Effective fake news detection with deep diffusive neural network. In: 36th IEEE International Conference on Data Engineering, ICDE 2020, Dallas, TX, USA, April 20-24, 2020.pp. 1826–1829. IEEE (2020). https://doi.org/10.1109/ICDE48307.2020.00180, https://doi.org/10.1109/ICDE48307.2020.00180

24. Zhang, Y., Qi, P., Manning, C.D.: Graph convolution over pruned dependency trees improves relation extraction. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Meth-ods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018. pp. 2205–2215. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/d18-1244, https://doi.org/10.18653/v1/d18-1244

Thank You!

We are open to all relevant queries.