FinRED: A Dataset for Relation Extraction in Financial Domain

Soumya Sharma
soumyasharma20@gmail.com
IIT Kharagpur, India

Tapas Nayak
tnk02.05@gmail.com
IIT Kharagpur, India

Arusarka Bose*
arusarkabose@gmail.com
IIT Kharagpur, India

Ajay Kumar Meena*
ajaymeena3210@gmail.com
IIT Kharagpur, India

Koustuv Dasgupta
Koustuv.x.Dasgupta@gs.com
Goldman Sachs
India

Niloy Ganguly
niloy@cse.iitkgp.ac.in
IIT Kharagpur, India
LU, Hannover

Pawan Goyal
pawang@cse.iitkgp.ac.in
IIT Kharagpur
India

ABSTRACT

Relation extraction models trained on a source domain cannot be applied on a different target domain due to the mismatch between relation sets. In the current literature, there is no extensive open-source relation extraction dataset specific to the finance domain. In this paper, we release FinRED, a relation extraction dataset curated from financial news and earning call transcripts containing relations from the finance domain. FinRED has been created by mapping Wikidata triplets using distance supervision method. We manually annotate the test data to ensure proper evaluation. We also experiment with various state-of-the-art relation extraction models on this dataset to create the benchmark. We see a significant drop in their performance on FinRED compared to the general relation extraction datasets which tells that we need better models for financial relation extraction.

CCS CONCEPTS

• **Deep learning** → **Span extraction**; • **Natural language processing** → *Information extraction, Relation Extraction*; • **Domain** → Financial.

KEYWORDS

financial information extraction, financial relation extraction, financial dataset

ACM Reference Format:

Soumya Sharma, Tapas Nayak, Arusarka Bose*, Ajay Kumar Meena*, Koustuv Dasgupta, Niloy Ganguly, and Pawan Goyal. 2022. FinRED: A Dataset for Relation Extraction in Financial Domain. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3487553.3524637

1 INTRODUCTION

The task of relation extraction (RE) is defined as identifying triplets from text. Recently joint entity and relation extraction models [15, 16, 19] have been proposed to eliminate the dependency on an external NER module. These models have achieved remarkable performance in this task. Freebase-New York Times dataset (FB-NYT) [9] is obtained by mapping relational triplets from Freebase [3], WebNLG [7] is a natural language generation dataset created semi-automatically from DBPedia [2] triplets. Although KBs such as Freebase, Wikidata, DBpedia contain a lot of relational triplets from different domains such as finance, material science, legal, but that is not reflected in the existing RE datasets as they used very generic sources of text corpus. There have been some efforts to create datasets for particular domain-specific relation extraction such as legal [1], biomedical [6, 8, 12, 17], scientific articles [11, 13]. There have been some efforts to create datasets in the financial domain [10, 18, 21], however to the best of our knowledge, there is currently no standard dataset in English to understand financial relations. There has been some work on financial numeral understanding [4, 5], however, our paper focuses on textual relations and not numerical. To bridge this gap, we create a **finance domain-specific distance supervised relation extraction dataset**. We obtain the finance relations from Wikidata KB and use text from the finance domain to create the FinRED dataset.

2 DATASET SOURCES AND CREATION

The FinRED dataset [1] has been created using two sets of documents: 1) Webhose Financial News, and 2) Earning Call Transcript (ECT).

Webhose Financial News: For the financial news articles, we use the freely available dataset created by Webhose. It contains 47,851 English financial news articles crawled from July 2015 - October 2015.

Earning Call Transcripts: We collect 4,713 ECTs dated from June 2019 to September 2019 from seekingalpha.com. We use the presentation as well as the questionnaire portion of the transcript and remove monologues with less than 200 characters. Here monologue refers to a monologue given by a company participant during the presentation or the response given by a company participant during the questionnaire. In the ECT corpus, we have about 200K monologues, 1.8M sentences with an average of 7.19 sentences per monologue. In total, we have 152K tokens in the corpus.

Knowledge Base (KB): We use a subset of the Wikidata KB as a source of relational triplets for financial domain containing manually filtered 29 financial relations. Using the idea of distance supervision [14], we align the relational triplets to the text corpus.

Dataset: In total we obtain about 21,000 sentences using the distance supervision method, however a lot of these sentences are

* Equal contribution.

[1] Repository https://github.com/soummyaah/FinRED/

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France.

ACM ISBN 978-1-4503-9130-6/22/04...$15.00
https://doi.org/10.1145/3487553.3524637

Table 1: Comparison between different RE datasets.

Dataset Name	Train	Test	#Relations	#Financial Relations
FB-NYT	56,196	5,000	24	4
WebNLG	5,519	703	216	12
FinRED (Ours)	**5,699**	**1,068**	**29**	**29**

Table 2: Examples in the dataset

Sentence	Head	Relation	Tail
Antony Jenkins has been sacked as chief executive officer of Barclays Plc.	Anthony Jenkins	chief executive officer	Barclays
MEXICO CITY — State-owned Mexican oil company Pemex is reporting second quarter losses of $US5.2 billion ($A7.16 billion) due mainly to lower petroleum prices	Pemex	product or material produced	petroleum
	Pemex	headquarters location	Mexico City

noisy. Eg: "Delhi-based National Housing Bank (NHB) is working to set up more than 80 new housing finance companies (HFCs), with a special emphasis on those that will focus on financing affordable houses." where the tuple is (More Than, industry, finance) since in Wikidata, "More Than" is listed as an insurance company of the financial industry. We study a small representation of the data in detail and remove datapoints with incorrect entities such as "More Than" and "industry". After reduction, we get 7,775 sentences which we divide into the train, dev, and test dataset. We observe that this dataset only contains 920 sentences from the earnings call transcript and we attribute this to the earning call transcripts containing primarily conversational data which often does not contain a triplet in a sentence. A comparison of *FinRED* with FB-NYT [9] and WebNLG [7] along with a few examples from the dataset has been showcased in Table 1.

Annotation of test data: We annotate the dataset with the help of 2 non-native fluent English speakers annotators, who go through the dataset twice, and all incorrect triplets are removed. All triplets marked as incorrect by both the annotators are removed from the dataset. The Cohen between the annotators is 82.1%.

3 EXPERIMENTS

To avoid the use of external NER module, we choose 3 joint entity and relation extraction models, namely, SPN [16], TPLinker [19], and CasRel [20] for our experiments. These models achieved state-of-the-art performance on standard relation extraction datasets Freebase-New York Times [9] and WebNLG [7]. We report precision, recall and F1 score for triplet extraction based on exact entity matching criterion and report the results in Table 3.

From Table 3, we can observe a drop of ~4% for SPN and ~25-30% for TPLinker and Casrel models in F1 as compared to FB-NYT and WebNLG datasets. Drop in the performance showcases that more research is required to better capture performance. For the SPN model, we observe that the relation F1 value is 88.37% and the entity F1 value is 96.36% showcasing that while the model can predict the entity properly, the lower model performance can be attributed to incorrect relation classification. We see similar trends in the performance of TPLinker and CasRel.

Table 3: Performance of the three baseline models on FinRED.

	FinRED (Ours)			FB-NYT	WebNLG
Model	Prec.	Rec.	F1	F1	F1
SPN	**89.22**%	**86.42**%	**87.80**%	92.30%	91.80%
TPLinker	76.59%	63.45%	69.40%	92.00%	86.70%
CasRel	69.71%	55.84%	62.01%	89.60%	93.40%

Table 4: Performance of SPN model on the sentence classes with different number triplets and with type of overlapping triplets in them.

Metric	Precision	Recall	F1
1 triplet	76.15%	86.46%	80.98%
2 triplets	86.36%	88.88%	87.60%
3 triplets	93.31%	85.24%	89.09%
4 triplets	95.79%	88.61%	92.06%
5 triplets+	95.89%	81.30%	88%
NEO	78.40%	89.61%	83.36%
EPO	93.07%	85.50%	89.16%
SEO	89.86%	86.18%	87.98%

Performance of SPN model on the sentence classes with different number of triplets is reported in Table 4. The performance improves with the number of triplets, but degrades marginally with more than 5 triplets. We hypothesise that a few more triplets (typically EPO and SEO) assist the machine make better predictions, but too many confuse it.

In Table 4 we also report the performance of SPN model on the sentence classes with the type of overlapping triplets in them. Based on the overlap of entities, the sentence can be divided into three types: 1) NEO (No entity overlap) 2) EPO (Entity Pair Overlap) and 3) SEO (Single Entity Overlap). Here, we can observe that the model shows a slightly higher performance for SEO (F1: 87.98%) and EPO (F1: 87.98%) as compared to NEO (F1: 83.36%). We surmise that this is because the entity prediction part of the model performs well so for multi-label and overlapping triplets, the performance is higher since it needs to identify less number of entities.

4 CONCLUSION

In this paper, we propose *FinRED*, a relation extraction dataset for the finance domain curated from earning call transcript corpus and financial news articles. It contains more finance domain-specific relations than the existing RE datasets. We experimented with three state-of-the-art joint entity and relation extraction models on this dataset and saw a significant drop in F1 score compare to general domain RE datasets showcasing that more research is required on the models for FinRED.

ACKNOWLEDGMENTS

This research was partially supported by Goldman Sachs sponsored research grant FTHS. (FinTalk: Research towards creating a platform for highlight generation and summarization of financial documents while taking into account user feedback).

REFERENCES

[1] Judith Jeyafreeda Andrew. 2018. Automatic Extraction of Entities and Relation from Legal Documents. In *Proceedings of the Seventh Named Entities Workshop*. Association for Computational Linguistics, Melbourne, Australia, 1–8. https://doi.org/10.18653/v1/W18-2401

[2] Christian Bizer, Jens Lehmann, Georgi Kobilarov, S. Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia - A crystallization point for the Web of Data. *J. Web Semant.* 7 (2009), 154–165.

[3] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. *Proc. Sigmod*, 1247–1250. https://doi.org/10.1145/1376616.1376746

[4] Chung-Chi Chen and Hen-Hsen Huang. 2019. Overview of the ntcir-14 finnum task: Fine-grained numeral understanding in financial social media data. In *Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies*.

[5] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2019. Numeracy-600K: Learning Numeracy for Detecting Exaggerated Information in Market Comments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 6307–6313. https://doi.org/10.18653/v1/P19-1635

[6] Sung-Pil Choi. 2018. Extraction of protein–protein interactions (PPIs) from the literature by deep convolutional neural networks with various feature embeddings. *Journal of Information Science* 44 (2018), 60 – 73.

[7] Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating Training Corpora for NLG Micro-Planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 179–188. https://doi.org/10.18653/v1/P17-1017

[8] Jinghang Gu, Longhua Qian, and Guodong Zhou. 2016. Chemical-induced disease relation extraction with various linguistic features. *Database: The Journal of Biological Databases and Curation* 2016 (2016).

[9] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 541–550. https://aclanthology.org/P11-1055

[10] Ali Jabbari, Olivier Sauvage, Hamada Zeine, and Hamza Chergui. 2020. A French Corpus and Annotation Schema for Named Entity Recognition and Relation Extraction of Financial News. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 2293–2299. https://aclanthology.org/2020.lrec-1.279

[11] Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A Challenge Dataset for Document-Level Information Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7506–7516. https://doi.org/10.18653/v1/2020.acl-main.670

[12] Fei Li, Meishan Zhang, Guohong Fu, and Dong-Hong Ji. 2017. A neural joint model for entity and relation extraction from biomedical text. *BMC Bioinformatics* 18 (2017).

[13] Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. 2017. Scientific Information Extraction with Semi-supervised Neural Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2641–2651. https://doi.org/10.18653/v1/D17-1279

[14] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics, Suntec, Singapore, 1003–1011. https://aclanthology.org/P09-1113

[15] Tapas Nayak and Hwee Tou Ng. 2020. Effective modeling of encoder-decoder architecture for joint entity and relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 8528–8535.

[16] Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, Xiangrong Zeng, and Shengping Liu. 2020. Joint entity and relation extraction with set prediction networks. *arXiv preprint arXiv:2011.01675* (2020).

[17] Ashok Thillaisundaram and Theodosia Togia. 2019. Biomedical relation extraction with pre-trained language representations and minimal task-specific architecture. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*. Association for Computational Linguistics, Hong Kong, China, 84–89. https://doi.org/10.18653/v1/D19-5713

[18] Mihaela Vela and Thierry Declerck. 2009. Concept and Relation Extraction in the Finance Domain. In *Proceedings of the Eight International Conference on Computational Semantics*. Association for Computational Linguistics, Tilburg, The Netherlands, 346–350. https://aclanthology.org/W09-3741

[19] Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. TPLinker: Single-stage Joint Extraction of Entities and Relations Through Token Pair Linking. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 1572–1582. https://doi.org/10.18653/v1/2020.coling-main.138

[20] Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 1476–1488.

[21] Haoyu Wu, Qing Lei, Xinyue Zhang, and Zhengqian Luo. 2020. Creating A Large-Scale Financial News Corpus for Relation Extraction. *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)* (2020), 259–263.