

---

# CONTEXT-NER: Contextual Phrase Generation at Scale

---

Himanshu Gupta<sup>1</sup> Shreyas Verma<sup>2</sup> Tarun Kumar<sup>4</sup> Swaroop Mishra<sup>1</sup>  
Tamanna Agrawal<sup>3</sup> Amogh Badugu<sup>4</sup> Himanshu Sharad Bhatt<sup>3</sup>

<sup>1</sup>School of Computing and AI, Arizona State University

<sup>2</sup>College of Computing, Georgia Institute of Technology

<sup>3</sup>American Express AI Labs

<sup>4</sup>Birla Institute of Technology & Science, Pilani

{hgupta35, srmishr1}@asu.edu, shreyas.verma@gatech.edu

{tamanna.agrawal, himanshu.s.bhatt}@aexp.com

{f20160025h, f2016005p}@alummi.bits-pilani.ac.in

## Abstract

NLP research has been focused on NER extraction and how to efficiently extract them from a sentence. However, generating relevant context of entities from a sentence has remained under-explored. In this work we introduce the task CONTEXT-NER in which relevant context of an entity has to be generated. The extracted context may not be found exactly as a substring in the sentence. We also introduce the EDGAR10-Q dataset for the same, which is a corpus of 1,500 publicly traded companies. It is a manually created complex corpus and one of the largest in terms of number of sentences and entities (1 M and 2.8 M). We introduce a baseline approach that leverages phrase generation algorithms and uses the pre-trained BERT model to get 33% ROUGE-L score. We also do a one shot evaluation with GPT-3 and get 39% score, signifying the hardness and future scope of this task. We hope that addition of this dataset and our study will pave the way for further research in this domain. <sup>1</sup>.

## 1 Introduction

Recent advances in NER have led to the development of several large pretrained models that achieved remarkable performance in its detection. However, previous studies are limited to performing NER task on sentences having low amount of context Zhang and Zhang [2022], Wang et al. [2014], Francis et al. [2019], Alexander and de Vries [2021], Wu et al. [2022], Wang and Wang [2022], Varshney et al. [2022], Shrimal et al. [2022] For example, asking the question "What is "Entity"?" to pre-trained Bert model gives the correct answer easily. This task becomes particularly arduous when a lot of numerical entities are involved in the sentence.

In pursuit of addressing the issue of *finding relevant phrases associated with entities*, we formulate the task called CONTEXT-NER; given a sentence and an entity, a relevant phrase describing the entity (its context) must be generated. The phrase may or may not be present in the sentence, but would be relevant to the entity. We also introduce EDGAR10-Q dataset, which is made from quarterly and annual financial reports of publicly traded LLCs. It is one of the largest in financial domain (Table 4a) consisting of complex sentences that are not prevalent in benchmark datasets, posing a new challenge for state of the art models. First, the sentences are very long and complex and second they also contain several numerical entities. 1 illustrates some examples from the EDAGR10-Q dataset.

---

<sup>1</sup>Dataset, the script to generate it, baseline approach and GPT-3 evaluation are freely available at <https://github.com/him1411/edgar10q-dataset>

Table 1: Example of the EDGAR10-Q data set.

Sentences	Entity	Entity Type	Associated Text
In October 2019, the Company increased the borrowing capacity on the revolving credit loan by \$33,000 increasing the available credit facility from \$60,000 to \$93,000	\$60,000 to \$93,000	Money	available credit facility
	\$33,000	Money	capacity on the revolving credit loan
If the loan is paid during months 13-24 or 25-36 and then a penalty of 2% and 1%, respectively, of the loan balance will be charged on the date of repayment.	13-24 or 25-36	Date	loan is paid during months
	2% and 1%	Percent	penalty of the loan balance
The weighted-average remaining lease term and discount rate related to the Company’s lease liabilities as of September 26, 2020 were 10.3 years and 2.0%, respectively	10.3 years	Date	remaining lease term
	2.0%	Percent	discount rate

Table 2: Illustration of the baseline approach based on Table 1.

Phrases extracted	Question	Answer
borrowing capacity, available credit facility borrowing capacity on revolving credit loan	What is borrowing capacity on revolving credit loan ?	\$60,000 to \$93,000
	How much is available credit facility ?	\$33,000
penalty of %, loan balance date of repayment	What is 13-24 or 25-36 ?	loan is paid during months
	What is 2% and 1% ?	2% (Wrong Answer)
lease liabilities, discount rate average lease term	What is average lease term ?	10.3 years
	What is discount rate ?	2.00%

Table 3: Statistics of the EDGAR10-Q Dataset

Entity Types	Counts	Other Statistics	Values	Number of Entities	Number of Sentences	Percentage of dataset	Sentence length
Floating Values (monetary and percent)	2143054	Entites per sentence	1.78	1	503433	49.8	31.7
number of Assets (Shares and Integers)	425850	Words per paragraph	113.14	2	331130	32.7	36.3
Ordinal Values	16891	Labels per entity	1.45	3	108103	10.7	42.4
Dates	195174	Words per sentence	40.23	4	47191	4.6	52.8
				5 or more	19855	1.9	73.9

(a) Types of entities and their descriptions

(b) Sentence and Paragraph wise Statistics

(c) Entity Wise Statistics. It is observed that more number of entities result in longer sentences.

It is evident that phrases are not always present in sentences and can be difficult to retrieve without adequate knowledge of the domain.

We also introduce a baseline method that leverages syntactic trees of the sentence to generate questions and find relevant phrases of the sentences. Table 2 show the extracted relevant phrases, the generated questions, and finally the retrieved answer. We describe the approach in more detail in Section 3.1. We also test GPT-3 Brown et al. [2020] for the same task using simple prompting techniques. The baseline method achieves 33% ROUGE-L score (f1) while GPT-3 gets 39%. Our experiments suggest CONTEXT-NER is a hard task for state of the art models. We identified this area for further research to enhance the learning capabilities for such complex tasks.

## 2 Dataset

The EDGAR10-Q dataset is made by scraping quarterly (10-Q) and annual (10-K) over the last two years (2019, 2020 and 2021). The reports are critical to the organization’s financial health and are prepared by in-house attorneys who are domain experts. All SEC filings are standardized with all entities in a sentence tagged with corresponding NER labels. Table 3a shows the four types of entities, namely money, time (duration), percent, and cardinal values (pure, shares, and integer) present in the data. Tables 3b and 3c further elucidate data richness through paragraph and sentence level statistics and Table 4a compares this dataset with benchmark NER Datasets. We observe that the EDGAR10-Q is the largest and richest in multiple parameters and a first-of-its-kind dataset in the financial domain.

Table 4: Comparison of dataset and Baseline results

(a) Comparison of the dataset with other NER datasets.

Dataset name	Documents	Sentences	Words	Entities
funsd	200	Not available	31485	9743
wikicoref	30	2229	59652	3557
scierc	500	Not available	Not available	8089
med mentions	4392	42602	1176058	352496
genia	2000	18545	436967	96582
conll 2003	1393	22137	301418	35089
<b>EDGAR10-Q</b>	<b>18752</b>	<b>1009712</b>	<b>77400425</b>	<b>2780969</b>

(b) Baseline approach ROUGE-L score on the dataset

Number of Entities	Precision	Recall	F1
1	0.44	0.30	0.32
2	0.45	0.32	0.34
3	0.41	0.28	0.30
4	0.37	0.26	0.28
5 or more	0.32	0.22	0.24
<b>Overall</b>	<b>0.43</b>	<b>0.29</b>	<b>0.32</b>

### 3 Experiments

#### 3.1 Baseline Approach

The Baseline approach is a simple yet efficient technique to extract the entities and their descriptions from the sentences. We start with data cleaning and entity extraction. Next, we present a comprehensive phrase generation method, and the phrases are further used to frame questions to the machine reading comprehension (MRC) or QA model. Finally, given the generated questions and sentences, the output from MRC model is used to associate the entities to their corresponding descriptions.

##### 3.1.1 Phrase generation

A noun phrase (NP) Stuart et al. [2013] includes a noun, a person, place, or thing, and the modifier that distinguishes it. We extract two types of phrases from the sentences, namely simple and complex. In simple phrase extraction, each sentence comprises subject-object and verb connecting them where Subject or Object is usually a noun or pronoun. After searching for a noun and pronoun, we check for any noun compound or adjective. On the other hand, for complex phrase extraction we first start with preposition extraction. We then follow similar steps as in simple phrase extraction to look for phrases in both left and right of the preposition. It has to be noted that simple phrases are not always found on both sides of the preposition. Algorithm 1 further summarizes the process of simple and complex phrase extraction from the sentences can be found in Appendix B. We obtain the following simple and complex noun phrases for the sentences mentioned in Table 2.

##### 3.1.2 MRC model

The process of using simple and complex phrases to generate questions and also using the MRC model to associate entities with their relevant phrases is described as follows. Phrases (described in Algorithm 1) and Entities are extracted on a sentence level for each paragraph. Based on the type of entity and the noun phrases, the questions are framed accordingly in a rule-based fashion. For instance, if the entity found out was of type date, then the question would be "when is" + NP?. Once these questions are generated, they are fed into the MRC Model, and the answers are checked for the relevant entity. In case of multiple questions with the same answer, we select the one with the highest confidence score.

There are instances where none of the generated questions returned an answer with the target entity or returned responses with a different entity. For those cases, we create the question "what is" entity?. Here its response would be considered as the relevant phrase. In case these questions return different entities as responses, all cases to identify the noun phrase fail and the algorithm does not return a response. A detailed description about the MRC model is present in appendix C.

#### 3.2 Experimental Setup

**Baseline Model Setup:** We run all our experiments using bert base model Devlin et al. [2018]. All experiments are done with Nvidia V100 16GB GPU.

**GPT Setup:** We evaluate GPT-3 (Text-DaVinci-002, max tokens = 256, top p = 1, frequency penalty = 0, presence penalty = 0) in one-shot setting.

Table 5: Performance Evaluation

(a) Sentence wise statistics of evaluation dataset				(b) Performance comparison between the proposed baseline and GPT-3 on evaluation dataset (ROUGE-L).						
Number of Entities	Number of Sentences	Percentage of dataset	Sentence Length	Number of Entities	Baseline Scores			GPT-3 Scores		
					Precision	Recall	F1	Precision	Recall	F1
1	2646	56.96	27.85	1	0.43	0.27	0.30	0.40	0.36	0.34
2	1239	26.67	29.76	2	0.46	0.32	0.34	0.43	0.45	0.40
3	600	12.91	31.21	3	0.44	0.30	0.33	0.44	0.50	0.42
4	105	2.26	42.17	4	0.40	0.28	0.30	0.36	0.41	0.33
5 or more	55	1.18	53.94	5 or more	0.48	0.27	0.32	0.43	0.48	0.41
				Overall	0.44	0.29	0.33	0.42	0.43	0.39

Table 6: Detailed Analysis of exact match and no match results.

(a) Sentence/Entity wise results for number of instances with exact match (ROUGE-L = 1).

Number of Entities	Baseline		GPT-3	
	Number of Instances	% of total Data	Number of Instances	% of total Data
1	128	5%	175	7%
2	169	14%	176	14%
3	103	17%	121	20%
4	20	19%	19	18%
5 or more	9	16%	18	33%

(b) Sentence/Entity wise results for number of instances with no match (ROUGE-L = 0)

Number of Entities	Baseline		GPT-3	
	Number of Instances	% of total Data	Number of Instances	% of total Data
1	911	34%	732	28%
2	598	48%	332	27%
3	354	59%	157	26%
4	81	77%	52	50%
5 or more	42	76%	23	42%

Table 3a shows the statistics of the evaluation dataset. The evaluation set is chosen to be a fraction of the dataset due to the high costs of GPT-3. As is evident from the table, the data distribution of the evaluation set is close to the original data set.

**Performance Evaluation metrics:** ROUGE-L score uses longest common sub sequence matching between the baseline and GPT-3 responses to compare output quality. We report precision, recall, and the F1 measure against the ROUGE-L Lin [2004] score. We also report the Exact Match Rajpurkar et al. [2016] which measure the ratio of the instances for which a model produces the exact same string as the gold labels.

### 3.3 Results

**Baseline results on complete dataset:** Table 4b gives the results of the baseline approach in the overall dataset. All metrics show a linearly decreasing trend with an increase in the number of entities.

**Evaluation set creation:** Due to the high costs of GPT-3 inference API, we had to perform its evaluation on a smaller sample whose statistics are given in Table 5a. We try to ensure that the evaluation set is a good representation of the original dataset to maintain homogeneity in performance.

**Comparison with GPT-3:** First 3 columns of Table 5b show the performance of the baseline approach on different metrics. Precision is uniformly around 40% while recall is ranging from 20 to 30% leading to overall F1 score in range from 30 to 34. The next 3 columns show the GPT-3 scores. Precision of GPT-3 is nearly the same w.r.t. baseline approach, but recall is significantly high (upto 10 points in some cases), resulting in an overall higher F1 compared to the baseline approach across all categories. The performance difference between the two increases as the length of sentences increases.

**Poor match performance:** The results of both GPT-3 and Baseline are consistently low highlighting complex sentence structures in the data set and demonstrating the need for more research in this task. Table 6 shows the wide gap between the proportion of sentences with exact and no matches, further strengthening the point.

## 4 Conclusion

In this work, we introduce the CONTEXT-NER task, which aims to "extract relevant phrases for entities" to bridge the gap between existing NER tasks. We also introduce the EDGAR10-Q dataset, which is one of the largest and relatively complex finance datasets. We further give a baseline approach to solve this problem and conduct extensive experiments using our baseline. We also compare its performance with GPT-3, highlighting the difficulty of the dataset. We hope that addition

of this dataset and our study will pave the way for further research on the task. Future work can include elaborate experimentation using other instruction-based / prompting techniques on this dataset.

## Acknowledgement

The authors thank Arizona State University’s Agave Research Computing cluster. The Cluster was used to create the dataset, run the baseline approach, and run the GPT-3 evaluation as well.

## Ethical Considerations

We have verified that all licenses of source documents used in this document allow their use, modification, and redistribution in a research context. There were no real-life names in the data set. No particular sociopolitical bias is emphasized or reduced specifically by our methods.

## References

- A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, 2019.
- D. Alexander and A. P. de Vries. "this research is funded by...": Named entity recognition of financial information in research papers. 2021.
- G. Angeli, M. J. J. Premkumar, and C. D. Manning. Leveraging linguistic structure for open domain information extraction. 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)*, pages 344–354, 2015.
- T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- O. Etzioni, M. Banko, S. Soderland, and D. S. Weld. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74, 2008.
- S. Francis, J. Van Landeghem, and M.-F. Moens. Transfer learning for named entity recognition in financial and biomedical documents. *Information*, 10(8):248, 2019.
- R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld. 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*, pages 541–550, 2011.
- M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd. spacy: Industrial-strength natural language processing in python, 2020. URL <https://doi.org/10.5281/zenodo.1212303>.
- Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.
- O. Levy, M. Seo, E. Choi, and L. Zettlemoyer. Zero-shot relation extraction via reading comprehension. *arXiv preprint arXiv:1706.04115*, 2017.
- Q. Li and H. Ji. Incremental joint extraction of entity mentions and relations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 402–412, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1038. URL <https://www.aclweb.org/anthology/P14-1038>.
- X. Li, F. Yin, Z. Sun, X. Li, A. Yuan, D. Chai, M. Zhou, and J. Li. Entity-relation extraction as multi-turn question answering. *arXiv preprint arXiv:1905.05529*, 2019.
- C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

- A. Liu, S. Soderland, J. Bragg, C. H. Lin, X. Ling, and D. S. Weld. Effective crowd annotation for relation extraction. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 897–906, 2016.
- E. Loper and S. Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002.
- C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.
- B. McCann, N. S. Keskar, C. Xiong, and R. Socher. The natural language decathlon: Multitask learning as question answering, 2018.
- S. Mishra, D. Khashabi, C. Baral, Y. Choi, and H. Hajishirzi. Reframing instructional prompts to gptk’s language. *arXiv preprint arXiv:2109.07830*, 2021a.
- S. Mishra, D. Khashabi, C. Baral, and H. Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*, 2021b.
- M. Miwa and M. Bansal. End-to-end relation extraction using lstms on sequences and tree structures. *arXiv preprint arXiv:1601.00770*, 2016.
- P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
- V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.
- A. Shrimal, A. Jain, K. Mehta, and P. Yenigalla. NER-MQMRC: Formulating named entity recognition as multi question machine reading comprehension. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 230–238, Hybrid: Seattle, Washington + Online, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-industry.26. URL <https://aclanthology.org/2022.naacl-industry.26>.
- G. Stanovsky, J. Michael, L. Zettlemoyer, and I. Dagan. Supervised open information extraction. In *NAACL-HLT*, 2018.
- L. M. Stuart, J. M. Taylor, and V. Raskin. The importance of nouns in text processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35, 2013.
- C. Sun, Y. Wu, M. Lan, S. Sun, W. Wang, K.-C. Lee, and K. Wu. Extracting entities and relations with joint minimum risk training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2265, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1249. URL <https://www.aclweb.org/anthology/D18-1249>.
- D. Varshney, A. Prabhakar, and A. Ekbal. Commonsense and named entity aware knowledge grounded dialogue generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1322–1335, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.95. URL <https://aclanthology.org/2022.naacl-main.95>.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.
- S. Wang, R. Xu, B. Liu, L. Gui, and Y. Zhou. Financial named entity recognition based on conditional random fields and information entropy. In *2014 International Conference on Machine Learning and Cybernetics*, volume 2, pages 838–843, 2014. doi: 10.1109/ICMLC.2014.7009718.
- X. Wang and Y. Wang. Sentence-level resampling for named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2151–2165, 2022.
- J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.
- T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.

- L. Wu, P. Xie, J. Zhou, M. Zhang, C. Ma, G. Xu, and M. Zhang. Robust self-augmentation for named entity recognition with meta reweighting. *CoRR*, 2022.
- Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.
- Y. Zhang and H. Zhang. Finbert-mrc: financial named entity recognition using bert under the machine reading comprehension paradigm. *arXiv preprint arXiv:2205.15485*, 2022.

## Appendix

### A Related Work

Etzioni et al. [2008] refers to a schemaless approach to extract facts from texts. This is traditionally done by extracting relations between the facts. The relations are typically verbs in the sentences and are used for the Open IE Relationship paradigm. However, relationships are based on the assumption that they connect two entities. In the case of financial data, we have an entity and its description and its relation, which is not helping us to link them.

We aim to extract key-value pairs using zero-shot Open Information Extraction. This scenario is more complicated than extracting key-value pairs using relations as done by Li et al. and Levy et al. [2019], Levy et al. [2017]. Levy et al. also uses a zero-shot approach to train his MRC model on templated questions and then uses that model on unseen relations. He has pointed out that generating natural questions is challenging and uses human-generated questions for his zero-shot scenario. Li et al.’s work was based on Levy et al. and formalizes relation extraction as multi-turn question answering. The author uses both natural and templated questions based on the relations. In both scenarios, the number of questions the author can ask is limited. Both types of questions would not perform well if the relation they encounter is unseen. Li et al. has also trained MRC models to return blank or no answer when they could not find any relevant context. We exercise this ability differently, as explained in Section 3

Miwa et al. [2016] approached relation extraction by extracting entities and relations together using neural network models (multiclass classification model based on tree LSTMs). The performance of all such approaches drops on unseen relations. Li and Ji [2014] use perceptron and efficient beam search to extract entities and relations. They also develop global features to capture the dependency between entity and relation extraction. Sun et al. [2018] build on the previously mentioned framework and uses a joint learning framework and a flexible global loss function to capture the interactions of the entities and their relationships.

Most of the prior work for entity relation extraction has relied upon template-based question-answer methods where a question was predefined based on the relation. We do not rely on templated questions which makes it scalable. Since the advancement of Bidirectional Attention mechanisms Vaswani et al. [2017], various improvements have been made in the field of machine reading comprehension by different models such as BERT, XLNet, ALBERT Devlin et al. [2018], Yang et al. [2019], Lan et al. [2019] that extract text from the passages given queries. Answer Extraction can be further decomposed into a multi-class text classification task that predicts starting and ending position of the answer.

McCann et al. [2018] introduced decaNLP, where various challenging tasks like relation extraction, question answering, semantic role labeling, etc., were presented as problems of spanning MRC over context. He further introduced a Multitask question answering network, which learns from all the challenges raised in the paper. Their framework showed improvements in NER and zero-shot capability in text classification. To extract entities, several state-of-the-art frameworks like Stanford CoreNLP’s NER, Spacy, NLTK, Flair are present as open source libraries. Manning et al. [2014], Honnibal et al. [2020], Loper and Bird [2002], Akbik et al. [2019]. Recently instruction based frameworks have also been used for multitask learning to jointly learn several NLP tasks. Mishra et al. [2021a,b], Wei et al. [2021].

A noun phrase (NP) [2013] includes a noun, a person, place, or thing, and the modifier that distinguishes it. Open IE is predicated on the idea that the relation (which are action verbs in most cases) is the central element of study, from which all other considerations flow. However,

in many instances, the verb is no help, particularly in financial data. Consider the sentence: "The Deferred revenue for 2020 is \$20 billion." Like most financial records are of the form "is, was, be," etc., the verb "is" in this sentence is an auxiliary verb and does not describe any particular event nor give any information about the entity.

Table 7 shows the instances where our pipeline achieves exact matches and incorrect matches. As seen in the 3rd and 4th row of the table, the sentences are fairly complex and its arduous to get the corresponding description of the entities from the sentences.

To compare our methods with existing Open IE model, we ran Stanford's Open IE and Stanovsky et al. Stanovsky et al. [2018] Open IE models on EDGAR10-Q dataset. We realize that Open IE models fail to perform when dealing with long range dependencies. Table 8 shows relation extraction on sentences given in Table 7.

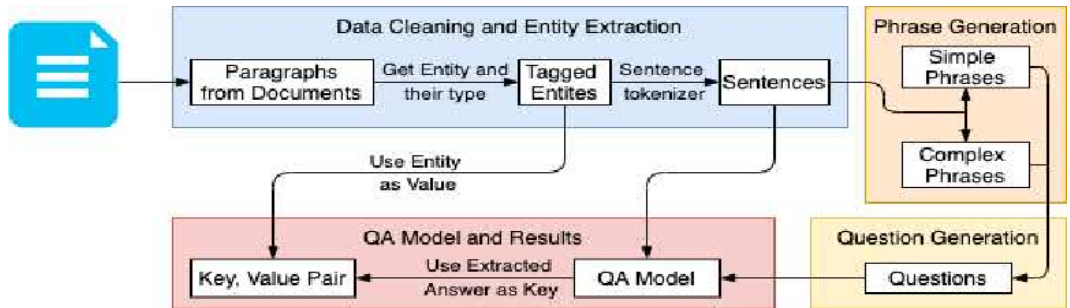


Figure 1: Illustrates the overall process flow of the proposed zero-shot open information extraction technique using question generation and reading comprehension.

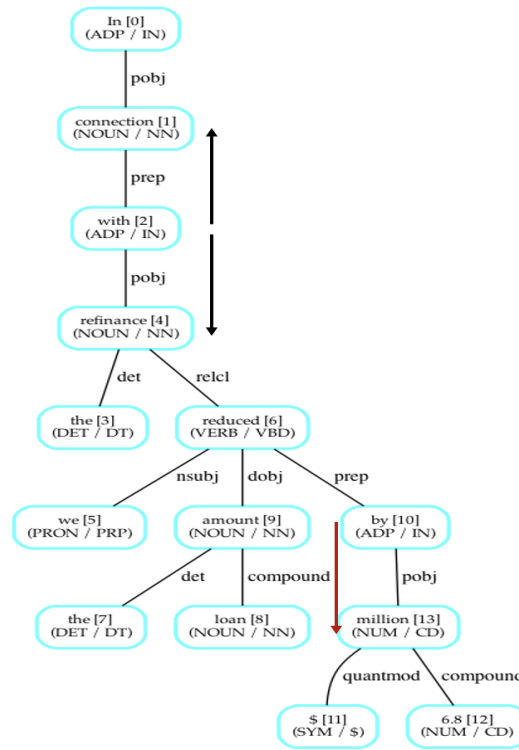


Figure 2: Syntactic tree structure for extraction of simple and complex noun phrases.



---

**Algorithm 1:** Phrase Generation *Pseudocode*

---

**Input:** Sentence**Output:** List of Phrases

```
1 Function simple_noun_phrase_extractor(Sentence):
2   doc = sequence_of_token(Sentence), phrase_list = []
3   for token in Doc:
4     phrase = ' '
5     if token.head.pos in [Noun, Pronoun] and token.dep in [Object, Subject]:
6       for subtoken in token.children:
7         if subtoken.pos is Adj or subtoken.dep is Comp: phrase += subtoken.text + ' '
8         if len(phrase) is not 0: phrase += token.text
9         if len(phrase) is not 0 and phrase doesnot have entities: phrase_list.append(phrase)
10  return phrase_list
11 Function complex_noun_phrase_extractor(Sentence):
12  doc = sequence_of_token(Sentence)
13  phrase_list = []
14  for token in Doc:
15    if token.pos is Preposition:
16      phrase = ' '
17      if token.head.pos in [Noun, Pronoun]:
18        for subtoken in token.head.children:
19          if subtoken.pos is Adj or subtoken.dep is Comp:
20            phrase += subtoken.text + ' '
21            phrase += token.head.text + ' ' + token.text
22          for right_tok in token.rights:
23            if right_tok in [Noun, Pronoun]:
24              for subtoken in right_tok.children:
25                if subtoken.pos is Adj or subtoken.dep is Comp:
26                  phrase += subtoken.text + ' '
27                  phrase += ' ' + right_tok.text
28            if len(phrase) is > 1 and phrase doesnot have entities:
29              phrase_list.append(phrase)
30  return phrase_list
```

---

## B Phrase generation

This paper presents a simple, yet efficient technique to extract entities and their descriptions from sentences. As shown in Figure 1, it starts with data cleaning and entity extractions. A noun phrase (NP) Stuart et al. [2013] includes a noun, a person, place, or thing, and the modifier that distinguishes it. Open IE is predicated on the idea that the relation (which is action verbs in most cases) is the central element of the study, from which all other considerations flow. However, in many cases, the verb is not helpful, particularly in financial data. Consider the sentence: "Deferred revenue for 2020 is \$20 billion." Like most financial records are of the form "is, was, be," etc., the verb "is" in this sentence is an auxiliary verb and does not describe any particular event or give any information about the entity.

We extract two types of phrases from the sentences, namely simple and complex. In simple phrase extraction, each sentence comprises subject-object and verb connecting them where Subject or Object is usually a noun or pronoun. After searching for a noun and pronoun, we check for any noun compound or adjective. On the other hand, for complex phrase extraction we first start with preposition extraction. We then follow similar steps as in simple phrase extraction to look for phrases in both left and right of the preposition. It has to be noted that simple phrases are not always found on both sides of the preposition. Algorithm 1 further summarizes the process of simple and complex phrase extraction from the sentences.

Now we demonstrate the extraction of simple and complex noun phrases for the sentence, '*In connection with the refinance we reduced the loan amount by \$6.8 million.*'. The syntactic tree for the

above sentence is shown in Figure 2. We search if the token's POS tag is a noun or pronoun as we are looking just for noun phrases. We also ensure that phrase lies either in the Subject or Object of the sentence to ensure we are skipping the relations. In this case, we got "amount" the first word of the phrase. After that, we iterate the node to see its children named subtoken in Algorithm 1. We search for subtoken's dependency relation with the token as a compound relation, or we search if the subtoken is an adjective. The intuition behind this is that if the subtoken and token have a compound relationship, they form a meaningful noun phrase. In this case, "amount" has a compound relationship with its subtoken "loan" so they together form "loan amount" as the meaningful noun phrase. Similar logic is followed for searching adjectives. Complex NPs are identified as series of noun phrases with a preposition separating them, so we start by identifying them. In this example, the preposition identified was "in". Then we iterate both up and down the node to find noun phrases that follow the same method mentioned above. The noun phrases identified from the top were "connection" and the bottom was "refinance". The entire complex NP was formed as NP from top + preposition in the middle and + NP from below. The resultant was "connection with refinance".

In this paper, we use Spacy<sup>2</sup> library for POS tags of the word which were leveraged in Algorithm 1. Similarly, using Algorithm 1, we obtain the following simple and complex noun phrases for sentences mentioned in Table 1.

## C Machine Reading Comprehension Model

This paper presents a zero-shot technique as we leverage the phrase generation to generate meaningful questions without further training of the machine reading comprehension (MRC) model. This allows our technique to be domain agnostic and thus can be easily expanded to newer domains. The process to leverage noun phrases to generate the questions and further using the MRC model to associate entities with their corresponding descriptions is described below:

- Each paragraph in the document is broken down into sentences. For each sentence, the following are extracted: Phrases (using simple and complex noun phrases described in Algorithm 1) and Entities using the Flair NER Model.
- On the basis of the entity type and the noun phrases, the questions are framed accordingly. For instance, if the entity found out was of type date, then the question would be "when is" + NP?. In our example, the question for the first sentence of Table 1 would be "how much is borrowing capacity on revolving credit loan ?".
- In instances where the entity type is of integer, float, or percent where appending "when is" or "how much is" does not give an advantage. For such cases, to keep the question generic we append "what is" to the noun phrase. For example, in the second sentence in Table 1, the question "What is the loan balance?" was created based on the entity type of 2% and 1%.
- Once these questions are generated, they are fed into the MRC Model, and its answer is checked if it contains the entity. To give an example, in the 1st sentence of Table 1, the following questions are created, and the model returns their corresponding answers and their confidence values:
  - "How much is borrowing capacity on revolving credit loan?" answer: "\$33,000", confidence score: 0.946
  - "How much is borrowing capacity ?" answer: "\$33,000", confidence score: 0.824
  - "How much is revolving credit loan ?" answer: "\$33,000", confidence score: 0.856
  - "How much is available credit facility ?" answer: "\$60,000 to \$93,000", confidence score: 0.5762

If there are multiple questions whose answer has the entity, we select the question whose answer is of the highest confidence value. In the above example, "borrowing capacity on revolving credit loan" is chosen as the key for \$30,000, and "revolving credit loan" is chosen as the key for both \$60,000 to \$93,000.

- If the entity is not present in the response of the MRC model, the question is discarded. In the 2nd Sentence of Table 1, the following questions are created :

---

<sup>2</sup>Spacy POS Tagging Library link: <https://spacy.io/usage/linguistic-features>.

- "What is penalty of % ?"
- "What is loan balance ?"

None of them are returning "13-24 or 25-36", so the phrases "penalty of %" and "loan balance" are discarded.

- There are instances where none of the generated questions returned an answer with the target entity or returned responses with a different entity as shown above. For those cases, we create the question "what is" entity?. Here, its response would be considered as the key (opposite to the case above). In the 2nd sentence of the Table, none of the questions returned relevant answers, So the following questions were created:
  - "What is 13-24 or 25-36 ?"
  - "What is 2% and 1% ?"
- In the above cases, where questions are formed based on entities, the answers are checked if they have given any other entity as the answer. For instance, the questions, "what is 2% and 1% ?" return "2" as the answer to the second sentence of Table 1. If the cases mentioned above hold, then the response is discarded. Here all the cases to identify the noun phrase associated with the entity fail, so no answer is returned.
- If they do not fail, then the response is also considered a viable answer. For instance, In the 2nd sentence, the question was framed: "What is 13-24 or 25-36 ?" which returned "loan is paid during months" as the answer.

Using the rules stated above, the entity and its associated noun phrases are identified. The last two columns of Table 1 show the questions which were generated and their responses from the MRC model. Inspired by the success of the pre-trained transformer model, we employ distilled BERT Sanh et al. [2019] by Hugging Face Wolf et al. [2020] trained on SQuAD dataset Rajpurkar et al. [2016] as the MRC model for our zero-shot question answering <sup>3</sup>.

## D Examples of Exact and Incorrect Matches

Table 7 shows the instances where our pipeline achieves exact matches and incorrect matches. As seen in the 3rd and 4th row of the table, the sentences are fairly complex and its arduous to get the corresponding description of the entities from the sentences.

## E Comparison with Open IE

Traditionally, information extraction approaches from textual documents assume prespecified relations for a domain and use crowd-sourcing or distant supervision approaches Hoffmann et al. [2011], Liu et al. [2016] to gather examples and train models for every type of relation. One of the limitations for such approaches is their inability to extract unseen relations that were not observed or specified during training and thus are not pragmatic. On the other hand, in Open information extraction (Open IE) Etzioni et al. [2008], the relations are not pre-defined and are extracted as and when they are encountered. To compare our methods with existing Open IE model, we ran Stanford's Open IE and Stanovsky et al. Stanovsky et al. [2018] Open IE models on EDGAR10-Q dataset. We realize that Open IE models fail to perform when dealing with long range dependencies. Table 8 shows relation extraction on sentences given in Table 7.

## F GPT-3 Prompts

The following prompt was provided for GPT-3 learning:

" Based on the example given below, generate entity-phrase pairs.

Sentence:

Issuance of common stock in May 2019 public offering at \$243.00 per share, net of issuance costs of \$15.

<sup>3</sup>Hugging Face's Model Link: <https://huggingface.co/transformers/v2.8.0/usage.html>.

Table 7: Instances of Exact and Incorrect Match by the model

Sentence	Entity	Labels	Predicted Answer
<b>Instances of Exact Match</b>			
Premium receivables are reported net of an allowance for doubtful accounts of \$250 and \$237 at September 30, 2020 and December 31, 2019, respectively.	\$250 and \$237	premium receivable	premium receivable
The fair value of the collateral received at the time of the transactions amounted to \$1,019 and \$351 at September 30, 2020 and December 31, 2019, respectively.	\$1,019 and \$351	fair value of collateral	fair value of collateral
<b>Instances of Incorrect Match</b>			
During the nine months ended September 30, 2020, we granted approximately 0.3 restricted stock units that are contingent upon us achieving earnings targets over the three year period from 2020 to 2022.	0.3	grants in period	restricted stock units
Certain selling equity holders elected to receive deferred, variable earn out consideration with an estimated value of \$21,500 over the rollover period of three years.	\$21,500	earn out consideration	estimated value

Table 8: Results given by OpenIE Models on instances of Sentences given in Table 6

Subject	Relation	Object
<b>Stanford Open IE Angeli et al. [2015]</b>		
Premium receivables	are reported	net of allowance
fair value	received at	time of transactions
we	granted	approximately 0.3 stock units
variable	earn out	consideration
<b>Stanovsky et al. Stanovsky et al. [2018]</b>		
Not Found	are	Not Found
the collateral	received at	at the time of the transactions
Not Found	restricted	stock units
Not Found	estimated	value

Entity:

\$15

Phrase:

Common stock public offering issuance costs

Generate Entity-phrase pairs based on the sentence below:

Sentence:

Sentence 1:

Entity:

GPT-3 Response "