

# Democratizing Financial Knowledge Graph Construction by Mining Massive Brokerage Research Reports

Zehua Cheng<sup>1</sup>, Lianlong Wu<sup>1</sup>, Thomas Lukasiewicz<sup>1</sup>, Emanuel Sallinger<sup>1,2</sup> and Georg Gottlob<sup>1,2</sup>

<sup>1</sup>Department of Computer Science, University of Oxford, UK

<sup>2</sup>Institute of Logic and Computation, TU Wien, Austria

## Abstract

This work presents a novel automatic financial knowledge graph (KG) construction framework by mining massive brokerage research reports without explicit financial expertise guidance and intensive manual rules. We propose a semantic-entity interaction module to construct the interaction feature between the entity and semantic context in the research reports and build a KG from scratch according to a predefined schema based on the obtained interaction features. We train the semantic-entity interaction module using a pre-extracted entity set as a remote supervision-based approach. We further introduce entity augmentation over this entity set from the inference samples of the semantic-entity interaction module to maintain the entity set.

## Keywords

Knowledge Graph, Language Model, Financial Research Report, Entity Resolution

## 1. Introduction

Knowledge graphs (KGs) have emerged as one of the most popular knowledge representation technologies for massive information processing tasks. Financial intelligence analysis is one of the most important works in intelligence analysis, which is facing large volumes of documents and tabular data. KGs have already helped financial analysts to process large amounts of data and cooperate with state-of-the-art trading systems [1, 2] to achieve a high volume return in the market. Such tools are usually monopolized by large companies and are very costly to maintain. To democratize such technologies, we need a framework that can automatically build a financial KG from scratch.

In the financial area, research reports contain a wealth of high-quality data collected by professional agencies that can be treated as an ideal resource for constructing a reliable knowledge graph. Financial research reports are professional documents with in-depth research on macroeconomics, finance, industries, industry chains, and companies by various financial research institutions

and brokerages. Such reports often cover a wide range of areas and comprehensive data. Therefore, it is reasonable to build a reliable KG based on financial research reports.

However, there are still some challenges in constructing KGs in the financial area from research reports, among which the most hardest ones are listed below:

- Entity-relationships are highly coupled to context. Entities are not explicitly represented in research reports but have a complex interaction with their text passages.
- The overall structure of different research reports are highly complicated. The structures of different research reports can contradict each other. As the research reports accommodate a wide range of data and knowledge, and much professional knowledge, different research structures and professional understandings may express the same content slightly differently.

Such features make it difficult to automatically construct a knowledge graph based on research reports from scratch. A solution should involve an in-depth interaction from inter-pipeline interactions to address such a challenge. The high coupling between entities and their context makes the rule-based approach challenging to intervene, and we find that it is more challenging to exploit this part of the features due to the inconsistency of wording in unstructured documents. Therefore, we believe that to deal with such highly coupled features, we need to consider them as a whole. Decoupling entity and contextual information and processing entity features and contextual features to different models separately is not ideal.

We use a language model to extract contextual semantic features and bridge the feature connections with a con-

Published in the Workshop Proceedings of the EDBT/ICDT 2022 Joint Conference (March 29-April 1, 2022), Edinburgh, UK

✉ zehua.cheng@cs.ox.ac.uk (Z. Cheng); lianlong.wu@cs.ox.ac.uk (L. Wu); thomas.lukasiewicz@cs.ox.ac.uk (T. Lukasiewicz); emanuel.sallinger@cs.ox.ac.uk (E. Sallinger); georg.gottlob@cs.ox.ac.uk (G. Gottlob)

🌐 <https://www.cs.ox.ac.uk/people/zehua.cheng/> (Z. Cheng);

<https://www.cs.ox.ac.uk/people/lianlong.wu/> (L. Wu);

<https://www.cs.ox.ac.uk/people/thomas.lukasiewicz/>

(T. Lukasiewicz);

<https://www.cs.ox.ac.uk/people/emanuel.sallinger/> (E. Sallinger);

<https://www.cs.ox.ac.uk/people/georg.gottlob/> (G. Gottlob)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

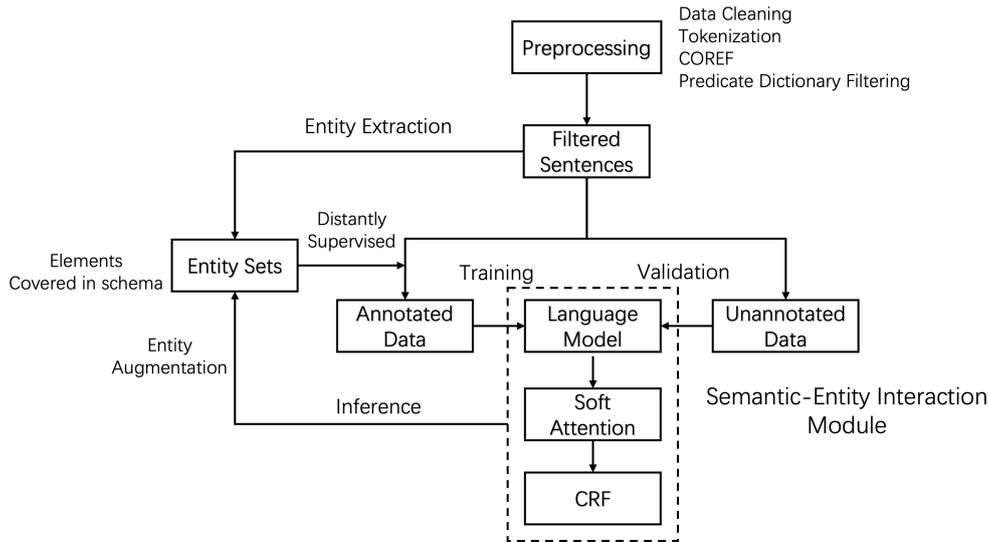


Figure 1: Overall Structure

ditional random field [3]. Language models like BERT [4] and GPT [5] have proven their performance in many challenging natural language processing tasks [6]. Application in Question Answering [7] has proved that language models are capable of dealing with complicated semantic language features. Therefore, BERT is an ideal solution for this semantic feature extraction. Based on the language model, introducing a downstream specific module can further improve the semantic features obtained by the language model. In named entity recognition (NER), there are successful applications combining BERT with conditional random fields (CRFs) [8, 9]. [10] formulate NER as machine reading comprehension (MRC) task by introducing an MRC module at the end of the BERT model.

Updating the entity set on the fly can further improve the reliability of the constructed knowledge graph. The entity set could be easily affected by the noise in the raw data. Under such circumstances, we do not want to put all the eggs into one basket. Filtering raw data is the first and the most crucial step for building a reliable knowledge graph. The most significant budget of constructing a knowledge graph is data cleaning [11]. By introducing a statistical supervision of raw data, such as domain-specific dictionaries and regularization of word frequencies, human intervention in data cleaning can be significantly reduced [12]. Therefore, we create an automated data cleaning pipeline to preprocess the raw data with various filtering methods. Scholars have also found that using semantic information can also reduce human effort in data cleaning [13, 14, 15]. We thus simultaneously use the inference entities of the language

model to extend the entity set.

In this work, we develop an automatic knowledge graph construction pipeline tailored to the financial domain based on research reports. We achieved an  $F_1$  score 73.5% based on a predefined schema over research reports. Our framework is highly scalable, since the overall structure is entirely automatic. We designed an entity augmentation to extend the entity set and construct a distant supervision over the training process. We also conduct ablation studies to examine the effects of the different components of the pipeline.

## 2. Related Works

### 2.1. Knowledge Graph Construction

Traditional KG construction is based on a manually specified ontology and intensive human efforts to learn the extraction for each relation in the ontology.

More specifically, supervised methods are learning from sample input and output pairs, like hidden Markov models (HMMs) [16], maximum entropy-based models, such as the MENE system [17] and ME Tagger [18]. Models based on support vector machines (SVMs) [19] and CRFs [3] are also common supervised methods. In addition, semi-supervised methods require less training data. For example, a binary AdaBoost classifier [20] was proposed for NER. NELL [21] has introduced a semi-supervised bootstrapping approach with a predefined ontology of categories and relations that involve human-in-the-loop cooperation, fully using human labour, and

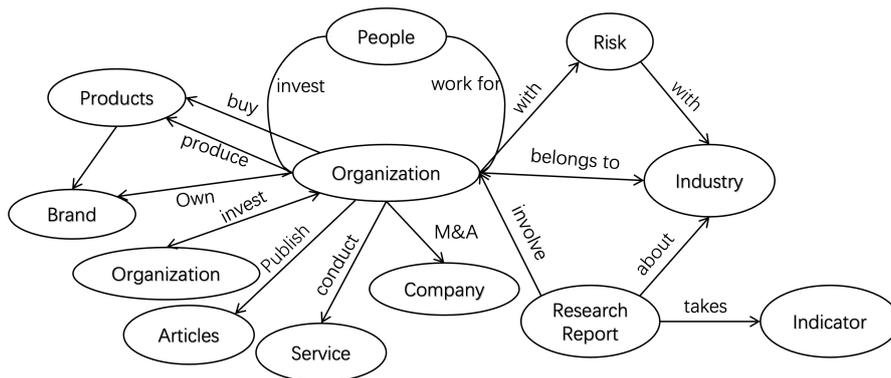


Figure 2: Schema of the Knowledge Graph

existing data. Specifically, Snorkel [22] provides a weakly supervised learning model, with handwritten rules, building a generative model based on the overlapping or even conflicting results of rules. Most recently, unsupervised methods, e.g., KNOWITALL [23], emerged for knowledge base construction.

## 2.2. Named Entity Recognition with Language Models

By using different types of heads, BERT [4] can tailor for a wide range of natural language processing tasks. BERT also has successful applications on named entity recognition [24]. [8] proposed to combine CRFs with BERT on the challenging NER in mining medical documents. The same model structure is also applied in NER for Portuguese documents [9]. [25] further introduced an additional BiLSTM in the BERT-CRF structure and further achieved better a performance in Chinese electronic health records NER. Some researchers [26] challenge the BiLSTM in [25], considering it redundant, since BERT and BiLSTM have the same function.

## 3. Automatic Knowledge Graph Construction Pipeline

This section introduces each component of our automated financial KG construction pipeline. We first present the overall structure and then the semantic-entity interaction module.

### 3.1. Overall Structure

The overall structure of our proposed framework is presented in Figure 1; its main ingredients are described as follows.

**Preprocessing.** We follow the standard data cleaning in NLP by removing brackets, parentheses, quotes, and other punctuation. Before the pipeline, we filtered the noisy text spans in sentence-level. We then use the coreference resolution system (COREF) [27] to the same entity in the filtered text. We filter out the domain-irrelevant entity structure for the output of COREF with a domain-specific predicate dictionary and then tokenize the filtered samples. Sense-disambiguated predicates construct this dictionary from the corpus with the highest frequency relevant to the financial domain. We extracted entities from the filtered data to obtain entity sets based on elements covered in the schema. The details of the schema is presented in Figure 2 and discussed in Section 4.

**Entity Augmentation.** We perform entity augmentation with the inference results of the semantic-entity interaction module, since the extracted entities are collected based on the manually designed schema from analysts’ interest. For scalability concerns, we merge the inference results of the semantic-entity interaction module to augment the entity set.

**Distant Supervision.** We maximise the utility of the extracted entities by constructing a distant supervision [28] to the semantic entity interaction module.

Finally, we score the predicate-argument to reflect our confidence in precision and conciseness.

### 3.2. Semantic-Entity Interaction Module

The overall structure of the semantic-entity interaction module is presented in Figure 1. Our proposed semantic entity interaction module is composed of a BERT language model with a CRF [3]. The input sequence is encoded by BERT into an intermediate representation with hidden dimension  $H$ . A soft attention is then applied to the intermediate representation to learn the interaction

better. The output of the soft attention is then fed to the CRF layer. We follow the notation in [29], and have the following scoring function:

$$s(\mathbf{X}, \mathbf{y}) = \sum_{i=0}^n A_{y_i, y_{i+1}} + \sum_{i=1}^n P_{i, y_i}, \quad (1)$$

where  $\mathbf{A}$  denotes the parameters of the CRF layer,  $A_{i,j}$  represents the score of transitioning from entity  $i$  to entity  $j$ , and  $P_i$  is the output score of the classification head of the BERT model. We train the semantic-entity interaction module with log-probability loss.

As presented in Figure 1, we perform entity augmentation during the inference phase of the semantic-entity interaction module to extend the entity sets. Practically, we use the pre-trained model, with fixed parameters of the transformer layers and the embedding layer, and only allow the classification head and the CRF to update according to backpropagation.

## 4. Data Resource

The original research reports and the annotations are collected by [30], which includes 1,200 research reports and annotated 5,131 entities for evaluation. The details of the dataset are shown in Table 1.

**Table 1**  
Knowledge Graph Dataset Statistics

Knowledge Graph	Entities	Relational Triples	Property Triples
Seeding KG	5,131	6,091	354
Evaluation KG	12,668	20,707	974

The task is to construct a knowledge graph according to the schema presented in Figure 2. Each element in the schema is explained as follows:

- **Research Report** indicates the resource origin, represented as the title of the research report.
- **Indicator** indicates the financial indicators in research reports, such as roe, eps, and gross margin.
- **People** indicates the actual natural persons.
- **Organization** indicates that the companies, businesses, governments, etc. are all institutional types of entities.
- **Product** refers to items produced by companies that can be bought and sold, and also includes software products. Usually, they involve ownership transition during the transaction.
- **Service** refers to actual service, which usually does not involve ownership transition during the transaction.

- **Risk** indicates the risk warning in the research report.
- **Article** indicates publications cited in the research report.
- **Industry** indicates the industry to which the company belongs.
- **Brand** indicates the brand that the company owns. Some companies may have overlapping brand names, so it is necessary to disambiguate the reference brand and the company name based on the context.

## 5. Experiment Setup

We implemented our framework and trained over an 8 NVIDIA V100 GPU cluster. The batch size is 32 per GPU. We use the BERT-base model as the pre-trained weights of the language model by setting the learning rate as  $1e^{-3}$  with the Adam optimiser for 10 epochs.

We use HanLP [31] to extract the entities from the filtered data.

## 6. Evaluation

We follow the evaluation of the Cold Start evaluation task in the TAC KBP [32]. The scoring metrics are based on the official evaluation toolkit<sup>1</sup>. The evaluation starts with a predefined schema (see the details of the schema in Figure 2) and a small number of seed knowledge graphs to build knowledge graphs from unstructured text data. The evaluation automatically extracts entities, relationships, and attribute values from the text of research reports that match the mapping schema, enabling the automated construction of financial knowledge graphs.

We use a  $F_1$  score to evaluate the model’s overall performance. The experimental results of the language model with different components are presented in Table 2. To fully present the novelty of the semantic entity interaction module, we present the ablation study by comparing the downstream specific module in the overall structure under the same preprocessing setup. We also perform an ablation study between BERT with CRF and BERT with MRC [10]. Similarly to BERT with CRF, [10] also involved an interaction between the language model and an additional downstream specific module.

We can infer from Table 2 that our proposed language model and CRF with soft attention has achieved the highest performance. The MRC module is not designed for this case, while CRF would be more suitable for processing such tasks. By introducing soft attention, the performance of the overall structure has been further

<sup>1</sup><https://github.com/wikilinks/nelval>

**Table 2**

Experimental results for different modules in precision, recall and  $F_1$  score (%). SA refers to a soft attention module.

Method	$F_1$	Precision	Recall
BERT w/CRF	72.5	83.2	<b>64.23</b>
BERT w/MRC	68.57	79.55	60.25
BERT w/SA w/CRF	<b>73.5</b>	<b>86.69</b>	63.79
BERT w/SA w/MRC	69.29	81.55	60.23

improved by 1%. Soft attention can also improve BERT with MRC by 0.68%.

## 7. Conclusion

We proposed a novel knowledge graph construction framework based on the brokerage research reports in this work. Our proposed method has achieved 73.5% in  $F_1$  score. We expect that our proposed method is also extensible and reliable where we expect the overall performance of our model can be further improved by using a more complicated language model like RoBERTa [33] or GPT-2 [5].

## References

- [1] X. Fu, X. Ren, O. J. Mengshoel, X. Wu, Stochastic optimization for market return prediction using financial knowledge graph, in: 2018 IEEE International Conference on Big Knowledge (ICBK), 2018, pp. 25–32.
- [2] S. Deng, N. Zhang, W. Zhang, J. Chen, J. Z. Pan, H. Chen, Knowledge-driven stock trend prediction and explanation via temporal convolutional network, in: Companion Proceedings of The 2019 World Wide Web Conference, 2019, pp. 678–685.
- [3] J. Lafferty, A. McCallum, F. C. N. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proceedings of the 18th International Conference on Machine Learning, ICML '01, 2001, pp. 282–289.
- [4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
- [5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, arXiv preprint arXiv:2005.14165 (2020).
- [6] A. Rogers, O. Kovaleva, A. Rumshisky, A primer in bertology: What we know about how bert works, Transactions of the Association for Computational Linguistics 8 (2020) 842–866.
- [7] C. Alberti, K. Lee, M. Collins, A bert baseline for the natural questions, arXiv preprint arXiv:1901.08634 (2019).
- [8] J. Mao, W. Liu, Hadoken: A bert-crf model for medical document anonymization, in: IberLEF@SEPLN, 2019, pp. 720–726.
- [9] F. Souza, R. Nogueira, R. Lotufo, Portuguese named entity recognition using bert-crf, arXiv preprint arXiv:1909.10649 (2019).
- [10] X. Li, J. Feng, Y. Meng, Q. Han, F. Wu, J. Li, A unified MRC framework for named entity recognition, arXiv preprint arXiv:1910.11476 (2019).
- [11] M. Muller, I. Lange, D. Wang, D. Piorkowski, J. Tsay, Q. V. Liao, C. Dugan, T. Erickson, How data science workers work with data: Discovery, capture, curation, design, creation, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–15.
- [12] M. Mahdavi, F. Neutatz, L. Visengeriyeva, Z. Abedjan, Towards automated data cleaning workflows, Machine Learning 15 (2019) 16.
- [13] E. Rahm, H. H. Do, Data cleaning: Problems and current approaches, IEEE Data Eng. Bull. 23 (2000) 3–13.
- [14] W. L. Low, M. L. Lee, T. W. Ling, A knowledge-based approach for duplicate elimination in data cleaning, Information Systems 26 (2001) 585–606.
- [15] Z. Kedad, E. Métais, Ontology-based data cleaning, in: International Conference on Application of Natural Language to Information Systems, Springer, 2002, pp. 137–149.
- [16] D. M. Bikel, R. Schwartz, R. M. Weischedel, Algorithm that learns what’s in a name, Machine Learning 34 (1999) 211–231.
- [17] A. Borthwick, A maximum entropy approach to named entity recognition, PhD thesis (1999).
- [18] J. R. Curran, S. Clark, Language independent NER using a maximum entropy tagger (2003) 164–167.
- [19] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (1995) 273–297.
- [20] X. Carreras, L. Màrquez, L. Padró, Named entity extraction using AdaBoost, 2002, pp. 1–4.
- [21] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, et al., Never-ending learning, Communications of the ACM 61 (2018) 103–115.
- [22] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, C. Ré, Snorkel: Rapid training data creation with weak supervision, Proceedings of the VLDB Endowment 11 (2017) 269–282. arXiv:1711.10160.
- [23] O. Etzioni, M. Cafarella, D. Downey, A. M. Popescu, T. Shaked, S. Soderland, D. S. Weld, A. Yates, Unsupervised named-entity extraction from the Web: An experimental study, Artif. Intell. 165 (2005) 91–134.
- [24] J. Vamvas, Bert for ner, Von <https://vamvas.ch/bert->

- for-ner (2019).
- [25] Z. Dai, X. Wang, P. Ni, Y. Li, G. Li, X. Bai, Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records, in: 2019 12th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics (CISP-BMEI), 2019, pp. 1–5.
  - [26] Z. Liu, Ner implementation with bert and crf model, 2020.
  - [27] M. Honnibal, I. Montani, spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing, *To appear* 7 (2017) 411–420.
  - [28] M. Mintz, S. Bills, R. Snow, D. Jurafsky, 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, 2009, pp. 1003–1011.
  - [29] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, *arXiv preprint arXiv:1603.01360* (2016).
  - [30] Biendata, Ccks 2020: Evaluation of automated construction of financial knowledge graph based on ontology, 2020.
  - [31] H. He, J. D. Choi, The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 5555–5577.
  - [32] H. Ji, J. Nothman, H. T. Dang, S. I. Hub, Overview of tac-kbp2016 tri-lingual edl and its impact on end-to-end cold-start kbp, *Proceedings of TAC* (2016).
  - [33] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* (2019).