Relation Extraction from Financial Reports

Tianda Sun

Master by Research

University of York

Computer Science

March 2022

Abstract

This project mainly focuses on using deep learning methods to extract relations from the so called 10-K SEC financial reports, and adds them to an ontology for further use. A 10-K report is a comprehensive report submitted by public companies each year to publish their financial performance. In the US, the 10-K reports are required by the U.S Security and Exchange Commission (SEC) to provide the investors with information on a company on which they can base their decisions to invest. It is far more detailed than the annual report where it describes the company's potential to succeed so it is useful for investors to refer to. In this research, we mainly focus on the distant supervision method to construct the dataset from the Financial Industry Business Ontology(FIBO) [2] and evaluate the performance of two distant supervision relation extraction models. Additionally, we discuss the potential flaws of distant supervision method on this task and investigate some possible improvements such as anaphora resolution to enhance the knowledge base, and point out further research direction for the domain-specific relation extraction area. In addition, this research provides results to Can Erten, a PhD student at the University of York, who will use the ontology from the reports in his research.

List of Content

List of Figures

List of Tables

Acknowledgements

Throughout the writing of this thesis, I have received a great deal of support and assistance from the people around me. I would first like to thank my supervisor, Dr Dimitar Kazakov whose expertise was invaluable in the formulation of the research topic and methodology in the financial area and Natural Language Processing, also the writing of the thesis. Also, I would like to acknowledge Can Erten who gives me great help with the data source for the financial knowledge base. I want to thank both of you for all the opportunities I was given to conduct my research and further my thesis at the Department of Computer Science in the University of York.

Also, I would like to thank my parents and my friends who had given me strong support when I was studying in York for the whole year. You are always there for me.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, university. All sources are acknowledged as references.

Chapter 1

Introduction

Ontologies have become a powerful tool in various area including finance, which also attracted a lot of research and applications on it. Since 2008, the Financial Industry Business Ontology (FIBO) has been established by the EDM Council with a number of companies and people [2]. Of course, FIBO is not the only financial technology (Fintech) organization. Based on the power of knowledge graph and ontology, some Fintech companies, such as Revolut or Yewno, have been established in recent years. Most of them focus on combining machine learning techniques to provide infrastructure and solutions to the financial community, as well as to extend the applications of financial ontologies in areas, such as stock trading or investment analysis [2].

Thus, the main motivation for this research project starts with the 10-K reports published by the US Securities and Exchange Commission(SEC) Department. The reason why we choose the 10-K report is that 10-K contains a range of useful information which covers the running situation of a company in detail. Therefore, ontologies based on 10-K reports can potentially be a powerful tool for investors as well as researchers to understand the management situation or the potential risk factor. In addition, constructing ontologies from 10-K reports is also helpful for researchers to automatically analyse the situation across one or multiple companies, e.g. through the use of deep learning-based techniques, in order to provide useful opinions to investors such as risk identification or stock prediction.

However, compared to structured data such as tables and charts, unstructured text, which represents the majority of the 10-Ks, is more difficult to process by the computer because of the complexity of human language. So the aim of our project is focusing on how to extract ontologies automatically from a large amount of unstructured text, which corresponds to a broader open research topic in natural language processing called relation extraction.

Relation extraction (RE) is a sub-task of information extraction. With the development of ontologies and knowledge graphs, a clear obstacle is how to produce these from a large corpus of natural language, most of whichis unstructured. To deal with this problem, researchers proposed a series of techniques based on hand-written features or traditional machine learning methods such as SVM or ANN. Beginning in 2009, a novel method called distant supervision was introduced by Mint et al. [18], and gradually became a popular and efficient method for relation extraction . Some models, including PCNN+ATT [14], RESIDE [30], BRE [37] have shown significant performance in the general knowledge area such as Wikipedia. But, in most domain-specific areas, such as finance, there still exists a gap in relation extraction methods research. Therefore, it is valuable to investigate the possible models and solutions for the relation extraction task in those areas, and identify potential directions for further research [2].

Chapter 2

Literature review

2.1 Fintech ontologies

Ontology is a term that describes a data model that represents knowledge as a set of concepts. In computer science, the basic form of ontology is presented by the form of a triple:

$$(subject, relation, object)$$

While the subject and object are the entities from the natural environment and the triple indicate the relation between them. As a basic component of the knowledge graph, the form of ontology is reading friendly to the computer program as well as human readers. Through extracting natural concepts from the environment, we can build various knowledge graphs which fulfil the semantic, logical and rules information in the application domain. Figure 2.1 shows an instance of a small knowledge graph on Apple Inc. In the knowledge graph, it is easy for computers to recognised the triples and understand the complex attributes of the entities in the real world.

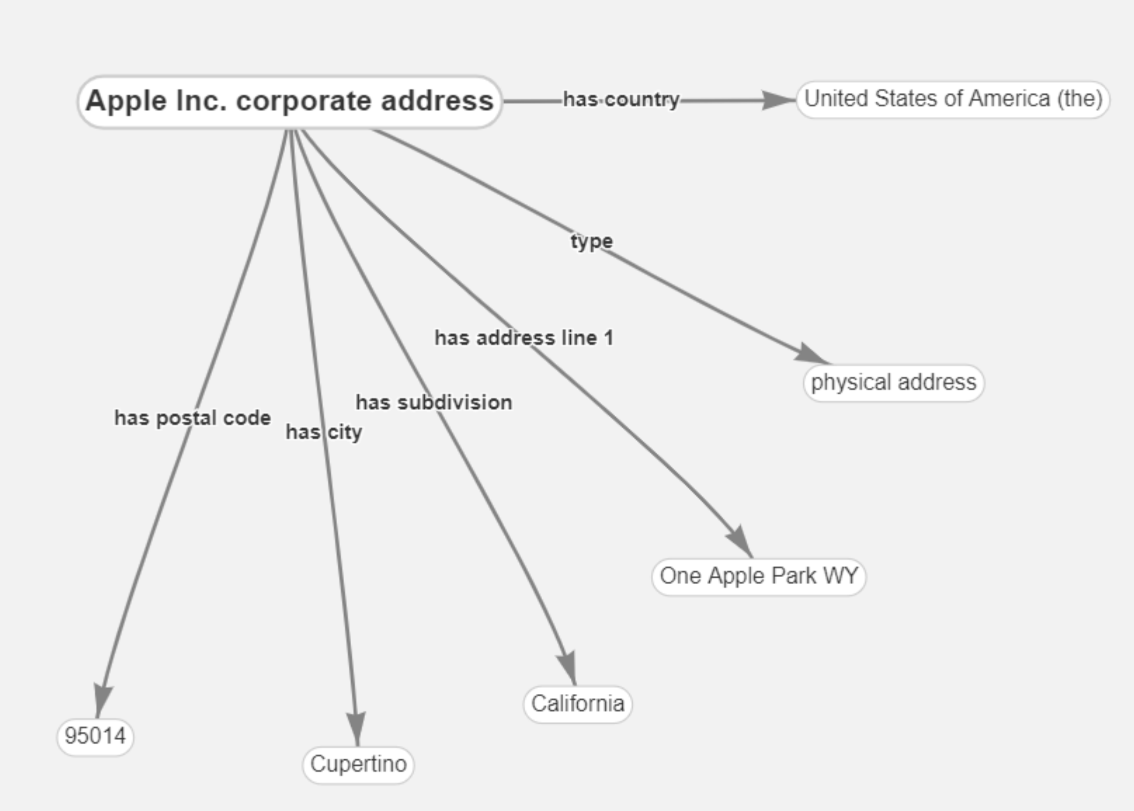


Figure 2.1: Financial Knowledge Graph of Apple.

Nowadays, the combination of AI and finance has attracted a number of companies and researchers working in the Financial Technology area. With the existing knowledge graph, it is also possible to combine and extend them to achieve new relations between the entities as shown in figure 2.2 below. Some Fintech companies such as Ontotext or Deloitte have focused on this area and provide services for financial companies. In this research, we will mainly use The Financial Industry Business Ontology(FIBO) published by EDM council as it is totally free for researchers to use.

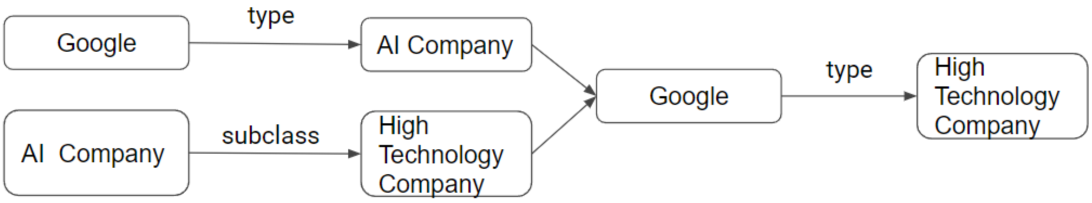


Figure 2.2: From Existing Knowledge to New Information

2.2 Deep learning based relation extraction

The purpose of RE is to identify semantic relations between entities from the corpus. There are various kinds of RE methods using deep learning technologies that have been developed in these years. With the wide use of deep learning techniques, traditional RE methods such as kernel-based or pattern-based methods have exposed to a variety of obstacles which especially rely on manually designed features so that the DNN-based methods have become the majority in RE tasks, especially the DNN-based supervised RE and distant supervision RE. Figure 2.3 describes the diagram of Deep learning-based relation extraction model structure:

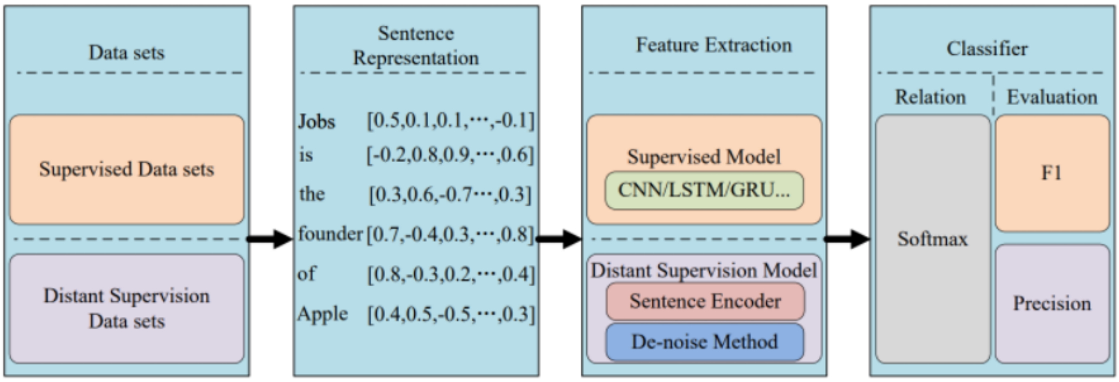

Data sets
Supervised Data sets
Distant Supervision Data sets
Sentence Representation
Jobs [0.5,0.1,0.1,···,-0.1]
is [-0.2,0.8,0.9,···,0.6]
the [0.3,0.6,-0.7···,0.3]
founder [0.7,-0.4,0.3,···,0.8]
of [0.8,-0.3,0.2,···,0.4]
Apple [0.4,0.5,-0.5,···,0.3]
Feature Extraction
Supervised Model
CNN/LSTM/GRU...
Distant Supervision Model
Sentence Encoder
De-noise Method
Classifier
Relation
Evaluation
Softmax
F1
Precision

Figure 2.3: General structure of the deep-learning-based relation extraction model [31]

In general, researchers mostly focus on supervised or distant supervision to define the relation extraction task model, while both of their process are formed by the same 4 parts:

1. **Dataset construction.** It is an essential part of most deep-learning-based models to achieve enough data and pre-process it to satisfy the requirements of the model. As a sub-task under the Natural Language Processing area, researchers of relation extraction take the same steps to pre-process the corpus data such as sentence/token split and tags/stop words removal. In the next step, researchers take different strategies to build the dataset for supervised and distant supervision models. Details on the mechanisms will be discussed in the next two sections.

2. **Word embedding.** Word embedding is the basic technique to transform natural language into high dimensional vectors which can be understood by the computer. Since the Word2Vec model was proposed by Mikolov et al. [17], the word embedding process has exceeded the one-hot vector representation and can be described as using the context in a sentence to predict the word(the CBOW model) or use the word to predict the context(the Skip-gram model). Based on the achievement of Mikolov et al., the GloVe model has been proposed by Pennington et al. which calculates the covariance matrix of the corpus to train the model and output the word vector which mostly contains the information from context and semantic [21].

   With the development of deep neural network, the BERT model proposed by Devlin et al. with bi-direction transformers comprehensively change the function of traditional NLP task including relation extraction [5]. Based on the Mask Language Model mission which randomly mask some tokens to predict the word, the BERT model can learn the context information for the whole paragraph but not limited to sentence-level [22].

   Position embedding.(PE) is another kind of feature that can enhance the sentence representation for the most encoder such as CNN-based model as they are hard to process the word location information. Usually the position embedding will indicate the relative distance for each entities between the remaining words in the sentence. Some RNN-based datasets such as SemEval 2010-task 8 [10] also use PE or position indicators (PI) to enhance the performance.

3. **Feature extraction model chosen.** To replace the manual feature extraction step in traditional methods, a series of models such as CNN/LSTM/GRU has been applied to automatically extract features from sentence embedding. A difference between supervised and distant supervision models is that the distant supervision dataset contains more noisy data. Therefore, a de-noise method is always needed for this type of methods (which are discussed further in section 2.4).

4. **Classifier.** For both supervised and distant supervision model it describes the relation extraction task as a classification mission, which aims to classify the input sentence as a specific relation type. So that most models use softmax as the classifier to output the result. For the evaluation methods, the supervised learning usually calculates F1 score:

$$F1 = 2 * \frac{precision * recall}{precision + recall} = \frac{TP}{TP + 1/2(FP + FN)}$$

while distant supervision only takes Precision or AUC (area under the precision-recall curve) to measure the result.

2.3 Supervised deep learning model for relation extraction

Various types of deep neural network models based on supervised learning such as CNN, RNN, LSTM has shown a significant performance on relation extraction task through appropriate variation and modifications. In these methods they formulate the relation extraction task as a multi-class classification problem. Those methods rely on human-annotated data and mostly are trained on general datasets such as TACRED [40] and SemEval 2010 Task-8 [10].

2.3.1 CNN-based model

A traditional choice for relation extraction is to use a CNN-based model as the sentence encoder to extract information from the text. Starting with Liu et al. in 2013 [15], they proposed the CNN model with synonym dictionary as the input to get a classification probability. In the following research, researchers try to reduce the work with NLP toolkits such as Part-of-speech(POS) tagging, tokenization and syntatic analysis, but directly extract the lexical features from the text. Based on Zeng et al. [39] who creatively consider the position embedding information, CNN-based models solve the problem of heavy pre-processing and the error propagation problem between multiply NLP toolkits. Other CNN-based models tend to combine with short dependency path(SDP) [34] or attention mechanism [32] to enhance the sentence representation.

2.3.2 RNN and LSTM-based models

To tackle the problem of the CNN models, which find it hard to consider global features and sequence information, the RNN-based models firstly used by Miwa et al.(2016) [20] to enhance the overall performance, especially for sentences with long-distance dependency between entity pairs. Most RNN-based methods use a Bi-LSTM network with an attention mechanism to enhance the representation for a long context. Figure 2.4 shows the basic structure of Bi-LSTM models proposed by Zhou et al.(2016) [41] As in the CNN-based model, the RNN-based model also tries to involve SDP information to support the decision making of the neural network. Both of them have achieved a high score on general datasets.

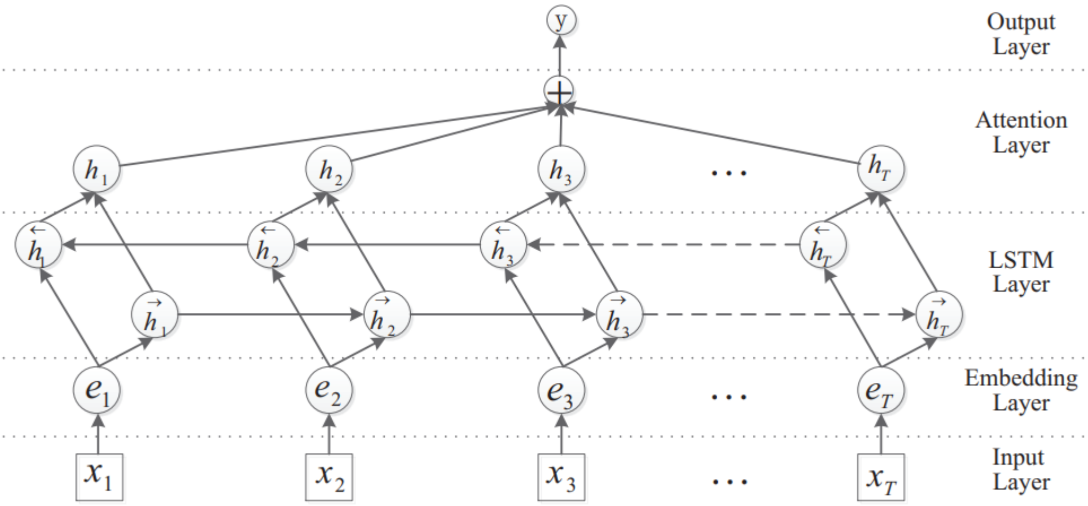


Figure 2.4: Diagram of the Bi-LSTM with Attention mechanism [41]

However, an obvious shortage of supervised learning for RE is the need for a large number of tagged datasets which may prove very costly indeed. By contrast, distant supervision has significant advantages to mitigate the requirement of the high accuracy of the tagged datasets, which has become an important research direction.

2.4 Distant supervision for relation extraction

Distant supervision is one possible way to implement relation extraction when large amounts of annotated data are not available. It is an approach to generate a large amount of tagged data from an existing knowledge base. In distant supervision, we make use of an existing database, such as Freebase or DBpedia, to collect examples for the relationship we want to extract. We then use these examples to automatically generate our training data. For example, Freebase contains the fact that Barack Obama and Michelle Obama are married. We take this fact, and then label each pair of “Barack Obama“ and “Michelle Obama“ that appear in the same sentence as a positive example for our marriage relation, and tag the entity pairs which does not present any relations as a negative example. This way we can easily generate a large amount of (possibly noisy) training data.

In this project, we focuses on the RE task for the 10-K report, and there is no open relation dataset in the financial area to satisfy the need of the supervised model. Therefore, distant supervision methods are more suitable for us to utilize.

The concept of distant supervision was first proposed by Mintz et al. [18] who pointed out the assumption: “If two entities participate in a relation, all sentences that mention these two entities can express that relation.” With the extension of the following researchers, this assumption has evolved into the assumption: "A relation holding between two entities can be either expressed explicitly or inferred implicitly from all sentences that mention these two entities." The first assumption [18] is too strong which may be due to a heavy noisy data problem so that it has been extended by researchers to the second assumption, which can capture more sentence features while it has been widely used in most of the recent research.

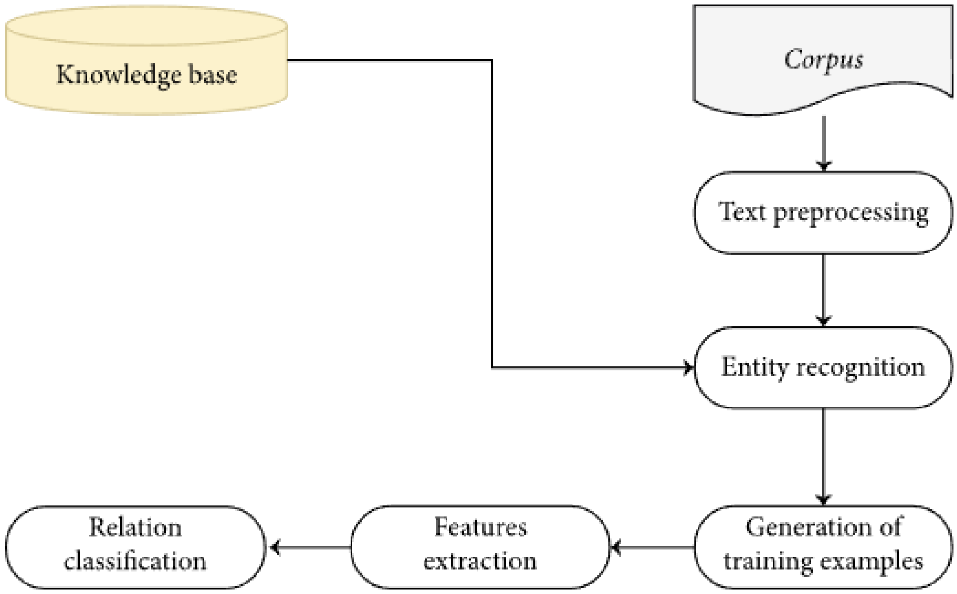


Figure 2.5: Diagram of the Distant Supervision Model [18]

Generally, a distant supervision relation extraction model is shown in figure 2.5. In the first step, the model needs to choose the appropriate knowledge base which can reflect the ontologies that we want to extract from the corpus. Then, through pre-processing the corpus to remove the non-text content, a named entity recognition task will be utilised to recognize the entities and match with the subject-object pairs from the knowledge base to generate the distant supervision dataset. At present, various state-of-the-art frameworks like Stanza's NER [23], Spacy [11], and NLTK [3] are presented as open-source libraries and easy to use by the researchers. Following the process, an important part of the relation extraction model is the feature extraction method. Classical methods mostly use hand-written rules or pre-defined features. For example, in the method proposed by Mintz et al. [18], there are two types of features considered:

1. Syntactic features: such as part of speech(POS) tags, dependency paths which link the pair of entities.

2. Lexical features: the context words which before or after the entity pairs, including their POS tags

With the development of artificial neural networks and deep learning techniques, this step has been replaced by neural models such as CNN, LSTM or BERT model. But in most cases, the final step of the distant supervision model still be defined as a classification problem to classify the relation to a certain type.

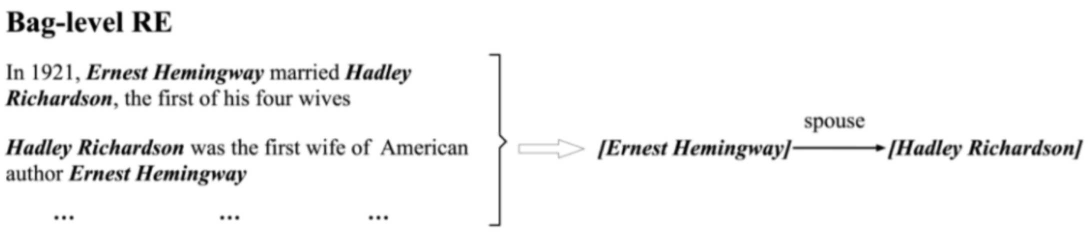
2.5 Bag-level models

Riedel et al. [27] point out the now familiar assumption: “A relation holding between two entities can be either expressed explicitly or inferred implicitly from all sentences that mention these two entities”, and propose a learning strategy known as Bag-level Relation Extraction. They found that following Mintz et al.’s initial assumption and tagging all sentences matching the entity pairs with the same assumed relation can cause a serious noise label problem. For example, if we have two sentences in the corpus:

1. Steven Jobs is the CEO of Apple.

2. Steven Jobs really like to eat the apple.

With the assumption of Mintz et al. [18], the model will tag those two sentences as the same relation as the knowledge base has even if they actually present irrelevant relations. To improve this, the bag-level RE model has been proposed by Riedel et al. [27] that is based on multi-instance learning. The key idea in multi-instance learning is to construct the training dataset by a series of bags with labels for classification, where each bag contains the instances without any tags on it. If at least one instance in the bag presents the correct relation between the entity pairs then the whole bag will be tagged as positive. On the contrary, if all the instances are negative then the bag will be tagged as negative either. Compare with supervised and unsupervised learning, this method effectively improves the accuracy of distant supervision. Figure 2.6 shows an example of the bag-level RE dataset.



Bag-level RE

In 1921, ***Ernest Hemingway*** married ***Hadley Richardson***, the first of his four wives

Hadley Richardson was the first wife of American author ***Ernest Hemingway***

Figure 2.6: Bag-level RE Dataset Construction [8]

At present, even if the bag learning strategy significantly enhances the performance of the distant supervision RE model, the de-noise method is still a key research direction in this area. In recent years, the main aspects of the distant supervision RE model can be divided into 3 different types:

1. **Sentence representation enhancement.** In general NLP task, the model needs to use word embedding to encode natural language text as a high dimensional vector which can be processed by the computer, then extract features from the vector representation. Apart from the general word embedding methods, such as Word2vec [17] or GloVe [21], some researchers recommend optimized embedding which adapts to the relation extraction tasks. In the DS-Joint model by Ren et al., they demonstrate the joint embedding process for both relation and entity and loss functions related to the modelling for relation type, entity type and the mutual information between relation and entities [26]. Another research by Su et al. demonstrates a novel method that uses global relation embedding generated from the knowledge graph to replace

the classic word embedding method [28]. Experiments from their research shows that their methods bring a significant improvement on an open domain dataset, such as NYT-10.
On the other hand, since application of the encoder-decoder models has been widely applied, researchers also focus to enhance the encoder structure to achieve better sentence encoding information. Most of these models are based on a variant of CNN or LSTM networks such as the PCNN model [38] or the Bi-LSTM model [16]. Using the BERT model introduced in 2018, Christou et al. proposed the REDSandT model which combined the ability of the transformer model and attention mechanism to capture the context information, and achieve the SOTA performance on the NYT-10 dataset [4].

2. **External knowledge involvement**. A possible direction to enhance the distant supervised relation extraction(DSRE) is to involve prior knowledge to improve the sparse features extracted from the natural language. In the early research, Zeng et al. [38] used position embedding as external information to construct the feature vector. In recent years, researchers focus on the work to enrol entity or knowledge graph related information, such as the entity description or the aliases of the relation.(e.g. “founded” and “co-founded” are aliases for the relation “founderOfCompany”) [30].

3. **Plug-and-play component**. The Plug-and-play component is a way to clean the dataset before the training step. With the development of GAN and reinforcement learning, some models such as DSGAN [24] and +RL [25] have been proposed and show an improvement in performance. These methods can be seen as an external part that separates from the main relation extraction pipeline.

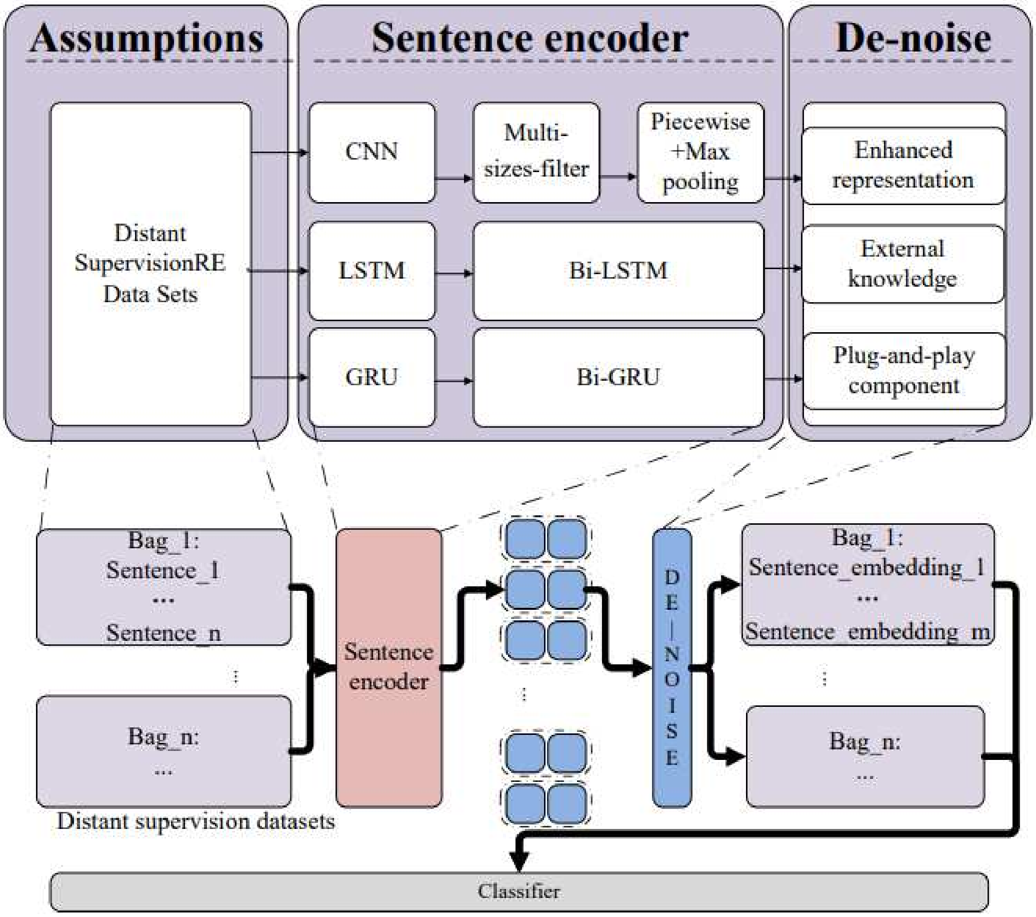


Figure 2.7: Diagram of the Distant Supervision Model [31]

Figure 2.7 shows the general architecture of the sentence-level(bag-level) distant supervision model. The Distant supervision RE datasets are generated by the alignment process between database and corpus. Various sentence encoders are the next step to encode the dataset as word embedding and position embedding (some methods may use more features). Then, use the de-noise methods to denoise the dataset and output it to the classifier and then output the correct result. Based on this structure, we initially choose 2 kinds of models for our project: PCNN+ATT [14] model and the BERT-based ATT model [8]. The PCNN+ATT model is a classical distant supervision model which is easy to reconstruct and usually used as a baseline, The BERT model is the improved version that replaces the PCNN encoder as BERT. Considering some pre-trained BERT models such as FinBERT [35] may suitable for us to enhance the result, the BERT-based model is our main research focus.

According to Mintz et al. [18], the evaluation method will use the held-out evaluation. In more details, the dataset will be separated into two different parts for training and testing. Then the Precision-Recall curve will be used to evaluate the performance of the model. In addition, considering that our dataset is not fully supervised and contains noise, the evaluation will also combine with the human evaluation step. Specifically, the top 100, 200 or 300 instances that have the highest score need to have their accuracy manually checked, after which the Top-N evaluation table is produced.

2.6 Document-level relation extraction

Sentence-level RE mainly focuses on extracting relations between entities in a sentence. However, a large amount of relation facts are hidden across multiple sentences, which is hard to extract by the sentence-level model. Statistical research from Yao et al. [36] demonstrates that at least 40.7% relation facts can only be extracted from multiple sentences. Figure 2.8 shows an example in the document-level RE task: the model needs to combine the relation in sentence 1 and sentence 8 to extract the relation semantic "located_in" between "Akron" and "St. Vincent–St.Mary High School".

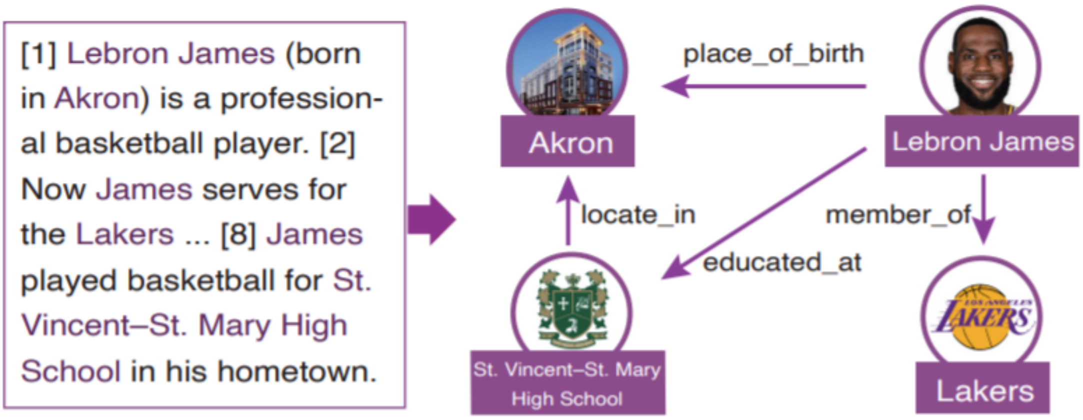

[1] Lebron James (born in Akron) is a professional basketball player. [2] Now James serves for the Lakers ... [8] James played basketball for St. Vincent–St. Mary High School in his hometown.
place_of_birth
Akron
Lebron James
locate_in
member_of
educated_at
St. Vincent–St. Mary High School
Lakers

Figure 2.8: Example of document-level RE [33]

Currently, most document-level RE models rely on high-quality tagged training data, which is costly and time-consuming. Thus, it is a valuable research direction to extenddistant supervision methods to the document-level. A important work in this area is the DocRED dataset and the models for document-level RE which published by Xiao et al [33]. To enhance the document-level performance, they used BERT as the document encoder to encode the input document into representation into entity mentions, entities and relational instances. Then they designed 3 pretrain tasks, which include: (1) **Mention-Entity Matching**, which aims to capture useful information from multiple mentions to produce informative representations for entities. (2) **Relation Detection**, which focuses on denoising "Not-A-Relation (NA)" and incorrectly labelled instances by detecting the entity pairs with relations. (3)**Relational Fact Alignment**, which requires the model to produce similar representations for the same entity pair from diverse expressions.

In addition, they proposed a rank model with the Relation Detection task on a human-annotated training set. They then use the rank model to give high scores to positive instances and low scores to NA instances. After that, they rank all the entity pairs in each document against their positive scores and keep the top entity pairs for pre-training, fine-tuning and evaluation. The evaluation result (the use of F1 score and IgnF1 score) on the DocRED dataset outperforms all the baselines they used, even exceeding the supervised methods such as HIN-BERT [29].

However, we may face difficulties if we want to construct the model to extract the relations in the 10-K report based on document-level, as the document-level dataset construction in the financial area involves both distant supervision and manual work, which is a challenge for a master-level project due to its limited duration. Thus, we leavethe document-level RE as our future research.

2.7 Anaphora resolution

Anaphora resolution, also known as pronoun resolution, is a NLP task that focuses on finding references to the front or the back items in the discourse. Figure 2.9 shows an example of anaphora resolution. Here the model needs to recognize the nouns and pronouns in the content and link those indicating the same entity. In the 2000s Ruslan Mitkov addressed the anaphora resolution question in detail, thus advancing the development in this area [19]. For our needs, the anaphora resolution task can mostly be seen as a sub-task belonging to co-reference resolution, which can be handled by dependency parsing-based neural model.

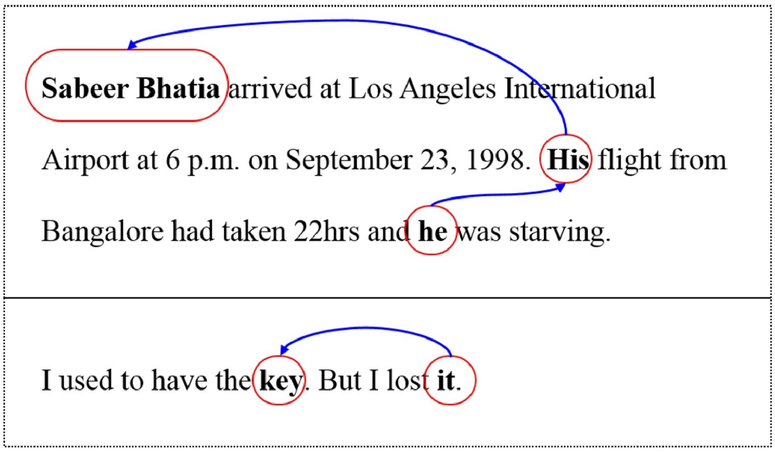

Sabeer Bhatia arrived at Los Angeles International
Airport at 6 p.m. on September 23, 1998. His flight from
Bangalore had taken 22hrs and he was starving.
I used to have the key. But I lost it.

Figure 2.9: Task of Anaphora Resolution [13]

The coreference function in Stanford Core NLP package provides a simple interface for the users to analyse the possible coreference chains (including anaphora relations) in paragraphs. We decide not to re-implement the anaphora model by ourselves but focus on extracting extra ontologies via the anaphora resolution algorithm.

Chapter 3

Problem analysis

In this chapter we discuss the choices for the source of data, methodology and evaluation used in this work.

3.1 Data source

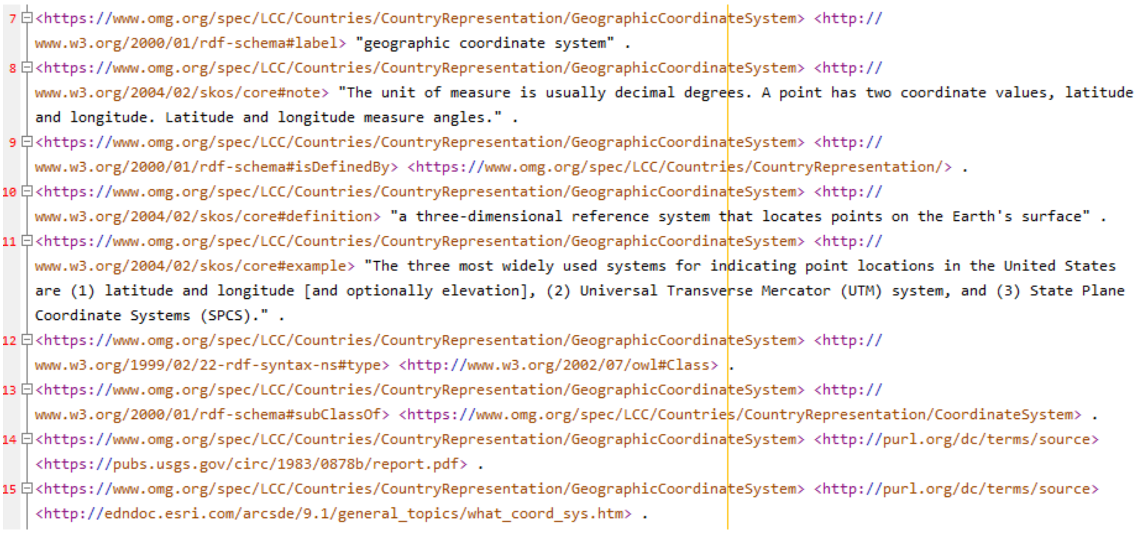
In this project, our aim is to evaluate the performance of the distant supervision model when applied to financial reports. We choose the 10-K reports between 2013 to 2016 as the corpus for our experiment. This data was collected by Eric He [9] and freely available online.

To find an appropriate knowledge base, we choose FIBO ontology from EDM Concil to map the corpus for potential entity pairs. For the anaphora resolution task, we will use the data collected and provided by Can Erten as part of his PhD research, which contains the list between companies and important employees from the SEC reports. In contrast with the whole FIBO ontology, this dataset allows an easier and more precise identification of the named entities present in it. Figures 3.1, 3.2 and 3.3 show an example data from the FIBO ontology, company - employee list, and the 10-K reports from SEC website. In figures 3.1 and 3.3, we use N-Triples form file that downloads from FIBO website. Each line of the file contains a triple in the form of (subject, predicate, object). For example, the first row in figure contains a triple (GeographicCoordinateSystem, label, geographic coordinate system). For the prefix, which is the HTML-style string before the last words of entities or relations, we directly remove them as this task only consider extract the relations but not merge with an existed knowledge graph.

3.2 Model selection and preparation

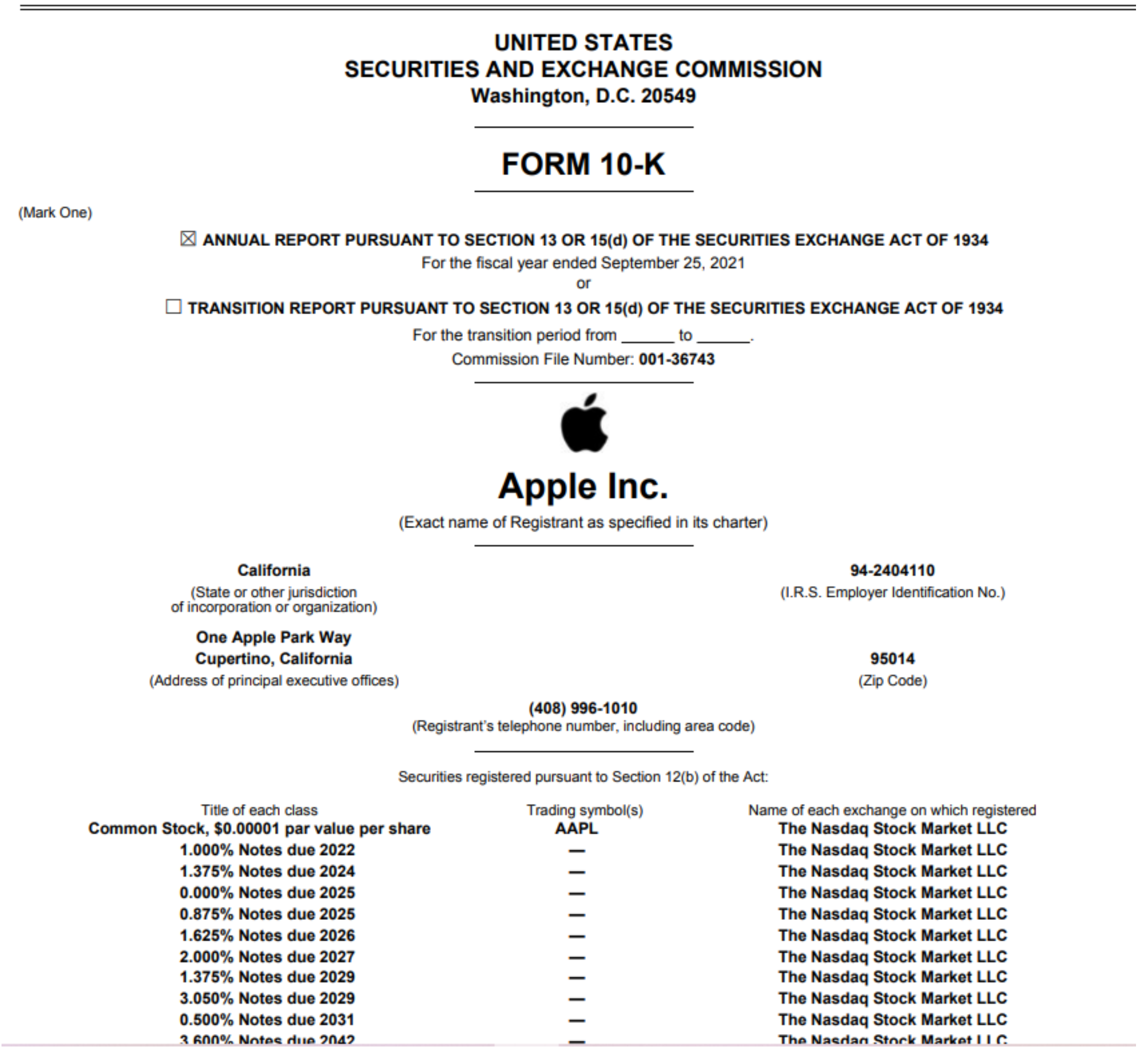
3.2.1 Text preprocessing

As the example data shows above, the original data contains a certain amount of irrelvant information such as HTML tags, tables or graphs. We pre-process the FIBO ontology to remove the prefix and form pure triples. For the 10-K reports we use the dataset which published by Eric He at data.world that contains all the cleaned/parsed report between 2013 to 2016 [9]. For the company and employee list we also take the same step to re-organise the data as the triple form. For example, in the first two rows of figure 3.3 indicate that ¡http://sec.com/0001455142¿ is a type of person and ¡http://sec.com/0001600125¿ is a type of company. Then from the 6th row and 14th row we achieve the name of the person is LaGreca Carl and



```
<https://www.omg.org/spec/LCC/Countries/CountryRepresentation/GeographicCoordinateSystem> <http://
www.w3.org/2000/01/rdf-schema#label> "geographic coordinate system" .
<https://www.omg.org/spec/LCC/Countries/CountryRepresentation/GeographicCoordinateSystem> <http://
www.w3.org/2004/02/skos/core#note> "The unit of measure is usually decimal degrees. A point has two coordinate values, latitude
and longitude. Latitude and longitude measure angles." .
<https://www.omg.org/spec/LCC/Countries/CountryRepresentation/GeographicCoordinateSystem> <http://
www.w3.org/2000/01/rdf-schema#isDefinedBy> <https://www.omg.org/spec/LCC/Countries/CountryRepresentation/> .
<https://www.omg.org/spec/LCC/Countries/CountryRepresentation/GeographicCoordinateSystem> <http://
www.w3.org/2004/02/skos/core#definition> "a three-dimensional reference system that locates points on the Earth's surface" .
<https://www.omg.org/spec/LCC/Countries/CountryRepresentation/GeographicCoordinateSystem> <http://
www.w3.org/2004/02/skos/core#example> "The three most widely used systems for indicating point locations in the United States
are (1) latitude and longitude [and optionally elevation], (2) Universal Transverse Mercator (UTM) system, and (3) State Plane
Coordinate Systems (SPCS)." .
<https://www.omg.org/spec/LCC/Countries/CountryRepresentation/GeographicCoordinateSystem> <http://
www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.w3.org/2002/07/owl#Class> .
<https://www.omg.org/spec/LCC/Countries/CountryRepresentation/GeographicCoordinateSystem> <http://
www.w3.org/2000/01/rdf-schema#subClassOf> <https://www.omg.org/spec/LCC/Countries/CountryRepresentation/CoordinateSystem> .
<https://www.omg.org/spec/LCC/Countries/CountryRepresentation/GeographicCoordinateSystem> <http://purl.org/dc/terms/source>
<https://pubs.usgs.gov/circ/1983/0878b/report.pdf> .
<https://www.omg.org/spec/LCC/Countries/CountryRepresentation/GeographicCoordinateSystem> <http://purl.org/dc/terms/source>
<http://edndoc.esri.com/arcsde/9.1/general_topics/what_coord_sys.htm> .
```

Figure 3.1: Example of FIBO Ontology [2]



UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549

FORM 10-K

(Mark One)

☒ **ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934**
For the fiscal year ended September 25, 2021
or
☐ **TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934**
For the transition period from ______ to ______.
Commission File Number: **001-36743**

Apple Inc.
(Exact name of Registrant as specified in its charter)

California
(State or other jurisdiction
of incorporation or organization)

94-2404110
(I.R.S. Employer Identification No.)

One Apple Park Way
Cupertino, California
(Address of principal executive offices)

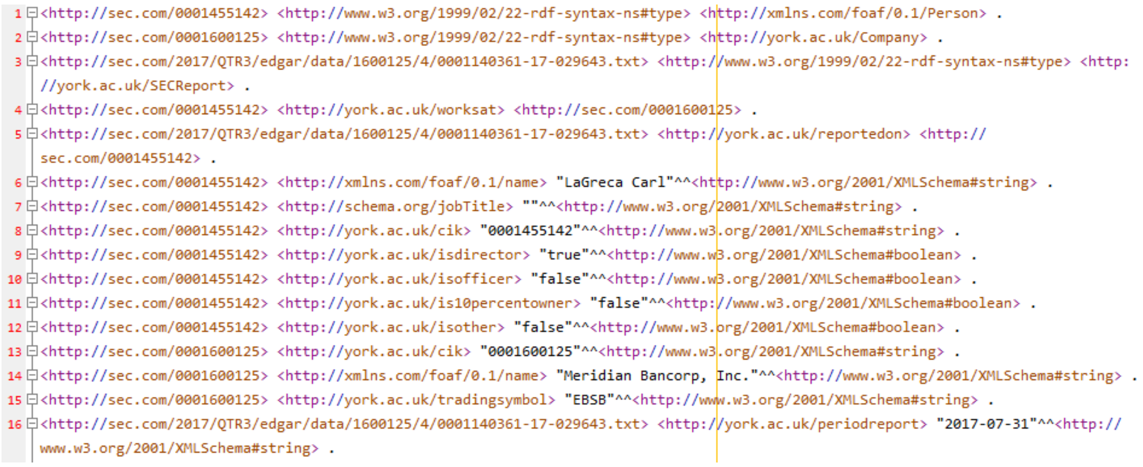
95014
(Zip Code)

(408) 996-1010
(Registrant's telephone number, including area code)

Securities registered pursuant to Section 12(b) of the Act:

Title of each class	Trading symbol(s)	Name of each exchange on which registered
Common Stock, $0.00001 par value per share	**AAPL**	**The Nasdaq Stock Market LLC**
1.000% Notes due 2022	**—**	**The Nasdaq Stock Market LLC**
1.375% Notes due 2024	**—**	**The Nasdaq Stock Market LLC**
0.000% Notes due 2025	**—**	**The Nasdaq Stock Market LLC**
0.875% Notes due 2025	**—**	**The Nasdaq Stock Market LLC**
1.625% Notes due 2026	**—**	**The Nasdaq Stock Market LLC**
2.000% Notes due 2027	**—**	**The Nasdaq Stock Market LLC**
1.375% Notes due 2029	**—**	**The Nasdaq Stock Market LLC**
3.050% Notes due 2029	**—**	**The Nasdaq Stock Market LLC**
0.500% Notes due 2031	**—**	**The Nasdaq Stock Market LLC**
3.600% Notes due 2042	**—**	**The Nasdaq Stock Market LLC**

Figure 3.2: Example of 10-K report from Apple Inc.



```
1 <http://sec.com/0001455142> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/Person> .
2 <http://sec.com/0001600125> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://york.ac.uk/Company> .
3 <http://sec.com/2017/QTR3/edgar/data/1600125/4/0001140361-17-029643.txt> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http:
  //york.ac.uk/SECReport> .
4 <http://sec.com/0001455142> <http://york.ac.uk/worksat> <http://sec.com/0001600125> .
5 <http://sec.com/2017/QTR3/edgar/data/1600125/4/0001140361-17-029643.txt> <http://york.ac.uk/reportedon> <http://
  sec.com/0001455142> .
6 <http://sec.com/0001455142> <http://xmlns.com/foaf/0.1/name> "LaGreca Carl"^^<http://www.w3.org/2001/XMLSchema#string> .
7 <http://sec.com/0001455142> <http://schema.org/jobTitle> ""^^<http://www.w3.org/2001/XMLSchema#string> .
8 <http://sec.com/0001455142> <http://york.ac.uk/cik> "0001455142"^^<http://www.w3.org/2001/XMLSchema#string> .
9 <http://sec.com/0001455142> <http://york.ac.uk/isdirector> "true"^^<http://www.w3.org/2001/XMLSchema#boolean> .
10 <http://sec.com/0001455142> <http://york.ac.uk/isofficer> "false"^^<http://www.w3.org/2001/XMLSchema#boolean> .
11 <http://sec.com/0001455142> <http://york.ac.uk/is10percentowner> "false"^^<http://www.w3.org/2001/XMLSchema#boolean> .
12 <http://sec.com/0001455142> <http://york.ac.uk/isother> "false"^^<http://www.w3.org/2001/XMLSchema#boolean> .
13 <http://sec.com/0001600125> <http://york.ac.uk/cik> "0001600125"^^<http://www.w3.org/2001/XMLSchema#string> .
14 <http://sec.com/0001600125> <http://xmlns.com/foaf/0.1/name> "Meridian Bancorp, Inc."^^<http://www.w3.org/2001/XMLSchema#string> .
15 <http://sec.com/0001600125> <http://york.ac.uk/tradingsymbol> "EBSB"^^<http://www.w3.org/2001/XMLSchema#string> .
16 <http://sec.com/2017/QTR3/edgar/data/1600125/4/0001140361-17-029643.txt> <http://york.ac.uk/periodreport> "2017-07-31"^^<http://
  www.w3.org/2001/XMLSchema#string> .
```

Figure 3.3: Example of company-employee list [6]

the company name is Meridian Bancorp, Inc. Finally, the 4th row indicates that LaGreca Carl, work at, work at Meridian Bancorp, Inc. and then get the triple: (LaGreca Carl, work at, Meridian Bancorp, Inc.)

3.2.2 Relation extraction algorithms

Following the result of our Literature review, although the researchers have published a series of models to enhance the performance of distant supervision relation extraction. Considering the limitation of our data source, it is hard to automatic involve suitable external knowledge for the financial RE model without domain knowledge or expert participation, while also hard to train a complex plug-and-play model such as DSGANs to de-noise the dataset in a short time period, we thereforeaim to build two simple models to verify the performance of distant supervision methods. The first method we chosed is the PCNN model + attention mechanism which is a classic model on the DSRE area, while the second one is the BERT-based model that replace the encoder as BERT and use attention mechanism to optimize the output. The model details will be discussed in the section 4.2. On the other hand, we also notice the fact that distant relation extraction method is limited by the existing knowledge base. So we also want to investigate whether anaphora resolution function can effectively extend our knowledge base or not.

3.3 Evaluation design

To evaluate the model performance, we will follow the held-out evaluation method performed Mintz et al. [18]. Specifically, the held-out method is to] randomly choose a part of data from the original dataset as the test dataset after it is created. Then use the AUC value (area under precision-recall curve) to measure the performance of the model. Consider the noise problem of distant supervision model]and the time limitation of this master project, we will not try to totally analysis the dataset or manually verify the whole dataset, but it is also necessary to check the precision for the top 100, 200, and 300 instances manually. For the anaphora resolution part, it is hard to define automatically whether the ontology is meaningful or not. It is however possible to conduct a subjective manual evaluation if the size of resulting dataset is not too large.

Chapter 4

Design and implementation

Based on the problem analysis above, our model structure for relation extraction on 10-K reports is described in Figure 4.1. Starting with the preprocessing of the 10-K report corpus downloaded from SEC website and the FIBO ontology, we construct our dataset in the form of dictionary in python which satisfied the requirement of our distant supervision model (see appendix), then use this dataset for the model training and evaluation and finally get the relations stored in a RDF format.

4.1 Dataset construction

Following the paradigm of the distant supervision method, we construct our dataset as algorithm 1 shows. We choose the ontology from the FIBO website between 2010 and 2020 as the knowledge base, and the SEC 10-K reports between 2013–2016 (22165 reports in total) as the corpus. Considering the relation distribution of FIBO ontology are not balanced, some relations with fewer frequencies have been removed from the knowledge base to avoid negative effect on the model. As table 4.1 shows, 3 relation types "is Defined by", "Subclass of", "identifies" have been chosen from our knowledge base to map to the financial corpus, while some high-frequency relations such as "type" or "label" are not chosen because the ontology based on these relations are meaningless.We also add the "NA" relation as a negative instance, meaning the two entities in the pair have no relation between them.

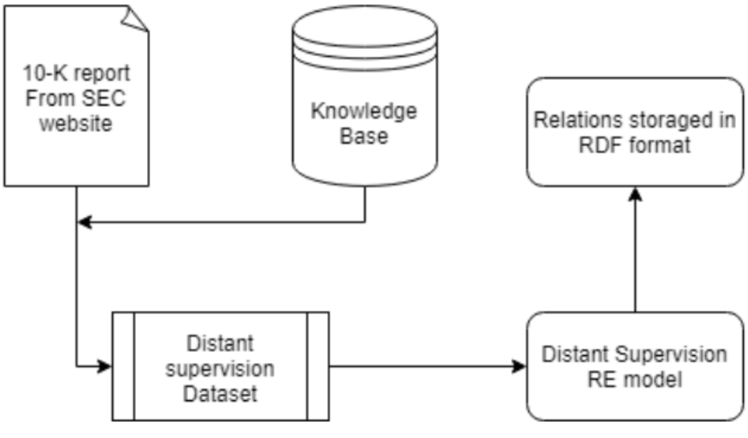

10-K report From SEC website
Knowledge Base
Relations storaged in RDF format
Distant supervision Dataset
Distant Supervision RE model

Figure 4.1: Diagram of relation extraction from 10-K SEC reports

Table 4.1: Relation frequency calculation for FIBO ontology

Relation	Frequency	Relation	Frequency	Relation	Frequency
type	21240	*identifies*	2676	operates In Municipa	2038
is Defined By	12058	is Member Of	2461	has Exchange Name	1956
label	4975	has Tag	2271	range	1306
definition	4231	has Website	2068	domain	1072
subClass Of	4048	operates in Country	2038	is Constituent Of	883

In the following step, we split the reports into sentences by using the NLTK package. For each sentence, we use the 'ne.chunk' function to recognize the entities and match them with subject and object of each triple in knowledge base. If they matched then we tag the sentence with the relation from that triple. Continue repeating this process for all the reports then the original dataset can be generated. Finally, we split the dataset in 60:30:10 ratio to produce the training, validation and test parts of the dataset.

Algorithm 1: Dataset construction by distant supervision

```
Input: Report corpus R; Triple list T From FIBO Ontology
Output: Dictionary D indicated the sentence and relation between entities
for Sentence $S \in R$ do
    Extract entities $E \in S$ with NLTK
    for Triple $t \in T$ do
        for Entities $e1, e2 \in t$ do
            if $e1, e2 \in E$ then
                D add Sentence S, relation $r \in t$
            end
        end
    end
end
```

4.2 Experiment design

To verify the performance of distant supervision methods on financial relation extraction, two models have been built on the dataset generatedfrom the 10-K reports. The first is the PCNN model developed by Zeng et al. [38] with the structure shown in figure 4.2. the input sentence vector combines both word embedding and position embedding. The context information around the entities are captured in to the convolution layer. Through the convolution process to max pooling the features and finally the relation is classified after the softmax layer. The second BERT-based model replaces the PCNN sentence encoder with BERT encoder, while both models use the instance-level attention mechanism to enhance the performance of relation extraction. The structure of the BERT encoder is shown on Figure 4.3:

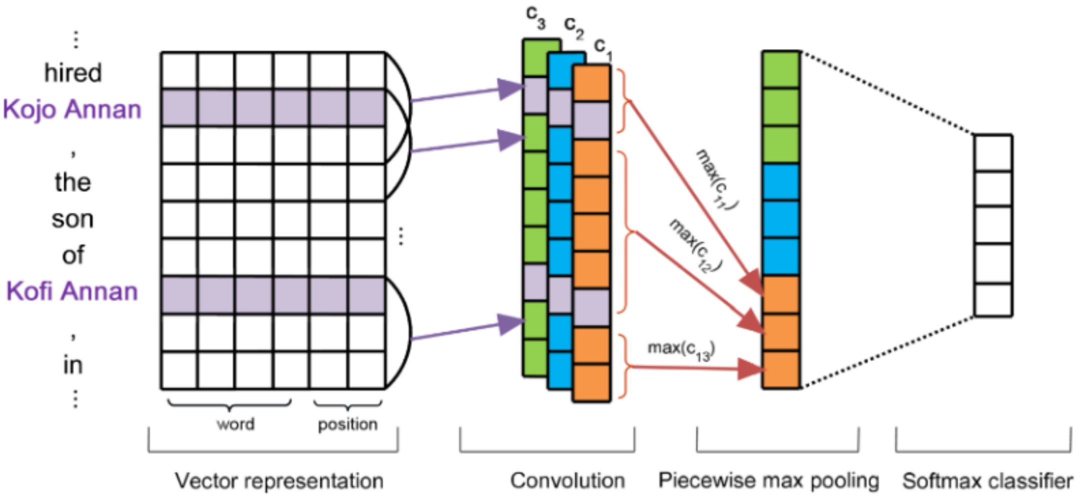


Figure 4.2: Structure of PCNN model [38]

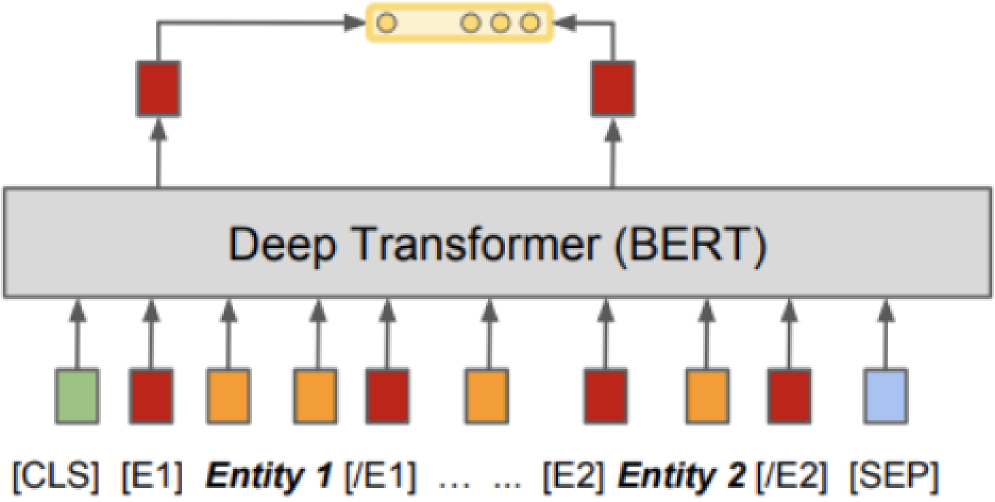


Figure 4.3: Structure of BERT Encoder [1]

We use the same architectural parameters as the original model proposed by Zeng et al. [38]in PCNN as follow: we set the windows size is 3 with 230 feature maps; the word dimension and position dimension is 50 and 5; and with the dropout probability 0.5. For the BERT-based model we construct it with the OpenNRE platform and set the BERT encoder as same as Soares et al.'s model [1].

We construct the model based on the OpenNRE package [8] and use the NLTK package as our NER toolkit. Because of the limitation of open source NER tools such as NLTK, Spacy or Stanza, some entities from the FIBO ontology cannot be recognized correctly to the pre-defined type so we use string matching as a replacement. Both models ran on the Viking cluster server at the University of York. We use one GPU node that contains NVIDIA Tesla V100 with 20GB RAM and train both models to converge.

Chapter 5

Experiment and results

This section presents the result after evaluating the efficacy of the proposed technique and lists the key findings and observations from the experimental evaluations. In addition, we introduce another data enhancement experiment we designed to extend our knowledge base.

5.1 Data preparation

Starting with the algorithm and model structure we described above, we construct our dataset on 42519 rows in total, divide it for training (25200 rows) , validation (12600 rows) and test set (4719 rows), as shown in table 5.1. We also calculate the structure of other datasets based on the common knowledge base such as Wikidata or DBpedia to compare the data distribution, which ensure our datasets follow the same percentage of train/vaild/test set and appropriate instances number per relation. In table 5.1, the NYT10 dataset is a popular dataset created by Riedel et al. which links the New York Times corpus to the Freebase ontology [27]. The dataset NYT10m and Wiki20m were proposed by Han et al. which manually verified the test set [8]. The GIDs dataset was proposed by Jat et al. who re-balanced the frequency of 5 relations from NYT-10 dataset [12]. Since the dataset construction algorithm takes days to run through, and our Master by research project only takes one year, this limits our potential to repeat experiments or to involve more relations or to manually re-balance the dataset as NYT10m or GIDs does. Thus, we decide not to change the current 10-K dataset for our evaluation of the performance of the distant supervision model.

Table 5.1: Statistics on the existing datasets. All relation types include the NA relation

Dataset	Total	Train	Val	Test	Percentage	Relations	Relation percentage
GIDS	18824	11297	1864	5663	60:30:10	6	3171 per relation
NYT10	695059	522611	0	172448	75:25	58	12194 per relation
NYT10m	475401	417893	46422	11086	88:9:2	25	19016 per relation
Wiki20m	901314	598721	137986	64507	78:15:7	81	11127 per relation
10-Ks(Ours)	42519	25200	12600	4719	60:30:10	4	10629 per relation

5.2 Evaluation of the distant supervision model

As the work described in the previous section, we trained the PCNN+ATT model with a 0.5 learning rate, 160 batch size and 0.00001 weight decay rate. We also try both word2vec and GloVe as the word embedding based on the general text without significant differences in the performance. For the BERT-based model, we train it with a learning rate of 0.00002 and 16 batch size. The evaluation of the result is done by the held-out evaluation. We split the training and testing datasets in the dataset construction step and make the instances mutually exclusive. We use a Precision-Recall curve to evaluate our result and manually evaluate the Top 100, 200 and 300 instances that get the highest score. Figure 5.1 shows the curve of Precision-Recall of our experiment and table 5.2 shows the AUC and Precision for 100, 200, 300 values.

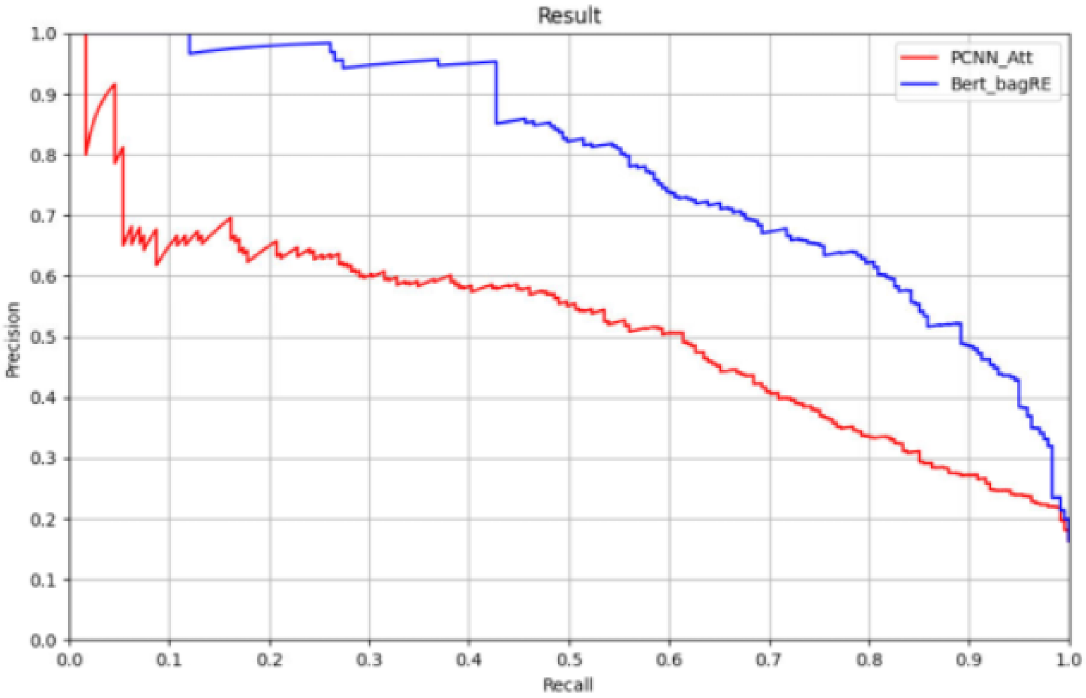


Figure 5.1: Precision-Recall curve of PCNN+ATT and BERT-based model [38]

Table 5.2: AUC and P@N evaluation results. P@N represents precision calculated for the top N rated relation instances

RE method	AUC	P@100	P@200	P@300
PCNN+ATT	0.59595	76.0	61.0	48.7
BERT+ATT	0.81033	91.0	82.0	66.0

Based on the result above, the experiments demonstrate that distant supervision method can extract the relations from the 10-K report, while the BERT-based model shows better performance than the baseline model on either the AUC score and P@N value. Since the limitation of the financial corpus and knowledge base, it is hard to formulate other de-noise method that involve the external knowledge such as the alias of relation or description of entities. It is also a valuable research direction to investigate how to combine multiple de-noise methods to enhance the performance of relation extraction tasks in the financial area.

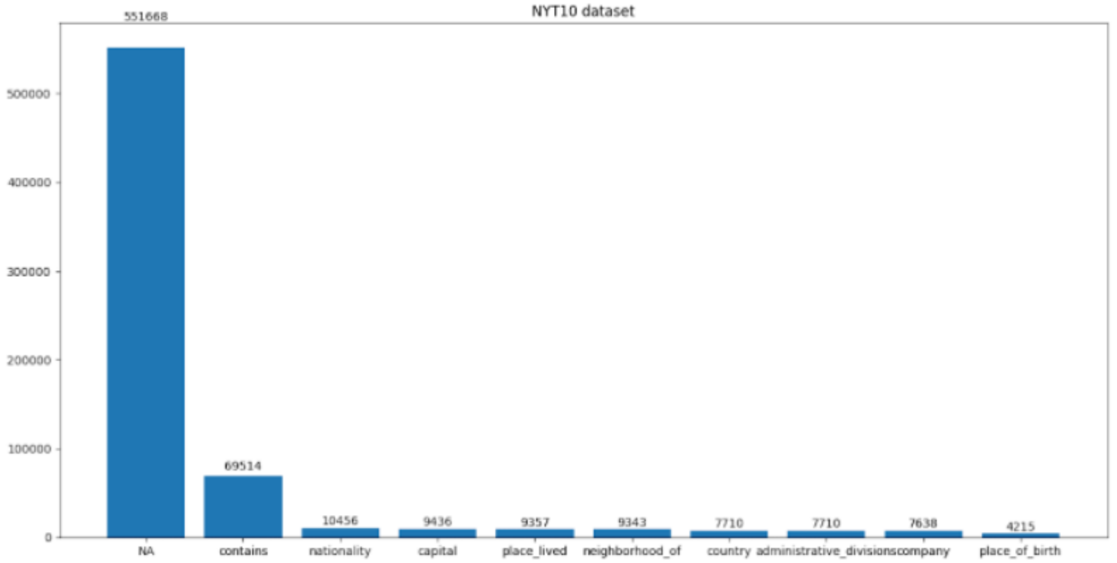


Figure 5.2: Relation distribution on NYT-10 dataset for top 10 training instances rank by frequency, other instances are quite a few appeared

The result we have achieved (0.81 AUC score) demonstrates the efficiency of the BERT-based model performance in the small amount dataset. However, considering we only involve 4 different relation types which are significantly less than the general dataset such as NYT-10, it is necessary to extend our current dataset to involve more relations and make a comprehensive test on our model. Also, since the relation distribution on the FIBO ontology is not balanced, constructing a large dataset may also lead to the relation distribution in a long-tail situation. According to the research of Han et al. [7], the long-tail means: in a dataset, most training instances are related to the common relations while the other relations only have very few instances or sentences. Figure 5.2 shows the relation distribution on the NYT-10 dataset, most of the instances are the “NA” type (means no relation between entities) and the relation “contains”, while the other relation types are significantly less frequent.

Table 5.3: 10-K Dataset relation distribution

is Defined By	**NA**	**subClass Of**	**identifies**
20171	13061	7179	2108

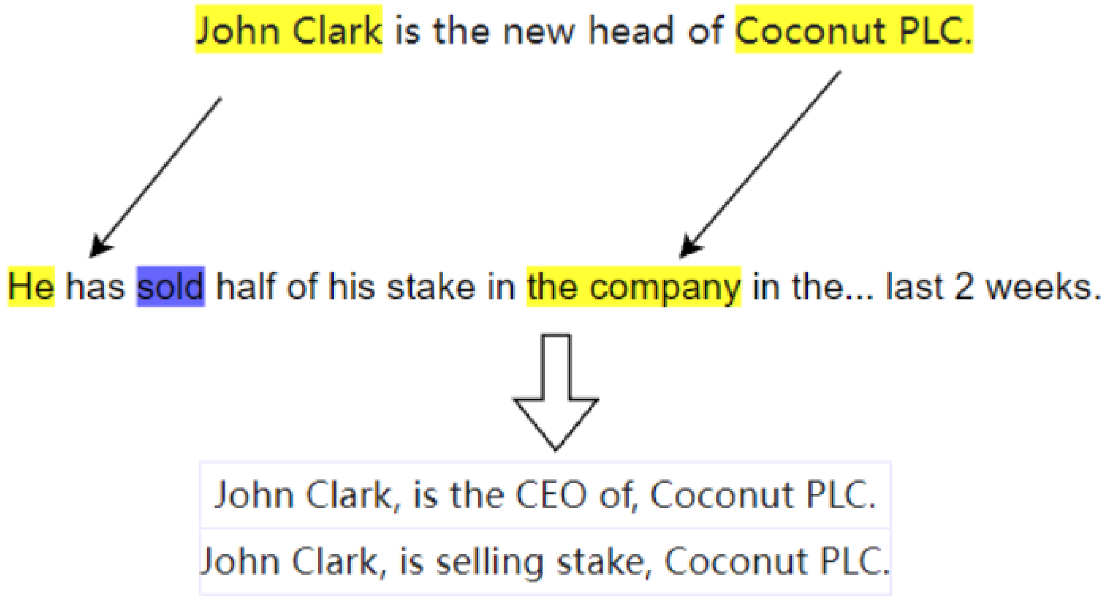
Table 5.4: Result distribution for PCNN and BERT-based model. The table shows the precision value for each relation type

	isDefinedBy	**identifies**	**subClassOf**
PCNN	0.6876	0.4681	0.2481
BERT	0.8287	0.4597	0.6164

In addition to the statistic shown in table 5.3, our dataset also shows a long-tail distribution, with the instances of the relation “is Defined By” significantly exceeding the rest. The result reflects a disbalance in the distribution in table 5.4. For further research, we expect to enrol more relations to our dataset and investigate the improvement of long-tail problems in our model. Strategies such as few-shot learning (Train representations of instances that can adapt from existing large-scale data [7]) or Meta-learning (Grasp the way of parameter initialization and optimization through the experience gained on the meta-train data [7]) appear to be possible research directions to solve this problem.

5.3 Enhancing the knowledge base with anaphora resolution

With regards to the result above, the distant supervision model demonstrates the ability to extract the relations from the financial corpus effectively. However, the experiment also exposes some flaws which affect the application in the financial area. A key problem is that the distant relation extraction highly relies on the existing knowledge base such as FIBO, which means that the distant supervision model does not have the ability to tackle the unseen relationship from the text. To relieve this problem, we propose a method combined with the anaphora resolution step to extract potential relations which can enhance the performance of the knowledge base. For a given knowledge base and corpus, the anaphora resolution can help to explore new relations. First, we will find the sentences which contain the entity pair and assume that the other sentences may also contain the pronoun of person and company entity, which can indicate some new relations between them. In addition, it is also possible that not two entities are both involved in the content, in the experiment we also try to match the situation that only one entity is involved.



John Clark is the new head of Coconut PLC.

He has sold half of his stake in the company in the... last 2 weeks.

John Clark, is the CEO of, Coconut PLC.

John Clark, is selling stake, Coconut PLC.

Figure 5.3: Example of anaphora resolution

Figure 5.3 shows an example that we want to achieve from this task. To start with a triple: [John Clark, Is the CEO of, Coconut PLC.], we shall find the same name entity in the sentence which includes the entity pair. Then, in the remaining content of the corpus, the model shall find the entity on the same conference chain that indicates “John Clark” and “he”, “Coconut PLC” and “the company” are the same entities. So that some potential relations such as “selling stake” can be mined from the corpus. Additionally, it is possible that not two entities are both involved in the content, in the experiment we also tried to match the situation that one entity appeared.

The pseudocode of the whole process are shown in Algorithm 2 below:

Algorithm 2: Use Anaphora resolution to enhance the exist knowledge base

```
Input: Report corpus R; Triple list T From Company-person Ontology
Output: New Triples Tn
Dataset preprocessing:
for Report $r \in R$ do
    for triple $t \in T$ do
        for Sentence $s \in R$ do
            if Company $c \in t$ in R, Person $p \in t$ in s then
                Package s and 3 sentences after s as a paragraph P.
                Add P into dataset D.
            end
        end
    end
end
Anaphora relation extraction:
for Search paragraph P in D do
    Extract entities $e \in P$ with Stanford CoreNLP
    for paragraph $P \in D$ do
        Extract coreference chain $C \in P$ with Stanford CoreNLP
    end
    for entity $e \in C$, person $p \in t$ do
        if entity $e = p$ then
            Use Stanford OpenIE to extract the ontology o from sentence
             $s \in c$ if ontology o in C then
                Return o
            end
        end
    end
end
```

Table 5.5: Illustration of result examples from the anaphora resolution method

Original knowledge base	**Extract triples**
{'ARC Group, Inc.', 'CEO', 'Kasturi Seenu G.'}	{'He', 'also serves as', 'President of DWG Acquisitions'}
{'CABOT CORP', 'Director', 'Keohane Sean D'}	{'He', 'was appointed', 'President of Reinforcement Materials'}
{'DXP ENTERPRISES INC', 'Employee', 'Jeffery John Jay'}	{'He', 'oversees', 'strategic direction'}

In this experiment. we choose the company-employee ontologies from Erten and Kazakov's paper [6] as the knowledge base, whose structure is easier to analyse than the FIBO ontology. The knowledge base involves all the companies and their employee's relationships from the SEC report between 2017Q1 and 2018Q4. As algorithm 2 shows, we firstly match all the 22165 reports to the specific company names, then from the filtered reports (11532 matched), run the Stanza NER tools to recognize the person entity for each sentence. In the following process, the algorithm tries to identify if the employee has appeared in the sentence and then use Stanford CoreNLP to identify the coreference chain from the entity sentence to the following 3 sentences. Finally, the algorithm will use the Open relation extraction model in Stanza for the sentence which including the pronoun on it. Table 5.5 demonstrate our anaphora resolution algorithm's result. Starting with 44544 ontologies between company and employees. our model extract 4748 extra ontologies where 177 directly indicates a pronoun and a new relation. Others who do not find the pronouns also shows extended relations which may useful to the financial analysis.

Through our experiment, the anaphora resolution shows the potential to become a data enhancement method that can involve more information in the knowledge base. However, we also found some of the relation results are repeated or meaningless. For example, the OpenIE model recognizes the ontologies repeatedly: ('he','held','management positions'), ('he','held','management positions including President'), ('he','held','management positions including President of Merrill Lynch Consumer Markets'). As the example shows above, the OpenIE model can extract new relations from the text but need further processing for detailed ontology information. Another problem is the Stanford CoreNLP consuming over 24 hours on annotating text of length 25,000 characters. It also finds it hard to deal with the long dependency, which limits our algorithm in handling long paragraphs when searching for relations. In our future research, we will investigate other models for long dependency anaphora resolution to extract ontology relations from this type of text.

Chapter 6

Conclusion

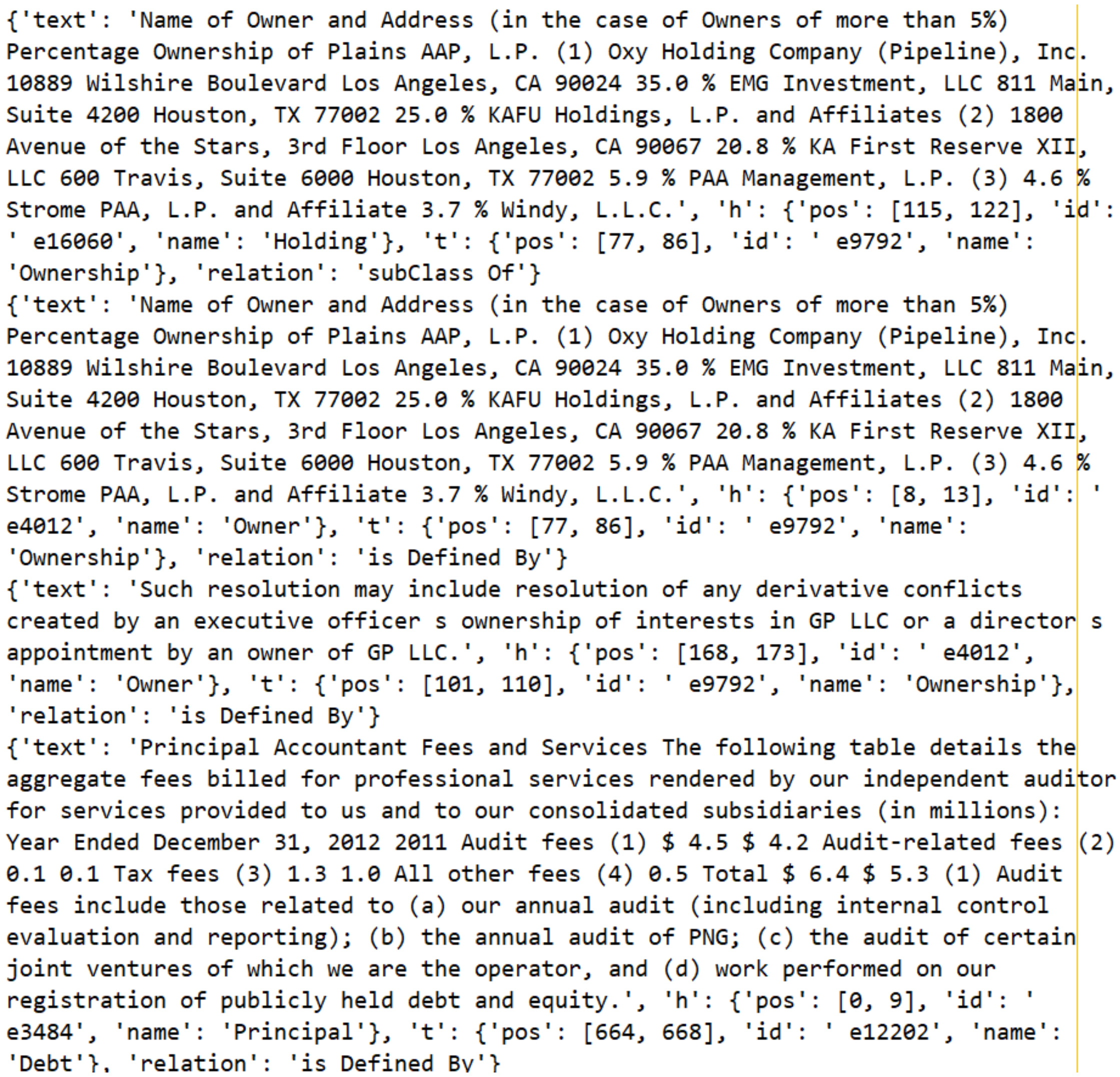
In this research, our main goal was to investigate the combination of distant supervised relation extraction techniques with financial reports from the SEC website. We formulated the PCNN and BERT model + attention mechanism to test the performance on the dataset which links the 10-K reports to the FIBO ontology. We demonstrate the effectiveness of the BERT-based model on financial text datasets – it achieved a 0.81 AUC score on the 10-K dataset that we have selected. However, the experiment also exposes the shortage of distant supervision models that cannot extract any new relations out of the knowledge base. To solve the question above, we also introduce dataset enhancement methods based on the anaphora resolution. We design our algorithm based on the company-employee dataset from Erten and Kazakov's paper and extract 4748 new relations while 177 directly indicate the new relation based on the employees. The time limit of this master project did not permit to experiment with more datasets from the SEC reports to further test the performance, but the current result shows the potential of this algorithm, which could be applied as a preprocessing step to enhance the knowledge base before applying the distant supervision method.

This work mainly achieved the relation extraction task with the distant supervision method on the 10-K reports for the first time. It demonstrates the performance of the PCNN/BERT-based distant supervision model on our dataset. In addition, this work also explore a novel method to extend the existing knowledge base and achieve new relations based on existing one with the anaphora resolution method, and indicated that this method still has large room for improvement. With respect to this dataset, we expect to extend our relation types so that they can reflect the complex distribution condition in the real world to evaluate the performance of our model comprehensively. In addition, we will investigate methods to combine our anaphora resolution algorithm with the distant supervision model. Not only extract the relations based on existing knowledge base but can explore new relation types efficiently.

For further research, we will investigate more relation extraction methods including plug-to-play methods to enhance our model performance, also investigate the combination of few-shot learning or meta-learning to mitigate the long-tail problem on the dataset. Considering the supplement of the original knowledge base can include more ontologies as the external knowledge to improve the model performance, it is valuable to extend the anaphora method to more relations but not limited to the company-person pairs. Also, reliable evaluation criteria need to be explored in further work.

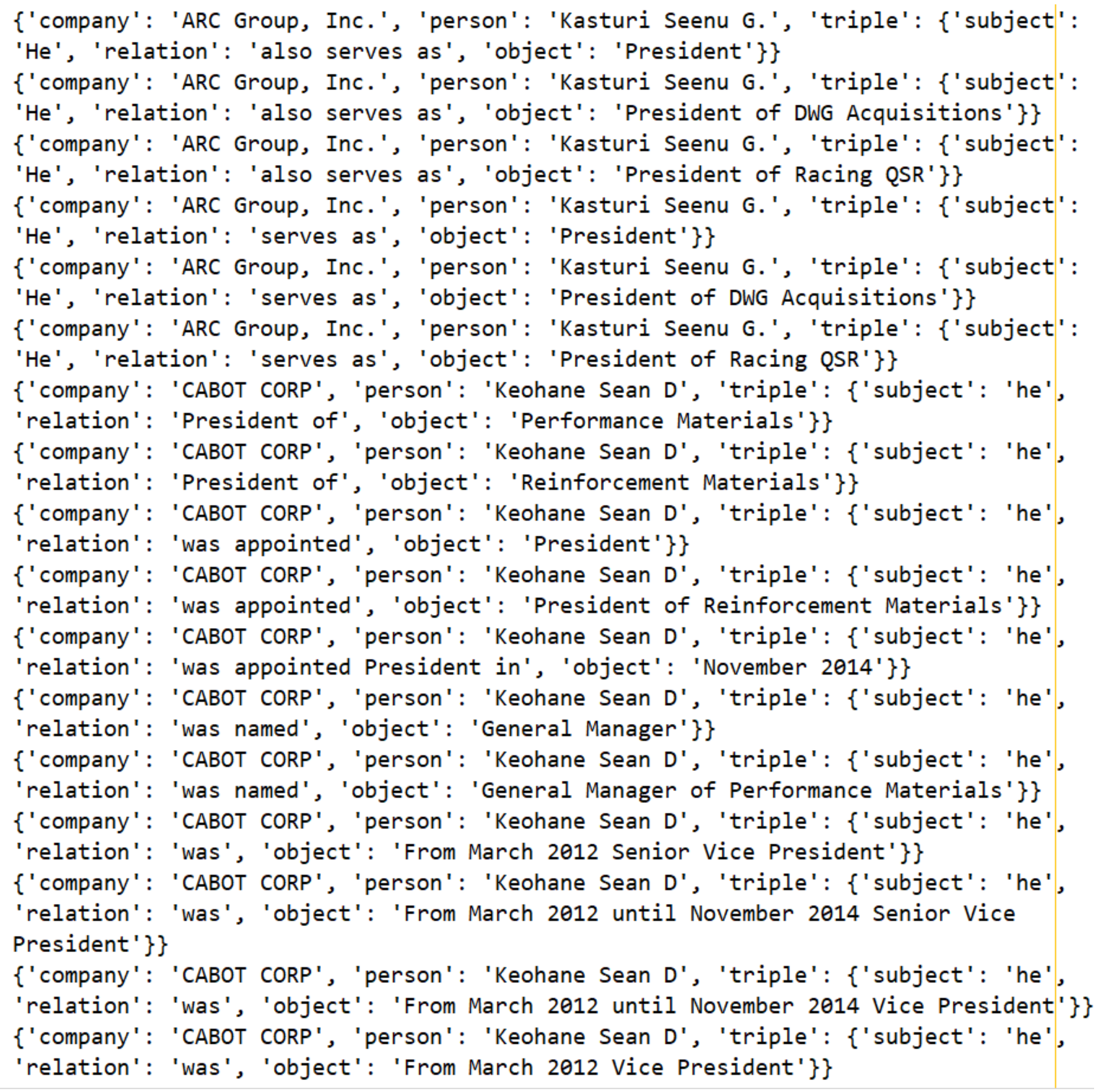
Appendix

Introduction of NYT10 dataset:
https://paperswithcode.com/dataset/new-york-times-annotated-corpus



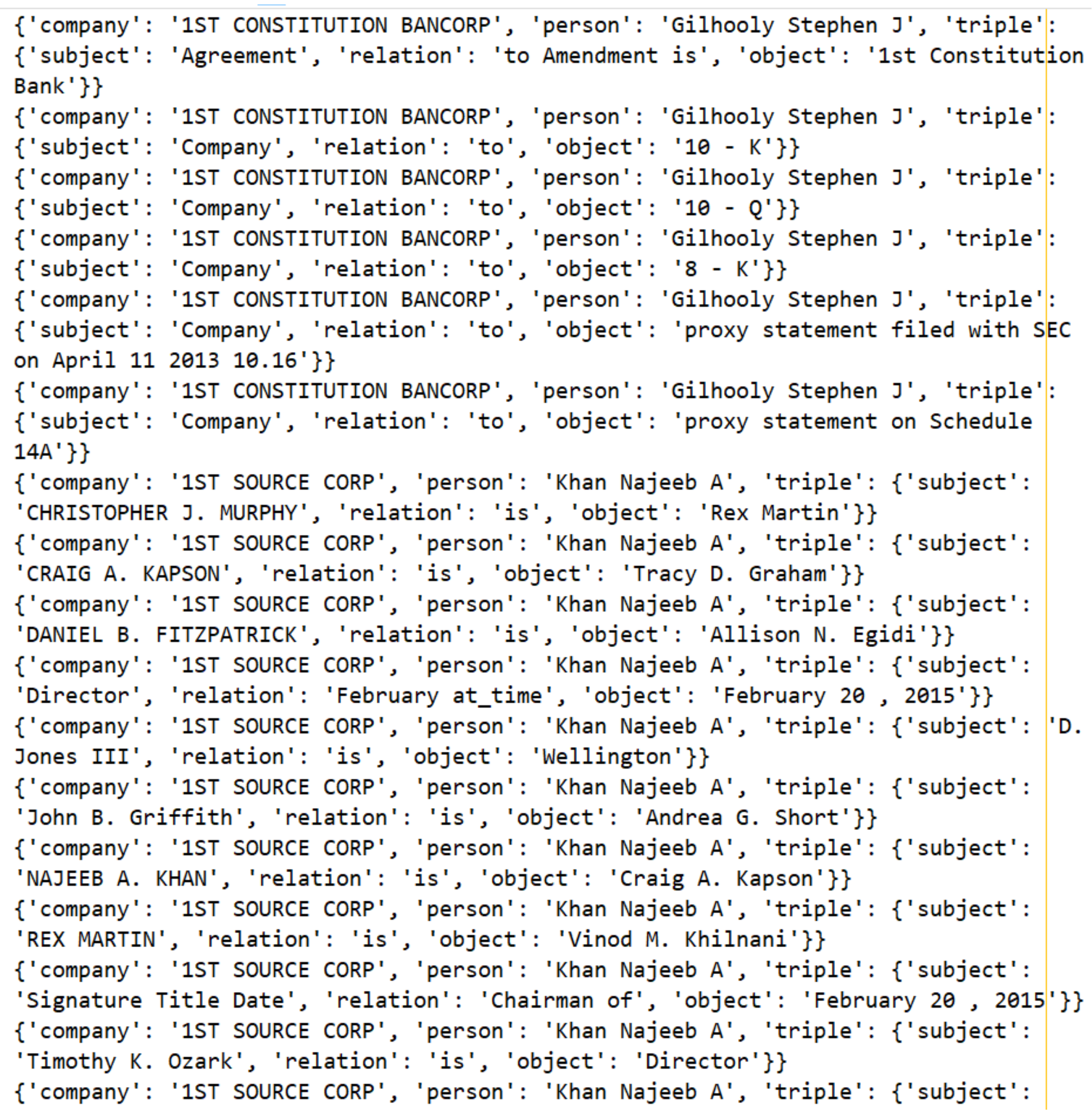
```
{'text': 'Name of Owner and Address (in the case of Owners of more than 5%)
Percentage Ownership of Plains AAP, L.P. (1) Oxy Holding Company (Pipeline), Inc.
10889 Wilshire Boulevard Los Angeles, CA 90024 35.0 % EMG Investment, LLC 811 Main,
Suite 4200 Houston, TX 77002 25.0 % KAFU Holdings, L.P. and Affiliates (2) 1800
Avenue of the Stars, 3rd Floor Los Angeles, CA 90067 20.8 % KA First Reserve XII,
LLC 600 Travis, Suite 6000 Houston, TX 77002 5.9 % PAA Management, L.P. (3) 4.6 %
Strome PAA, L.P. and Affiliate 3.7 % Windy, L.L.C.', 'h': {'pos': [115, 122], 'id':
' e16060', 'name': 'Holding'}, 't': {'pos': [77, 86], 'id': ' e9792', 'name':
'Ownership'}, 'relation': 'subClass Of'}
{'text': 'Name of Owner and Address (in the case of Owners of more than 5%)
Percentage Ownership of Plains AAP, L.P. (1) Oxy Holding Company (Pipeline), Inc.
10889 Wilshire Boulevard Los Angeles, CA 90024 35.0 % EMG Investment, LLC 811 Main,
Suite 4200 Houston, TX 77002 25.0 % KAFU Holdings, L.P. and Affiliates (2) 1800
Avenue of the Stars, 3rd Floor Los Angeles, CA 90067 20.8 % KA First Reserve XII,
LLC 600 Travis, Suite 6000 Houston, TX 77002 5.9 % PAA Management, L.P. (3) 4.6 %
Strome PAA, L.P. and Affiliate 3.7 % Windy, L.L.C.', 'h': {'pos': [8, 13], 'id': '
e4012', 'name': 'Owner'}, 't': {'pos': [77, 86], 'id': ' e9792', 'name':
'Ownership'}, 'relation': 'is Defined By'}
{'text': 'Such resolution may include resolution of any derivative conflicts
created by an executive officer s ownership of interests in GP LLC or a director s
appointment by an owner of GP LLC.', 'h': {'pos': [168, 173], 'id': ' e4012',
'name': 'Owner'}, 't': {'pos': [101, 110], 'id': ' e9792', 'name': 'Ownership'},
'relation': 'is Defined By'}
{'text': 'Principal Accountant Fees and Services The following table details the
aggregate fees billed for professional services rendered by our independent auditor
for services provided to us and to our consolidated subsidiaries (in millions):
Year Ended December 31, 2012 2011 Audit fees (1) $ 4.5 $ 4.2 Audit-related fees (2)
0.1 0.1 Tax fees (3) 1.3 1.0 All other fees (4) 0.5 Total $ 6.4 $ 5.3 (1) Audit
fees include those related to (a) our annual audit (including internal control
evaluation and reporting); (b) the annual audit of PNG; (c) the audit of certain
joint ventures of which we are the operator, and (d) work performed on our
registration of publicly held debt and equity.', 'h': {'pos': [0, 9], 'id': '
e3484', 'name': 'Principal'}, 't': {'pos': [664, 668], 'id': ' e12202', 'name':
'Debt'}, 'relation': 'is Defined By'}
```

Figure 6.1: Example of a distant supervision dataset. “h” and “t” represent the head and tail of the entity (subject or object), “pos” indicate the position of entities inside the text, “id” is a randomly generated unique identifier



```
{'company': 'ARC Group, Inc.', 'person': 'Kasturi Seenu G.', 'triple': {'subject':
'He', 'relation': 'also serves as', 'object': 'President'}}
{'company': 'ARC Group, Inc.', 'person': 'Kasturi Seenu G.', 'triple': {'subject':
'He', 'relation': 'also serves as', 'object': 'President of DWG Acquisitions'}}
{'company': 'ARC Group, Inc.', 'person': 'Kasturi Seenu G.', 'triple': {'subject':
'He', 'relation': 'also serves as', 'object': 'President of Racing QSR'}}
{'company': 'ARC Group, Inc.', 'person': 'Kasturi Seenu G.', 'triple': {'subject':
'He', 'relation': 'serves as', 'object': 'President'}}
{'company': 'ARC Group, Inc.', 'person': 'Kasturi Seenu G.', 'triple': {'subject':
'He', 'relation': 'serves as', 'object': 'President of DWG Acquisitions'}}
{'company': 'ARC Group, Inc.', 'person': 'Kasturi Seenu G.', 'triple': {'subject':
'He', 'relation': 'serves as', 'object': 'President of Racing QSR'}}
{'company': 'CABOT CORP', 'person': 'Keohane Sean D', 'triple': {'subject': 'he',
'relation': 'President of', 'object': 'Performance Materials'}}
{'company': 'CABOT CORP', 'person': 'Keohane Sean D', 'triple': {'subject': 'he',
'relation': 'President of', 'object': 'Reinforcement Materials'}}
{'company': 'CABOT CORP', 'person': 'Keohane Sean D', 'triple': {'subject': 'he',
'relation': 'was appointed', 'object': 'President'}}
{'company': 'CABOT CORP', 'person': 'Keohane Sean D', 'triple': {'subject': 'he',
'relation': 'was appointed', 'object': 'President of Reinforcement Materials'}}
{'company': 'CABOT CORP', 'person': 'Keohane Sean D', 'triple': {'subject': 'he',
'relation': 'was appointed President in', 'object': 'November 2014'}}
{'company': 'CABOT CORP', 'person': 'Keohane Sean D', 'triple': {'subject': 'he',
'relation': 'was named', 'object': 'General Manager'}}
{'company': 'CABOT CORP', 'person': 'Keohane Sean D', 'triple': {'subject': 'he',
'relation': 'was named', 'object': 'General Manager of Performance Materials'}}
{'company': 'CABOT CORP', 'person': 'Keohane Sean D', 'triple': {'subject': 'he',
'relation': 'was', 'object': 'From March 2012 Senior Vice President'}}
{'company': 'CABOT CORP', 'person': 'Keohane Sean D', 'triple': {'subject': 'he',
'relation': 'was', 'object': 'From March 2012 until November 2014 Senior Vice
President'}}
{'company': 'CABOT CORP', 'person': 'Keohane Sean D', 'triple': {'subject': 'he',
'relation': 'was', 'object': 'From March 2012 until November 2014 Vice President'}}
{'company': 'CABOT CORP', 'person': 'Keohane Sean D', 'triple': {'subject': 'he',
'relation': 'was', 'object': 'From March 2012 Vice President'}}
```

Figure 6.2: Example of the result of anaphora resolution algorithm: these triples do not contain the pronoun, but some of them are meaningful. Note that the “company“ and “person“ are indicate the original entity pair we used.



```
{'company': '1ST CONSTITUTION BANCORP', 'person': 'Gilhooly Stephen J', 'triple':
{'subject': 'Agreement', 'relation': 'to Amendment is', 'object': '1st Constitution
Bank'}}
{'company': '1ST CONSTITUTION BANCORP', 'person': 'Gilhooly Stephen J', 'triple':
{'subject': 'Company', 'relation': 'to', 'object': '10 - K'}}
{'company': '1ST CONSTITUTION BANCORP', 'person': 'Gilhooly Stephen J', 'triple':
{'subject': 'Company', 'relation': 'to', 'object': '10 - Q'}}
{'company': '1ST CONSTITUTION BANCORP', 'person': 'Gilhooly Stephen J', 'triple':
{'subject': 'Company', 'relation': 'to', 'object': '8 - K'}}
{'company': '1ST CONSTITUTION BANCORP', 'person': 'Gilhooly Stephen J', 'triple':
{'subject': 'Company', 'relation': 'to', 'object': 'proxy statement filed with SEC
on April 11 2013 10.16'}}
{'company': '1ST CONSTITUTION BANCORP', 'person': 'Gilhooly Stephen J', 'triple':
{'subject': 'Company', 'relation': 'to', 'object': 'proxy statement on Schedule
14A'}}
{'company': '1ST SOURCE CORP', 'person': 'Khan Najeeb A', 'triple': {'subject':
'CHRISTOPHER J. MURPHY', 'relation': 'is', 'object': 'Rex Martin'}}
{'company': '1ST SOURCE CORP', 'person': 'Khan Najeeb A', 'triple': {'subject':
'CRAIG A. KAPSON', 'relation': 'is', 'object': 'Tracy D. Graham'}}
{'company': '1ST SOURCE CORP', 'person': 'Khan Najeeb A', 'triple': {'subject':
'DANIEL B. FITZPATRICK', 'relation': 'is', 'object': 'Allison N. Egidi'}}
{'company': '1ST SOURCE CORP', 'person': 'Khan Najeeb A', 'triple': {'subject':
'Director', 'relation': 'February at_time', 'object': 'February 20 , 2015'}}
{'company': '1ST SOURCE CORP', 'person': 'Khan Najeeb A', 'triple': {'subject': 'D.
Jones III', 'relation': 'is', 'object': 'Wellington'}}
{'company': '1ST SOURCE CORP', 'person': 'Khan Najeeb A', 'triple': {'subject':
'John B. Griffith', 'relation': 'is', 'object': 'Andrea G. Short'}}
{'company': '1ST SOURCE CORP', 'person': 'Khan Najeeb A', 'triple': {'subject':
'NAJEEB A. KHAN', 'relation': 'is', 'object': 'Craig A. Kapson'}}
{'company': '1ST SOURCE CORP', 'person': 'Khan Najeeb A', 'triple': {'subject':
'REX MARTIN', 'relation': 'is', 'object': 'Vinod M. Khilnani'}}
{'company': '1ST SOURCE CORP', 'person': 'Khan Najeeb A', 'triple': {'subject':
'Signature Title Date', 'relation': 'Chairman of', 'object': 'February 20 , 2015'}}
{'company': '1ST SOURCE CORP', 'person': 'Khan Najeeb A', 'triple': {'subject':
'Timothy K. Ozark', 'relation': 'is', 'object': 'Director'}}
{'company': '1ST SOURCE CORP', 'person': 'Khan Najeeb A', 'triple': {'subject':
```

Figure 6.3: Example of the result of anaphora resolution algorithm: these are triples that not contain the pronoun; most of them indicate new information based on the original company-person list. Note that the "company" and "person" indicate the original entity pair we used.

Bibliography

[1] Baldini Soares, L., FitzGerald, N., Ling, J., and Kwiatkowski, T. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy, July 2019), Association for Computational Linguistics, pp. 2895–2905.

[2] Bennett, M. The financial industry business ontology: Best practice for big data. *Journal of Banking Regulation 14*, 3-4 (2013), 255–268.

[3] Bird, S., Klein, E., and Loper, E. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[4] Christou, D., and Tsoumakas, G. Improving distantly-supervised relation extraction through bert-based label & instance embeddings.

[5] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, Minnesota, June 2019), Association for Computational Linguistics, pp. 4171–4186.

[6] Erten, C., and Kazakov, D. L. Ontology graph embeddings and ilp for financial forecasting. In *Inductive Logic Programming, Proceedings of the 30th International Conference:* (2021), Springer.

[7] Han, X., Gao, T., Lin, Y., Peng, H., Yang, Y., Xiao, C., Liu, Z., Li, P., Sun, M., and Zhou, J. More data, more relations, more context and more openness: A review and outlook for relation extraction.

[8] Han, X., Gao, T., Yao, Y., Ye, D., Liu, Z., and Sun, M. OpenNRE: An open and extensible toolkit for neural relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations* (Hong Kong, China, Nov. 2019), Association for Computational Linguistics, pp. 169–174.

[9] He, E. 2013-2016 cleaned/parsed 10-k filings with the sec - dataset by jumpyaf. https://data.world/jumpyaf/2013-2016-cleaned-parsed-10-k-filings-with-the-sec, Jul 2017.

[10] Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., and Szpakowicz, S.

SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation* (Uppsala, Sweden, July 2010), Association for Computational Linguistics, pp. 33–38.

[11] Honnibal, M., and Montani, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017.

[12] Jat, S., Khandelwal, S., and Talukdar, P. P. Improving distantly supervised relation extraction using word and entity based attention. *ArXiv abs/1804.06987* (2017).

[13] Lee, C., Jung, S., and Park, C.-E. Anaphora resolution with pointer networks. *Pattern Recognition Letters 95* (2017), 1–7.

[14] Lin, Y., Shen, S., Liu, Z., Luan, H., and Sun, M. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Stroudsburg, PA, USA, 2016), Association for Computational Linguistics.

[15] Liu, C., Sun, W., Chao, W., and Che, W. Convolution neural network for relation extraction. In *Advanced data mining and applications*, H. Motoda, Ed., vol. 8347 of *LNCS sublibrary: SL 7 - Artificial intelligence.* Springer, Heidelberg, 2013, pp. 231–242.

[16] Liu, T., Zhang, X., Zhou, W., and Jia, W. Neural relation extraction via inner-sentence noise reduction and transfer learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels, Belgium, Oct.-Nov. 2018), Association for Computational Linguistics, pp. 2195–2204.

[17] Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR 2013* (01 2013).

[18] Mintz, M., Bills, S., Snow, R., and Jurafsky, D. 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: Volume 2 - Volume 2* (United States, 2009), K.-Y. Su, Ed., ACM Digital Library, Association for Computational Linguistics, p. 1003.

[19] Mitkov, R., and Sb, W. W. Anaphora resolution: The state of the art. Tech. rep., 1999.

[20] Miwa, M., and Bansal, M. End-to-end relation extraction using lstms on sequences and tree structures.

[21] Pennington, J., Socher, R., and Manning, C. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha, Qatar, Oct. 2014), Association for Computational Linguistics, pp. 1532–1543.

[22] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (New Orleans, Louisiana, June 2018), Association for Computational Linguistics, pp. 2227–2237.

[23] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (2020).

[24] Qin, P., Xu, W., and Wang, W. Y. DSGAN: Generative adversarial training for distant supervision relation extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Melbourne, Australia, July 2018), Association for Computational Linguistics, pp. 496–505.

[25] Qin, P., Xu, W., and Wang, W. Y. Robust distant supervision relation extraction via deep reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Melbourne, Australia, July 2018), Association for Computational Linguistics, pp. 2137–2147.

[26] Ren, X., Wu, Z., He, W., Qu, M., Voss, C. R., Ji, H., Abdelzaher, T. F., and Han, J. Cotype: Joint extraction of typed entities and relations with knowledge bases. In *Proceedings of the 26th International Conference on World Wide Web* (Republic and Canton of Geneva, CHE, 2017), WWW '17, International World Wide Web Conferences Steering Committee, p. 1015–1024.

[27] Riedel, S., Yao, L., and McCallum, A. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, J. L. Balcázar, F. Bonchi, A. Gionis, and M. Sebag, Eds., vol. 6323 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 148–163.

[28] Su, Y., Liu, H., Yavuz, S., Gür, I., Sun, H., and Yan, X. Global relation embedding for relation extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (New Orleans, Louisiana, June 2018), Association for Computational Linguistics, pp. 820–830.

[29] Tang, H., Cao, Y., Zhang, Z., Cao, J., Fang, F., Wang, S., and Yin, P. Hin: Hierarchical inference network for document-level relation extraction. *Advances in Knowledge Discovery and Data Mining 12084* (2020), 197 – 209.

[30] Vashishth, S., Joshi, R., Prayaga, S. S., Bhattacharyya, C., and Talukdar, P. RESIDE: Improving distantly-supervised neural relation extraction using side information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels, Belgium, Oct.-Nov. 2018), Association for Computational Linguistics, pp. 1257–1266.

[31] Wang, H., Qin, K., Zakari, R. Y., Lu, G., and Yin, J. Deep neural network-based relation extraction: an overview. *Neural Computing and Applications 34*, 6 (2022), 4781–4801.

[32] Wang, L., Cao, Z., de Melo, G., and Liu, Z. Relation classification via multi-level attention CNNs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Berlin, Germany, Aug. 2016), Association for Computational Linguistics, pp. 1298–1307.

[33] Xiao, C., Yao, Y., Xie, R., Han, X., Liu, Z., Sun, M., Lin, F., and Lin, L. Denoising relation extraction from document-level distant supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Online, Nov. 2020), Association for Computational Linguistics, pp. 3683–3688.

[34] Xu, K., Feng, Y., Huang, S., and Zhao, D. Semantic relation classification via convolutional neural networks with simple negative sampling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (Lisbon, Portugal, Sept. 2015), Association for Computational Linguistics, pp. 536–540.

[35] Y Yang, K.Z. Zhang, P. K., Uy, M. C. S., and Huang, A. Finbert: A pretrained language model for financial communications. *ArXiv abs/2006.08097* (2020).

[36] Yao, Y., Ye, D., Li, P., Han, X., Lin, Y., Liu, Z., Liu, Z., Huang, L., Zhou, J., and Sun, M. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy, July 2019), Association for Computational Linguistics, pp. 764–777.

[37] Yu, H., Cao, Y., Cheng, G., Xie, P., Yang, Y., and Yu, P. Relation extraction with bert-based pre-trained model. In *2020 International Wireless Communications and Mobile Computing (IWCMC)* (June 2020), pp. 1382–1387.

[38] Zeng, D., Liu, K., Chen, Y., and Zhao, J. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (Lisbon, Portugal, Sept. 2015), Association for Computational Linguistics, pp. 1753–1762.

[39] Zeng, D., Liu, K., Lai, S., Zhou, G., and Zhao, J. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (Dublin, Ireland, Aug. 2014), Dublin City University and Association for Computational Linguistics, pp. 2335–2344.

[40] Zhang, Y., Zhong, V., Chen, D., Angeli, G., and Manning, C. D. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (Copenhagen, Denmark, Sept. 2017), Association for Computational Linguistics, pp. 35–45.

[41] Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., and Xu, B. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Berlin, Germany, Aug. 2016), Association for Computational Linguistics, pp. 207–212.