Delegates Relation Extraction : NLP based approach

Computation and Linguistic Analysis

Rajesh Sharma
Computer Science
Indiana University

Abstract— **Merger and Acquisition (M & A) activity is exponentially increasing with time and has increased from $1400 billion to $2600 billion in the USA in the past 15 years. Information of M&A deals between two organizations is readily available on the internet. However, to get further insights into the data, one must go further into the press release or blogs to get information about the delegations involved. In this work, we consider extracting the delegate's information like name and job position of the people involved in the process using various NLP techniques.**

Keywords— Relation Extraction, Named Entity Recognition, Mergers and Acquisition

I. INTRODUCTION

The terms "merger" and "acquisition" are frequently and synonymously used in the business world. However, both actually mean different things. While an acquisition occurs when one corporate organization totally or partially takes over the target company, a merger refers to the joining of two businesses or organizations into one. Even while Merger and Acquisition (M&A) requires a lot of work and significant input from all departments of the organization, executives like the CEO, CFO, and vice president plays a crucial role in formulating strategies, making final deal decisions, and even anticipating even post-deal happenings.

Generally, to gain media and public attention, companies put press releases with acquisition announcements with all the essential details. Most of the deals information are also kept in publicly accessible U.S. Securities and Exchange Commission (SEC)[1] database.

The USA signed most of the M&A deals in 2021, at f $2549 billion followed by China and UK with the value of $528 billion and $347 billion respectively[2]. Information related to these deals are readily available in the internet in the form of press releases from the companies or blog posts on famous websites (BusinessWire, PR Newswire etc.) ,but getting names and job title of all delegates (CEO, CFO, VP etc.) from both sides of the deal, we have to read one or multiple articles manually. Collecting and storing all the information manually is a very tedious and time-consuming process.

So in this work, we have automated this task of identifying the executives mentioned in company press releases information using web scraping and NLP techniques like relationship extraction and Named Entity Recognition.

II. RELATED WORK

In NLP, Relation Extraction (RE) has received much of attention, and using GNN for RE tasks is becoming increasingly common.

KGPool[3]: In this research, relation extraction (RE) from a single sentence is used to map a single sentence and the two given entities to a canonical fact in a knowledge graph. Also in this study, sentence structure information is dynamically selected and transformed into a representation, supplementing the latent representation of sentential context that was learnt using a neural model.

LUKE[4]: In this paper, new contextualized word representations based on bidirectional transformers are introduced. The aim at hand in this research is to predict randomly masked words and entities from a large entity-annotated corpus acquired from Wikipedia.
Additionally, this work extends the transformer's self-attention mechanism, which accounts for the type of tokens (words or items) while determining attention scores.

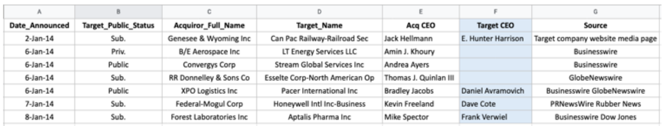
Relation of the Relations[5]: It takes many resources to predict the relationship between two items at once, but this work suggests a new paradigm for connection extraction that considers all relationship's predictions of all relationships in the same context.
Furthermore, they used a data-driven strategy that uses GNN to learn new rules rather than relying on hand-written ones.

Enriching Pre-trained Language Model with Entity Information for Relation Classification [6]:
Relationship categorization has already seen a lot of development, such as the pre-trained BERT model.
The state-of-the-art BERT model, which was pre-trained, is used in this study for relation classification tasks whereas target entity tasks are included for RE tasks.
This work aims to identify the target entities and transfer the data through an embedded model that has already been trained.

III. DATASET

Data of around 1800 deals from 2014 to 2020 was manually collected(outsourcing). The 2014-year delegate data has been manually labeled and saved [Figure 1]. Our dataset has three labels: Person Labels (or Person Name), Job Label(or Job title) and Organization label(or Organization).

For example, ‘*Genesee & Wyoming I*nc’(or acquirer) acquired ‘*Can Pac Railway-Railroad Sec’* (or target) where ‘*Jack Hellma*nn’(Person Label) was the *‘CEO’*(Job Label) of the acquirer and ‘*E. Hunter Harrison’* was the ‘*CEO’* of target company.



A	B	C	D	E	F	G
Date_Announced	Target_Public_Status	Acquiror_Full_Name	Target_Name	Acq CEO	Target CEO	Source
2-Jan-14	Sub.	Genesee & Wyoming Inc	Can Pac Railway-Railroad Sec	Jack Hellmann	E. Hunter Harrison	Target company website media page
6-Jan-14	Priv.	B/E Aerospace Inc	LT Energy Services LLC	Amin J. Khoury		Businesswire
6-Jan-14	Public	Convergys Corp	Stream Global Services Inc	Andrea Ayers		Businesswire
6-Jan-14	Sub.	RR Donnelley & Sons Co	Esselte Corp-North American Op	Thomas J. Quinlan III		GlobeNewswire
6-Jan-14	Public	XPO Logistics Inc	Pacer International Inc	Bradley Jacobs	Daniel Avramovich	Businesswire GlobeNewswire
7-Jan-14	Sub.	Federal-Mogul Corp	Honeywell Intl Inc-Business	Kevin Freeland	Dave Cote	PRNewsWire Rubber News
8-Jan-14	Sub.	Forest Laboratories Inc	Aptalis Pharma Inc	Mike Spector	Frank Verwiel	Businesswire Dow Jones

Figure 1: Snapshot from initial data

And for the rest of the years, from 2015 to 2020, we only have the Target company name, the Acquirer company name and the date when deal was signed.

IV. METHODOLOGY

To get to the outcome, we need to perform a couple of steps, starting with selecting the websites from which to fetch the data, then implementing Named Entity Recognition and relation extraction from the text. The project's flow is shown by the steps below (which apply to every deal):

- Collect all website links
- Scrutinize the major websites
- Get textual data from all scrutinized websites
- Named Entity Recognition
- Relation Extraction

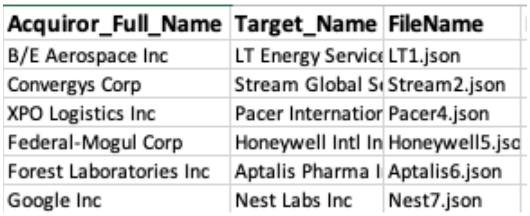
Only 2014 data (177 trades) is used to check and confirm the outcomes before constructing the flow and modules. The remaining data were then subjected to that code's execution, but because some labels were absent, there was no subsequent validation process. Additionally, the primary objective right now is to identify the CEO, their position (Chief Executive Officer), and the associated company.

A. Dataset Collection and Cleaning

More data about all the deals needs to be gathered. To get all the additional data required, web scraping approach is used. The steps below are carried out to gather information about each deal:

- Website names scraping- Scraping website names is a two stage process. In the first stage, the names of the target and acquired companies are passed to Google as search parameters, and all of the links are obtained from the Google search result page. In the second stage websites linked are scrutinized to keep only company press releases, data from sec.gov website or information from well known websites (such as Business Wire, PRNewswire, GlobeNewswire, NYTimes, biz journal, mergr and yahoo finance) that have information about the deal. Last step will be to remove all websites starting with address ‘https://google. ‘ as these all tied to useless(for this scenario) services related to google services like maps, policies, shopping etc. which are of no use. All this data is updated on the original data file.
- Gathering websites data – After getting the websites, collect all the information provided within, for time being collect only the text contained within each website’s paragraph tags. For each deal we have numerous websites, thus we should store all text that was gathered from each of them in a JSON file using the dictionary data type, where the key is the website link and the value was the information collected from it.

After these steps we have JSON files with a count equal to number of deals in the dataset. The names of the Acquired and Target companies are listed individually in a Metadata file (new file created after data gathering step), along with the name of the JSON file that contains details on the deal. Following [Figure 2] is the snapshot from metadata file.



Acquiror_Full_Name	Target_Name	FileName
B/E Aerospace Inc	LT Energy Servic	LT1.json
Convergys Corp	Stream Global S	Stream2.json
XPO Logistics Inc	Pacer Internatior	Pacer4.json
Federal-Mogul Corp	Honeywell Intl In	Honeywell5.jso
Forest Laboratories Inc	Aptalis Pharma I	Aptalis6.json
Google Inc	Nest Labs Inc	Nest7.json

Figure 2: Snapshot of metadata file

B. Data Pre-Processing

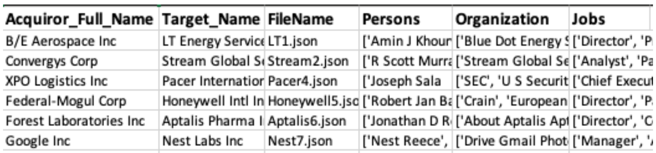
The textual data collected above is raw information taken directly from website that contains extraneous information like stop words, digits, multiple white spaces, punctuations etc. that needs treatment. We also need to handle Upper- and Lower-case words otherwise algorithm will consider the same word differently depending on the case it is written in. However, case sensitivity is treated only for vector space embedding and not for Named Entity Recognition (or NER) as it requires some terms that are case sensitive like Person names. A person name always starts with capital character. The textual pre-processing also shortens the text which accelerates processing.

C. Names Recognition

After gathering and finishing the dataset next step is to locate all Person Names, all job titles and organization names. For each deal all website data (values of JSON file) is combined into one textual data, and after pre-processing (stop word, punctuation, multispacer, number removal) the data following fetching is done:

- Person Names: The Named Entity Recognition(NER) approach of Stanford NLP library[7] used to obtain all person's name. Combined textual data is tokenized, and NER algorithm is implemented. After NER, person name data is stores in a list before being stored.
- Organization Names: The initial dataset already gives organization names. However, NER is still used to tag organizations for further future learning.
- Job Titles: We started by web-crawling all of the commonly used job titles names from the internet[2]. The key word matching technique is then used to identify every job title in our textual data, and all matching job titles are then stores in a list format.

The output of the aforementioned steps is saved in metadata csv file[Figure 3] with some additional columns namely Persons, Jobs and Organization. Later, it is employed for analysis and knowledge acquisition.



Acquiror_Full_Name	Target_Name	FileName	Persons	Organization	Jobs
B/E Aerospace Inc	LT Energy Servic	LT1.json	['Amin J Khour	['Blue Dot Energy S	['Director', 'P
Convergys Corp	Stream Global S	Stream2.json	['R Scott Murra	['Stream Global Se	['Analyst', 'Pa
XPO Logistics Inc	Pacer Internatior	Pacer4.json	['Joseph Sala	['SEC', 'U S Securit	['Chief Execut
Federal-Mogul Corp	Honeywell Intl In	Honeywell5.jsc	['Robert Jan Ba	['Crain', 'European	['Director', 'P
Forest Laboratories Inc	Aptalis Pharma I	Aptalis6.json	['Jonathan D R	['About Aptalis Apt	['Director', 'C
Google Inc	Nest Labs Inc	Nest7.json	['Nest Reece',	['Drive Gmail Phot	['Manager', '

Figure 3: Snapshot after performing NER

D. Word Embedding

To learn and examine relationship between person's name, their jobhs and their organization, Word2Vec, a vector spacve embedding technique is utilized. Before implementing the algorithm, pre-processing is done on the combined textual data. After pre-processing list of all sentences is created and after tokenization all words of each sentence, It is passed as a training data to the Word2Vec model. Parameters used to train the model are: min_count=2, window=10, sample= 6e-5, alpha=0.03, sg=1, min_alpha=0.0007.

V. Results and Discussion

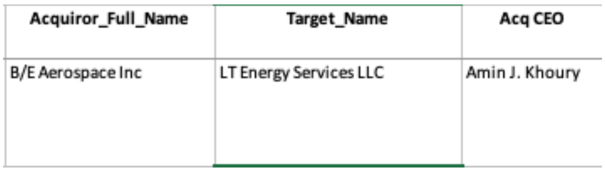
A. NER

Before determinig how names(Person names and Job title) relate to one another, we must compare the results of NER, to see if we successfully gathered the necessary names. Using the binary classification method we can accomplish this task. Like checking whether a person's label is present in the people list that has been fetched,

Example: for the deal between B/E Aerospace Inc and LT Energy Services LLC following are the results of NER process (only Persons and Jobs results)

Persons	Jobs	Acq CEO
['Amin J Khoury', 'Greg Powell', 'Jake Williams', 'Pharr Lee', 'Amin Khoury', 'Houlihan Lokey', 'Kevyn DeMartino']	['Director', 'Principal', 'Managing Director', 'Partner', 'Founder', 'Chairman', 'President', 'Chief Executive Officer', 'Executive', 'CEO', 'Vice President']	Amin J Khoury

And following[Figure 4] is the snapshot from the intial dataset (2014 labeled) with Companies name and Acquired Company CEO name:



Acquiror_Full_Name	Target_Name	Acq CEO
B/E Aerospace Inc	LT Energy Services LLC	Amin J. Khoury

Figure 4: Snapshot from initial data

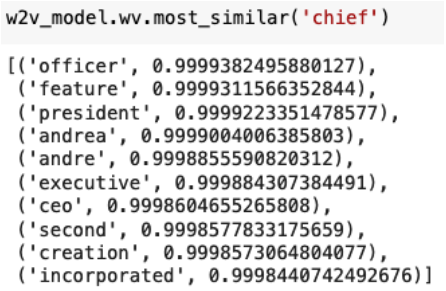
Here can see Acq CEO(actual labele) is in the Persons(NER fetched) list.

After checking the Acquired CEO column, Target CEO column goes through the same process. 96% of CEO names were retrieved by our algorithm, meaning that 96% of person labels were in the list of people we had retrieved for each deal seperately. For the job titles, all job tile list fetched using NER have CEO names.

B. Word Embedding

The main motive to implement Vector Space Embedding was to determine whether a person's name and job title have a striong cosine similarity on the vector space or not. However, it did produce some desired reuslts like the word was closely embedded with CEO name.

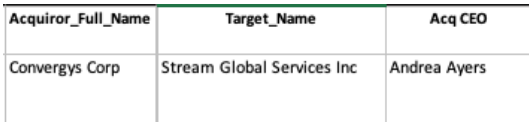
Following[Figire 5] is an example where collected textual information of deal between Convergys Corp acquired and Stream Global Services Inc is passed for training to Wrod2Vec model



```
w2v_model.wv.most_similar('chief')

[('officer', 0.9999382495880127),
 ('feature', 0.9999311566352844),
 ('president', 0.9999223351478577),
 ('andrea', 0.9999004006385803),
 ('andre', 0.9998855590820312),
 ('executive', 0.999884307384491),
 ('ceo', 0.9998604655265808),
 ('second', 0.9998577833175659),
 ('creation', 0.9998573064804077),
 ('incorporated', 0.9998440742492676)]
```

Figure 5: Prediction result of Word2Vec model



Acquiror_Full_Name	Target_Name	Acq CEO
Convergys Corp	Stream Global Services Inc	Andrea Ayers

Figure 6: Snapshot from initial data

The information in the above figure [Figure 6] indicates that Andrea Ayers was the CEO of Convergys Corp. when the purchase was completed. Moreover, our Word2Vec algorithm produced the same outcome that is, Peroson name and Job title are closely embeded. As we can see in Figure 5, Andera(or Person name) is still the fourth-nearest word to the chief(or job name), so a way to extract Person Name from the result is still required.

VI. Future Scope

A. *Word Embedding*

All the project was build and executed on a local system. In futre with additional hardware, we can try to Hyper Parameter tuning the Word2Vec model or use another with a bigger pre-trained model such as BERT, GloVe or Transformers.

Finding a method to compare only a person's name and job title while obtaining the most similar term after training the data to the job title might be considered.

B. *Relation Extraction*

Finding a relationship between a person's name and their job title can be done using semi-supervised or unsupervised relation extraction techniques, which can then be applied to any job role not only CEO. I am still working on this project and exploring ways to extract relations, like we might want to ensemble Dependecy Tree with word embedding technique.

C. *Dataset*

More data from earlier is available now, with all M&A deals with manually labeled delegates information. Using this dataset for training models was notr feasible due to time and resource limitations. Using complex algorithms, this dataset can be used as a knowledge foundation to produce more sophisticated rules.

VII. References

[1] https://www.sec.gov/about

[2] https://www.statista.com/topics/1146/mergers-and-acquisitions/

[3] Abhishek Nadgeri, Anson Bastos, Kuldeep Singh, Isaiah Onando Mulang', Johannes Hoffart, Saeedeh Shekarpour, and Vijay Saraswat. 2021. KGPool: Dynamic Knowledge Graph Context Selection for Relation Extraction. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 535–548, Online. Association for Computational Linguistics.

[4] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6442–6454, Online. Association for Computational Linguistics.https://arxiv.org/abs/2006.03719

[5] Jin, Zhijing & Yang, Yongyi & Qiu, Xipeng & Zhang, Zheng. (2020). Relation of the Relations: A New Paradigm of the Relation Extraction Problem. https://arxiv.org/abs/2006.03719

[6] Wu, S., & He, Y. (2019). Enriching Pre-trained Language Model with Entity Information for Relation Classification. Proceedings of the 28th ACM International Conference on Information and Knowledge Management. https://arxiv.org/abs/1905.08284

[7] https://www.joblist.com/b/all-jobs