Thomson Reuters' Submission to the FEIII 2017 Challenge Non-scored Tasks

Elizabeth Roman
Thomson Reuters Labs
Boston, MA 02210
elizabeth.roman@tr.com

Brian Ulicny
Thomson Reuters Labs
Boston, MA 02210
brian.ulicny@tr.com

Yilun Du
MIT
Cambridge, MA 02138
yilundu@mit.edu

Srijith Poduval
MIT
Cambridge, MA 02138
spoduval@mit.edu

Allan Ko
MIT
Cambridge, MA 02138
allanko@mit.edu

ABSTRACT

In this paper we describe a machine learning approach to predict roles of extracted SEC triples for the non-scored task of the 2017 FEIII Challenge.

In addition, we describe a graph and data analysis derived from SEC triples.

CCS CONCEPTS

•Information systems → Data analytics;

KEYWORDS

Thomson Reuters, SEC, Information Extraction, Machine Learning, Graph Analytics, Cytoscape

1 INTRODUCTION

The main task of the FEIII 2017 challenge [1] was to predict the relevance of triples extracted from 10-K and 10-Q SEC filings. The provided datasets containing the following information for a triple: i) filer entity, ii) mentioned company, iii) role that describes the relationship between filer entity and mentioned entity and iv) snippet of three sentences in the SEC filing from which the triple was mined. Table 1 shows the 10 role types and their descriptions. The non-scored tasks then asked us to consider additional tasks built on top of the initial task: 1) apply NLP and Machine Learning to further embellish the relationship found and 2) create a graph from the relationships and derive measures such as centrality of filer entity or the exposure of a filing entity to some mentioned entity.

Table 1: Roles

Role	Description
Affiliate	Company or organization related through common ownership, common control of management or owners, or through some other control mechanism, such as long-term lease
Agent	A person that is designated by another person (the principal) to act on behalf of the principal in specified activities
Counterparty	Given a person or party of interest (in a trade), the party faces a counterparty; there can be more than one counterparty
Guarantor	A legal arrangement involving a promise by a person (guarantor) to perform the obligations of a second person (or many persons), in the event that the latter person fails to meet their obligations
Insurer	A person who, through a contractual agreement, undertakes to compensate specified losses, liability, or damages incurred by the person of interest
Issuer	This term refers to an issuer of securities which are (1) registered under Section 12 of the Securities Exchange Act of 1934, or (2) required to file reports under Section 15(d) of that Act, or (3) has filed a registration statement with the SEC
Seller	Exchanges a good or service in exchange for a payment
Servicer	Typically a person that collects payments from one party and makes payments to another party
Trustee	Person who is given legal title to, and management authority over, the property placed in a Trust
Underwriter	A person that assumed the risk of purchasing securities from the issuing entity and reselling them to the public, either directly or through dealers

2 IMPLEMENTATION

2.1 Role Prediction

As a first step towards addressing the task of producing a more embellished relationship, we developed a role predictor that could discern features in the text that relate to the role other than explicit mention (keywords). Our approach to predicting roles of companies in SEC triples relied upon training a classifier on the automatically extracted (and therefore noisy) triples. We trained an off-the-shelf gradient boosting regression model, which achieved an accuracy (fraction of correctly predicted role labels) of 89%. A confusion matrix of this model can be seen below (Figure 1). In the following sections we describe the features of the role predictor.

2.1.1 Word Vectors. For each triple context text, we converted each word of the context text to its corresponding word vector (300 dimensions) in the pre-trained Google News word embedding. We then took the weighted average (by TF-IDF) of the word vectors to get an overall word vector for the overall context text.

2.1.2 Thomson Reuters Business Classification. The Thomson Reuters Organization Authority (OA) database was used to fetch relevant information about entities. The OA database contains wide-ranging types of information such as geographical location, industry, parental entities and domicile country. After mapping each Mentioned and Filing Entity to its corresponding Thomson Reuters Permanent Identifier (PermID) using Thomson Reuters OpenPermID API, we then mapped each entity in the triple pair to

DSMM'17, Chicago, IL, USA

978-1-4503-5031-0/17/05…$15.00
DOI: http://dx.doi.org/10.1145/3077240.3077244

its corresponding industry using the Thomson Reuters Business Classification (TRBC). PermID [2] is a unique and permanent identifier for that allows integration of information around entities such as people, organizations, and quotes. TRBC is the most comprehensive, detailed, and up-to-date sector and industry classification available with a granularity of up to five levels.

The PermID mapping yielded matches for 72% of the mentioned entities in the training data, with the matcher rating 69% of the matches as "Excellent", 5% as "Good", and 26% as "Possible". For the testing set, 71% of the mentioned entities were matched to PermID, with 73% of matches deemed "Excellent", 7% deemed "Good", and 20% deemed "Possible".

2.1.3 Data Fusion Features. Thomson Reuters Data Fusion [3] was used to explore the graphical relationship between filer entity and mentioned entity in each triple. Data Fusion uses news articles and other sources of information to form a connection graph between different entities. We queried the number of different length (2 to 4 edges) paths from the filer to the mentioned entity provided the paths went through at least one person mention. We hypothesized that strong connections between filer and mentioned company would suggest i) an affiliate type of relationship, and ii) help to filter out erroneous roles.

2.1.4 Calais Features. Thomson Reuters Open Calais [4] was used to obtain rich metadata from the three sentence contexts extracted from the SEC filings. Open Calais uses machine learning and natural language processing to analyze and tag text. The API tags entities, relationships, and values such as companies, people, industries, events, and currencies. For each entity, Open Calais outputs a confidence level as well. Open Calais also does "about-ness" tagging: detecting certain phrases and outputting a numerical score determining that phrase's relevance in the context of the entire document. In addition to the above, Open Calais produces social tags, which are topics based on Wikipedia folksonomy, and associates an importance value with that social tag. For example, a news article about the Apple Watch would have social tags like IOS, Smartwatches, Wearable Computers, Human-computer interaction, Apple Inc., and others attached to it. Another set of topic tags outputted by Open Calais is based on a list of topics defined by the Thomson Reuters Coding Schema (TRCS) and/or by the International Press Telecommunications Council (IPTC) news taxonomy. Finally, Open Calais can also detect possible industries discussed in the text as Thomson Reuters Business Classification (TRBC) taxonomy codes.

For each triple, the three sentence context collected by the IBM SystemT tools was inputted into Open Calais. The JSON returned from the Open Calais API was parsed to extract all relevant metadata. Features from Calais metadata were: the number of companies mentions, number of unique companies, the average confidence score of companies detected, the average TRCS topic score, the average industry tag relevance score, the average social tag importance score, the number of tags, the number of each tag type, the sum of the scores (relevance/importance/confidence) for each tag type, and number of times currency was mentioned in the three sentences.

2.2 Graph of Company Connections

For the second unscored task, we decided to look at the central vertices of a graph created from the SEC triples. In the initial graph each financial entity and mentioned entity triple was represented as an edge. Every mentioned entity was mapped as a directed edge to its filing entity. We created a visualization of the graph using Cytoscape [5], shown in Figure 2. The graph was connected across all input data. To determine the centrality of each company, Eigenvalue Centrality, Vertex Centrality, Betweenness Centrality, and Closeness Centrality were used. In general, Bank of America, Federal Home Loan Mortgage, and Fannie Mae were the most central companies in the network (see Figure 2).

3 CONCLUSION

We presented our approach to predicting roles using NLP and Machine Learning and leveraging various technologies and datasets from Thomson Reuters. The features and method could be adapted to determine the directionality of the link as well as determine other properties of the relationship (embellishment). We have provided analytics showing the graph created from mining SEC filings is a starting point for gleaning useful insights.

4 ACKNOWLEDGMENTS

We would like to thank Brian Romer for helping with graph visualization in Cytoscape.

REFERENCES

[1] Louiqa Raschid, Doug Burdick, Mark Flood, John Grant, Joe Langsam, Ian Soboroff, and Elena Zotkina. Financial entity identification and information integration (FEIII) challenge 2017: The report of the organizing committee. In *Proceedings of the Workshop on Data Science for Macro-Modeling (DSMM@SIGMOD)*, 2017.
[2] Open permid. http://developers.thomsonreuters.com/open-permid.
[3] Data fusion community edition. http://developers.thomsonreuters.com/data-fusion.
[4] Bring structure to unstructured content. http://www.opencalais.com.
[5] Cytoscape: An open source platform for complex network analysis and visualization. http://www.cytoscape.org.

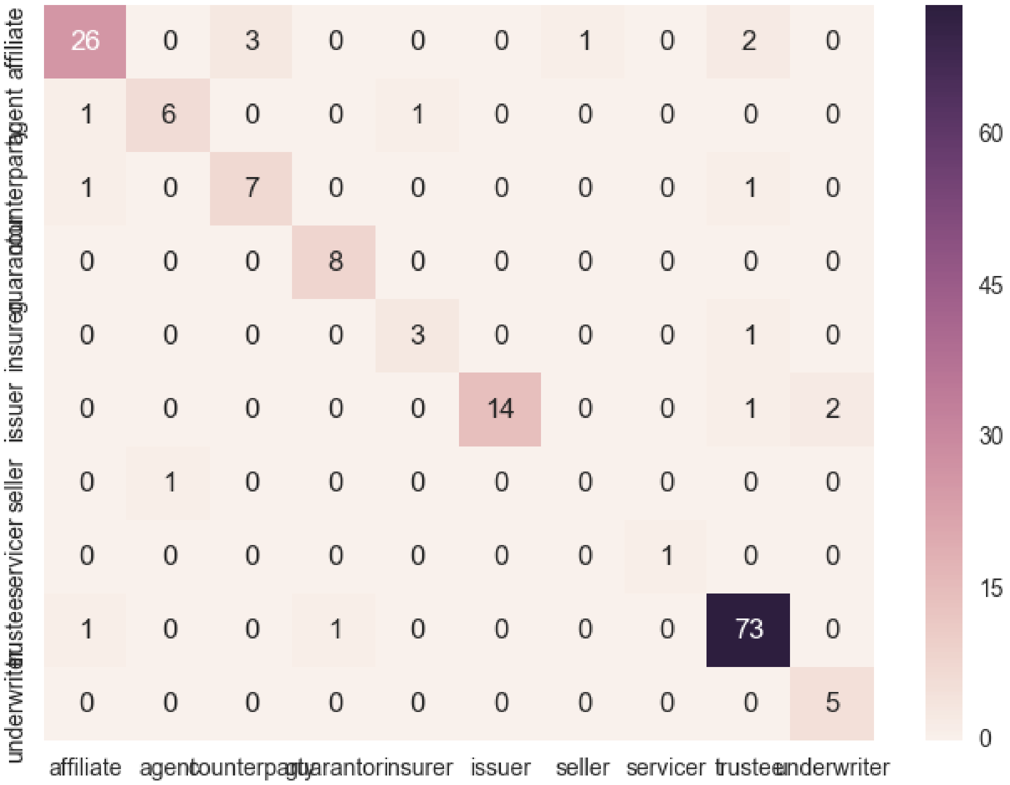

Figure 1: Confusion Matrix of Role Predictor

Table 2: Centrality Rankings for SEC Triples

	1st	2nd	3rd	4th
Eigenvalue	Federal Home Loan Mortgage	Bank of America	Equity Securities	Fannie-Mae
Vertex	Federal Home Loan Mortgage	Bank of America	Fannie-Mae	Goldman Sachs
Betweeness	Bank of America	Morgan Stanley	Citigroup	America Express
Closeness	Federal Home Loan Mortgage	Fannie-Mae	Bank of America	Goldman Sachs

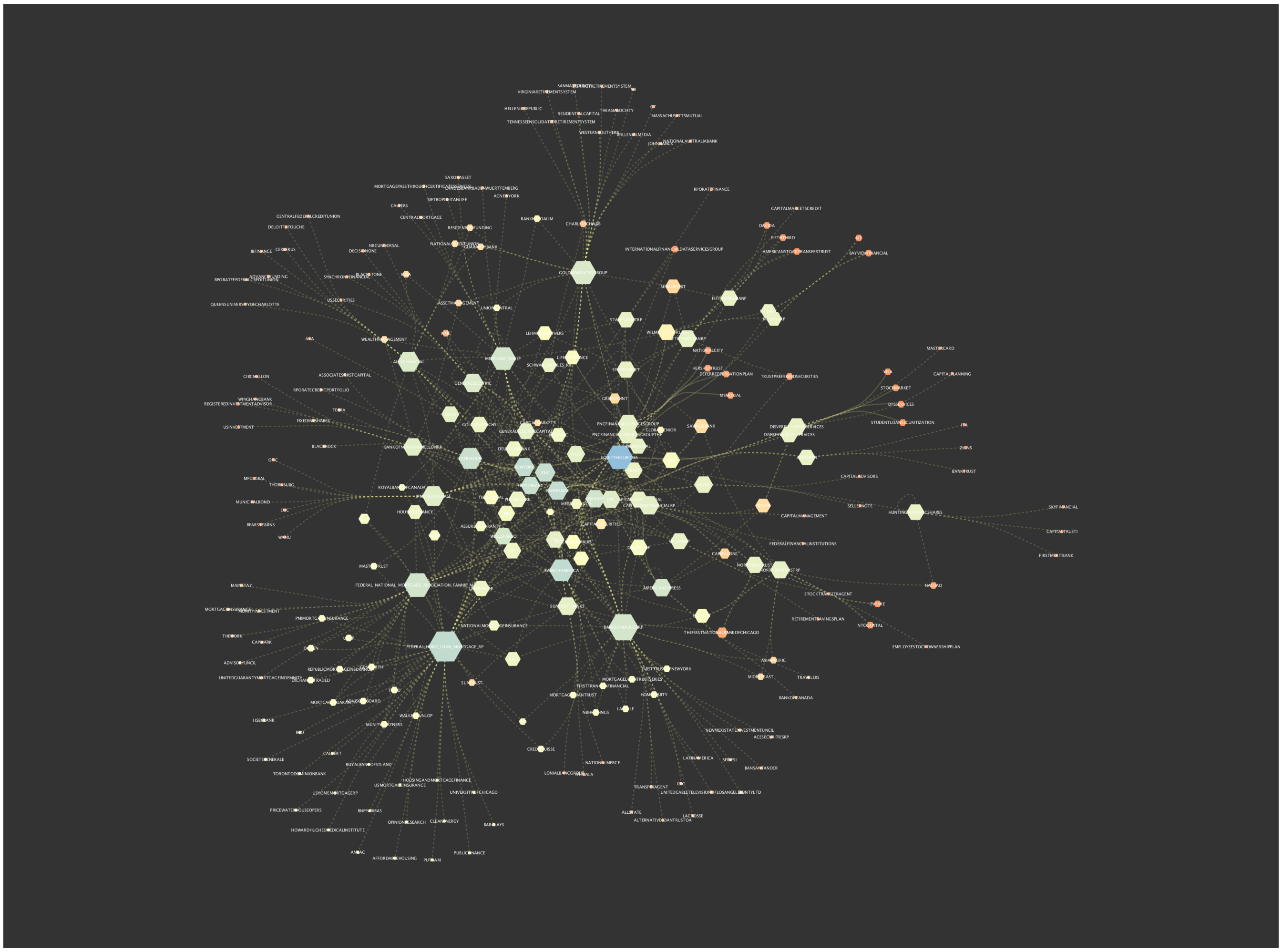

Figure 2: Graph of Entity Connections from SEC Triples