Thomson Reuters' Solution for Triple Ranking in the FEIII 2017 Challenge

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

Brian Ulicny
Thomson Reuters
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 our approach to the triple ranking task of the FEIII 2017 challenge. Our method leveraged different machine learning classifiers in an ensemble as well as Thomson Reuters knowledge bases and information services to bring in external world knowledge of mentioned entities and extract information from the contextual sentences. Internal evaluation of our method was done by computing the Normalized Discounted Cumulative Gain (NDCG) as tracked by the challenge and classification accuracy. The official FEIII Challenge evaluation showed our system performed highly in single ranking of all triples, placing in 2nd or 3rd place out of 17 participants for 4 of 6 scoring variants; the system also performed above average in per role ranking for 4 of 6 average role NDCG scoring variants.

CCS CONCEPTS

•Computing methodologies → Ranking; Supervised learning by classification;

KEYWORDS

Thomson Reuters, SEC, Information Extraction, Machine Learning

1 INTRODUCTION

The FEIII 2017 challenge [1] aims to identify and understand relationships among financial entities found in unstructured SEC filings. A dataset of triples containing the following information was provided:

- a financial entity reference as a text string containing the company name of the Mentioned company
- Role keyword: string that describes the relationship between Filer company and the Mentioned company. The ten roles extracted were: Affliate, Agent, Counterparty, Guarantor, Insurer, Issuer, Seller, Servicer, Trustee, and Underwriter.
- Context of three sentences which gives evidence of the relationship

The scored task was to rank the triples within each role by relevancy, defined as ranking triples in the order by which they best support the role assignment and provide financially relevant knowledge. In evaluation, a test set contains triples that have been rated by domain experts with the following labels: Highly Relevant, Relevant, Neutral, Irrelevant. The absolute ranking produced by our system is then compared to a perfect ranking based on the relevancy ratings of domain experts. The metric applied to a ranking is the Discounted Cumulative Gain (DCG) :

$$\mathrm{DCG}_p = \mathrm{rel}_1 + \sum_{i=2}^{n} \frac{\mathrm{rel}_i}{\log_2(i)}$$

A gain (rel_i in the above formula) of 2 points is given for Highly Relevant triples, 1 point for Relevant triples, and 0 points otherwise. The DCG of our system's ranking is normalized by the DCG of the perfect ranking that is based on the expert labels to give the NDCG score between 0 and 1. NDCG scores are computed among all triples ("single NDCG") and as the average of per role NDCGs.

2 DATASETS

The challenge provided datasets of triples extracted in an automated fashion from 10-K and 10-Q SEC filings from the time period 2011-2016. A labeled dataset of 975 triples containing relevance ratings by domain experts was provided; the labels used by the annotators were Highly Relevant, Relevant, Neutral, and Irrelevant. Additionally, an unlabeled working dataset of 9597 triples was also given.

3 METHODOLOGY

We experimented with different supervised machine learning approaches to rank the triples. Due to the small size of the our training set (814 triples), we combined all triples into a single training set rather than create models for each role. We combined the different ratings given to a triple by annotators by averaging them. In the feature engineering phase, we relied upon incorporating external world knowledge of companies from Thomson Reuters to supplement the features based purely on the contextual sentences. In the

DSMM'17, Chicago, IL, USA

978-1-4503-5031-0/17/05…$15.00
DOI: http://dx.doi.org/10.1145/3077240.3077253

classification phase, we combined off-the-shelf classifiers into an ensemble to obtain a high performance. The following sections describe the Thomson Reuters datasets and services we used and our final model.

3.1 Thomson Reuters Data and Features

Powered by the award-winning Reuters News Agency, Thomson Reuters delivers critical information, analytics, and solutions to clients in the financial and risk, legal, tax and accounting, and media markets, enabling them to make the decisions that matter the most. Here we describe the knowledge bases and information services from Thomson Reuters that our solution builds upon. All of the Thomson Reuters resources described here are currently available in a community edition via APIs.

3.1.1 Thomson Reuters Organization Authority and Thomson Reuters Business Classification. In order to rank context sentences based on role assignment accuracy and financial relevance, it would help to know what industries the entities are known to participate in. We extracted this information as Thomson Reuters Business Classification (TRBC) codes from Thomson Reuters Organization Database (OA). The OA database contains information about companies such as headquarter addresses and parent company. Company records can be looked up with SQL queries through different identifiers, such as CIK and Thomson Reuters Permanent Identifier (PermID). After mapping all the filer and mentioned entities to PermID via a programmer-friendly RESTful API [2], we used the primary TRBC industries of both entities as features.

3.1.2 Thomson Reuters Data Fusion. Another source of external knowledge came from Thomson Reuters Data Fusion [3]. Data Fusion is a powerful linked data platform which enables access to the Thomson Reuters knowledge graph through an API and graphical web user interface. The knowledge graph connects entities, such as people and companies, through links mined from sources such as news and supply chain data. Users can also upload their own data to be "stitched" into the knowledge graph.

Using the Data Fusion API, we extracted counts of paths from filer entity to mentioned entity of different lengths as features. The paths were required to contain at least one vertex of type person.

3.1.3 Thomson Reuters Open Calais. We utilized Thomson Reuters Open Calais [4] to tag the contextual sentences of each triple and obtain features from the rich metadata. Open Calais can extract various pieces of information from an unstructured document such as people, organizations, locations, industries, relations (e.g. position), events (mergers and acquisitions), values (e.g. currency type), and topics. The topics outputted by Calais come from Wikipedia folksonomy (social tags) as well as Thomson Reuters Coding Schema (RCS codes) and International Press Telecommunications Council (IPTC) taxonomy. The following information gleaned from Open Calais was used as input into the classifiers we developed:

- number of company detections
- average company detection confidence (a number between 0 and 1)
- the average importance score of social tags detected (a number between 0 and 1)
- the average relevance score of industries detected (a number between 0 and 1)
- the number of unique companies
- number of each tag type
- for each tag type, the sum of the confidence scores
- for each tag type, the average of the confidences scores
- the number of currency detections, added to number of currency detections from regular expressions
- the average score of TRCS topics detected

3.2 Other Features

3.2.1 Word Embeddings. We incorporated word embedding features from the contextual sentences into the final model as well. Initially training our own word vectors on the unlabeled triple dataset, we found the performance of the model using pre-trained Google News word vectors to be much better. This is likely due to the small size of the unlabeled triple dataset. A vector for each set of contextual sentences was created by summing the tf-idf weighted Google word vector for each word in the contextual sentences.

3.2.2 Role. The role asserted by the triple was also used and treated as a categorical feature.

3.2.3 Text Features. We experimented with a number of other features, which ultimately were not incorporated in the final model. These features included: contextual sentences character length, contextual sentences word length, and average word size in contextual sentences.

3.3 Models

We randomly split the annotated triples into a training set (814 triples) and a testing set (161 triples). We trained an ensemble of classifiers on the training set and combined by averaging the scores produced by the ensemble. Four-fold cross-validation was used during training. The models in the final ensemble were: i) gradient boosting regression trained with TRBC, Data Fusion, word embedding, and role features and hyperparametrized, ii) a Support Vector Regression model trained with Open Calais features and iii) a bagging model trained with word embedding and role features.

In addition to an ensemble of classifiers, we also experimented with simpler baseline models and state-of-the-art libraries. The first viable model was created by training a random forest classifier with Google word embedding features. We also attempted to use word embeddings from fastText [5], a library for text representation and classification with a focus on scalability and efficiency, however our results were not as competitive. Finally, we trained a Support Vector machine with term frequency-inverse document frequency features.

An initial analysis showed that a variety of features from TRBC, word embeddings, Open Calais, and Data Fusion were the most informative, with no one feature type being the most informative overall. Future work in this area would yield more intuition into why features together may be discriminating.

4 EVALUATION AND RESULTS

4.1 Internal

Prior to the challenge evaluation, we internally evaluated our performance based on the NDCG and classification accuracy on a held-out set of annotated triples (80/20 train/test split; 161 test triples) selected at random, so that the test set was unbalanced with respect to role. In evaluating accuracy, we round scores/ratings and compute the fraction of correct label predictions. Figure 1 shows a Confusion Matrix for the classifier. The NDCG was computed per role and then averaged across the roles. Our final results using an ensemble of classifiers, were .978 for NDCG and 82.5% for accuracy.

Model	Accuracy	NDCG
fastText	25%	.88
SVM/Tf-Idf	75%	.93
Random Forest/Word Embeddings	81%	.96
Ensemble of Classifiers	82.5%	.98

4.2 Challenge

The official challenge evaluation contained 900 triples. The NDCG was computed among all triples and also by role and then averaged. A major difference between labels in the training set and evaluation set was the annotation of role validation (correctness). In the single NDCG evaluation, our performance was very good, placing in second or third place for 4 of the 6 scoring variants and above average for one other variant. In evaluations computed per role, we performed above average for 4 of the 6 scoring variants. In both the single and average role NDCG evaluations, our worst performances were on scoring variants with the greatest emphasis on detecting the triples with roles validated by the contextual sentences; because our system was trained on noisy data that did not contain this information, this is not altogether surprising.

5 CONCLUSION

We have presented our supervised machine learning method to rank triples extracted from SEC filings by relevance. The system achieves a high performance using standard classifiers and relying on external world knowledge about entities referenced in the triple. Potential future work on our system includes re-training the system with the extended role validation ground truth and refining the features to better capture role validation.

6 ACKNOWLEDGMENTS

We would like to thank Amit Shavit and Omar Bari for reviewing our approach early on and giving us helpful suggestions, and Ian Soboroff for answering our questions.

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] Fasttext. https://research.fb.com/projects/fasttext.

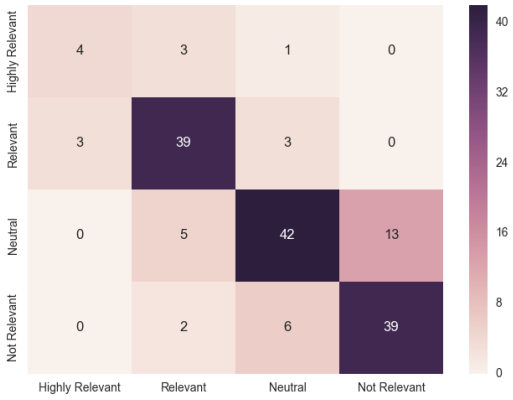

Figure 1: Confusion Matrix of Relevancy Labeling