Financial Entity Identification and Information Integration (FEIII) 2017 Challenge: The Report of the Organizing Committee

Louiqa Raschid
University of Maryland
louiqa@umiacs.umd.edu

Douglas Burdick
IBM Research
drburdic@us.ibm.com

Mark Flood
Office of Financial Research
mark.flood@ofr.treasury.gov

John Grant
University of Maryland
grant@cs.umd.edu

Joe Langsam
University of Maryland
joe@langsam.org

Ian Soboroff
NIST
ian.soboroff@nist.gov

ABSTRACT

This report presents the goals and outcomes of the 2017 Financial Entity Identification and Information Integration (FEIII) Challenge. We describe the dataset and challenge task and the protocol to create labeled data. The report summarizes the process, outcomes and plans for the 2018 Challenge.

Keywords

FEIII, financial entities, relationships, text extraction, sentence ranking.

1. MOTIVATION AND INTRODUCTION

The 2016 Financial Entity Identification and Information Integration (FEIII) Challenge was to develop technologies to automatically align diverse financial entity identification schemes and identifiers [3]. This included data from the Federal Financial Institutions Examination Council (FFIEC), the Securities and Exchange Commission (SEC), and the Global Legal Entity Identifier Foundation (GLEIF).

In 2017, we focus on identifying and understanding relationships among financial entities that reflect activities, e.g., participation within a financial contract, and that have an impact on the behavior and performance outcomes of the financial entities. The FEIII 2017 dataset was thus created from 10-K and 10-Q filings retrieved from the Securities and Exchange Commission (SEC) EDGAR website and additional resources from the National Information Center (NIC) of the FFIEC. We extracted sentences (context sentences) from the filings that provide evidence for a specific role / relationship between the filing financial entity and another mentioned financial entity.

The challenge task was a ranked classification task, to identify relevant and interesting sentences in the filing that provide evidence for a specific relationship, described by a role keyword, between the filing financial entity and a mentioned financial entity. The scored evaluation task was to identify those sentences that (1) validated the relationship and / or (2) provided highly relevant and interesting knowledge that further describes financial activities, behavior and performance outcomes. The FEIII Challenge also required us to develop protocols appropriate to the financial domain to generate labeled data.

A total of eight organizations representing academia and industry in the US, Europe and Asia, generated seventeen submissions.

The report is organized as follows: Section 2 presents details of the task, the dataset and the labeling protocol. Section 3 summarizes the (self reported) approaches taken by the participants and the scores. Section 4 discusses lessons learned and plans for the 2018 challenge.

2. 2017 CHALLENGE DESCRIPTION

2.1 Dataset

The dataset was drawn from the filings of 25+ holding companies with assets exceeding US $ 10 Billion that have provided a resolution plan (Living Will). Using IBM System-T tools [1, 2], we extracted the following triples from SEC 10-K and 10-Q filings:

- Filing Entity (FE): The text string containing the name of the filing entity.
- Mentioned Entity (ME): The text string containing the name of the mentioned entity.
- Role keyword: A keyword that represents the relationship between the filing entity and the mentioned entity.

We also captured contextual text – in particular, three context sentences – around the ME reference and Role keyword. The context sentences provide evidence of the relationship. In addition, these sentences may provide highly relevant information about financial activities, behavior and performance outcomes. In addition to the triples and three context sentences, we shared identifier data (LEI, CIK, RSSD ID lists) as well as relationship information, e.g., from NIC, and corresponding metadata.

DSMM'17, May 14, 2017, Chicago, IL, USA
© 2017 ACM. ISBN 978-1-4503-5031-0/17/05…$15.00
DOI: http://dx.doi.org/10.1145/3077240.3077248

Dataset	Count of filings	Count of triples
Training	25	975
Test	26	900
Working	(536 + 25)	(8622 + 975)
Total	587	10497

Table 1: Data Summary Statistics.

Role	Count of triples
affiliate	129
agent	40
counterparty	108
guarantor	28
insurer	47
issuer	98
seller	49
servicer	57
trustee	304
underwriter	40
Total	900

Table 2: Distribution of Testing Dataset Triples over Roles.

Table 1 provides summary statistics for the dataset and Table 2 provides a distribution of the triples of the testing dataset over the roles.

2.2 Task and Labeling Protocol

Recall that our challenge task was to identify those sentences that (1) validated the relationship and / or (2) provided highly relevant and interesting knowledge that further describes financial activities, behavior and performance outcomes.

Our protocol for labeling triples, and the corresponding three context sentences, from the training dataset, was fairly informal and involved two experts. We developed a more robust protocol for labeling of the test dataset as follows: 1. Each set of three context sentences was reviewed by three student labelers. They chose a rating for the triple and context sentences from the following set: Highly Relevant [H]; Relevant [R]; Neutral [N]; Irrelevant [I]. 2. The labelers also determined if the three context sentences validated the triple. 3. Those triples and sentences that did not have high inter-annotator agreement for the rating, but received at least one [H] rating, was examined by an expert who made a judgement on the rating. 4. Those triples and sentences that received a rating of [H] or [R] but did not have high inter-annotator agreement for the validation was examined by a second expert who made a judgement on the validation.

We provided the following guidelines to the student labelers and the experts for rating the triple and the corresponding three context sentences:

- Highly Relevant [H]: One group of highly relevant sentences will identify potential sources of significant (large) expenses and/or significant business opportunities. Examples of the source of the expenses or opportunities include litigation, spin-offs, acquisitions, etc. Most of these sentences describe a change from the status quo or current situation. Another group of highly relevant sentence will identify corporate character, e.g., the compensation of senior executives or commentary about business activities.
- Relevant [R]: One group of relevant sentences will identify existing assets, liabilities, revenues, or expenses. They may be very specific, e.g., interest rate expenses. Another group of relevant sentences will also identify the size and nature of current business activities, e.g., retail division, underwriting, investment banking, etc.
- Neutral [N]: These sentences may describe the type of business activity, the location of some business entity or activity. They are informative sentences but convey less information value compared to the highly relevant or relevant sentences.
- Irrelevant [I]: This is boilerplate text that is not informative.

Label	Count of triples
Highly Relevant [H]	285
Relevant [R]	225
Neutral [N]	268
Irrelevant [I]	52
Ambiguous	145
Total	975

Table 3: Distribution of Training Dataset Triples over Labels.

Note: It was often difficult to differentiate highly relevant and relevant sentences.

Table 3 provides a distribution of the training dataset triples (and context sentences) over the four labels. We did not consider validation for the training dataset due to time limitations. Table 4 provides a distribution of the testing dataset triples over the six labels; we get six labels when we consider both the rating of the triples and sentences and the validation of the triple by the sentences.

We note that during the labeling of the test dataset, we realized that the ratings of the triples and context sentences were not always aligned or correlated with the validation of the triples. A significant percentage of the highly relevant sentences did not unfortunately validate the triple. We will discuss this in a later section.

Table 5 provides details of the relevance values associated with the following six labels: H+V; H; R+V; R; N; I.

- Sentences that received a highly relevant rating and validated the triple received the highest relevance value.
- The labeled training data did not include labels [H+V] or [R+V], and did not provide training data to differentiate between (highly relevant or relevant) sentences that did / did not validate the triple. Thus, we also gave a high relevance value to triples and sentences rated [H].

We considered multiple combinations of relevance values and we report on gt_1 through gt_5 as well as (gt_1 500) which only considers the Top 500 scores.

Label	Count of triples
Highly Relevant and Validating [H+V]	149
Highly Relevant [H]	160
Relevant and Validating [R+V]	215
Relevant [R]	154
Neutral [N]	142
Irrelevant [I]	80
Total	900

Table 4: Distribution of Testing Dataset Triples over Labels.

	gt$_1$	gt$_2$	gt$_3$	gt$_4$	gt$_5$
Highly Relevant / Valid [H+V]	4	4	4	4	4
Highly Relevant [H]	3	3	0	3	0
Relevant / Valid [R+V]	0	2	0	3.5	3.5
Relevant [R]	0	2	0	0	0
Neutral [N]	0	1	0	0	0
Irrelevant [I]	0	0	0	0	0

Table 5: Relevance Scores for Labeled Ground Truth Triples.

	gt$_1$	gt$_1$ 500	gt$_2$	gt$_3$	gt$_4$	gt$_5$
Min	0.71	0.34	0.88	0.62	0.86	0.76
Max	0.92	0.82	0.96	0.79	0.94	0.86
Mean	0.82	0.62	0.93	0.70	0.90	0.81

Table 6: Minimum, Maximum and Mean values for Normalized Discounted Cumulative Gain for Different Ground Truth Relevance Scores.

Role	Min	Max	Mean
affiliate	0.50	0.72	0.59
agent	0.53	0.89	0.71
counterparty	0.70	0.90	0.82
guarantor	0.63	0.90	0.72
insurer	0.73	0.96	0.82
issuer	0.69	0.88	0.76
seller	0.59	0.94	0.76
servicer	0.77	0.94	0.84
trustee	0.68	0.81	0.74
underwriter	0.48	0.76	0.60
overall	0.69	0.79	0.74

Table 7: Minimum, Maximum and Mean values for Normalized Discounted Cumulative Gain with Ground Truth gt$_5$ for Different Roles.

3. SOLUTION APPROACHES

A total of eight organizations representing academia and industry in the US, Europe and Asia, generated seventeen submissions [4, 5, 6, 7, 8, 9, 10]. We briefly summarize the highlights.

- **P1 (corporate):** The team determined the differential affinity of keywords in the context sentences to each of the roles, and computed the cumulative relevance over all keywords.
- **P2 (corporate):** Made extensive use of a proprietary database of financial entity identifiers to match entity mentions in the triples and context sentences.
- **P3 - P10 (academic):** This team explored a combination of features including a Bag-of-words that was used to extract N-grams and a deep-learning based word embedding model. They also exploited inter-annotator agreement (Cohen's Kappa) to select or discard labels.
- **P11, P17 (corporate):** Addressed the original task of identifying sentences that validated the triple. Identified multiple scenarios where there is partial validation of the triple or where the triple is in fact, ambiguous or incorrect. Specified rules using regular expressions and text features to encode invalid scenarios.
- **P12 (corporate):** This participant made extensive use of (proprietary) databases, knowledge graphs, and text extraction utilities.
- **P13, P14 (academic):** This participant used a deep-learning based word embedding model. They made good use of the unlabeled data and they constructed a classifier for each role.
- **P15 (corporate):** This participant used a syntactic and semantic parser and obtained a range of linguistic features (POS tagging, subject and object modifiers, etc.). Specialized text extraction utilities were also customized for the financial domain.
- **P16 (academic):** This participant used an external resource to determine concept probabilities for context sentences. Highly relevant sentences appeared to be associated with more concepts. They also used a deep-learning based word embedding model. They further studied the differences between the first and last context sentence.

Table 6 reports on the minimum, maximum and mean Normalized Discounted Cumulative Gain (NDCG) scores, over all triples/roles, for the various combinations of relevance values (see Table 5), over all seventeen submissions. We note that the NDCG is generally pretty high; this reflects the large count of [H+V], [H] and [R+V] triples and context sentences in the dataset.

Table 7 reports on the minimum, maximum and mean NDCG scores for each role. We report on gt$_5$ since it has a focus on validation and gives relevance values to [H+V] and [R+V] but not to [H]. We make the following observations:

- Some roles and relationships appear to be more clearly described within the context sentences. For example, the roles counterparty, insurer and servicer appear to yield higher NDCG scores. We note that there are a smaller count of triples for the roles insurer and servicer, in the dataset. It is possible that such infrequent roles are more clearly understood.
- Other roles, e.g., trustee and affiliate, appear to be more complex, and the corresponding relationships may not be well described within the context sentences; they appear to yield lower NDCG scores.

4. LESSONS LEARNED, IMPACT AND FUTURE PLANS

The lessons learned focus on the orthogonality of the rating of the triples and context sentences versus the validation of the triple by the context sentences. Our initial hypothesis was that a sentence that contains highly relevant financial knowledge will have a high(er) probability of validating a triple. We were somewhat surprised to find that the converse also was true, i.e., a sentence that validated a triple had a high(er) probability of containing highly relevant financial knowledge.

Despite all of these (positive) observations and correlations, the ratings of the triples and context sentences were not always aligned or correlated with the validation of the triples. A significant percentage of the highly relevant sentences did not unfortunately validate the triple.

We briefly summarize the following scenarios where highly relevant sentences failed to validate the triple:

- A common case was that the role was correctly associated with the mentioned entity. However, the sentences did not fully validate that the relationship was with the filing entity. They also did not invalidate the relationship.
- In some cases, the role was correctly associated with the mentioned entity, but the sentences showed that the relationship was with a different mentioned entity and not with the filing entity. Here the context sentences definitely invalidated the triple.
- The role was incorrectly associated with a mentioned entity, i.e., the triple was not constructed correctly.
- There were several cases where the mentioned entity was identical to the filing entity; this was almost always associated with an error in constructing the triple.
- As mentioned earlier, ambiguous or complex financial roles, e..g, trustee or affiliate, more easily lead to a lack of validation.
- Long and complex context sentences often covered multiple relationships. They may also have been incomplete or lacking in detail about some relationships, making it difficult to fully validate a triple.

FEIII 2018 will continue the challenge of sentence ranking and triple validation. We will expand on the 2017 Challenge as follows:

- Identifying context sentences that provide significant details to embellish the relationship.
- Identifying and resolving the identical relationship instance across the dataset.
- Inferring new knowledge from multiple triples and context sentences.

5. ACKNOWLEDGMENTS

This research was partially supported by NIST award 70NANB15H194 and NSF grants CNS1305368 and DBI1147144 and the Smith School of Business.

6. ADDITIONAL AUTHORS

Elena Zotkina (University of Maryland, email: `ezotkina@umiacs.umd.edu`).

7. REFERENCES

[1] D. Burdick, M. A. Hernández, H. Ho, G. Koutrika, R. Krishnamurthy, L. Popa, I. Stanoi, S. Vaithyanathan, and S. R. Das. Extracting, linking and integrating data from public sources: A financial case study. *IEEE Data Eng. Bull.*, 34(3):60–67, 2011.

[2] L. Chiticariu, R. Krishnamurthy, Y. Li, F. Reiss, and S. Vaithyanathan. Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1002–1012. Association for Computational Linguistics, 2010.

[3] M. Flood, J. Grant, H. Luo, L. Raschid, I. Soboroff, K. Yoo, and E. Zotkina. Financial entity identification and information integration (FEIII) challenge: The report of the organizing committee. In *Proceedings of the Workshop on Data Science for Macro-Modeling (DSMM@SIGMOD)*, pages 1:1–1:4, 2016.

[4] R. Hicks. Factset - the advantage of scored data. In *Proceedings of the Workshop on Data Science for Macro-Modeling (DSMM@SIGMOD)*, 2017.

[5] M. Kejriwal. Predicting role relevance with minimal domain expertise in a financial domain. In *Proceedings of the Workshop on Data Science for Macro-Modeling (DSMM@SIGMOD)*, 2017.

[6] J. Park, H. Cho, and S. Hwang. Understanding relations using concepts and semantics. In *Proceedings of the Workshop on Data Science for Macro-Modeling (DSMM@SIGMOD)*, 2017.

[7] J. Perez, C. Proux, A. Sandor, and D. Proux. Hybrid feature factored system for scoring extracted passage relevance in regulatory filings. In *Proceedings of the Workshop on Data Science for Macro-Modeling (DSMM@SIGMOD)*, 2017.

[8] T. Repke, M. Loster, and R. Krestel. Ranking sentences describing relationships between financial entities by relevance. In *Proceedings of the Workshop on Data Science for Macro-Modeling (DSMM@SIGMOD)*, 2017.

[9] C. Wang, H. Zhu, Y. Li, L. Chiticariu, R. Krishnamurthy, and D. Burdick. Towards re-defining relation understanding in financial domain. In *Proceedings of the Workshop on Data Science for Macro-Modeling (DSMM@SIGMOD)*, 2017.

[10] P. Yadav and R. Naini. Entity relationship ranking using differential keyword-role affinity. In *Proceedings of the Workshop on Data Science for Macro-Modeling (DSMM@SIGMOD)*, 2017.

DISCLAIMER: The views and opinions expressed in this paper are those of the authors and do not represent the policy or position of any of the author-affiliated organizations including the Office of Financial Research and the Department of the Treasury.