Comparing Features for Ranking Relationships Between Financial Entities based on Text

Tim Repke
Hasso Plattner Institute
Potsdam, Germany
tim.repke@hpi.de

Michael Loster
Hasso Plattner Institute
Potsdam, Germany
michael.loster@hpi.de

Ralf Krestel
Hasso Plattner Institute
Potsdam, Germany
ralf.krestel@hpi.de

1 INTRODUCTION

Evaluating the credibility of a company is an important and complex task for financial experts. When estimating the risk associated with a potential asset, analysts rely on large amounts of data from a variety of different sources, such as newspapers, stock market trends, and bank statements. Finding relevant information in mostly unstructured data is a tedious task and examining all sources by hand quickly becomes infeasible.

An important aspect of risk management are the relations of a company of interest to other financial entities. Automatically extracting such relationships from unstructured text files, such as 10-K filings, significantly reduces the amount of manual work. Such structured knowledge enables experts to quickly gain insight into a company's relationship network. However, not all extracted relationships may be important in a given context. In this paper, we propose an approach to rank extracted relationships based on text snippets, such that important information can be displayed more prominently.

2 DATASET

The dataset used for this work was provided in the context of the FEIII Challenge 2017[4], which contains *triples* extracted from 10-K and 10-Q filings, describing a relationship (*role*) between the *filing company* and a *mentioned financial entity*. Text snippets of three sentences provide the context a relation appeared in. Relationships are limited to ten predefined roles (see table 1). Judging from their respective text snippets, triples were labelled by experts according to their relevance from a business perspective as *irrelevant*, *neutral*, *relevant*, or *highly relevant*. There are 975 training samples from 25 10-K filings, and 900 triples for testing from 25 disjunct filings.

Task Description. The challenge is aimed to explore methods that automatically produce a ranking of triples with the same role by relevance. This complements last year's challenge to identify financial entities in free text[1].

Inter Annotator Agreement. The Inter Annotator Agreement, measured by Cohen's Kappa[2], has a weighted average of $\overline{\kappa} = 0.45$, which indicates a high level of disagreement. About 40% of training triples were rated by more than one expert, in the test set all triples received at least three ratings.

3 OUR APPROACH

We rank the snippets for each role based on multi-class classifiers, which are trained on the experts' labels. As input to a classifier we compare three feature sets to represent snippets, namely *bag-of-words* (BOW), *embeddings* (EMB), and *syntax features* (SYN).

We use ensembles of four one-versus-rest Support Vector Classifiers (SVC) with sigmoid kernel, as well as random forests with 20 trees to classify snippets. The confidence scores in an ensemble are normalised using the softmax function. From that we derive the ranking score as the maximum probability weighted by its corresponding label. Class imbalance is adjusted for by weights.

In our experiments, the SVC model has proven to be a good choice for BOW and EMB, but has shown unsatisfactory performances on syntax features. Therefore, we chose random forests, which perform much better in this case.

Although a model could learn role specific characteristics and improve its performance, we found that due to the limited number of training samples better results are achieved by learning one model from all samples disregarding the role.

3.1 Bag-of-Words

Our first model uses a simple bag-of-words representation of the snippets to classify them. N-grams are extracted for $n = 1$ to 3 and are weighted based on information gain between the classes. In order to reduce the feature space and guard against over-fitting, the most and least frequent terms are removed from the index.

3.2 Sentence Embeddings

Difficulties with previously unseen examples might arise from the limited training size. Word embeddings can alleviate this problem by representing words in a 50- to 300-dimensional vector space. These representations are learned by using unsupervised deep learning. Internally, a neural network is trained to predict the following word in a sequence of words based on the word's context window.

We learned paragraph embeddings[1] from 25 of the original full text 10-K filing documents containing 60k sentences (2m words). Previous research has shown, that such embeddings manage to outperform BOW approaches[3]. We use a window size of 10 and a paragraph vector of size 50, which is trained for 10 epochs over the sentences in all filings.

DSMM'17, Chicago, IL, USA

DOI: http://dx.doi.org/10.1145/3077240.3077252

[1] Using Gensim https://radimrehurek.com/gensim/models/doc2vec.html

Table 1: Averaged experimental results for each role using BOW+EMB+SYN

	affiliate	agent	counterpart	guarantor	insurer	issuer	seller	servicer	trustee	underwriter
# samples (train/eval)	185	61	64	34	19	129	20	21	420	21
# samples (test)	129	40	108	28	47	98	49	57	304	40
NDCG (5-fold-cv)	0.93	0.92	0.93	0.93	0.97	0.89	0.91	0.91	0.97	0.94
Baseline (random)	0.89	0.87	0.88	0.89	0.92	0.83	0.84	0.88	0.92	0.89

Table 2: Experimental results for bag-of-words (BOW), embedding (EMB), syntax (SYN) features, and ensemble

Approach	NDCG	σ(NDCG)	F1-Score	σ(F1)
Baseline (random)	0.88	0.03	-	-
Baseline (worst)	0.72	0.06	-	-
BOW	0.88	0.05	0.34	0.13
EMB	0.89	0.04	0.24	0.18
SYN	0.94	0.04	0.44	0.11
BOW+EMB+SYN	**0.95**	0.04	0.43	0.12

To build the EMB representation for the text snippet associated with a triple, the embedding is used to induce a vector for each of the three sentences in the snippet, which are then concatenated.

3.3 Syntax Features

Additionally, to provide a language independent approach, we created a set of syntax-based features. Following the Gini impurity metric, features, such as the ratio of upper-case words and numbers, or the number of dollar signs and word repetitions, appear to be most meaningful for classification. In total we chose 20 features describing the amount or presence of different syntactical characteristics.

3.4 Ensemble

Each of the numerical representations and their resulting models have individual strengths and weaknesses. For example, the language independence of SYN can tolerate a changing vocabulary to a certain extent, but misses the advantage to identify key phrases which may prove useful for classification. As a conclusion, we combined the three models by summing the individual predictions to form a soft vote.

4 EVALUATION

The system's performance is measured by normalised discounted cumulative gain (NDCG)[2]. We perform 5-fold cross-validation *(5-fold-cv)* by leaving out training triples based on the documents they were extracted from. Those triples form the evaluation set *(eval)*. Table 2 lists the mean NDCG scores and the standard deviation (σ), which are calculated for each role's ranking as shown for the ensemble in table 1. For comparison, we consider a baseline of the worst possible ranking (inverse order of the ideal ranking) and the average of multiple random rankings.

The BOW model performs best on evaluation data (NDCG@0.98) but the feature selection shows seemingly very specific terms which are likely to negatively affect the model's ability to classify unseen samples, which is proven in by the test data. Training a model on text usually requires a reasonably large corpus which we hope to counteract by using embeddings based on 10-K filings. However, even the EMB model barely beats the baseline (NDCG@0.92 on eval). Best and most stable results are achieved with the SYN model and the ensemble, which perform the same on the evaluation data as on test data.

Looking at the performance of the classification task itself (measured by the F1-Score) we observe stable results for the ensemble with low deviation on evaluation data, which is supported by same scores on test data. Contrary to that is the BOW model, which showed very high deviations and a significant drop from F1-Score 0.73 on evaluation data to test data.

5 CONCLUSION

Overall, we managed to achieve good NDCG scores of around 0.95 using an ensemble of models. We have shown, that BOW is very sensitive in changing vocabulary used in 10-K filings used in training data to 10-Q filings as in the test data. Our assumption that paragraph embeddings may be more robust to such changes by training them on a significantly larger set of text and is able to reflect phrase similarities did not hold. In combination with syntax features in an ensemble, ranking triples describing the relationship between financial entities based on text snippets yields most stable outcomes.

As this work only focuses on a small textual context, for future work we are interested in additional external data, e.g. the impact of a business relationship may be judged by comparing revenues of the involved companies. Thus, triples could be enriched by adding (historical) revenue of the two involved financial entities.

REFERENCES

[1] 2016. *DSMM'16: Proceedings of the Second International Workshop on Data Science for Macro-Modeling.* ACM, New York, NY, USA.
[2] Bruce Croft, Donald Metzler, and Trevor Strohman. 2009. *Search Engines: Information Retrieval in Practice.* Pearson.
[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Conference on Neural Information Processing Systems 2013.*
[4] Louiqa Raschid, Doug Burdick, Mark Flood, John Grant, Joe Langsam, Ian Soboroff, and Elena Zotkina. 2017. 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).*