Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5412–5417
Marseille, 20-25 June 2022

A Graph-Based Method for Unsupervised Knowledge Discovery from Financial Texts

Joel Oksanen*, Abhilash Majumder†, Kumar Saunack†, Francesca Toni*, Arun Dhondiyal†
*Imperial College London
South Kensington Campus, London SW7 2AZ, UK
joel.oksanen17, f.toni@imperial.ac.uk
†MSCI Inc.
7 World Trade Center, New York 10007, USA
abhilash.majumder, kumar.saunack, arun.dhondiyal@msci.com

Abstract

The need for manual review of various financial texts, such as company filings and news, presents a major bottleneck in financial analysts' work. Thus, there is great potential for the application of NLP methods, tools and resources to fulfil a genuine industrial need in finance. In this paper, we show how this potential can be fulfilled by presenting an end-to-end, fully unsupervised method for knowledge discovery from financial texts. Our method creatively integrates existing resources to construct automatically a knowledge graph of companies and related entities as well as to carry out unsupervised analysis of the resulting graph to provide quantifiable and explainable insights from the produced knowledge. The graph construction integrates entity processing and semantic expansion, before carrying out open relation extraction. We illustrate our method by calculating automatically the environmental rating for companies in the S&P 500, based on company filings with the SEC (Securities and Exchange Commission). We then show the usefulness of our method in this setting by providing an assessment of our method's outputs with an independent MSCI source.

Keywords: Financial Applications, Knowledge Discovery, Information Extraction

1. Introduction

Knowledge discovery from various structured and unstructured data sources has become a popular research topic in many applied fields where data is used extensively, including financial services. A financial analyst's work involves manually reviewing lengthy SEC (Securities and Exchange Commission) filings and financial news articles in order to extract relevant pieces of information. This presents a major bottleneck which could be alleviated using automated knowledge discovery and information extraction methods. At the same time, the financial services industry is heavily regulated, which means the knowledge from such systems must be accurate and explainable. This makes *black-box* models an unfavourable solution to this problem. Another problem is the lack of publicly available training datasets for financial knowledge discovery, which makes it difficult to use supervised learning methods in general.

In this paper, we tackle these problems by proposing a novel end-to-end method for unsupervised knowledge discovery from financial texts. We focus specifically on index creation. Currently, this involves analysis of filings and other sources by several analysts. Our proposed method allows any analyst to have extracted information ready for consumption from an independent source, without actually having to go through the filings, which usually contain not less than 50 pages. Our method creatively uses various Natural Language Processing (NLP) methods to extract a structured Knowledge Graph (KG) from the unstructured textual data. The KG is centred around a user-defined *topic*, for example *sustainability*. The KG consists of *nodes* for companies and topic-related entities as well as edges indicating *relations* between the nodes. Our method can automatically analyse the resulting KG to produce numerical insights about companies in relation to the chosen topic, similar to the results of a human review of the articles. These figures are based on a legible graph instead of a neural network model, which means that their origins are fully explainable.

The paper is organised as follows. In Section 2 we briefly discuss related work and existing methods we build upon. In Section 3 we provide our methodology. In Section 3.6 we describe a case study for assessing companies' sustainability, specifically showing a comparison between our results and MSCI ESG scores for the same companies. Finally, in Section 5 we conclude, pointing in particular to future work.

2. Related Work

The use of NLP for financial settings is an important application domain nowadays. Existing work includes question answering (Liu et al., 2020), numerical reasoning on financial reports (Chen et al., 2021) and portfolio selection (Liang et al., 2021), amongst others.

In general, knowledge graph learning (KGL) is an exciting field of research, as it creates knowledge that is readily explainable to human users. Most state-of-the-art KGL methods rely on some degree of human intervention (Ji et al., 2021), but there have been some attempts to build completely unsupervised KGL meth-

Group	
Co, Company	Inc, Incorporated
Grp, Group	Corporation, Corp
Ltd, Limited	Bank, BanCorp,
and, &	Bancorp

Table 1: Company name expansion groups.

ods, for example in the domain of medicine (Frisoni et al., 2020). Although KGs are already being used to facilitate financial systems (Cheng et al., 2020), to our knowledge, no attempts for unsupervised KGL have been made in the financial domain, which has the unique characteristic of being centered around companies and related aspects.

Rather than defining our method from scratch, we opted for creatively reusing a number of publicly available methods, tools and resources, in the spirit of *not re-inventing the wheel*. In particular, we use a Named Entity Recognition (NER) model from Stanza (Qi et al., 2020) trained on the OntoNotes dataset (Hovy et al., 2006), CoreNLP models for co-reference resolution (Recasens et al., 2013) and Open Information Extraction (OpenIE) (Angeli et al., 2015), and NumberBatch word embeddings (Speer et al., 2017) to embed pre-trained term semantics in our KG. Our method combines these existing tools alongside novel domain-specific methods in a fully automatic KGL system capable of extracting relevant company information from unstructured texts.

3. Methodology

In the first section (3.1), we describe a novel method to map organisational entities to a specific company identifier, that is effective in the given domain. We focused on combining several distinct methods to create a potentially useful tool, instead of developing an end-to-end model from scratch for data extraction and scoring. Since the data to be extracted depends on the context, a versatile tool which can extract the relevant information just by changing the keywords, instead of having to go through the long process of model training and fine-tuning, is applicable to a larger set of scenarios. Our method takes as input a set of financial texts and a few hand-selected seed terms that help it define the KG topic. We describe the four-stage process to construct the KG from these inputs: named entity extraction (Section 3.2), semantic expansion (Section 3.3), open relation extraction (Section 3.4), and KG construction (Section 3.5). Section 3.6 outlines a novel method for the automatic analysis of the resulting KG, in relation to the chosen topic, using the semantic knowledge obtained in Section 3.3. Figure 1 shows an overview of the pipeline underpinning the entire methodology.

3.1. Company Name Expansion

The ability to map company names to unique identifiers (ISINs) is crucial for building a concise knowledge graph with a single node for each company. While it is a trivial task to obtain an official company name for each identifier (e.g. *Consolidated Edison Inc.*), in natural language, companies are often referred to with several different variations of this name (e.g. *Con Edison*). Therefore, we developed an automatic method to expand each official company name to a set of unique names by which a company might be referred to, which will be used in Section 3.2 to extract company entities from texts.

The method for company name expansion takes as input a single (ISIN, name) tuple for each company. We first normalize the names by removing all character casing, after which we expand each name through the three steps outlined below.

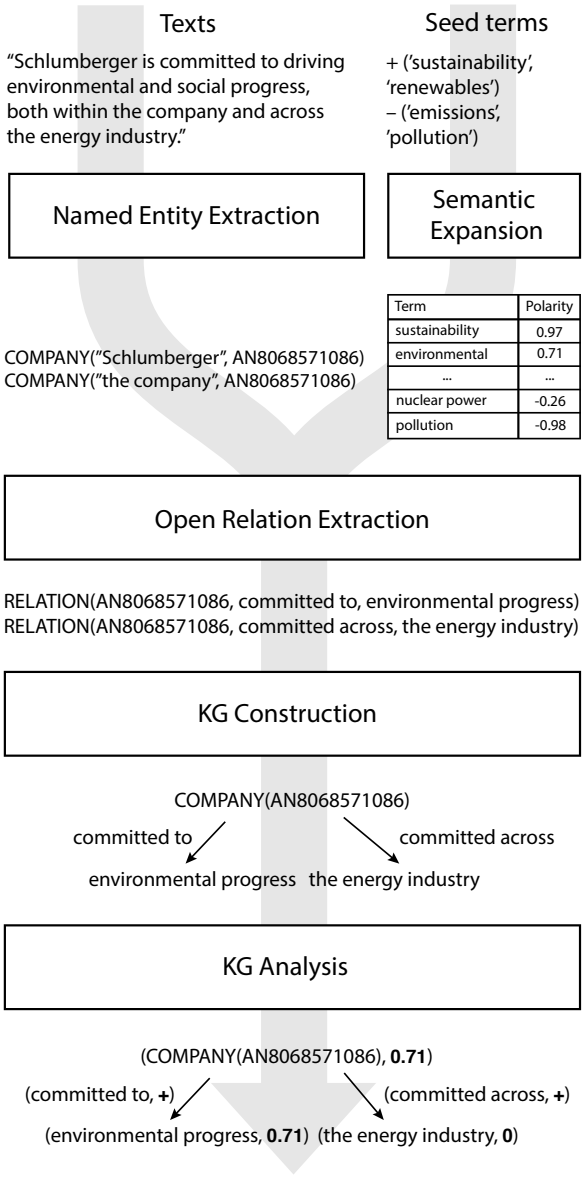

Figure 1: An overview of our methodology exemplified for the sustainability topic.

Subsets We divide each name into words, and create

new names out of all combinations of those words. This covers cases where a company is referred to by a subset of its full name, e.g. *Apple* instead of *Apple Inc.*

Prefixes For each name, we include versions where one or multiple of its words have been abbreviated into just a prefix. This covers abbreviated company names, e.g. *Con Edison* instead of *Consolidated Edison Inc.*

Hand-selected features Finally, we check if each name contains any instances from a set of hand-selected words that commonly occur in company names, such as *Ltd.* in *Garmin Ltd.*, and include versions with other common forms of the word, e.g. *Garmin Ltd* and *Garmin Limited*. The full list of such word groups is shown in Table 1.

The expansions are likely to include names that are duplicated across several identifiers (e.g. *Company* for *McCormick & Company* and *McKinsey & Company*). For our final mapping, we only include unique names that map to exactly one identifier.

3.2. Named Entity Extraction

The named entity extraction stage involves finding and extracting the *named entities* of interest from the input texts, which will become nodes in the KG. Named entities are entities belonging to a certain class – in this paper, we focus on entities belonging to the COMPANY and PERSON classes. To find instances of these classes in the texts, we use an out-of-the-box Named Entity Recognition (NER) model from Stanza (Qi et al., 2020) trained on the OntoNotes dataset (Hovy et al., 2006). The dataset includes the PERSON and ORG (*organization*) classes, the latter of which is a superclass of the COMPANY class.

After obtaining the entities belonging to the ORG and PERSON classes, we check for each ORG entity if it corresponds to a company identifier to obtain the COMPANY entities. This is done by finding matches for the ORG entities in the expanded company name mapping obtained via the process detailed in Section 3.1.

Finally, we use the CoreNLP (Recasens et al., 2013) model for co-reference resolution to obtain any named entity co-references missed by the NER model.

3.3. Semantic Expansion

The semantic expansion stage derives from a small hand-selected set of *seed terms* a large set of possible topic entities for the KG. The seed terms define the KG topic and are divided into two groups of + and −, which defines a linear topic polarity scale used in the KG analysis (Section 3.6). We obtain NumberBatch (Speer et al., 2017) word embeddings for each of the seed terms and average the vectors in both groups, obtaining a positive $\boldsymbol{P}_+$ and a negative pole $\boldsymbol{P}_-$ for the topic. We then calculate the cosine similarities between each word embedding $\boldsymbol{w}_i$ and the two poles $\boldsymbol{P}_+$ and $\boldsymbol{P}_-$, which we define as $tr_+(\boldsymbol{w}_i)$ and $tr_-(\boldsymbol{w}_i)$ respectively. Using these, we calculate an overall topic relatedness measure $tr(\boldsymbol{w}_i) = tr_+(\boldsymbol{w}_i) + tr_-(\boldsymbol{w}_i)$ for each $\boldsymbol{w}_i$, and select the words corresponding to the top 500 $\boldsymbol{w}_i$ with the highest $tr(\boldsymbol{w}_i)$ as our topic entities e_i. For each of the selected entities, we calculate a scaled polarity measure $p_{scaled}(e_i)$ as

$$p_{scaled}(e_i) = \frac{tr_+(\boldsymbol{w}_i) - tr_-(\boldsymbol{w}_i)}{\max_j |tr_+(\boldsymbol{w}_j) - tr_-(\boldsymbol{w}_j)|}.$$

In the final entity polarity measure $p(e_i)$, we set small polarities as neutral:

$$p(e_i) = \begin{cases} p_{scaled}(e_i) & |p_{scaled}(e_i)| \geq 0.1 \\ 0 & |p_{scaled}(e_i)| < 0.1. \end{cases}$$

3.4. Open Relation Extraction

We use the CoreNLP Open Information Extraction (OpenIE) annotator (Angeli et al., 2015) to extract open-domain (*subject*, *relation*, *object*) triples from the texts. For each triple, we check if the *subject* and *object* correspond to a named entity or a topic entity: if a match is found for both, we include the relation in our KG. Named entity matches must be unique (a single company or a person per subject/object), whereas there can be multiple topic entities in one subject/object (for example, *solar* and *energy* in *solar energy*).

3.5. Knowledge Graph Construction

The subjects and objects of the extracted relations form the nodes of the KG, connected by the relations between them. In order to construct a useful KG, we must aggregate the nodes such that in the final graph there is only one node corresponding to an entity or entity combination. The aggregation methods used for each entity class are detailed below.

Company We have already mapped COMPANY entities to a unique company identifier in the named entity extraction stage (Section 3.2), so these are aggregated directly based on the identifier.

Person To aggregate PERSON entities, we use partial fuzzy string matching. When comparing two entities, the shorter entity is compared with each substring of the longer entity by calculating the Levenshtein distance; if the smallest Levenshtein distance is below a certain threshold, the two entities are joined. This accounts for spelling errors as well as cases where a person is referred to by only their last name.

Topic TOPIC nodes are aggregated by the topic entities that were identified in them: nodes containing the same set of topic entities are joined as one.

3.6. Knowledge Graph Analysis

The knowledge graph analysis stage uses the KG to calculate a polarity figure $p(c_i)$ for each company node c_i representing the company's position with regards to the chosen topic. The figures are derived from the company nodes' relations to the topic nodes t_i, for which we can define a polarity measure

$$p(t_i) = \frac{\sum_{e_j \in t_i} p(e_j)}{|e_j \in t_i|}.$$

Let $r(n_i, n_j)$ if there is a relation from node n_i to node n_j. We define a polarity for each relation: either $p(n_i, n_j) = +1$, meaning a positive correlation between the two nodes, or $p(n_i, n_j) = -1$, meaning a negative correlation. This is obtained through a simple bag-of-words approach, checking for words from a hand-selected list of common negations (*not*, *stop*, *deny*, etc.) in the relation. If the number of negations in the relation is odd, $p(n_i, n_j) = -1$, else $p(n_i, n_j) = +1$.

The polarities for the remaining nodes for companies and persons n_i are calculated recursively starting from the topic nodes as

$$p(n_i) = \frac{\sum_{n_j \in R_{n_i,n_j}} p(n_j) \times p(n_i, n_j)}{|R_{n_i,n_j}|},$$

where $R_{n_i,n_j} = \{n_j | r(n_i, n_j), rel(n_i, n_j)\}$ is the set of related object nodes n_j relevant for the polarity calculation of n_i, where

$$rel(n_i, n_j) = \begin{cases} False & p(n_j) = 0 \\ True & n_j \text{ is topic} \\ n_i \text{ is not person} & n_j \text{ is person} \\ False & n_j \text{ is company.} \end{cases}$$

In the end, each company node c_i will have a polarity figure $p(c_i)$ that quantifies its relation to the given topic.

4. Case Study

Here, we explore the usefulness of our tool for gaining insights when creating indices, and evaluate it against a standard index.

We took the annual 10-K filings for all companies listed in the SP 500 (as of September 01, 2021, from Wikipedia) for the previous 5 years (2016-2021). We removed all tabular data from the filings and all content prior to Item 1 in the filing before running the resulting text through the pipeline. Scores were generated using the resulting graph for each company, using the following seed terms:

$+ = \{\text{environmental, sustainability, renewables}\};$
$- = \{\text{emissions, pollution, fossil fuel, regulation}\}.$

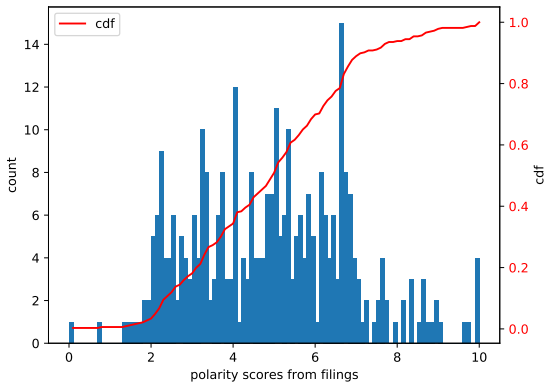

Figure 2: Distribution of scores generated from the (texts in the) 10-K filings. The red line represents the cumulative distribution function (cdf) for the scores.

We ignored all the companies with less than 15 facts and topic nodes combined: out of the 500 companies considered, this led to 358 companies only (which we could extract sufficient data for). We then obtained the final scores by calculating the mean of all polarity figures from all 5 years. Figure 2 shows the distribution the generated scores, along with the cumulative distribution function. As shown in the plot, the polarity scores obtained from our pipeline are skewed towards the lower ranges and peak around 7.5.

Finally, we compared this output against the MSCI ESG (Environmental, Social and corporate Governance) ratings for the corresponding companies. Results from the real estate sector are presented in Figure 3. The 10-K filings for this sector usually contain specific sections related to ESG, so datapoints can be extracted for a larger number of companies within this sector. As can be seen from the plot, the scores automatically obtained from the filings using our pipeline follow the general trend of the actual scores assigned to the company by MSCI. However, there are variations from the trend, which are mainly due to the way that the company reports its performance in the annual reports. Indeed, some companies tend to focus more on the compliance and their achievements towards sustainability goals, while others tend to highlight this much less, and this causes fluctuations in the scoring.

5. Conclusion

We presented a flexible, unsupervised method, combining different areas of and resources in NLP, to extract information from financial texts and applied it to ESG-related information from company filings data. The results indicate that, within industries, the filings can provide a way to roughly gauge manually-defined ESG ratings. However, trends across industries are distinctly different. Indeed, for the software industry filings are mostly populated by environmental regulations and restrictions, requiring further work to generate reliable

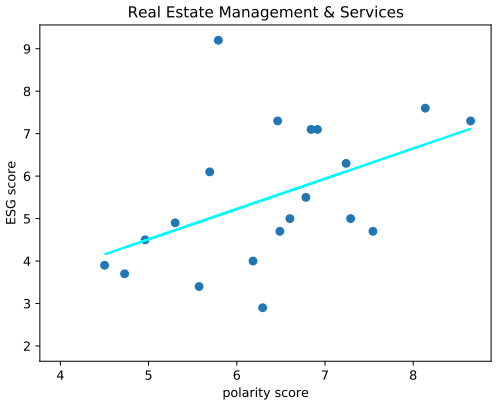

Figure 3: Scatter plot of scores from filings (x-axis) and MSCI ESG scores (y-axis) of companies in the real estate sector (scores have been rescaled to the range [0,10]). The blue line indicates the best fit line for the scores. A correlation score of 1 is represented by the line y=x.

scores.

Our method generates scores from knowledge graphs, automatically extracted from text in an unsupervised manner. The knowledge graph creation itself provides a promising way to extract information (other than the mere scores), which may eventually eliminate the need for manual analysis of financial texts.

Going forwards, there are several avenues which can be explored to improve the method's performance:

- It would be interesting to explore changing the scoring mechanism to incorporate numerical performance: currently, pledges of a million or 10 million dollars are treated in the same way, although there is a difference in the degree of commitment in both approaches. By adding this information, the scores can be a better representation of the efforts of a company towards its (environmental) goals.

- It would be useful to add additional news sources: by covering a neutral third party's review of a company's performance, the scores would be more reflective of the actual (ESG) performance of the company. Additionally, since reporting across industries tends to have a similar format for the same news outlet, by considering additional sources the problem of non-standard self-reporting will hopefully be resolved.

- Our method was evaluated on annual 10-K filings from the American financial information environment, which are presented in a somewhat more standardised format compared to reports in other contexts (for example the British English or European contexts). To further improve performance across different contexts, one could investigate the integration of domain adaptation methods into the pipeline, the importance of which has previously been demonstrated in the financial domain in e.g. (Loughran and McDonald, 2011) and (El-Haj et al., 2014).

6. Acknowledgements

This research was facilitated by the Imperial Business Partners programme at Imperial College London. The authors would like to thank Christopher J. Parker from the Imperial Business Partners programme and Akbar Hussain from MSCI for their support throughout.

7. Bibliographical References

Chen, Z., Chen, W., Smiley, C., Shah, S., Borova, I., Langdon, D., Moussa, R., Beane, M., Huang, T., Routledge, B. R., and Wang, W. Y. (2021). FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 3697–3711.

Cheng, D., Yang, F., Wang, X., Zhang, Y., and Zhang, L. (2020). Knowledge graph-based event embedding framework for financial quantitative investments. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 2221–2230.

El-Haj, M., Rayson, P., Young, S., and Walker, M. (2014). Detecting document structure in a very large corpus of UK financial reports. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1335–1338.

Frisoni, G., Moro, G., and Carbonaro, A. (2020). Unsupervised descriptive text mining for knowledge graph learning. In *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, pages 316–324.

Ji, S., Pan, S., Cambria, E., Marttinen, P., and Yu, P. S. (2021). A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, page 1–21.

Liang, Q., Zhu, M., Zheng, X., and Wang, Y. (2021). An adaptive news-driven method for cvar-sensitive online portfolio selection in non-stationary financial markets. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2708–2715.

Liu, Z., Huang, D., Huang, K., Li, Z., and Zhao, J. (2020). Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4513–4519.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66:35–65.

8. Language Resource References

Angeli, G., Johnson Premkumar, M. J., and Manning, C. D. (2015). Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Recasens, M., de Marneffe, M.-C., and Potts, C. (2013). The life and death of discourse entities: Identifying singleton mentions. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Speer, R., Chin, J., and Havasi, C. (2017). ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pages 4444–4451.