Learning Company Embeddings from Annual Reports for Fine-grained Industry Characterization

Tomoki Ito[1*], **Jose Camacho Collados**[2], **Hiroki Sakaji**[1], **Steven Schockaert**[2]

[1]Graduate School of Engineering, The University of Tokyo, Japan
[2]School of Computer Science & Informatics, Cardiff University, UK

Abstract

Organizing companies by industry segment (e.g. artificial intelligence, healthcare or fintech) is useful for analyzing stock market performance and for designing theme base investment funds, among others. Current practice is to manually assign companies to sectors or industries from a small predefined list, which has two key limitations. First, due to the manual effort involved, this strategy is only feasible for relatively mainstream industry segments, and can thus not easily be used for niche or emerging topics. Second, the use of hard label assignments ignores the fact that different companies will be more or less exposed to a particular segment. To address these limitations, we propose to learn vector representations of companies based on their annual reports. The key challenge is to distill the relevant information from these reports for characterizing their industries, since annual reports also contain a lot of information which is not relevant for our purpose. To this end, we introduce a multi-task learning strategy, which is based on fine-tuning the BERT language model on (i) existing sector labels and (ii) stock market performance. Experiments in both English and Japanese demonstrate the usefulness of this strategy.

1 Introduction

Investing in individual companies carries a high risk to investors, as stock prices can move in highly unpredictable ways. A popular alternative is to reduce this idiosyncratic risk by instead investing in funds that track the performance of a particular index (i.e. weighted set of companies). While most indices have traditionally been designed to capture particular geographic regions (e.g. S&P 500 for the US market), in recent years, funds that track a particular industry segment have been gaining in popularity. For instance, such funds allow investors who believe that technology companies will continue to outperform to specifically target that segment of the economy. However, investment companies who want to offer such industry-specific funds are faced with the problem of choosing or defining a suitable index to track. Currently, this is predominately achieved by relying on standardized sets of sector or industry labels, such as those from the Global Industry Classification Standard (GICS). However, such labels are often not sufficiently fine-grained. For instance, while they allow us to define an Information Technology or Health Care index, they do not allow us to do the same for more specific domains such as Fintech or Artificial Intelligence (AI). Moreover, such labels are hard assignments, whereas the exposure of a given company to a domain such as AI tends to be a matter of degree. Finally, these labels do not allow us to quickly adapt to changes in the market (e.g. a major company deciding to create a high-profile AI-lab). Similar problems arise when we want to analyze stock market performance. While the existing categorization of companies allows us to analyze which sectors and geographic regions have performed well or poorly over a given time period, doing such analysis at a fine-grained industry level is currently not straightforward.

To address these limitations, in this paper we introduce a method for automatically developing vector representations of companies that can be useful for searching or categorizing companies at a fine-grained industry level. While there has been some previous work on predicting industry segments of companies [Chen *et al.*, 2018; Lamby and Isemann, 2018], in this paper we specifically focus on annual reports of companies as a source of information. Compared to the use of news stories [Lamby and Isemann, 2018], this has several advantages. First, the information captured in news stories can be heavily biased. For instance, a company such as Facebook is frequently mentioned in the news in relation to their AI research, whereas from an economic perspective, the performance of the AI sector might only be weakly correlated to that of Facebook. Moreover, representations learned from news stories can capture what companies have focused on in the past, but it is more difficult to capture changes in company strategy from such sources. In contrast, annual reports are authoritative documents, which explicitly describe the sectors in which the company is currently active.

In particular, we consider the problem of learning company embeddings from annual reports, such that the embedding of a given company characterizes the industries in which it is active. Learning such embeddings from annual reports is challenging, however, since only a small fragment of these reports is typically devoted to describing the industry in which

*Contact Author: m2015titoh@socsim.org

the company is active. This clearly differentiates our problem from the general problem of learning entity embeddings from descriptions [Jameel and Schockaert, 2016; Xie *et al.*, 2016; Wang *et al.*, 2019b], as approaches for the latter problem have focused on learning representations that reflect the entire description. To solve this challenge, we propose a method for fine-tuning a pre-trained neural language model [Devlin *et al.*, 2019]. Since we do not have access to annotations of fine-grained industry labels, we rely on a number of distant supervision signals, including the broad industry sector label of the company as well as its recent stock market performance. Our main contributions are as follows:

1. We introduce a new dataset[1] for the problem of characterizing industries from annual reports.
2. We propose several evaluation tasks for quantitatively evaluating the performance of these embeddings. On the one hand, these tasks aim to analyze to what extent companies with similar vector representations are active in similar industries. On the other hand, we also focus on a zero-shot learning setting, where the aim is to use company vectors to find companies that are active in a given industry, given only the name of that industry.
3. We introduce a model for learning company embeddings from annual reports, which is based on a multi-task learning set-up for fine-tuning the BERT language model [Devlin *et al.*, 2019].
4. We analyze the effectiveness of the considered distant supervision signals, as well as the general strengths and weaknesses of the learned embeddings.

2 Related Work

The general problem of learning vector space representations of entities has been widely studied in recent years. Such methods can be categorized based on the kind of information that is used. For instance, a large number of graph embedding methods have been proposed [Perozzi *et al.*, 2014; Tang *et al.*, 2015; Grover and Leskovec, 2016], which learn vector representations of entities based on their local neighbourhood in an associated graph, e.g. a social network of users or a citation graph of academic papers. As another example, there is a popular line of research which learns embeddings of knowledge graphs [Bordes *et al.*, 2013; Trouillon *et al.*, 2017; Balazevic *et al.*, 2019], which can for instance be useful to inject knowledge from such resources into Natural Language Processing (NLP) architectures. Most closely related to this paper, there have been several models that aim to combine entity descriptions with structured information, and knowledge graphs in particular [Wang *et al.*, 2014; Camacho-Collados *et al.*, 2016; Xie *et al.*, 2016; Wang *et al.*, 2019b]. One important difference with our setting is that these models assume that all the information in the given text descriptions is relevant, whereas our main challenge is to distill the relevant information for characterizing industries.

While there is considerable work on learning general entity representations, and recent work leveraging NLP techniques to model financial market dynamics [Xing *et al.*, 2018], the specific problem of learning company embeddings has only received limited attention thus far. Chen *et al.* [2018] introduced a model called Company2Vec, which learns company embeddings based on the intuition that companies are likely to be similar if employees tend to transition from one to the other. For our setting, this approach has two drawbacks. First, it relies on proprietary and sensitive personal data from the LinkedIn platform. Second, the corresponding notion of similarity is clearly skewed by factors such as geographic location. In [Lamby and Isemann, 2018], an analysis is carried out to assess to what extent industry sectors can be predicted from standard word embeddings, finding that such embeddings indeed capture a non-trivial amount of industry information. While achieving promising results, the method falls short when considering low-frequency companies and it does not consider ambiguity (e.g. the word *apple* can be a company but it also has other meanings).

3 Method

In this section we describe our method for learning vector representations of companies (i.e. company embeddings).

3.1 Fine-tuning BERT

Let x_i be a document about company i. In our experiments x_i will be the latest annual report of the company, although the same model could be used with other kinds of financial documents. The main strategy is to learn a mapping from such documents x_i to a corresponding company vector $\mathbf{h_i}$ by fine-tuning the BERT [Devlin *et al.*, 2019] language model.

Neural language models such as BERT are deep neural networks, which have been pre-trained in an unsupervised way on large amounts of text, typically by learning to predict masked words in sentences. Because of this pre-training process, they capture a large amount of knowledge about the meaning of words and phrases, and the typical syntactic structure of sentences. We can exploit this knowledge in applications, by fine-tuning a pre-trained language model on task-specific training data, rather than learning a neural network from scratch. This strategy has led to substantial performance gains across a wide range of Natural Language Processing (NLP) tasks [Wang *et al.*, 2019a]. Most closely related to our work, BERT has been shown to be effective for learning embeddings of entities from their description [Logeswaran *et al.*, 2019; Wang *et al.*, 2019b]. Formally, we have:

$$\boldsymbol{h}_i = \mathrm{BERT}(x_i)$$

where x_i is the textual description of company i and $\boldsymbol{h}_i \in \mathbb{R}^e$ is the resulting embedding. Specifically, $\mathrm{BERT}(x_i)$ refers to the mean vector of the token-level embeddings that are predicted by BERT. We also tried using the vector predicted for the [CLS] token, as is common in the literature [Wang *et al.*, 2019a], but found this to be less effective.

3.2 Multitask Learning

As we will see in the experiments, without fine-tuning the embeddings learned by BERT are not particularly useful for our setting. Among others, this is because annual statements

[1] Our preprocessed dataset and code is available at https://github.com/itomoki430/Company2Vec.

contain information beyond descriptions of the industry in which the company is operating. For this reason, we consider a multi-task learning set-up, allowing us to fine-tune BERT on three tasks: predicting the sector labels, capturing stock market performance and modelling sector names. The overall loss thus takes the following form:

$$\mathcal{L} = \mathcal{L}^{sec} + \mathcal{L}^{stock} + \mathcal{L}^{sn} \quad (1)$$

We now discuss each of these components in more detail.

Sector Category Loss

As a first supervision signal, we consider the task of predicting which sector of the economy company i belongs to. Note that while sector labels are coarser-grained that the industry segments which we want to model, the assumption is that by learning to extract sector-level information from the document x_i, the model will also extract finer-grained information. The advantage of using sector labels is that they are readily available. Let us write $\boldsymbol{d}_i^{sec}$ for the one-hot encoding of the sector of company i. We define $\mathcal{L}^{sec}$ as follows:

$$\boldsymbol{y}_i^{sec} = \text{Softmax}(\boldsymbol{W}^{sec}\boldsymbol{h}_i + \boldsymbol{b}^{sec})$$
$$\mathcal{L}^{sec} = \sum_{i \in \Omega} CE(\boldsymbol{d}_i^{sec}, \boldsymbol{y}_i^{sec})$$

where we write *CE* for the cross-entropy and Ω is the set of all considered companies.

Stock Performance Loss

Research has shown that companies from the same industry tend to exhibit similar stock price fluctuations [Gopikrishnan *et al.*, 2000]. Inspired by this finding, we also consider the following component in the loss function:

$$\mathcal{L}^{stock} = \sum_{i \in \Omega} \sum_{j \in \Omega} \|\boldsymbol{h}_i^T \cdot \boldsymbol{h}_j - Sim(\boldsymbol{v}_i^{stock}, \boldsymbol{v}_j^{stock})\|$$

where *Sim* is the cosine similarity and $\boldsymbol{v}_i^{stock}$ denotes a vector containing the monthly stock return value for company i, for the last five years, i.e.:

$$\boldsymbol{v}_i^{stock} := [r_i(t_1), r_i(t_2), \cdots, r_i(t_{60})],$$

with $r_i(t_j)$ the monthly stock return value for company i for j months ago. In case stock price data is not available for both companies over the full period, the longest period for which data is available is used instead.

Sector Name Loss

The third component of the loss function is aimed at fine-tuning BERT such that it maps the name of a given sector onto the correct index of that sector. Let us write $\mathcal{S}$ for the set of all sectors. For $j \in \mathcal{S}$, we write sn_j for the name of sector j (e.g. "Healthcare"). As before, we write $\boldsymbol{d}_j^{sn}$ for the corresponding one-hot encoding. Then we have:

$$\boldsymbol{s}_j = \text{BERT}(sn_j)$$
$$\boldsymbol{y}_j^{sn} = \text{Softmax}(\boldsymbol{W}^{sn}\boldsymbol{s}_j + \boldsymbol{b}^{sn})$$
$$\mathcal{L}^{sn} = \sum_{j \in \mathcal{S}} CE(\boldsymbol{d}_j^{sn}, \boldsymbol{y}_j^{sn})$$

The reason why we include this component is because we want to be able to use the trained model to identify companies that belong to a given industry given only a text description of that industry. For instance, if the input is *Artificial Intelligence* then the model should be able to predict what part of the vector space contains AI companies, despite not having seen any training examples of such companies. The loss $\mathcal{L}^{sn}$ encourages the model to make such predictions for sector names, where the assumption is that this ability will also transfer to descriptions of more specific industry segments.

4 Experimental Setting

To evaluate the performance of our method we propose two tasks for which we construct a dataset. In the following we describe the construction of the datasets and the experimental evaluation details.

4.1 Datasets

To test our methodology we constructed two datasets: one English-language dataset about the US stock market and one Japanese-language dataset about the Japanese stock market. The US dataset includes the financial annual reports, stock return data, and sector label data for 2,462 US companies (see below for details). The Japanese dataset includes the same information for 3,016 Japanese companies. We split the datasets into train, validation, and test fragments, containing respectively 1800, 262, and 400 companies for the US dataset, and 2200, 316, and 500 companies for the Japanese dataset. For companies in the test splits, stock return data and sector labels are not used during training.

Text Corpora. For the US dataset, we used the financial annual reports (i.e., Form 10-K documents) of listed companies in the US stock market, focusing in particular on those that were published in 2019. We were able to obtain 2,462 such reports from http://www.annualreports.com in September 1st, 2019. For the Japanese dataset, we used the financial annual reports of listed companies in the Tokyo Stock Exchange, focusing on those that were published in 2018 (which is the most recent year for which reports were available). These documents are written in Japanese. We were able to obtain 3,016 reports from https://github.com/chakki-works/CoARiJ. For these datasets we make use of the business description section that contain a summary of the activities of the company, and thus typically contains the most relevant information for learning the embeddings.

Stock Data. For the $\mathcal{L}^{stock}$ loss , we need monthly return data. For both datasets we used data from a period of five years. In particular, for the US companies, we used data for April 2014 to March 2019, while for the Japanese companies, we used data for April 2013 to March 2018.

Sector Labels. For the sector loss $\mathcal{L}^{sec}$ and sector name loss $\mathcal{L}^{sn}$, we utilized the sector labels provided by annualreports.com[2] in the case of the US dataset. For the Japanese companies, we used the 17 sector labels that were assigned by the Tokyo Stock Exchange (TSE)[3].

[2] http://www.annualreports.com/Browse?type=Industry

[3] https://www.jpx.co.jp/markets/statistics-equities/misc/01.html

4.2 Training

For the US companies, we used the BERT-base-uncased model[4] [Devlin *et al.*, 2019], whereas we used a Japanese BERT pre-trained model[5] for the Japanese companies. An important difference between these two models is that the English BERT model was trained on general purpose text (i.e. Wikipedia and the Books and Movie Corpus [Zhu *et al.*, 2015]), whereas the Japanese BERT model was trained on three million Japanese business news articles[6]. In both cases we utilized the first 512 tokens of the business description section in each report as textual data for the embedding. To adapt both models to the language that is used in the annual reports, we first fine-tuned them on our text corpus, using the standard masked word and next sentence prediction tasks [Devlin *et al.*, 2019]. After this step, we trained our model on the loss function (1) using the Adam optimizer [Kingma and Ba, 2015] for 30 epochs with early stopping.

4.3 Evaluation Tasks

We evaluated our method on two tasks, namely a related company extraction test and a theme-based extraction test.

Task 1: Related Company Extraction Test

The aim of this task is to assess to what extent companies with similar vectors are similar in terms of the industries to which they belong. To this end, for each company X, we first obtain the K most similar companies, in terms of the cosine similarity between their embeddings. Then we evaluate to what extent the categories to which these companies belong are the same as the category of X. Following the work in [Yu *et al.*, 2012], we used the Mean Average Precision at K (MAP@K) evaluation metric, where K = 5,10,50.

For the US companies, we use two types of categories, corresponding to the sector labels and the industry labels provided by annualreports.com. Out of the 11 sector labels, only 9 appeared in the test data. The industry labels are essentially a finer-grained version of the sector labels. In the test set, a total of 140 different industry labels appeared, all of which were used for this evaluation. For the Japanese companies, we used the TOPIX-17 sector labels and TOPIX-33 sector labels, as defined by TSE[7], as the categories. TOPIX-33 sector labels are a refinement of the TOPIX-17 sector labels. For example, companies of "ENERGY RESOURCE" sector in TOPIX-17 are divided into "Mining" or "Oil and Coal Products" in TOPIX-33. The US sector labels and TOPIX-17 labels are the same ones that were used for training, which clearly makes the task easier than if previously unseen categories were used. Therefore, we will also report results for configurations of our model in which only a small number of sector labels are used during training. This will allow us to analyze to what extent the model is able to capture categories which it has not seen during training.

[4]Available at https://github.com/huggingface/transformers

[5]Available at https://drive.google.com/drive/folders/1iDlmhGgJ54rkVBtZvgMlgbuNwtFQ50V-

[6]https://qiita.com/mkt3/items/3c1278339ff1bcc0187f

[7]https://www.jpx.co.jp/english/markets/indices/line-up/files/e_fac_13_sector.pdf

Task 2: Theme-Based Extraction Test

In this task, we evaluate to what extent our method is able to find companies that are relevant to a given theme, given only the name of that theme. As theme names, for the US dataset, we used the same 140 industry labels from Task 1. For the Japanese dataset, we used a finer-grained classification involving 274 themes, which we extracted from https://minkabu.jp/screening/theme. Note that while each US company has a unique industry label, companies in the Japanese dataset may belong to multiple themes. We believe the latter setting is more realistic, but we were not able to obtain a similar dataset for the US stock market. We again treat this problem as a ranking task. In particular, for each theme Y, we first determine the K most relevant companies, by comparing the company vectors to the vector that was predicted by our fine-tuned BERT model for the theme name Y.

4.4 Baselines

To our knowledge, there are no previous models that have specifically been proposed for learning company vectors from annual reports. As baselines, we thus use two standard document representation methods. First, we consider the bag-of-words representation of the annual report (BOW), using term frequency weights.[8] For Task 2, we similarly use a BOW representation of the theme descriptions. For both tasks, companies are ranked based on cosine similarity.

As a second baseline, we used the mean vector of the skip-gram Word2Vec word embedding (SG) [Mikolov *et al.*, 2013] that was trained on all financial documents. To learn this skip-gram embedding, we utilized the 200-dimensional word embedding vectors that were trained on the corpus of US annual reports and Japanese annual reports, respectively, using a window size of 5. For Task 2, Hirano *et al.* [2019] already proposed an approach based on word vectors for Japanese, which we use as an additional baseline. This baseline first searches for synonyms of each theme name, using both the similarity based on word embeddings and the similarity based on co-occurrence in annual reports. Then, it extracts the companies related to the theme using the frequency of the theme name, and each of its discovered synonyms, in each annual report. For this method, we rely on the same skip-gram embedding as for the SG baseline. We also tried the same method for English but could not obtain any meaningful results.

5 Results

In this section we present the results in Task 1 (i.e. Related Company Extraction) and Task 2 (i.e. Theme-Based Extraction) and a qualitative analysis of the results provided by our model.

5.1 Related Company Extraction

The results for Task 1 are shown in Table 1 for the US dataset and in Table 2 for the Japanese dataset. In addition to the results of our full model and the baselines, the tables contain an ablation analysis, showing results for configurations where some components were removed from the loss function. The

[8]To allow for a direct comparison, for the baselines we used the same 512 tokens as for the BERT-based methods.

	US (SECTOR)			US (INDUSTRY)		
	MAP@5	MAP@10	MAP@50	MAP@5	MAP@10	MAP@50
BOW	0.177	0.127	0.066	0.184	0.177	0.182
SG	0.216	0.167	0.084	0.179	0.174	0.173
BERT_{CLS}	0.115	0.083	0.041	0.152	0.144	0.143
BERT	0.324	0.270	0.152	0.243	0.242	0.238
BERT + Stock	0.471	0.419	0.242	0.325	0.328	0.338
BERT + Sector	0.569	0.544	0.501	0.313	0.324	0.349
BERT + Stock + Sector	0.590	0.567	0.509	0.328	0.337	0.365
BERT + Sector + Sector Name	**0.613**	**0.582**	**0.545**	0.331	0.337	0.369
BERT + Stock + Sector + Sector Name	**0.613**	0.578	0.530	**0.349**	**0.359**	**0.388**
BERT + Sect. (2 labels) + Sect. Name	0.459	0.412	0.260	0.290	0.288	0.294
BERT + Sect. (5 labels) + Sect. Name	0.540	0.499	0.389	0.326	0.330	0.350
BERT + Stock + Sect. (2 labels) + Sect. Name	0.485	0.435	0.259	0.322	0.327	0.337
BERT + Stock + Sect. (5 labels) + Sect. Name	0.531	0.487	0.379	0.319	0.327	0.349

Table 1: Results for Task 1 (Related company extraction) on the US dataset.

	JAPAN (TOPIX-17)			JAPAN (TOPIX-33)		
	MAP@5	MAP@10	MAP@50	MAP@5	MAP@10	MAP@50
BOW	0.368	0.302	0.220	0.295	0.243	0.188
SG	0.281	0.228	0.150	0.199	0.153	0.101
BERT_{CLS}	0.128	0.097	0.058	0.081	0.056	0.032
BERT	0.202	0.156	0.098	0.145	0.108	0.068
BERT + Stock	0.405	0.330	0.216	0.338	0.274	0.199
BERT + Sector	0.654	0.618	0.568	0.542	0.503	0.448
BERT + Stock + Sector	**0.675**	**0.636**	**0.577**	0.557	0.521	0.458
BERT + Sector + Sector Name	0.660	0.622	0.556	0.547	0.508	0.445
BERT + Stock + Sector + Sector Name	0.672	0.633	0.561	**0.576**	**0.534**	**0.464**
BERT + Sect. (2 labels) + Sect. Name	0.420	0.360	0.268	0.337	0.282	0.221
BERT + Sect. (5 labels) + Sect. Name	0.462	0.389	0.310	0.387	0.318	0.262
BERT + Stock + Sect. (2 labels) + Sect. Name	0.486	0.418	0.335	0.410	0.354	0.294
BERT + Stock + Sect. (5 labels) + Sect. Name	0.472	0.405	0.325	0.396	0.338	0.272

Table 2: Results for Task 1 (Related company extraction) on the Japanese dataset.

	US			JAPAN		
	MAP@5	MAP@10	MAP@50	MAP@5	MAP@10	MAP@50
BOW	0.165	0.172	0.189	0.116	0.099	0.088
SG	0.030	0.032	0.043	0.066	0.054	0.050
BERT_{CLS}	0.004	0.003	0.008	0.019	0.015	0.013
BERT	0.094	0.108	0.124	0.024	0.020	0.019
[Hirano *et al.*, 2019]	-	-	-	0.118	0.101	0.093
BERT + Stock	0.164	0.177	0.196	0.030	0.025	0.027
BERT + Sector	0.188	0.208	0.238	0.114	0.100	0.099
BERT + Stock + Sector	0.174	0.192	0.221	0.106	0.090	0.087
BERT + Sector + Sector Name	0.215	0.238	0.268	0.175	0.150	0.133
BERT + Stock + Sector + Sector Name	0.194	0.210	0.241	0.160	0.143	0.136
BERT + Sect. (2 labels) + Sect. Name	0.141	0.151	0.166	0.101	0.089	0.085
BERT + Sect. (5 labels) + Sect. Name	0.190	0.208	0.238	0.161	0.148	0.136
BERT + Stock + Sect. (2 labels) + Sect. Name	0.199	0.220	0.238	0.125	0.122	0.120
BERT + Stock + Sect. (5 labels) + Sect. Name	**0.234**	**0.254**	**0.279**	**0.176**	**0.161**	**0.144**

Table 3: Results for Task 2 (Theme-based extraction).

Company	Sector	Industry	Company	Sector	Industry
US Lime & Minerals	Industrial Goods	General Building Mater.	Whiting Petroleum	Basic Materials	Oil & Gas Drill. & Explr.
Freeport-McMoRan Copper&Gold	Basic Materials	Copper	Halcon Resources	Basic Materials	Oil & Gas Drill. & Explr.
United State Antimony	Basic Materials	Industrial Metals & Minerals	Callon Petroleum Company	Basic Materials	Independent Oil & Gas
Approach Resources	Basic Materials	Oil & Gas Drill. & Explr.	Cimarex Energy Co.	Basic Materials	Independent Oil & Gas
Xenia Hotels & Resorts	Financial	REIT - Hotel/Motel	Viking Therapeutics	Healthcare	Biotechnology
Ashford Hospitality Prime	Financial	REIT - Hotel/Motel	Adaptimmune Therapeutics	Healthcare	Biotechnology
LaSalle Hotel Properties	Financial	REIT - Hotel/Motel	Sage Therapeutics	Healthcare	Biotechnology
RLJ Lodging Trust	Financial	REIT - Hotel/Motel	Celldex Therapeutics	Healthcare	Biotechnology
Telephone & Data Systems	Technology	Wireless Comms.	Tandy Leather Factory	Consumer Goods	Textile-Apparel Footw.&Acc.
Verizon Communications	Technology	Telecom Services - Domestic	Steve Madden	Consumer Goods	Housewares & Accessories
Sprint Corp	Technology	Wireless Comms.	Vince Holdings	Consumer Goods	Textile - Apparel Clothing
U.S. Cellular	Technology	Telecom Services - Foreign	Vera Bradley	Consumer Goods	Textile - Apparel Footw. & Acc.

Table 4: Three nearest neighbours for selected companies in the test set in the vector space resulting from our full BERT multitask model.

full method is shown as BERT + Stock + Sector + Sector Name. On the last four rows, we furthermore show results for a more challenging setting where only 2 or 5 sector labels were used during training, instead of the full set of sector labels from the dataset (see Section 4.1).

As can be seen in Table 1, BERT already outperforms the BOW and SG baselines on the US dataset, even without incorporating any of the three supervision signals. For comparison, we also show results of BERT when using the [CLS] output vector instead of averaging the token-level vectors, which performs substantially worse. Incorporating stock performance and sector labels clearly helps, with further performance gains being achieved when incorporating the sector name loss. When only 2 or 5 sector labels are available for training, as expected, the performance drops. However, for the industry labels, the drop is surprisingly small, which shows that the model learns to identify which parts of the annual reports contain the most relevant information, rather than simply learning to predict particular sector labels. The Japanese results in Table 2 broadly follow a similar pattern, although a larger drop in performance is seen for the configurations in which only 2 or 5 sector labels are used during training. Moreover, the BOW and SG baselines are also stronger in this case, outperforming the BERT configuration.

5.2 Theme-Based Extraction

Table 3 summarizes the results for Task 2. This task is clearly more challenging than Task 1, especially considering the fine-grained nature of the considered themes, which is reflected in the overall scores. The BOW baseline performs surprisingly well on this task. In terms of our model, the sector name component of the loss function now clearly plays an important role, which is not surprising, given that this component specifically trains BERT to map category names onto the embedding space. Surprisingly, the variant where only 5 sectors are used during training actually leads to the best results for the US and Japan. This reflects the fact that learning a mapping from sector names to the embedding space is most important for this task; including fewer sector names allows the model to focus more on the segment name component.

5.3 Qualitative Analysis

To analyze the company embeddings qualitatively, Table 4 shows the nearest neighbours for selected companies from the US test set (for the full BERT multitask model). As can be observed, in some cases, the neighbors have the same sector and industry labels (*Xenia Hotels & Resorts* and *Viking Therapeutics*). The case of *Viking Therapeutics* provides an example where the industry segment captured by the embedding is finer-grained than the pre-defined industry labels, given that all neighbors are specifically concerned with therapeutics. Even in cases where the industry labels are different, the nearest neigbors are often meaningful. For instance, the neighbors of *Tandy Leather Factory* are all focused on products made with leather (i.e. shoes for *Steve Madden* and *Vince Holdings* and handbags for *Vera Bradley*). This shows the potential of our vectors for capturing themes that cut across the traditional classification of industry segments. In the case of *US Lime & Minerals*, the nearest neighbors belong to a different sector. However, *US Lime & Minerals* is clearly related to the Basic Materials sector, as they focus on the processing of limestone. This illustrates the potential benefit of vector representation in identifying borderline cases, or more generally, for estimating the degree to which a company is exposed to a given sector or industry segment.

6 Conclusion

This paper addresses the problem of learning company embeddings from annual reports, such that the embedding of a given company characterizes the industries in which it is active. To achieve this end, we introduce a multi-task learning strategy, which is based on fine-tuning the BERT language model on (i) existing sector labels and (ii) stock market performance. Experiments in a newly constructed dataset of US and Japanese companies (in English and Japanese language, respectively) demonstrated the usefulness of this strategy. The proposed distant supervision signals were effective to improve the performance in several tasks. Finally, given the flexibility of our multitask model framework, in future work, it would be interesting to incorporate other sources of business information, such as Price Earnings Ratio (PER) and Price Book-value Ratio (PBR). Similarly, it would be useful to analyze how the authoritative information that is contained in annual reports can be complemented with more informal sources, such as news stories and company websites.

Acknowledgements. Jose Camacho Collados and Steven Schockaert have been supported by ERC Starting Grant 637277. Tomoki Ito was Supported by JSPS KAKENHI Grant Number JP17J04768.

References

[Balazevic *et al.*, 2019] Ivana Balazevic, Carl Allen, and Timothy M. Hospedales. TuckER: Tensor factorization for knowledge graph completion. In *Proceedings of EMNLP*, pages 5184–5193, 2019.

[Bordes *et al.*, 2013] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *Proceedings of NIPS*, pages 2787–2795, 2013.

[Camacho-Collados *et al.*, 2016] José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64, 2016.

[Chen *et al.*, 2018] Xi Chen, Yiqun Liu, Liang Zhang, and Krishnaram Kenthapadi. How linkedin economic graph bonds information and product: applications in linkedin salary. In *Proceedings of SIGKDD*, pages 120–129, 2018.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[Gopikrishnan *et al.*, 2000] Parameswaran Gopikrishnan, Bernd Rosenow, Vasiliki Plerou, and H Eugene Stanley. Identifying business sectors from stock price fluctuations. *arXiv preprint cond-mat/0011145*, 2000.

[Grover and Leskovec, 2016] Aditya Grover and Jure Leskovec. Node2vec: Scalable feature learning for networks. In *Proceedings of SIGKDD*, pages 855–864, 2016.

[Hirano *et al.*, 2019] Masanori Hirano, Hiroki Sakaji, Shoko Kimura, Kiyoshi Izumi, Hiroyasu Matsushima, Shintaro Nagao, and Atsuo Kato. Related stocks selection with data collaboration using text mining. *Information*, 10, 2019.

[Jameel and Schockaert, 2016] Shoaib Jameel and Steven Schockaert. Entity embeddings with conceptual subspaces as a basis for plausible reasoning. In *Proceedings of ECAI*, pages 1353–1361, 2016.

[Kingma and Ba, 2015] D.P. Kingma and L.J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[Lamby and Isemann, 2018] Martin Lamby and Daniel Isemann. Classifying companies by industry using word embeddings. In *International Conference on Applications of Natural Language to Information Systems*, pages 377–388, 2018.

[Logeswaran *et al.*, 2019] Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. Zero-shot entity linking by reading entity descriptions. In *Proceedings of ACL*, pages 3449–3460, 2019.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[Perozzi *et al.*, 2014] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of SIGKDD*, pages 701–710, 2014.

[Tang *et al.*, 2015] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of WWW*, pages 1067–1077, 2015.

[Trouillon *et al.*, 2017] Théo Trouillon, Christopher R. Dance, Éric Gaussier, Johannes Welbl, Sebastian Riedel, and Guillaume Bouchard. Knowledge graph completion via complex tensor factorization. *J. Mach. Learn. Res.*, 18:130:1–130:38, 2017.

[Wang *et al.*, 2014] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph and text jointly embedding. In *Proceedings of EMNLP*, pages 1591–1601, 2014.

[Wang *et al.*, 2019a] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*, 2019.

[Wang *et al.*, 2019b] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhiyuan Liu, Juanzi Li, and Jian Tang. Kepler: A unified model for knowledge embedding and pre-trained language representation. *arXiv preprint arXiv:1911.06136*, 2019.

[Xie *et al.*, 2016] Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, and Maosong Sun. Representation learning of knowledge graphs with entity descriptions. In *Proceedings of AAAI*, 2016.

[Xing *et al.*, 2018] Frank Z Xing, Erik Cambria, and Roy E Welsch. Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50(1):49–73, 2018.

[Yu *et al.*, 2012] Kuifei Yu, Baoxian Zhang, Hengshu Zhu, Huanhuan Cao, and Jilei Tian. Towards personalized context-aware recommendation by mining context logs through topic models. In *Proceedings of PAKDD*, 2012.

[Zhu *et al.*, 2015] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 19–27, 2015.