

Master's Thesis

Learning to Explain Causal Rationale of
Stock Price Changes in Financial Reports

Ye Eun Chun

Department of Computer Science and Engineering

Graduate School of UNIST

2020

Learning to Explain Causal Rationale of Stock Price Changes in Financial Reports

Ye Eun Chun

Department of Computer Science and Engineering

Graduate School of UNIST

Learning to Explain Causal Rationale of Stock Price Changes in Financial Reports

A thesis submitted to the Graduate School of UNIST
in partial fulfillment of the
requirements for the degree of
Master of Science

Ye Eun Chun

06/17/2020

Approved by



Advisor

Kwang In Kim

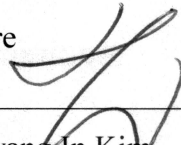
Learning to Explain Causal Rationale of Stock Price Changes in Financial Reports

Ye Eun Chun

This certifies that the thesis of Ye Eun Chun is approved.

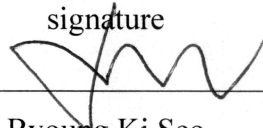
06/17/2020

signature



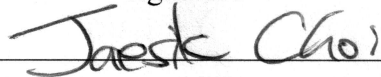
Advisor: Kwang In Kim

signature



Byoung Ki Seo

signature



Jaesik Choi

signature

Abstract

When a critical event occurs, it is often necessary to provide appropriate explanations. Previously, several theoretical and empirical foundations which discover causes and effects in temporal data have been established. However, for textual data, a simple causality modeling is not enough to handle variations in natural languages. To address the challenges in textual causality modeling, we annotate and create a large causality text dataset, called ‘Causal Rationale of Stock Price Changes’ (CR-SPC) to fine-tune pre-trained language models. Our dataset includes 283K sentences from the 10-K annual reports of the U.S. companies, and sentence-level labels, from which we observe diverse patterns of causality from each industrial sector for stock price changes. Because of this diversity and an imbalance in training data across sectors, BERT+fine-tune baseline on Sector-only data shows a biased performance. We propose to transfer from related sectors, implemented as a two-stage fine tuning framework. First-stage fine tuning transfers from related sector, to overcome the limited training resource, then the second stage follows to fine tune for the given sector. Our proposed framework yields significantly improved results for detecting causal rationale from industrial sectors with low amounts of data. Furthermore, we generate labels for 382K unlabeled sentences and augment the size of the dataset by self-training on CR-SPC dataset.

Contents

I. Introduction	1
II. Background	3
2.1 Bidirectional Encoder Representations from Transformers (BERT).....	3
2.2 Semi-supervised Learning	3
2.3 Local Interpretable Model-Agnostic Explanations (LIME)	4
III. Related Work.....	5
IV. Causal Rationale of Stock Price Changes Dataset (CR-SPC)	6
4.1 Management’s Discussion and Analysis (MD&A).....	6
4.2 Sentence Extraction	6
4.3 Data Annotation of Industrial Categories	7
4.4 Annotator Sensitivity	11
V. Methods to Train Causal Rationale.....	12
5.1 Two-stage Fine-tuning.....	12
5.2 Semi-supervision using Self-training.....	13
VI. Experimental Results	14
6.1 Experimental Settings	14
6.2 Few Shot Training.....	14
6.3 Baseline Models.....	15
6.4 Two-stage Fine-tuning.....	16
6.5 Semi-supervision.....	17
6.6 Turing Test.....	17
6.7 Interpretation of Causal Rationale Detection.....	20
VII. Discussions	233
VIII. Conclusion and Future Work.....	244
References	25
Acknowledgements	28

List of Figures

1	Overview of a two-stage fine-tuning framework.	12
2	Flow of self-training process.....	13
3	AUC and Average Precision of BERT Base by the number of training data from our CR-SPC dataset.....	15
4	Comparison between ‘Sector-only’, ‘1st Fine-tune’ and ‘2nd Fine-tune’ model.	16
5	Responses of participants in Turing Test.	18
6	Interpretation of predictions from (A) ‘Sector-only’, (B) ‘1st-fine-tune’ and (C) ‘2nd-fine-tune’ models with LIME.	21

List of Tables

1	Dataset (CR-SPC) Composition: Number of sentences and documents in each sector.	8
2	Examples of causal rationale sentences by each sector.	9
3	Examples of non-causal rationale sentences by each sector.	10
4	Classification performance of baseline models in supervised learning with our CR-SPC dataset.....	15
5	Test performance (AUC, AP) of Sector-only, 1st-Finetune, 2nd-Finetune, and Semi-supervised models on the test data of each sector.	16
6	Test performance of supervised-low learning (low-quality + high-quality data (CR-SPC)) and semi-supervised learning (pseudo labeled + high-quality data (CR-SPC)).....	17
7	Generated summary samples used in Turing Test.	19
8	Top 13 words contributing to causal sentences for each sector.	22

List of Abbreviations

ALBERT A Lite BERT . 1

AP Average Precision 2, 14, 15, 16, 17

Attn. Bi-LSTM Attentional Bidirectional Long Short-Term Memory. 15

Attn. LSTM Attentional Long Short-Term Memory. 15

AUC Area Under the ROC Curve. 14, 15, 16, 17, 233

BERT Bidirectional Encoder Representations from Transformers. 1, 2, 3, 5, 7, 14, 15, 16

Bi-LSTM Bidirectional LSTM 15

CR-SPC Causal Rationale of Stock Price Changes' 1, 2, 8, 12, 14, 15, 17, 18, 23, 244

EDGAR the Electronic Data Gathering, Analysis, and Retrieval system 6

LIME Local Interpretable Model-Agnostic Explanations 4, 20, 21, 23, 24

LSTM Long Short-Term Memory 14, 15

MD&A Management's Discussion and Analysis (MD&A). 6, 11, 17

NLM Neural language models 1, 12, 13, 14, 15, 18

SEC the Securities and Exchange Commission 1, 6

SIC Standard Industrial Classification (SIC) 7

Chapter I

Introduction

Many critical decisions on future events may require appropriate explanations of decisions based on accurate predictions. Justifying the predictive statements is directly related to identifying the temporal causes of the events. That is, when one could observe an event which is an apparent cause of a desired outcome, one can decide with confidence on future events.

There has been extensive research in extracting causes of events in numerical data. As an example, Granger cause finds linear temporal dependence between two (or more) temporal sequences (Bressler & Seth, 2011). Shapley values discover the contributions of individual input attributes when a decision is made by a complex function or system (Shapley, 1971). Such techniques can also be used to explain (numerical) causes of decisions made by automated systems, e.g. Robo-advisers in financial services (Hwang et al., 2016; Karuna, 2019; Lloyd et al., 2014).

However, human uses various types of information such as numerical and textual inputs when making an important decision. As an example, analysts write summarized reports by extracting important (causal) information from multiple textual sources such as conference calls, annual reports, earning statements and markets reports.

In this paper, we consider fine-tuning of pre-trained Neural language models (NLMs), which have been state-of-the-arts for many Natural Language Processing (NLP) tasks. For example, pre-trained NLMs such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) and A Lite BERT (ALBERT) (Lan et al., 2019) demonstrate outstanding performance in some tasks such as answering questions and computing conditional probabilities of a masked word in a sentence. However, NLMs require large training datasets to achieve human level performance. Currently, there is not enough data to detect causality from textual information.

Our first contribution is to collect sufficient annotations to achieve reasonable performance of BERT+fine-tune, which represents fine-tuning of pre-trained BERT Base model.

1) We collect 283K sentences from the 10-K annual reports of the U.S. companies maintained by the Securities and Exchange Commission (SEC) and manually label individual sentences, whether the sentences explain the cause of certain stock price changes (increase or decrease). We name the 283K pairs of a sentence and a corresponding label as ‘Causal Rationale of Stock Price Changes’ (CR-SPC). Our CR-SPC dataset is built on an unprecedented scale with guides of experts in the financial field. This dataset is useful to extract main causes of its financial events, from annual reports written officially from most U.S. public companies. Thus, individual investors can save efforts to read a huge amount of reports by themselves.

However, we find collecting the dataset alone does not solve the problem. One challenge we observe in the process of annotation is a **diverse causality**, that, we find diverse causalities over different industrial sectors. Another is an **imbalanced training**, where the number of data for each industrial sector varies, and so thus Average Precision (AP) of BERT+fine-tune on Sector-only data which means data of a single sector from the CR-SPC dataset. Thus, we need to build a model carefully, as applying a common model to all sectors does not work in our problem setting.

2) We propose a two-stage fine-tuning, where first stage aims to distill knowledge from related sectors, followed by the second stage fine tuning for the specific sector. Table 5 shows that, we can **overcome the lack of data** in a specific class (Sectors 2 and 5) with a two-stage fine-tuning. The proposed method yields 83.78% and 90.19% in AP for Sectors 2 and 5 respectively, which are significant increases compared to 63.82% and 66.05% in AP of BERT+fine-tune on Sector-only data.

3) Our last contribution is to augment the size of data to reduce the annotating costs and overcome **annotator sensitivity** by adding 382K pseudo labeled data. The quality of annotated labels varies over who annotators are, from the experts in financial fields to students who read an annual report on the first time. We overcome this issue by selecting matched labels with high-quality annotators. We compare the test performance between models for a specific sector, in which one is trained with pseudo labels and high-quality labeled data, and the other one is trained with low-quality and high-quality labeled data together. As a result, in the case of Sector 10, we achieve 87.27% in AP from the pseudo labels aggregated model, compared to 73.58% in AP from low-quality labels aggregated model.

Chapter II

Background

2.1 Bidirectional Encoder Representations from Transformers (BERT)

Traditional word embedding representation such as word2vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014) is generated by context-free models, therefore a word that has two different meanings still have the same representation in the traditional word embeddings. BERT is an unsupervised language representation, which is pre-trained on a large corpus, including the entire English Wikipedia and the book corpus (Y. Zhu et al., 2015). Given a sentence, random words are masked out and BERT looks words before and after the masked word to help predict what the word is. The bidirectionality helps BERT to understand the true meaning of a language. As a result, pre-trained BERT shows the state-of-the-art performances on many Natural Language Processing (NLP) tasks. Thus, for a specific NLP task, we only have to fine-tune the pre-trained BERT by adding just few additional output layers.

2.2 Semi-supervised Learning

Language modeling tasks require large training datasets to achieve human level performance. Typically, those datasets require human annotators for labeling, thus they are difficult and expensive to obtain. Semi-supervised learning has invaluable benefits in many neural language modeling tasks. It can utilize labeled and unlabeled data together to achieve better performance than using labeled data alone. In other words, we can replace some part of human annotation with unlabeled data. Therefore, semi-supervised learning reduces the annotation effort.

Semi-supervised learning has several different methods. Those different techniques are self-training, probabilistic generative models, co-training, graph-based models, semi-supervised support vector machines, and so on (X. Zhu & Goldberg, 2009). In this paper, we will focus on self-training as a semi-supervised learning technique.

Self-training is defined as the learning process that uses its own predictions to teach itself. Therefore, it is often called as self-teaching or bootstrapping. The major advantages of self-training are its simplicity and the fact that the choice of teacher model is open. However, if there exist early mistakes made by the initial teacher model, those mistakes are reinforced by generating incorrect labeled data repeatedly. To overcome this problem, various heuristics (e.g., adding noise) have been proposed.

2.3 Local Interpretable Model-Agnostic Explanations (LIME)

LIME is an explanation method that makes the predictions of a classifier interpretable by learning an interpretable model locally around the prediction (Ribeiro et al., 2016).

In this section, we describe the explanation system, LIME. Before we explain details about LIME, we define terms for future reference. In many NLP tasks, words or sentences are transcribed into vector representations. A vector is composed of a fixed size of numbers and it makes a machine understand human languages. However, those vector representations are not easy to understand for human users. Thus, LIME needs to explain classifiers in human understandable representations such as words, instead of lists of numbers. The term, x represents the original features (e.g., vector), whereas x' is a human understandable version of the original features (e.g., presence or absence of words in a sentence). When z' is given as a perturbed sample, we recover the sample back to the original representation z .

We first look into the way of obtaining explanation, then we will describe how to obtain elements in detail. Let g be an explanatory model and G be a class of possible interpretable models (e.g., linear models or decision trees). f stands for an original predictor model. $\Omega(g)$ is a measure of complexity of the explanation model where $g \in G$. $\pi_x(z)$ denotes as a proximity measure between instance z to x . In order to ensure interpretability and local fidelity of LIME, LIME produces an explanation by minimizing the following:

$$\xi(x) = \operatorname{argmin}_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \quad (2.3.1)$$

Since we want a model-agnostic explainer, the locality-aware loss $\mathcal{L}(f, g, \pi_x)$ is minimized without making any assumptions on f . Thus, we approximate the loss by drawing samples which are weighted by $\pi_x(z)$. Distance function D can be either cosine distance or L2 distance, where σ stands for width. The loss and weight are defined as follows:

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in Z} \pi_x(z) (f(z) - g(z'))^2 \quad (2.3.2)$$

$$\pi_x(z) = \exp(-D(x, z)^2 / \sigma^2) \quad (2.3.3)$$

Chapter III

Related Work

Rationale is a reason that causes a particular belief or phenomena. Extracting rationale is invaluable when decision has to be made. Research on extracting rationale from text has been tried with various types of text documents (Blanco et al., 2008; Girju, 2003; Ittoo & Bouma, 2011; Khoo et al., 2000). A model detecting and identifying rationale from chat messages was suggested in (Alkadhi et al., 2017). Bug reports from Chrome web browser were used as main sources to extract rationale (Rogers et al., 2012). Also, patent documents were utilized to discover design rationale (Liang et al., 2012). To extract causal textual structures, one may consider a rule based system where specific words such as ‘due to’, ‘owing to’ and ‘affects’ are listed to identify sentences including causal information for prediction (Chang & Choi, 2006; Girju et al., 2002; Sakai et al., 2015). Girju et al. (2002) used the inter-noun phrase causal relation to improve the question answering performance. To extract inter-noun phrase causal relations, they used the cue phrase filter. Chang and Choi (2006) used lexical patterns as a filter to find causality candidates and proposed cue phrase confidence score for a better causality extraction. However, such a rule-based system is vulnerable to lexical and syntactic variations of natural languages.

Pre-trained language models such as BERT have achieved the state-of-the-art performance in many NLP tasks. In addition, there has been many researches on distilling knowledge from pre-trained models for a specific task (Jiao et al., 2019; Sun et al., 2019; Tang et al., 2019).

Training a deep learning model by weakly annotated data is an active research area in both image (Papandreou et al., 2015) and natural language domain (Lin et al., 2012). Semi-supervised learning combining labeled and unlabeled data has been tried in text classification. A semi-supervised method has been proposed to learn embedding of small text areas in unlabeled data (Johnson & Zhang, 2015). In (Dai & Le, 2015), authors showed using unlabeled data from related tasks improved the generalization of a supervised model. Self-training is also widely used in text classification when only a small set of labeled data is available (Ko & Seo, 2009; Pavlinek & Podgorelec, 2017).

Chapter IV

Causal Rationale of Stock Price Changes Dataset (CR-SPC)

According to SEC, “the Electronic Data Gathering, Analysis, and Retrieval system (EDGAR) (Securities & Commission, n.d.), performs automated collection, validation, indexing, acceptance, and forwarding of submissions by companies and others who are required by law to file forms with the US Securities and Exchange Commission (SEC).”

Our goal is to collect sentences containing appropriate reasons of (possibly future) stock price changes. Future stock price is usually affected by the past performance and the future expectation of firms based on the history. Therefore, we assume that it is desirable to providing the textual causes of company's financial performance when predicting the future stock price.

4.1 Management's Discussion and Analysis (MD&A)

MD&A is located in item 7 at a 10-K report. The MD&A provides the company's perspective on its operations and financial results of the prior year. Therefore, this section is the primary source of information about what causes their financial results during the last year.

Thus, we focus on and extract the MD&A section from 10-K reports. We collect MD&A of reports filed in 1997 and 2017. MD&A of 1997 is gathered from the existing MD&A data repository (Kogan et al., 2009) and 2017's is downloaded directly from the SEC system.

4.2 Sentence Extraction

Our goal is to detect causal rationale from a given text in a sentence level. Thus, we need to annotate individual sentences for more than a thousand of reports. In order to reduce the cost of annotation, we extract possible causal sentences with causal expressions. Then, we annotate for individual industrial sectors.

Sentences that we want to collect should include causes and reasons of a certain financial performance. In the MD&A section, sentences explaining reasons often contain causal expressions like ‘due to’, ‘result in’, and ‘attributable to’. We assume that some causal sentences describe financial performance when they have performance-related words such as ‘increase’ and ‘decrease’. To extract almost possible explainable sentences inclusively, we use thirteen keywords and expressions; 1) result from, 2) result of, 3) increase, 4) contract, 5) because of, 6) decrease, 7) significant, 8) due to, 9) decline, 10) net sale, 11) caused, 12) negative and 13) impact.

4.3 Data Annotation of Industrial Categories

Different industries have their own fields of interests which affect changes in stock prices. Thus, we need to carefully select what the main reasons are among the extracted sentences. In the process of annotation, we find sentences containing specific causal factors for each sector. Those diverse causalities make ordinary BERT+fine-tune harder to learn causal rationales from individual sectors. In this section, how to divide annual reports into twelve categories, and unique causal factors we find for each sector is described.

The Standard Industrial Classification (SIC) classified the U.S. companies based on their primary business. We divide 10-K reports into twelve sub-categories which is predefined from (French, n.d.) with respect to the SIC codes. These twelve categories are 1) consumer non-durables, 2) consumer durables, 3) manufacturing, 4) energy, 5) chemicals, 6) business equipment, 7) telephone, 8) utilities, 9) shops, 10) health, 11) finance and 12) others. All sectors are numbered in this order.

Observation: Causal Diversity of Individual Sectors

The consumer non-durable sector includes food, tobacco, textiles, apparel, leather and toy companies. This business sector's revenue is mainly affected by **the fluctuation of materials costs**.

The consumer durable sector consists of cars, TV's, furniture and household appliance companies. This business sector's revenue is affected by **exchange rate** fluctuations as there are many companies export goods. **The sales volume of products** has a main impact on the profit.

The manufacturing sector is composed of machinery, trucks, planes, office furniture, paper, and printing companies. **The sales volume of products** is a main factor in this business sector. Many companies in this business sector sell their products to other companies, so **the accomplishments of contract** determine the sales volume.

The energy sector contains oil, gas producing, and coal extraction companies. **Oil and gas production** are important factors in explaining the main financial performance because it determines the volume of sales.

For the chemical business sector, **the raw material price and sale price** are main reasons affecting the change in financial performances.

The business equipment sector is composed of computer, software and electronic equipment companies. Research and development expenses are related to creating new products or services. In particular of high technology companies, they usually spend a lot of **expense in research and development**. Therefore, this should be considered as the main factor of the company's revenue.

The telephone sector contains telephone and television transmission companies. The Utilities sector contains gas and electric companies. Their revenues mainly depend on **subscription fee and usage**, so **the growth in the number of customers** is most important in these sectors.

The shop sector is composed of wholesale, retail and some service companies like laundries, repair. In this sector, **a change in the number of shops** is contributing the revenue.

The health sector consists of health care, medical equipment, and drug companies. Main factors of this sector are **research and development revenue**.

In the finance sector, **interest income or investment income**, which is not important in other sectors, is considered to be an important factor.

The last sector contains mines, construction, building maintenance, transportation, hotels, bus service and entertainment companies. Most of companies in this sector heavily depend on **service revenue, service related information like service income, service expense and contract** should be explained in this sector.

Explanation of changes in net income, net sales, operating income, revenue, and gross profit is typically common to all sectors. However, each business sector has its own main factor as described above. Therefore, labeling with different standards is desired. The number of sentences from each sector is shown in Table 1. Examples of sentence containing main factors are shown in Table 2. Examples of non-causal rationales are shown in Table 3.

	1	2	3	4	5	6	7	8	9	10	11	12	Total
Causes	1,145	346	1,528	497	283	2,505	721	490	1,424	563	558	1,072	11,132
Non-Causes	18,959	7,255	33,485	15,054	4,750	57,292	21,538	19,398	31,698	19,529	25,014	18,386	272,358
Total	20,104	7,601	35,013	15,551	5,033	59,797	22,259	19,888	33,122	20,092	25,572	19,458	283,490
# of Doc	140	61	248	79	36	379	76	55	138	127	119	126	1,584
Ratio of Causes (%)	5.7	4.55	4.36	3.2	5.62	4.19	3.24	2.46	4.3	2.8	2.18	5.51	3.93

Table 1. Dataset (CR-SPC) Composition: Number of sentences and documents in each sector.

Sector of Industry	Example [Document]
Consumer Non-Durables	The gross profit margin as a percentage of sales improved from 44.3% in Fiscal Year 1995 to 46.8% in Fiscal Year 1996, principally due to lower green coffee and material costs, and lower plant overhead costs. [Brothers Gourmet Coffees, Inc., July, 1997]
Consumer Durables	The sales increase for fiscal 1996 was principally due to improved sales of buses and ambulances. [Collins Industries, Inc., January, 1997]
Manufacturing	The increase in 1996 net sales was due primarily to increases in sales revenues recognized on the contracts to construct the first five Sealift ships, the Icebreaker and the forebodies for four double-hulled product tankers, which collectively accounted for 63% of the Company's 1996 net sales revenue. [Avondale Industries, Inc., March, 1997]
Energy	Gas revenue increased \$32.9 million or 81% because of a 39% price increase combined with a 30% increase in production. [Cross Timbers Oil Co., March, 1997]
Chemicals	Loss of margin was principally due to sales price decreases and raw material price increases in the pyridine and related businesses, and higher manufacturing costs due to weather related problems in the first quarter 1994. [Cambrex Corp., March, 1997]
Business Equipment	Research and development expenses increased from \$10.1 million (2% of net revenues) in fiscal 1995 to \$34.6 million (21% of net revenues) in fiscal 1996 due to the increase in Software development resulting from the acquisition of the three Software studios in calendar 1995. [Acclaim Entertainment, Inc., November, 1997]
Telephone	Revenue from cable television operations increased by \$90,713 or 24.6%, over the corresponding year ended May 31, 1996 as a result of regulated price increases, increases in the number of cable television subscribers and acquisitions. [Century Communications Corp., August, 1997]
Utilities	Gas operating revenues increased \$36.7 million, or 21.0%, due to increased volumes as a result of customer growth and higher gas costs. [WPS Resources Corp., March, 1997]
Shops	Aggregate sales generated by franchised stores increased by \$10,001,000, or 13.1%, to \$86,485,000 for calendar year ended December 31, 1995, as compared to \$76,484,000 for the same period in 1994, due to an increase of the number of franchised stores, as well as higher sales volume per store. [Sterling Vision, Inc., April, 1997]
Health	The increase in research and development expenses in 1996 and 1995 was due primarily to higher expenditures for the Actiq Cancer Pain Program, new product development and other expenditures for product development, including clinical trials. [Anesta Corp., March, 1997]
Finance	Mortgage investment income decreased for 1995 as compared to 1994 primarily due to the assignment to HUD of the mortgage on El Lago Apartments in June 1995. [American Insured Mortgage Investors Series 85 L P, March, 1997]
Others	The Company's largest revenue source is from the marketing and administration of extended vehicle service contracts ("VSCs") under the EasyCare(R) name, which provided 99% of revenues for 1996. [Automobile Protection Corp-APCO, March, 1997]

Table 2. Examples of causal rationale sentences by each sector.

Sector of Industry	Example [Document]
Consumer Non-Durables	Total general and administrative costs decreased by \$79,000 in 1995 due primarily to the absence of a management fee for 1995. [Highwater Ethanol, LLC, January, 2017]
Consumer Durables	Bank borrowings during 1995 were attributable to the Silver Furniture acquisition and the refinancing of Silver Furniture's bank indebtedness. [Chromcraft Revington, Inc., March, 1997]
Manufacturing	Because components are sold directly to the Company's manufacturing sources, the Company is not aware of the precise quantities sourced from particular suppliers. [Fossil, Inc., March, 1997]
Energy	Due to the apparent age of the material, no fine or enforcement action is expected. [Arabian Shield Development Co, March, 1997]
Chemicals	Due to personnel additions to the department, employee wages increased approximately \$56,700 in 1996. [American Vanguard Corp., March, 1997]
Business Equipment	Due to a variety of factors including differences in relational database product performance across wide area networks, differences in speed of various communication links, differences in hardware platform performance, and other factors, there is a limited ability to accurately predict product performance under certain of these environments. [Peoplesoft, Inc., March, 1997]
Telephone	Cost of services related to the wireless telephone operations during the year ended May 31, 1996 was \$26,129, an increase of \$3,977 or 18.0% as compared to the year ended May 31, 1995. [Century Communications Corp., August, 1997]
Utilities	The remainder of the increase was attributable to increases in ad valorem taxes, repair and maintenance expense mainly related to the WCLSF and the employee incentive plan which rewards certain of Tejas' employees with bonuses when the company achieves certain annual financial growth targets. [Tejas Gas Corp., March, 1997]
Shops	The Board may increase or decrease the number of shares under the Program or terminate the Program in its discretion at any time. [Boise Cascade Co., February, 2017]
Health	The 1995 results were also negatively impacted by a reduction of the Company's income tax benefit resulting from reserves established related to the expiration of certain state operating losses. [American White Cross Inc., April, 1997]
Finance	In the last three years, inflation has not had a significant impact on the Company because of the relatively low inflation rate. [Weeks Corp., March, 1997]
Others	In addition, the timing of revenue is difficult to forecast because the Company's sales cycle is relatively long. [Claremont Technology Group Inc., September, 1997]

Table 3. Examples of non-causal rationale sentences by each sector.

4.4 Annotator Sensitivity

Inspired by Kappa (Cohen, 1960), we use agreement as an indicator of annotator quality and keep cross-annotator agreement high to ensure annotation quality. We use a set of expert annotators and selectively keep annotations from annotators correlating with their judgement.

The classification performance, highly depends on the quality of a dataset we collected. If labeling is not consistent over the dataset, automatically trained models (e.g., deep neural network models) will suffer non-stable training. Since we collect annotations from experts and non-experts, there are some discrepancy between labels from annotators. In order to handle this issue, we select labels from a set of expert annotators. We regard these labels as a standard and train a model with these labels alone. We call this model as an initial teacher model. Then, we apply this teacher model to documents that are labeled from the other annotators. If all labels in a single MD&A document match with predictions of the teacher model, we add them to a training set. We build another teacher model with newly added training set again and apply this model to the rest of other documents. We repeat this process until no matched document is found. As a result, we collect 1,584 10-K reports and 283,490 sentences. Among sentences we collect, 11,132 sentences contain causes of financial performance and 272,358 sentences do not contain causes.

Chapter V

Methods to Train Causal Rationale

As described in the Section 4.3, CR-SPC dataset consists of twelve different industrial sectors. Each industrial sector possesses its own field of interests with respect to reasons of stock price changes. Thus, we need to build a model that predicts appropriate causal rationales from each industrial sector.

In this paper, we have following two research questions: (1) how can we utilize a set of dataset which belongs to other classes in order to accurately predict causal rationale for a certain class, (2) how can we exploit unlabeled data to improve the performance of Causal NLMs.

5.1 Two-stage Fine-tuning

This section tests a hypothesis if two-stage fine-tuning helps to predict different causal rationales of stock price changes for each sector better than using Sector-only data alone.

We observe that different industry sectors contain different causal rationales of stock price changes. Therefore, training each sector individually is required for a better classification. However, the number of data for each industry sector is between 5K and 60K, in which some sectors do not contain enough data to detect causal rationale precisely.

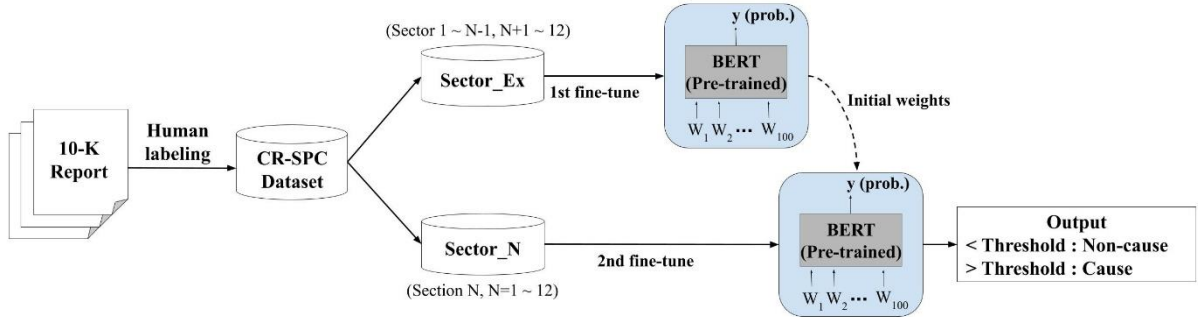


Figure 1. Overview of a two-stage fine-tuning framework.

Thus, we propose a two-stage fine-tuning network for a better prediction. In a two-stage fine-tuning network, we first fine-tune a pre-trained NLM with all sector data except for a class that we want to train eventually. Then we secondly fine-tune the NLM with a certain sector data. In this way, we can overcome the lack of data and make NLMs to learn global features and domain features together effectively. Figure 1 shows the overview of our two-stage fine-tuning framework.

5.2 Semi-Supervision using Self-Training

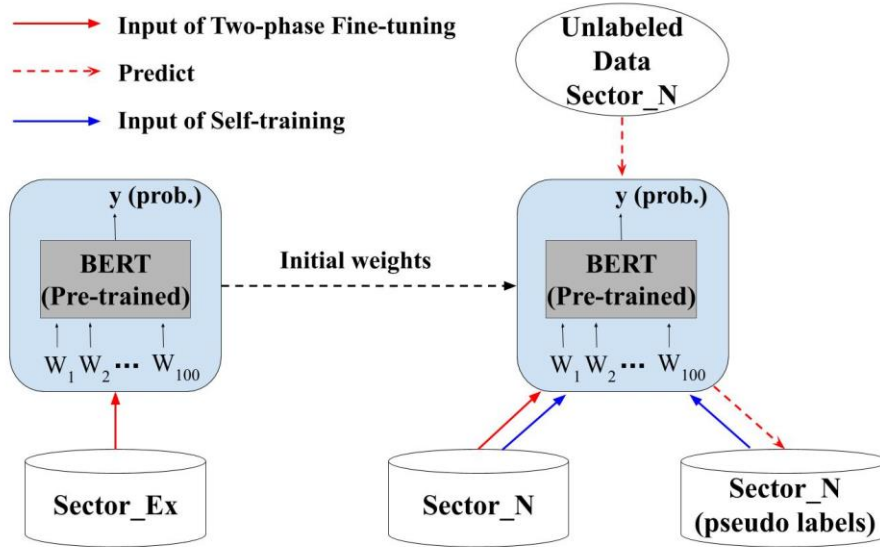


Figure 2. Flow of self-training process.

This section tests a hypothesis if pseudo-labeling unlabeled contributes to learning, by adding 382K to our 283K dataset. Collecting labeled data is expensive and time consuming. In particular, to annotate financial reports, annotators need to have the domain knowledge of financial fields. This makes annotating process harder and expensive compared to other task annotation. Thus, we present an auto-annotation method based on NLMs trained with self-training.

We hypothesize that additional unlabeled data may improve the performance of causal extraction of NLMs when the pseudo labels are carefully selected. To examine our hypothesis, we suggest a self-training network.

Our self-training network consists of three steps: 1) train a NLM with a two-stage fine-tuning on labeled sentences as described in Section 5.1, 2) generate pseudo labels on unlabeled sentences of a certain industry sector by using the NLM as a teacher model, and 3) train a student NLM that is initialized with weights from 1st fine-tuning on the combination of labeled sentences and pseudo labeled sentences.

In order to proceed with the self-training network, we need pseudo labels for unlabeled data. Given a Causal NLM trained with labeled data, predictions on unlabeled data is made. At this time, we regard this trained model as a teacher model. The teacher model gives us the probability of causality on each sentence. We use this probability of each sentence as a pseudo label for the unlabeled sentence. As described in Section 4.4, collecting additional labels at the same time as keeping cross-annotator high is hard and expensive. Thus, we can collect more consistent data by applying self-training on unlabeled data. Figure 2 shows the flow of the self-training process.

Chapter VI

Experimental Results

This section first describes the details of our experimental setting. Then we address the following research questions:

- 1) Is our CR-SPC dataset sufficient to achieve reasonable performance of Causal NLMs? How many annotations do we need to train Causal NLMs stable?
- 2) Which Causal NLMs work on the CR-SPC dataset best?
- 3) Does our two-stage fine-tuning framework outperform ‘Sector-only’ and ‘1st fine-tuning’ models on extracting causal sentences from a specific industrial sector?
- 4) Do additional pseudo labels work better than low quality labels from non-expert annotators?
- 5) Do people perceive generated summaries from Causal NLMs as ones written by human?
- 6) How does our proposed model predict sentences differently compared to other models? Does the two-stage fine-tuning framework have advantages of using it over other models?

6.1 Experimental Settings

We conduct experiments on our CR-SPC dataset consisting of sentences and corresponding labels in supervised learning, and the combination of labeled 283K and unlabeled 382K sentences in semi-supervised learning. In supervised learning, we split our 283K dataset into train, validation and test sets with the ratio of 81% / 9% / 10%. For training of individual sectors, the same ratio of train/validation/test sets is applied to each sector. We use Area Under the ROC Curve (AUC) and Average Precision (AP) of causal sentences as the evaluation metrics. As an input, a single sentence is tokenized at the length of 100, with Keras tokenizer (Chollet & others, 2015) and BERT tokenizer (Devlin et al., 2018) respectively for Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) based models and BERT Base. For LSTM based models, we use Glove for word embedding. All models are optimized for the validation dataset. A trained model gives us the probability of causality on each sentence. We set a threshold as 0.4 because the majority of models we used show the highest F1-score at this point.

6.2 Few Shot training

We use BERT Base for fine-tuning on our CR-SPC data. We increase the amount of training data and verify the causal extracting performance on each step. At each step, we randomly select a set of training data by increasing 10,000 sentences, then report the mean and the standard deviation of AUC and AP from 10-fold cross validation of the rest of the CR-SPC dataset.

AP and AUC of BERT Base increased with huge gaps (13.67% and 2.16%, respectively) from 10K to 20K. Then they increase gradually until 200K then slightly drop afterward. At 200K, we observe 83.29% and 98.97% for AP and AUC respectively (Figure 3).

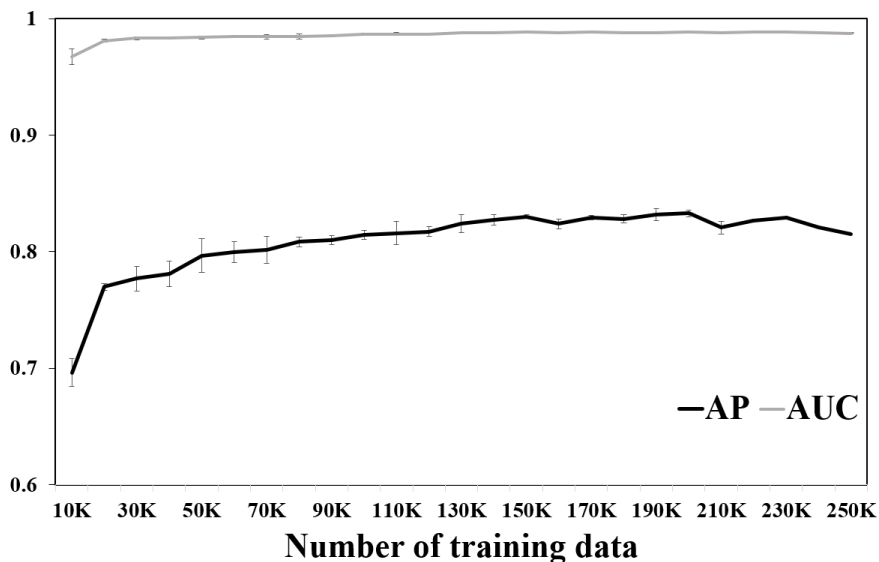


Figure 3. AUC (gray) and Average Precision (black) of BERT Base by the number of training data from our CR-SPC dataset.

6.3 Baseline Models

We use LSTM, Bidirectional LSTM (Bi-LSTM) (Graves et al., 2005), attentional LSTM (Attn. LSTM) (Zhou et al., 2016), attentional Bi-LSTM (Attn. Bi-LSTM), BERT Base as baselines to compare the performance of extracting rationale of stock price changes. For the baseline experiment, we only use CR-SPC dataset for training NLMs.

As shown in Table 4, models with an attention layer improved in AUC and AP compared to the models with no attention layer. BERT Base achieved the highest AUC (98.61%) and AP score (80.92%).

Model	AUC (%)	AP (%)
LSTM	98.31 ± 0.06	77.59 ± 0.66
Bi-LSTM	98.35 ± 0.08	77.92 ± 0.81
Attn. LSTM	98.53 ± 0.03	79.07 ± 0.36
Attn. Bi-LSTM	98.55 ± 0.03	79.48 ± 0.46
BERT Base	98.61 ± 0.02	80.92 ± 0.17

Table 4. Classification performance of baseline models in supervised learning with our CR-SPC dataset.

6.4 Two-stage Fine-tuning

We compare the test performances (AUC and AP) of three different models for each sector. Sector-only models are fine-tuned with BERT Base models on sector data only. 1st fine-tune models are BERT Base models fine-tuned on all sector data except a specific sector. 2nd fine-tuned models are secondly fine-tuned on the 1st fine-tuned models with the specific sector data. We use three fine-tuning layers for all models. The batch sizes are fixed at 32. We report the mean of AUC and AP from 10-K cross validations for each sector.

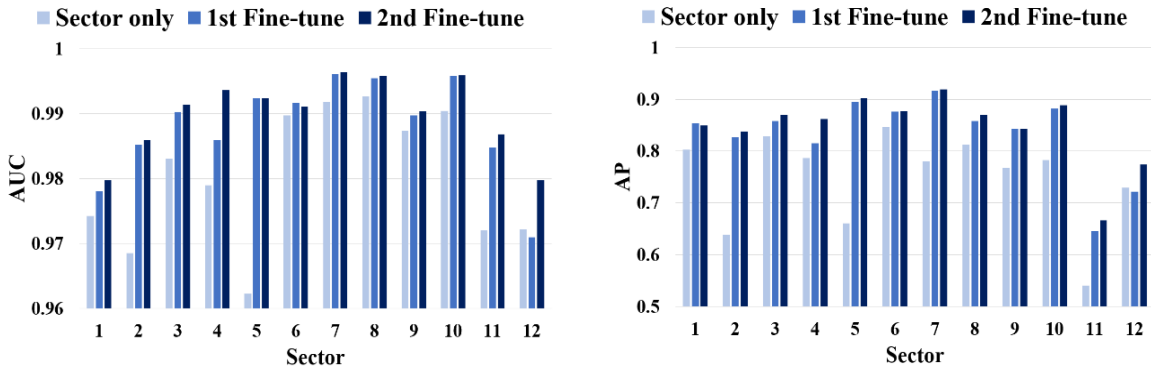


Figure 4. Comparison between ‘Sector-only’, ‘1st Fine-tune’ and ‘2nd Fine-tune’ model (left: Area Under the ROC Curve (AUC), right: Average Precision (AP)).

Section	Sector-only		1st-Finetune		2nd-Finetune		Semi-supervised	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP
1	97.42	80.30	97.80	85.32	97.97	85.00	98.16	84.97
2	96.85	63.82	98.52	82.64	98.59	83.78	98.50	84.08
3	98.31	82.84	99.03	85.79	99.14	86.97	98.97	86.48
4	97.90	78.69	98.59	81.53	99.36	86.16	99.02	81.92
5	96.22	66.05	99.24	89.51	99.24	90.19	99.14	89.15
6	98.97	84.69	99.16	87.61	99.11	87.69	99.14	87.76
7	99.18	77.98	99.61	91.65	99.64	91.94	99.52	91.70
8	99.26	81.20	99.54	85.84	99.58	87.00	99.64	88.95
9	98.74	76.77	98.98	84.26	99.03	84.31	99.09	84.39
10	99.04	78.30	99.59	88.20	99.59	88.88	99.57	87.27
11	97.20	54.00	98.47	64.50	98.68	66.62	98.84	66.89
12	97.22	72.92	97.10	72.17	97.97	77.45	98.00	78.33

Table 5. Test performance (AUC, AP) of Sector-only, 1st-Finetune, 2nd-Finetune, and Semi-supervised models on the test data of each sector (highest score in bold) (%).

As shown in Table 5 and Figure 4, all 2nd fine-tune models achieved improved AUC and AP scores compared to Sector-only models. The differences between 2nd fine-tune models and Sector-only models are significantly higher at Sectors 2 and 5 with 19.96% and 24.14% in AP, respectively. Sectors 1 and 6 show slightly decreased performances of 2nd fine-tune compared to 1st fine-tune. In Sector 12, 1st fine-tune models achieve the lowest performance of both AUC and AP compared to the other models

for the same sector. In addition, we observe that the micro average over the performance of 2nd fine-tune models is higher than 1st-fine-tune models with increases of .19% and 1.4% in AUC and AP, respectively. We conduct one-tailed t-tests to determine the statistical significance of the difference in performances of AP ($p < .05$) and AUC ($p < .01$).

6.5 Semi-supervision

For a semi-supervised learning, we use 2nd fine-tuned models as the teacher models to produce pseudo labels for unlabeled sentences of objective sector. Then, we combine labeled data and pseudo labeled data of the sector to fine-tune 1st fine-tuned models for a specific sector. To confirm the quality of pseudo labels, we train 1st fine-tuned model on the low-quality labeled data combined with high-quality labeled data (CR-SPC) of the sector, and we call this model as a supervised-low model. Then we compare the performances of two models. As described in Section 4.4, low quality data is annotations not passed the agreement with expert annotations. The low-quality dataset consists of 324,315 sentences and 1,403 reports.

Test performances of semi-supervised models show improved performances in AUC and AP for Sectors 4, 5, 10, 11 and 12 compared to supervised-low models (Table 6). In particular of Sector 10, the difference of AP (87.27%) in Semi-supervised and AP (73.58%) in supervised-low is huge.

Sector	Supervised-low		Semi-supervised	
	AUC (%)	AP (%)	AUC (%)	AP (%)
4	98.79	79.55	99.02	81.91
5	98.91	85.93	99.14	89.15
10	99.04	73.58	99.57	87.27
11	98.81	62.93	98.84	66.89
12	97.72	76.29	98.00	78.33

Table 6. AUC and AP of supervised-low learning (low-quality + high-quality data (CR-SPC)) and semi-supervised learning (pseudo labeled + high-quality data (CR-SPC)).

6.6 Turing Test

We generate automated summaries from unlabeled MD&A of 2017. The first two lines consist of the rule-based sentences containing companies' basic information. The latter part of the summaries consists of sentences that are classified as rationale of changes in stock prices from attentional Bi-LSTM trained on CR-SPC dataset. Those sentences are extracted from a document and listed as the order of appearance in the document.

We prepare three financial summaries written by analysts which are published at J.P. Morgan and twenty-five summaries generated by our program in similar length for both. Each participant reads five

summaries and decides the writer of each summary. Five summaries consist of either a set of two summaries written by analysts and three summaries generated by our model or vice versa. With 30 students who participated in this experiment, we collected 150 responses. Among 150 responses, 75 were generated by our model and the other 75 were written by analysts. Examples of summary generated from our program are shown in Table 7.

From the Turing Test, out of 75 summaries by our model, 24 were answered as analyst-written and 51 were perceived as software-generated. Among 75 summaries by analysts, 25 were thought as software-generated and 50 were perceived as analyst-written. That is, 32% of individual summaries generated from our NLM trained on CR-SPC dataset are perceived as human written. The responses of participants in Turing Test are presented in Figure 5.

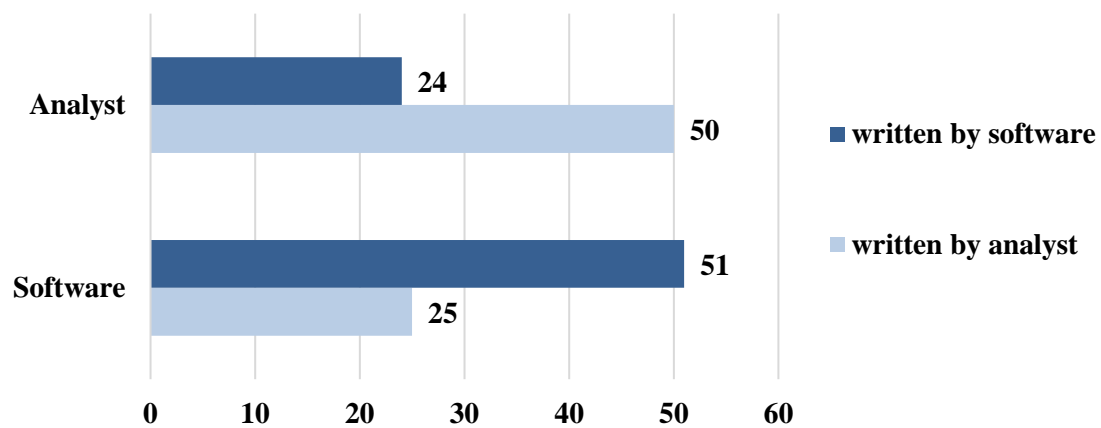


Figure 5. Responses of participants in Turing Test.

Summary ID	Summary [Company name, source, year]
1	<p>The results of 2017 are reported by New Age Beverages Corp. Changes in financial performance during the last year are described as follows: Their revenue for the period is primarily attributed to their acquisition of the Xing brands and the related increase in demand for Xing products, as well as expanded distribution on the Búcha Live Kombucha brand. The increase in gross margin was due to several factors, including (1) an increase in gross sales, (2) reduced freight costs and manufacturing labor, and (3) improved raw material and packaging supply costs including gaining the benefits of increased scale. The increase in the gross margin was due to several factors, including (1) a significant increase in gross and net sales, (2) significantly increased scale and efficiencies that led to lower freight costs and transportation costs, and (3) an improvement in the production processes of some of their key products that led to lower overall manufacturing costs. [New Age Beverages Corp, 10-K, 2017]</p>
2	<p>YIELD10 BIOSCIENCE, INC. posted the results of 2017. The significant changes in financial performance are explained in the following: The Company's technology sales, services and licensing revenues increased 5% in 2016, as compared to 2015, as a result of a strengthening in sales of the Company's digital authentication solution, partially offset by a decrease in the Company's IT hardware reselling business that resulted from the Company's decreased focus on this component of its digital business. Costs of revenue increased 4% in 2016 as compared to 2015 which was less than the 10% increase in the Company's revenue over the same period, which generally reflected the increase in sales of products that have a higher margin, such as security sales and technology card sales such that material costs, outside service costs and delivery costs decreased as a percentage of revenue during the 2016 period. Stock-based compensation costs decreased 66% in 2016 as compared to 2015 due to a general decrease in the number and value of equity compensation awards granted by the Company since 2014. [YIELD10 BIOSCIENCE, INC, 10-K, 2017]</p>
3	<p>Luvu Brands, Inc. announced 2017 results. Their financial record was affected by several reasons such as: The decrease in sales through the Wholesale channel was due to lower sales of Liberator products to retailers, offset in part by greater sales of Liberator, Jaxx and Avana products through and to Amazon. The improvement in gross profit was primarily due to greater sales of manufactured consumer products which have a higher average gross profit margin than products purchased for resale, and production improvements implemented this year which increased productivity and reduced cost of goods sold. Other income (expense) increased 16% from the prior year due to higher average borrowing balances and higher interest expense on those larger balances. [Luvu Brands, Inc., 10-K, 2017]</p>

Table 7. Generated summary samples used in Turing Test (extracted sentences are in bold).

6.7 Interpretation of Causal Rationale Detection

LIME is the abbreviation for Local Interpretable Model-Agnostic Explanations. With this explanation technique, we can visualize the most important features affecting predictions from any models. In order to see the difference on the predicted sentences from various models, we applied LIME on our models. We set the explainer to select top 13 contributing features. The results from LIME is visualized in Figure 6. In Figure 6, true labels for all sentences are all causal sentences.

In the case of a sentence from Sector 2, predicted answer from Sector-only model is incorrect. Predictions from the 1st-fine-tune model and 2nd-fine-tune model are correct answers which are improved by ‘sales’ and ‘increase’.

From the result of Sector 11, the word, ‘interest’ works as non-causal in Sector-only model. However, it disappears in the 1st and 2nd fine-tune models. The word ‘level’ becomes an opposite contributing factor in the 2nd fine-tune model. The highest probability of causal rationale is achieved by 2nd fine-tune model.

For a causal sentence from Sector 4, words including ‘oilfield’, ‘primarily’, ‘attributable’ and ‘whose’ become weaker contributing factors in the 2nd fine-tune model.

The probability of a causal sentence from Sector 5 is predicted as highest with the 2nd-fine-tune model, compared to Sector-only and 1st-fine-tune models. From 2nd-fine-tune's prediction, we observe that the word, ‘higher’ becomes a contributing factor, and the color gradient becomes darker on ‘sales’. On the other hand, ‘volume’ disappears in the 1st and 2nd fine-tune models. All probabilities on the sentence are above the threshold (0.4), so they are all classified as a causal sentence.

We also apply LIME to 2nd-fine-tuned models of each sector to test how each sector's model predicts differently when classifying causal sentences. First, we apply 2nd-fine-tuned models of each sector used in Chapter 6.4 to the test dataset which are used for baseline models, so that we can test on sentences from every sector. Then, we use LIME to analyze the top 13 words contributing to causal sentences for each model. In order to rank the words contributing to causal rationales for each sector, we are stemming all words appeared in 500 sentences of test dataset, and then, we sum the weights of each stem word. If the sum weight of a certain word is higher than others, it means the word is more contributing to the decision of causal rationales. Also, we remove the verb ‘be’ and prepositions from the rank to obtain meaningful words.

The results of comparing 13 words contributing to causal rationales for each sector are listed in Table 8. Words with the higher sum of weights are located at the top. We find that the words 'increase' and 'due' are selected as the most contributing words to causal rationales in all sectors. In addition, words such as 'revenue', 'income', 'primarily', 'lower' and 'attributable' are found in all sectors but their orders in the rank are slightly different each other. Moreover, words such as 'higher', 'cost', 'profit',

'promote' and 'margin' can be found in a set of specific sectors. According to those results, we find that each model classifies given sentences differently.

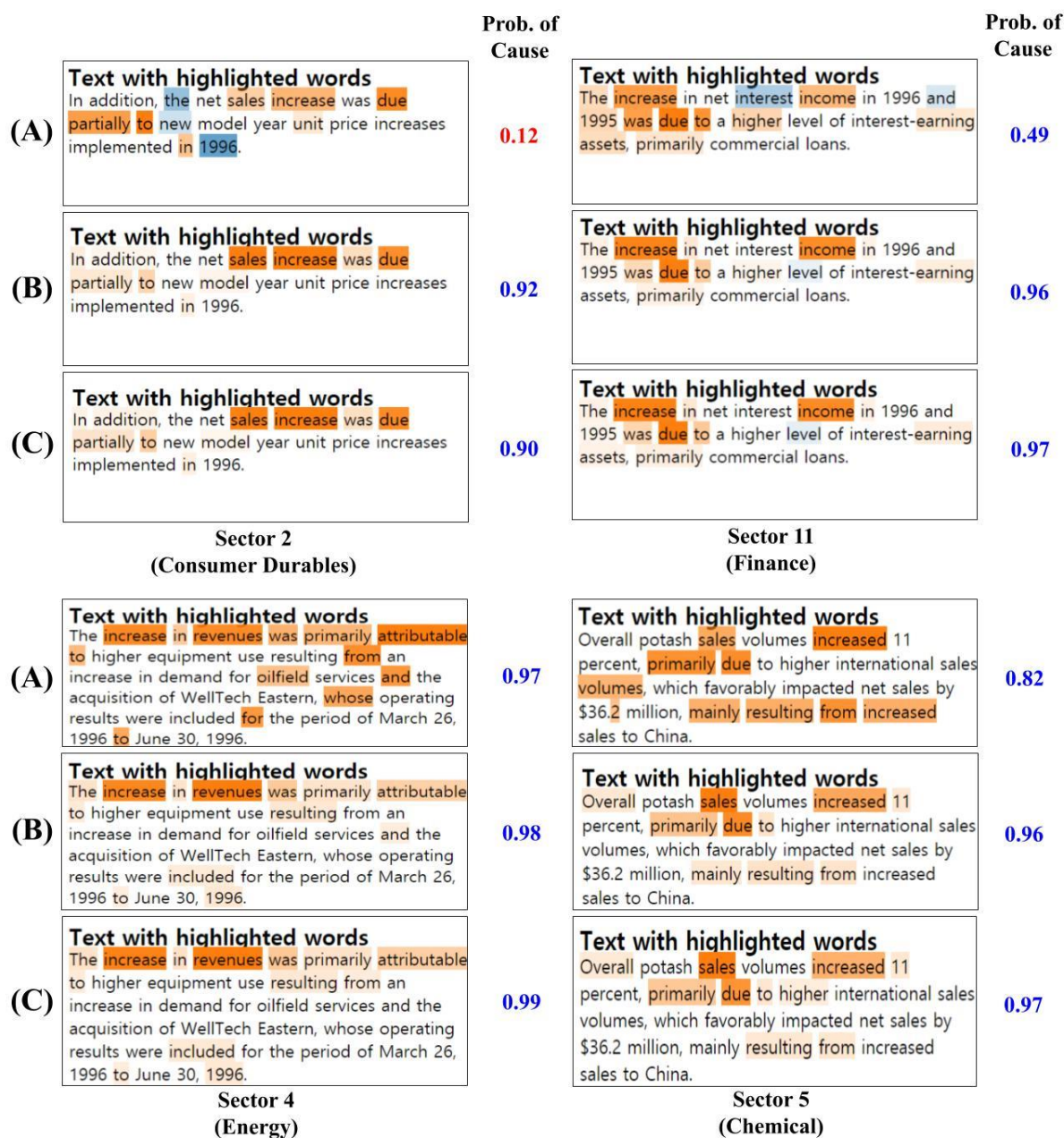


Figure 6. Interpretation of predictions from (A) 'Sector-only', (B) '1st-fine-tune' and (C) '2nd-fine-tune' models with LIME. (Features contributing on causal rationale are highlighted in orange, the opposites are in blue.)

Sector					
1	2	3	4	5	6
increas	increas	increas	increas	increas	increas
due	due	due	due	due	due
revenu	revenu	revenu	revenu	revenu	revenu
primarili	incom	incom	incom	sale	incom
incom	primarili	sale	result	primarili	primarili
sale	sale	result	primarili	result	attribut
result	result	primarili	sale	incom	result
decreas	lower	lower	attribut	attribut	sale
attribut	attribut	attribut	lower	decreas	higher
lower	decreas	decreas	decreas	lower	lower
cost	profit	declin	declin	higher	profit
higher	declin	profit	profit	declin	decreas
7	8	9	10	11	12
increas	increas	increas	increas	increas	increas
due	due	due	due	due	due
revenu	revenu	revenu	revenu	incom	revenu
incom	incom	incom	incom	revenu	sale
result	primarili	primarili	primarili	primarili	primarili
primarili	result	result	sale	result	result
sale	sale	sale	result	sale	attribut
attribut	attribut	lower	attribut	attribut	decreas
lower	lower	decreas	lower	lower	incom
decreas	decreas	attribut	higher	decreas	lower
cost	declin	declin	decreas	declin	profit
profit	margin	cost	profit	profit	declin
declin	profit	profit	declin	higher	rose

Table 8. Top 13 words contributing to causal sentences for each sector.

Chapter VII

Discussions

In Figure 3, we observe that AUC is high enough on the first step. Therefore, there is not much room for further improvements as we increase the number of training data. This is because of the imbalanced ratio of causes and non-causes in our CR-SPC dataset. In addition, we find that the test performances between individual sectors have a large distribution as shown in Figure 4. That is mainly because of the size of dataset for sectors and the characteristics of each sector.

The results of LIME on 1st and 2nd fine-tune models show that two models resemble each other, that is because of distilled knowledge of the first stage of two-stage fine-tuning. However, some words appear or disappear in 2nd fine-tune models, which shows the advantage of using the two-stage fine-tuning framework on extracting causal rationales.

From the semi-supervised learning, we observe a dramatic increase of test performances in some sectors. That means we can use this method for automatic annotation, instead of collecting annotations from non-experts. However, our tests highly depend on our dataset, and there is a possibility of any bias on the judgement of the main reasons of stock price changes.

Chapter VIII

Conclusion and Future Work

In this work, we create a large scale of Causal Rationale of Stock Price Changes (CR-SPC) dataset to extract causal rationales from financial reports in various sectors automatically. We propose a two-stage fine-tuning framework to overcome diverse causalities and imbalanced learning. In addition, we also augment the size of dataset by self-training. We found that our two-stage fine-tuning enhances the performance of causal extracting models trained on sectors with a small number of data. We also collected additional 382K pseudo labeled data and observed better performances on the pseudo labels from self-training compared to the performance on low-quality data in some sectors. Furthermore, we showed a possible application of our work to the real financial fields (e.g., automatic summary generation). Finally, we apply LIME to our two-stage fine-tuned model and other models to compare the improvements qualitatively.

At this moment, we choose causal rationales from 10-K reports with respect to typical interests of industrial sectors. However, we have not tested that those causal rationales have actual correlations with stock price changes. Therefore, we need to investigate their relationships in the following research.

References

- Alkadhi, R., Lata, T., Guzman, E., & Bruegge, B. (2017). Rationale in development chat messages: an exploratory study. *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*, 436–446.
- Blanco, E., Castell, N., & Moldovan, D. I. (2008). Causal relation extraction. *Lrec*.
- Bressler, S. L., & Seth, A. K. (2011). Wiener–Granger Causality: A well established methodology. *NeuroImage*, 58(2), 323–329.
- Chang, D.-S., & Choi, K.-S. (2006). Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities. *Information Processing & Management*, 42(3), 662–678.
- Chollet, F., & others. (2015). *Keras*.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- Dai, A. M., & Le, Q. V. (2015). *Semi-supervised Sequence Learning*.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). {BERT:} Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.0.
- French, K. R. (n.d.). *Kenneth R. French - Detail for 12 Industry Portfolios*.
- Girju, R. (2003). Automatic detection of causal relations for question answering. *Proceedings of Annual Meeting on Association for Computational Linguistics Workshop on Multilingual Summarization and Question Answering-Volume 12*, 76–83.
- Girju, R., Moldovan, D. I., & others. (2002). Text mining for causal relations. *FLAIRS Conference*, 360–364.
- Graves, A., Fernández, S., & Schmidhuber, J. (2005). Bidirectional LSTM networks for improved phoneme classification and recognition. *International Conference on Artificial Neural Networks*, 799–804.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Hwang, Y., Tong, A., & Choi, J. (2016). Automatic Construction of Nonparametric Relational Regression Models for Multiple Time Series. *Proceedings of the International Conference on Machine Learning*, 48, 3030–3039.
- Ittoo, A., & Bouma, G. (2011). Extracting explicit and implicit causal relations from sparse, domain-specific texts. *International Conference on Application of Natural Language to Information Systems*, 52–63.

- Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., & Liu, Q. (2019). Tinybert: Distilling bert for natural language understanding. *ArXiv Preprint ArXiv:1909.10351*.
- Johnson, R., & Zhang, T. (2015). Semi-supervised Convolutional Neural Networks for Text Categorization via Region Embedding. In *Proceedings of Neural Information Processing Systems conference* (pp. 919–927).
- Karuna, A. (2019). Here Is What’s Behind The AI Revolution. *Forbes Technology Council*.
<https://www.forbes.com/sites/forbestechcouncil/2019/06/03/here-is-whats-behind-the-ai-revolution/#72028b844343>
- Khoo, C. S. G., Chan, S., & Niu, Y. (2000). Extracting causal knowledge from a medical database using graphical patterns. *Proceedings of the Annual Meeting on Association for Computational Linguistics*, 336–343.
- Ko, Y., & Seo, J. (2009). Text classification from unlabeled documents with bootstrapping and feature projection techniques. *Information Processing & Management*, 45(1), 70–83.
- Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., & Smith, N. A. (2009). Predicting Risk from Financial Reports with Regression. *Proceedings of The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 272–280.
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*.
- Liang, Y., Liu, Y., Kwong, C. K., & Lee, W. B. (2012). Learning the “Whys”: Discovering design rationale using text mining — An algorithm perspective. *Computer-Aided Design*, 44(10), 916–930.
- Lin, C., He, Y., Everson, R., & Ruger, S. (2012). Weakly Supervised Joint Sentiment-Topic Detection from Text. *IEEE Transactions on Knowledge and Data Engineering*, 24(6), 1134–1145.
- Lloyd, J. R., Duvenaud, D., Grosse, R., Tenenbaum, J. B., & Ghahramani, Z. (2014). Automatic Construction and Natural-language Description of Nonparametric Regression Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 1242–1250.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv Preprint ArXiv:1301.3781*.
- Papandreou, G., Chen, L.-C., Murphy, K., & Yuille, A. L. (2015). Weakly- and Semi-Supervised Learning of a {DCNN} for Semantic Image Segmentation. *CoRR*, abs/1502.0.
- Pavlinek, M., & Podgorelec, V. (2017). Text classification method based on self-training and LDA topic models. *Expert Systems with Applications*, 80, 83–93.
- Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.
<http://www.aclweb.org/anthology/D14-1162>

- Ribeiro, M., Singh, S., & Guestrin, C. (2016). “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. 97–101. <https://doi.org/10.18653/v1/N16-3020>
- Rogers, B., Gung, J., Qiao, Y., & Burge, J. E. (2012). Exploring techniques for rationale extraction from existing documents. *Proceedings of the International Conference on Software Engineering*, 1313–1316.
- Sakai, H., Nishizawa, H., Matsunami, S., & Sakaji, H. (2015). Extraction of causal information from PDF files of the summary of financial statements of companies. *Transactions of the Japanese Society for Artificial Intelligence*, 30, 172–182.
- Securities, & Commission, E. (n.d.). Important Information About EDGAR. In *About EDGAR System*.
- Shapley, L. (1971). Cores of convex games. *International Journal of Game Theory*, 1(1), 11–26.
- Sun, S., Cheng, Y., Gan, Z., & Liu, J. (2019). Patient knowledge distillation for bert model compression. *ArXiv Preprint ArXiv:1908.09355*.
- Tang, R., Lu, Y., Liu, L., Mou, L., Vechtomova, O., & Lin, J. (2019). Distilling task-specific knowledge from BERT into simple neural networks. *ArXiv Preprint ArXiv:1903.12136*.
- Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 207–212.
- Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1), 1–130.
- Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *Proceedings of the IEEE International Conference on Computer Vision*, 19–27.

Acknowledgements

My deep gratitude goes first to my advisor, Professor Jaesik Choi for the continued support of research, for his patience and guidance throughout my time of study.

I am also indebted to the co-advisors and the rest of my thesis committee: Prof. Byoung Ki Seo, Prof. Junyeop Lee, Prof. Seungwon Hwang and Prof. Kwang In Kim for their encouragement and insightful comments.

My appreciation extends to all members of the Statistical Artificial Intelligence Lab for their support and helpful advice. I have learned a lot from them, from programming skills to constructive criticism.

Lastly, I would like to give my deepest gratitude to my parents and friends for their unconditional support and love.

This thesis wouldn't become a reality if there is not help of many individuals. I would like to express my sincere thanks to all of them.