The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Extraction of Business Contents from Financial Reports Using Recurrent Neural Network Model

Tomoki Ito*1 Hiroki Sakaji*1 Kiyoshi Izumi*1

*1Graduate School of Engineering, The University of Tokyo

To extract business contents automatically from financial reports is an important problem in the financial area. Especially, segment names and their explanations are important contents that should be extracted. However, the methods for extracting these types of information from financial reports have not been established. In this study, we aim to develop a practical solution for extracting these types of information. To solve this problem, we developed a manually annotated dataset for the task of extracting the segment names and their explanations of each company from financial reports and then developed a recurrent neural network model to solve this task. Our method using the manually annotated dataset outperformed the baseline methods without the dataset in the task of extracting segment names and their explanations of each company. This results demonstrated that our approach is useful for extracting the business contents of each company. This work is the first work for applying a machine learning method to the task of extracting segment names and their explanations. The insights from this work should be valuable in the industrial area.

1. Introduction

With the development of information and communication technology, interest from investors on the technology of financial text mining has been increasing. When investors conduct investment activities, it is indispensable to gather performance information on listed companies. Among them, "financial summary reports" in which listed companies are published quarterly is one of the useful information sources for making investment decisions.

To analyze the financial summary reports, Information on business segments name and segment explanation is important, because each company usually describes settlement information in business segment units, as shown in Figure 1. However, it is heard that some financial companies usually extract these segment information manually, and it makes costs. Therefore, to extract this information has a great demand in the financial area.

Several studies on extracting the important contents from financial reports have been conducted, [Sakaji 17, Kitamori 17, Isonuma 17]. However, the method for extracting the segment names and segment explanations has not been established.

In this research, we aim to develop a practical method for extracting "business segment name" and "the explanation of the segment name" from the financial summary report. To achieve our aim, we first developed manually annotated dataset including textual data for "overview of reporting segments" and "business segment name" for each financial summary reports. We then develop a recurrent neural network model (RNN) that can extract "business segment name" from financial summary reports using the annotated dataset. To demonstrate the practicality of our method, we demonstrate whether our appraoch can extract segment names using only small training dataset. This insight should be useful for the industrial area.

Contact: m2015titoh@socsim.org

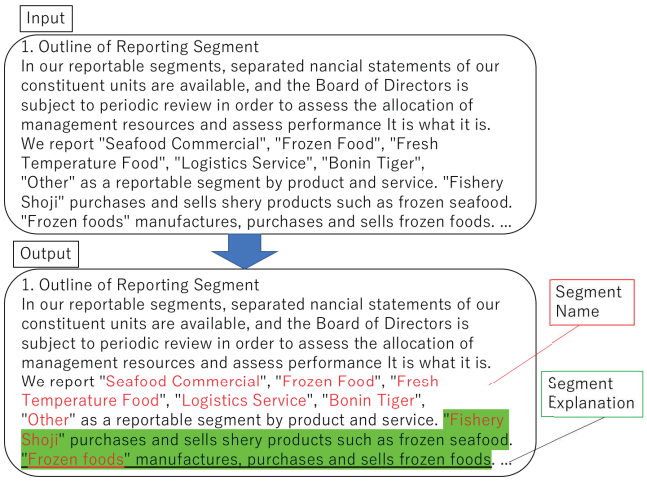

Figure 1: Goal Image: extraction of segment names and their explanations from a financial report

Our contributions are summarized as follows.

1) We present a crucial task setting: extraction of segment names and segment extractions from financial reports, and created an annotated dataset for this task.

2) We developed a practical method for extracting segment names and segment explanations from financial reports. This work is the first implementation for this task using a machine learning method, as far as we know.

2. Extraction of the segment information using RNN

This section introduces the proposed method for extracting segment names and their explanations from financial summary reports.

2.1 Task Setting

Let $\mathbf{Q} = \{w_t^{\mathbf{Q}}\}_{t=1}^{N}$ be a document. Our task is to extract

a) a set of segment names $\Omega^{\mathbf{Q}}$, and

b) a set of segment explanations $\{\{w_t^{\mathbf{Q}}\}_{t=s}^{f}\}_{s\in S^{\mathbf{Q}},f\in F^{\mathbf{Q}}}$ from $\{w_t^{\mathbf{Q}}\}_{t=1}^{N}$.

It should be noted that $|\Omega^{\mathbf{Q}}|and|S^{\mathbf{Q}}|, |F^{\mathbf{Q}}|$ are not constant values because the number of the segment differs between financial summary reports.

2.2 Structure of RNN

To solve the this task, we developed the RNN model as shown in Figure 2. This RNN model is constructed using the idea in the pointer network model [Vinyals 15].

2.2.1 Embedding

Given a comment $\mathbf{Q} = \{w_t^{\mathbf{Q}}\}_{t=1}^{n}$, this layer, first, converts all words in the comment to their respective word-level embeddings $\{\boldsymbol{e}_t^{\mathbf{Q}}\}_{t=1}^{n}$ where $\boldsymbol{e}_t^{\mathbf{Q}} \in \mathbb{R}^e$.

2.3 Segment Name Layer

This layer converts $\{\boldsymbol{e}_t^{\mathbf{Q}}\}_{t=1}^{n}$ to context representations $\{\boldsymbol{h}_t^{\mathbf{Q}}\}_{t=1}^{n}$ using a bi-directional long short-term memory, LSTM[Schuster 97]:

$$\boldsymbol{h}_t^{\mathbf{Q}} = \mathrm{LSTM}(\boldsymbol{e}_t^{\mathbf{Q}}) \tag{1}$$

where $\boldsymbol{h}_t^{\mathbf{Q}} \in \mathbb{R}^e$. This layer, then, converts $\boldsymbol{h}_t^{\mathbf{Q}}$ to the segment name layer$\{y_t^{\mathbf{Q}}\}_{t=1}^{n}$:

$$\boldsymbol{a}_t^{\mathbf{Q}} = \boldsymbol{W}^O \boldsymbol{h}_t^{\mathbf{Q}} + \boldsymbol{b}, \tag{2}$$

$$y_t^{\mathbf{Q}} = \mathrm{argmax} a_t^{\mathbf{Q}} \tag{3}$$

where $y_t^{\mathbf{Q}}$ represents word $w_t^{\mathbf{Q}}$ is included in a segment name set Ω^Q ($y_t^{\mathbf{Q}} = 1$) or not ($y_t^{\mathbf{Q}} = 0$), and $\boldsymbol{W}^O \in \mathbb{R}^{2\times e}$ and $\boldsymbol{b} \in \mathbb{R}^2$ are the parameter values.

2.4 Segment Explanation Start Layer

This layer converts $\{\boldsymbol{e}_t^{\mathbf{Q}}\}_{t=1}^{n}$ to context representations $\{\boldsymbol{u}_t^{\mathbf{Q}}\}_{t=1}^{n}$ using a LSTM:

$$\boldsymbol{u}_t^{\mathbf{Q}} = \mathrm{LSTM}(\boldsymbol{e}_t^{\mathbf{Q}}) \tag{4}$$

where $\boldsymbol{u}_t^{\mathbf{Q}} \in \mathbb{R}^e$. This layer, then, converts $\boldsymbol{h'}_t^{\mathbf{Q}}$ to the segment explanation start point layer$\{y_t^{\mathbf{Q}}\}_{t=1}^{n}$:

$$\boldsymbol{c}_t^{\mathbf{Q}} = \boldsymbol{W}_s^O \boldsymbol{h}_t^{\mathbf{Q}} + \boldsymbol{b}_s, \tag{5}$$

$$s_t^{\mathbf{Q}} = \mathrm{argmax}\ \boldsymbol{c}_t^{\mathbf{Q}} \tag{6}$$

where $y_t^{\mathbf{Q}}$ represents word $w_t^{\mathbf{Q}}$ is included in a start point set $S^{\mathbf{Q}}$ ($y_t^{\mathbf{Q}} = 1$) or not ($y_t^{\mathbf{Q}} = 0$), and $\boldsymbol{W}_s^O \in \mathbb{R}^{2\times e}$ and $\boldsymbol{b}_s \in \mathbb{R}^2$ are the parameter values.

2.5 Segment Explanation Finish Layer

This layer converts $\{\boldsymbol{e}_t^{\mathbf{Q}}\}_{t=1}^{n}$ to context representations $\{\boldsymbol{g}_t^{\mathbf{Q}}\}_{t=1}^{n}$ using a LSTM:

$$\boldsymbol{g}_t^{\mathbf{Q}} = \mathrm{LSTM}(\boldsymbol{e}_t^{\mathbf{Q}}) \tag{7}$$

where $\boldsymbol{g}_t^{\mathbf{Q}} \in \mathbb{R}^e$. This layer, then, converts $\boldsymbol{g}_t^{\mathbf{Q}}$ to the segment explanation finish point layer$\{y_t^{\mathbf{Q}}\}_{t=1}^{n}$:

$$\boldsymbol{d}_t^{\mathbf{Q}} = \boldsymbol{W}_f^O \boldsymbol{h}_t^{\mathbf{Q}} + \boldsymbol{b}_f, \tag{8}$$

$$f_t^{\mathbf{Q}} = \mathrm{argmax}\ \boldsymbol{d}_t^{\mathbf{Q}} \tag{9}$$

where $f_t^{\mathbf{Q}}$ represents word $w_t^{\mathbf{Q}}$ is included in a finish point set $F^{\mathbf{Q}}$ ($f_t^{\mathbf{Q}} = 1$) or not ($f_t^{\mathbf{Q}} = 0$), and $\boldsymbol{W}_f^O \in \mathbb{R}^{2\times e}$ and $\boldsymbol{b}_f \in \mathbb{R}^2$ are the parameter values.

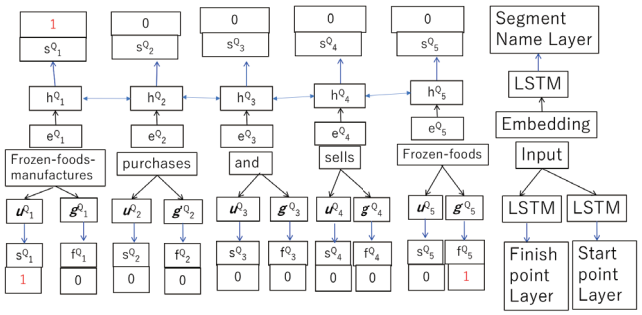

Figure 2: RNN Architecture

2.6 Learning

We can develop the RNN with a training dataset including $\{w_t^{\mathbf{Q}}\}_{t=1}^{N}$ and $\{\Omega^{\mathbf{Q}}, S^{\mathbf{Q}}, F^{\mathbf{Q}}\}$. We used the following L as a loss function

$$\begin{aligned} L &= CE(\{\boldsymbol{a}_t^{\mathbf{Q}}\}_{t=1}^{N}, \{w_t^{\mathbf{Q}} \in \Omega^{\mathbf{Q}}\}_{t=1}^{N}) \\ &+ CE(\{\boldsymbol{c}_t^{\mathbf{Q}}\}_{t=1}^{N}, \{w_t^{\mathbf{Q}} \in S^{\mathbf{Q}}\}_{t=1}^{N}) \\ &+ CE(\{\boldsymbol{d}_t^{\mathbf{Q}}\}_{t=1}^{N}, \{w_t^{\mathbf{Q}} \in F^{\mathbf{Q}}\}_{t=1}^{N}) \end{aligned}$$

where $CE(a, b)$ represents the softmac cross entropy between a and b.

3. Experimental Evaluation

This section introduces how we evaluated our method using a real textual dataset.

3.1 Dataset

To evaluate our method, we manually created a dataset including textual data for "overview of reporting segments", "business segment name," and "segment explanation" for each financial summary reports. We created dataset for 880 financial summary reports.

3.2 Segment information Extraction

We split the dataset into a training dataset and the remainder as a test dataset. We then developed the RNN model using the training dataset and extracted the segment names and the segment explanations from the test dataset.

3.2.1 Segment name

In extracting the segment names with the RNN, we extracted all the words $\Omega' = \{w_t^{\mathbf{Q}} | y_t^{\mathbf{Q}} = 1, 1 \le t \le n\}$ as the predicted segment names.

3.2.2 Segment explanation

In extracting the segment explanations with the RNN, we first extracted all the word positions $S' = \{t | s_t^{\mathbf{Q}} = 1, 1 \le t \le n\}$ and $F' = \{t | f_t^{\mathbf{Q}} = 1, 1 \le t \le n\}$ as the predicted start terms and finish terms. We then extracted segment explanations as shown in Algorithm 1.

After that, we evaluated the result using the F_1 score. To evaluate the practicality of our method, we evaluated our method in the case where the size of the training dataset is small and the size of the training dataset is sufficiently large. Table 1 summarizes the dataset organizations.

Table 1: Dataset details
(a) Training dataset

ID	1	2	3	4	5
number reports	50	100	200	300	400
number of segment names	116	250	543	836	1,134
number of non-segment names	9,115	20,892	40,263	59,026	79,167

(b) Test dataset

ID	1	2	3	4	5
number of reports	830	780	680	580	480
number of segment names	2,351	2,217	1,924	1,631	1,333
number of non-segment names	162,881	151,104	131,733	112,970	92,829

Algorithm 1 Extraction of Segment Explanation Part

$P \leftarrow \phi$, $m \leftarrow 0$, $d \leftarrow \{\}$;
for $l \in S' + F'$ **do**
  if m $= 0$ **then** m $\leftarrow l$;
  else P $\leftarrow \{w_t^{\mathbf{Q}}\}_{t=m}^{l}$, m $\leftarrow 0$;
  end if
end for
for $\{w_t^{\mathbf{Q}}\}_{t=m}^{l} \in P$ **do**
  for $t \in [m, m+1, \cdots l]$ **do**
    if $y_t^{\mathbf{Q}} = 1$ **then** $d[w_t^{\mathbf{Q}}] = \{w_t^{\mathbf{Q}}\}_{t=m}^{l}$;
    end if
  end for
end for
return d: dictionary of segment name (key) and explanation (value);

3.3 Comparison Method

3.3.1 Segment name

To evaluate our method for extracting segment names, we compared the results of our method with the results of extracting segment names using word embedding representations and the logistic regression model (baseline method).

3.3.2 Segment explanation

To evaluate our method for extracting segment explanations, we compared the results of our method with the results of the following baseline methods, namely, baseline (100), baseline (200), and baseline (300). In the baseline (100), baseline (200), and baseline (300) methods, we extracted the 100, 200, and 300 terms that existed after the predicted segment names as the explanation part.

3.3.3 Other Settings

Other experimental settings are summarized as follows: In developing RNN, we used the word embeddings calculated using the skip-gram method (window size = 5) [Mikolov 13] based on financial reports (between October 2002 and May 2018, 90,813 files). We set the dimensions of the RNNs' hidden and embedding vectors to 200, epoch to 40 with early stopping.

4. Results and Discussion

Table 2 summarizes the results, showing that the proposed method outperformed the baseline method.

In addition, the results demonstrated that the proposed method was practical because this method was able to extract segment names even with small training data. This practicality is considered to be caused by the consistent usage of words in the financial documents.

Table 2: Evaluation Result
(a) Segment Name

ID	1	2	3	4	5
Baseline	0.058	0.371	0.497	0.505	0.510
RNN	**0.827**	**0.851**	**0.873**	**0.869**	**0.843**

(b) Segment Explanation

ID	1	2	3	4	5
Baseline (100)	0.038	0.037	0.034	0.035	0.036
Baseline (200)	0.127	0.111	0.114	0.115	0.114
Baseline (300)	0.106	0.100	0.097	0.098	0.098
RNN	**0.587**	**0.570**	**0.619**	**0.594**	**0.652**

5. Related Works

Several studies have been done for extracting important information from financial documents[Sheikh 12, Pires 13, Sakaji 17, Kitamori 17, Isonuma 17]. In [Sheikh 12], the rule-based method for extracting up-date information form a news article was proposed. In [Pires 13], the method for extracting table contents from a financial document was proposed. Combining these works and our method, it can be possible to extract more valuable information from financial documents.

As for useful technique for information extraction, techniques used in Question Answering[Wang 17, Wang 18] can be useful. The application of these technique to our task can lead to the improvement of the extraction technique for extracting segment information.

6. Conclusion

In this study, we applied the RNN model to the task of extracting segment names and their explanations form financial reports. This work is the first implementation for these types of information using a machine learning method, as far as we know. We demonstrated that our method could extract segment names and their explanations with more higher F_1 scores than the baseline method. In addition, we experimentally demonstrated that we could extract seg-

ment names and their explanations with only small training dataset. This result should be the usual insight for the industrial area because this showed that our method was sufficiently practical. In the future, we will apply our method to the other similar tasks, and finally, develop a text-visualization system that visualizes the financial report contents of each company in a user-friendly manner.

References

[Isonuma 17] Isonuma, M., Fujino, T., Mori, J., Matsuo, Y., and Sakata, I.: Extractive Summarization Using Multi-Task Learning with Document Classification, in *EMNLP 2017* (2017)

[Kitamori 17] Kitamori, S., Sakai, H., and Sakaji, H.: Extraction of sentences concerning business performance forecast and economic forecast from summaries of financial statements by deep learning, in *IEEE CIFEr 2017* (2017)

[Mikolov 13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, in *NIPS 2013* (2013)

[Pires 13] Pires, F. M. and Abreu, S.: Automatic Selection of Table Areas in Documents for Information Extraction, in *EPIA 2003* (2013)

[Sakaji 17] Sakaji, H., Murono, R., Sakai, H., Bennett, J., and Izumi, K.: Discovery of Rare Causal Knowledge from Financial Statement Summaries, in *IEEE CIFEr 2017* (2017)

[Schuster 97] Schuster, M. and Paliwal, K.: Bidirectional Recurrent Neural Networks, *IEEE Transactions on Signal Processing*, Vol. 45, No. 11, pp. 2673–2681 (1997)

[Sheikh 12] Sheikh, M. and Conlon, S.: A rule-based system to extract financial information, *Journal of Computer Information Systems*, Vol. 52, (2012)

[Vinyals 15] Vinyals, O., Fortunato, M., and Jaitly, N.: Pointer Networks, in *NIPS 2015* (2015)

[Wang 17] Wang, W., Yang, N., Wei, F., Chang, B., and Zhou, M.: Gated Self-Matching Networks for Reading Comprehension and Question Answering, in *ACL 2017* (2017)

[Wang 18] Wang, W., Yan, M., and Wu, C.: Multi-Granularity Hierarchical Attention Fusion Networks for Reading Comprehension and Question Answering, in *ACL 2018* (2018)