Extraction of Causal Relationships across Multiple Sentences from Securities Reports

Takerou Aniya
Major in Computer and Information Sciences
Graduate School of Science and Engineering,
Ibaraki University
Email:21nm704f@vc.ibaraki.ac.jp
Nakanarusawa, Hitachi, Ibaraki, Japan

Minoru Sasaki
Dept. of Computer and Information Sciences
Faculty of Engineering, Ibaraki University
Email:minoru.sasaki.01@vc.ibaraki.ac.jp
Nakanarusawa, Hitachi, Ibaraki, Japan

***Abstract*— One of the most important sources of investment decisions in the stock market is the textual data contained in securities reports by companies. Investors consider investment strategies based on the information. However, since these text data are updated and published every day, it takes a great deal of time and money to read through and obtain information from all of them. In this study, we devised a method for extracting causal expressions from multiple sentences in securities reports. We extracted candidate causal expressions using clue expressions, and trained SVM (Support Vector Machine) by combining the similarity with the previous sentence and common phrases with the features obtained from a single sentence and verified how effective the method is. The effectiveness of the newly added features of inter-sentence similarity and common usage was confirmed.**

Keywords-causal relationship extraction; securities report; pattern recognition

I. INTRODUCTION

One of the most important sources of investment decisions in the stock market is the text data contained in news reports, newspaper articles, financial summaries, and securities reports released after the end of a company's fiscal year. These text data contain important information. The information has a significant impact on a company's performance, such as past and current business results and initiatives, future prospect, development of new products, anticipated impact from overseas and political events, and measures to deal with scandals that have been discovered. This information can be used by investors to formulate investment strategies.

Since these text data are updated and published daily, it takes a great deal of time and money for investors to read through and obtain all the information. In the past, it was difficult to structure such text data, so investors could only obtain the content by reading it directly.

In recent years, there has been a great deal of research in the fields of machine learning, artificial intelligence, and natural language processing to efficiently extract the desired information from text. These include research on automatically obtaining expressions that influence business performance by focusing on single sentences that indicate business performance factors in securities reports and using keywords that provide clues to sentence structure and causal relationships in Japanese[1], and research on extracting causal relationship sentences from newspaper articles by focusing on sentence structures that contain causal relationships.[2]

These studies mainly focus on extracting causal expressions consisting of single sentences, and when causal relationships are described in more than two sentences in securities reports, they are often excluded from the analysis, and there are few studies that extract causal expressions spanning two sentences. Therefore, this study devised a method for extracting causal expressions spanning two or more sentences from securities reports and verified the effectiveness of the inter-sentence similarity and common usage used in the features for SVM.

In Section 2, we introduce the research and related methods relevant to this study. In Section 3, we explain the flow of the method devised in this study and the features used for machine learning. In Section 4, we show the experimental method using the research method, and in Section 5, we present the results. Section 6 discusses the results, and Section 7 concludes the study.

II. RELATED WORKS AND METHODS

This section presents the related works and methods related to our research.

A. Related works

In the study by Sato et al. [1], a discriminant model of causal sentences for single sentences was constructed using text data contained in the securities reports of the companies that make up the TOPIX 1000 from 2008 to 2016. Here, for example, in the sentence “猛暑日が連続したため、飲料水の売上が伸びた。” (Sales of drinking water increased due to a series of extremely hot days.), “ため” (due to) is used as a clue expression as it is an important clue to indicate causality and is used to obtain candidate causal sentences from the target text data. Also, referring to the study by Sakaji et al. [2], we extracted features from candidate causal sentences using four features: particle pairs, clue expressions in sentences, morphological unigrams, and morphological bigrams, and constructed a discriminant model using SVM.

In the study by Sakaji et al. [2], based on the clue expressions obtained from the Nikkei Shimbun newspaper from 1990 to 2005, the expression patterns of causal

 ISBN: 978-1-61208-874-7

relationships and clue expressions were classified into five patterns, and a discriminant model was constructed using SVM while filtering the clue expressions (to remove the cases where the clue expressions have non-causal relationships).

In the study by Sato et al. [1], the objective was to extract from securities reports, but only single-sentence causal relationship was targeted, and causal relationships spanning multiple sentences were excluded from the target data. In addition, the study by Sakaji et al.[2] included up to two sentences in the sentence structure pattern, but the target data was the Nikkei Shimbun, not Securities reports and the feature extraction was based on the features that can be extracted from a single sentence. The results of their experiment were a fit rate of 0.68, a recall rate of 0.59, and an F-value of 0.63.

B. Morphological analysis

Methods such as splitting a sentence into words, generating a vector based on these words ,and using it as input for machine learning is often used in NLP(Natural Language Processing). Japanese is not divided into words like English. So it needs to be divided to the word level, and this is done by a morphological analyzer.

MeCab is an open source morphological analyzer developed through a joint research unit project between the Graduate School of Informatics, Kyoto University and the Communication Science Laboratories of Nippon Telegraph and Telephone Corporation. For example, the morphological analysis of" 当社グループにとって過去最高の結果となった "(The results were the best for our group.) by MeCab is shown in Figure 1.

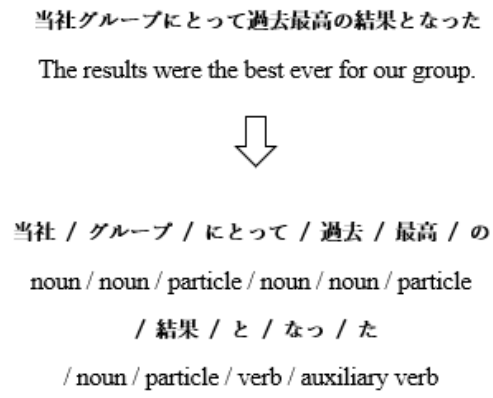

Figure 1. Examples of Morphological Analysis

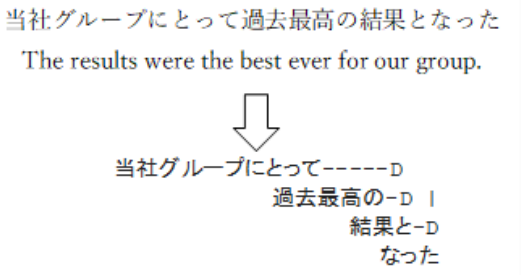

Figure 2. Example of Engagement Analysis

C. Engagement analysis

Engagement analysis (syntactic analysis), which is the process of analyzing modification relationships at the word level and in clauses, is the process of analyzing the structure of a sentence. In this study, we used CaboCha, a Japanese clause analyzer. CaboCha is based on SVM.

For example, Figure 2 below shows the result of using CaboCha to analyze the phrase "当社グループにとって過去最高の結果となった" (The results were the best ever for our group.")

Using CaboCha, the sentences are listed by clause with a set of clause IDs that indicate the clauses to be engaged. This method is used to obtain the relationships between sentences and clauses, which are used as part of the input to SVM.

To make the list, we used the tree structure obtained by analysing sentences with CaboCha.The tree structure is shown in Figure 3. From the chunk and its link in the tree structure, it is possible to obtain information on Japanese clauses and their destinations.

An example of a list obtained by the above method is shown in Figure 4. The "c" represents a clause, and the "to" indicates the ID of the destination. A value of -1 indicates that there is no target.

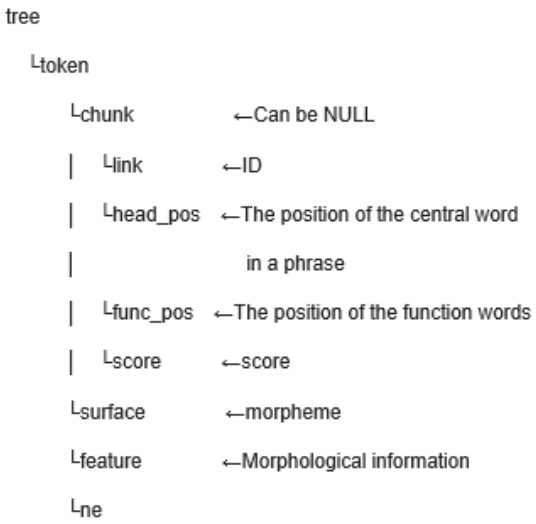

Figure 3. The tree structure of Cabocha

 ISBN: 978-1-61208-874-7

当社グループにとって過去最高の結果となった

The results were the best ever for our group.

[{'c':'当社グループにとって(for out group)','to':3},

{'c':'過去最高の(the best ever)','to':2},

{'c':'結果と(The results)','to':3},

{'c':'なった(were)','to':-1}]

Figure 4. Example of list that can be retrieved.

The list obtained by this method is used to extract syntactic features of sentences.

III. EXTRACTION OF CAUSAL RELATIONSHIPS USING CLUE EXPRESSIONS

In this paper, after extracting candidate causal sentences from securities reports issued by three companies (Nissan, Honda, and Toyota) using clue expressions, we propose a causal relationship extraction method that adds inter-sentence similarity as a feature and common usage as a new feature to the features used in existing research and uses them as input to SVM. In addition, we show the data used in this study.

A. Overview of the Proposed Method

The rough execution sequence of the proposed method in this study is shown in Figure 5 below.

The first step is to extract the items “業績等の概要” (Summary of business performance), “対処すべき課題” (Issues to be addressed), and “事業等のリスク” (Business risks) from the securities reports issued by three companies (Nissan, Honda, and Toyota), which contain sentences related to business performance, and then to extract candidate causal sentences containing clue expressions and the sentences immediately preceding them.

In the related study shown in Section 2, the clue expressions for extracting single sentence causal relationship candidates were determined, so they cannot be used directly. The clue expressions used in this study are those that were determined to be possible clue expressions in the case of a causal relationship spanning two sentences from the securities reports of the three companies mentioned above. Next, the obtained candidate causal sentences and the sentences immediately before them are assigned positive labels if they are causal relationships spanning two sentences, and negative labels if they are not. Finally, we obtain particle pairs, included clue expressions, morphological unigrams, and morphological bigrams from the sentences containing the clue expressions, and when all the obtained features are arranged, we assign 1 to the features included in the sentence and 0 to the features not included. The inter-sentence similarity with the immediately preceding sentence and the features of common usage are also added to the features and input to SVM.

By comparing the above method with the existing methods, we verify the effectiveness of the method for extracting causal relations across two sentences from the target securities report.

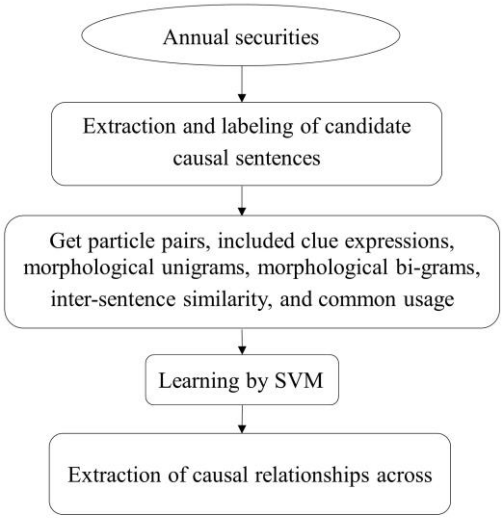

Figure 5. Flow of the proposed method

B. The data used in this study

The data used in this study are Nissan's securities reports for fiscal years 2000, 2004, 2005, 2007~2019, Honda's securities reports for fiscal years 2007~2019, and Toyota's securities reports for fiscal years 2003~2020. From these reports, the following items were included "業績等の概要 "(Summary of business performance), "対処すべき課題”(Issues to be addressed), and "事業等のリスク "(Business risks) including text that specifically discusses business results. The wording and structure of the securities reports of the same company do not change significantly from one year to the next. Therefore, the extraction of cause-and-effect expressions describing factors and results using clues is easier than in news articles.

C. Extraction of causal expression candidates

It is necessary to appropriately extract causal candidate sentences securities reports. Considering that clue expressions appear especially at the beginning and end of sentences in the case of causal candidate sentences that span two sentences, we surveyed securities reports and selected many expressions that appear at the beginning and end of sentences, referring to the clue expressions in the study by Sakaji et al.[2]

Figure 6 below shows the original clue expressions selected in this study.Based on these clue expressions, causal candidate sentences are extracted from the target part of the securities report.

 ISBN: 978-1-61208-874-7

Specifically, if a sentence contains a clue expression, the sentence and the immediately preceding sentence are extracted as causal candidates spanning two sentences.

このようななか (in this situation)	この結果、 (as a result)	その結果、 (as a result)
このような (such as this)	したがって (therefore)	従って (therefore)
これらの要因 (These factors)	要因は (The factor is)	要因に (The factor is)
主な (mainly)	主な成果として (The main results are)	によるものです (due to)
そのため (for the reason)	その為 (for the reason)	ためである (due to)
可能性がある (may be affected)	貢献して (contributing to)	が響いた (take its toll)
これらの (These)	これは主に (This is mainly because)	可能性があります (may be affected)
それらの要因 (Those factors)	によるものである (due to)	を反映 (reflect)

Figure 6. Selected clue expressions

D.Features used

In this study, we used particle pairs as syntactic features, and cue expressions in sentences, morphological unigrams, morphological bigrams, inter-sentence similarity, and common usage as other features.To obtain particle pairs, we used CaboCha to group morphemes into clauses, and extracted particle pairs based on the clause id of the target sentence.For morphological unigrams and morphological bigrams, those with a frequency of 100 or more were used at the stage of arranging the overall features.

The inter-sentence similarity is the value calculated for the similarity between two sentences. The calculation method was based on Yamamoto et al's study [3]. In the case of cause-and-effect sentences that span two sentences, the content (result or cause) of the immediately preceding sentence tends to be supplementary to the content of the following sentence. Therefore, the similarity between the two sentences was calculated by focusing on the nouns, adjectives, and verbs that appeared in the previous sentence, without considering the sentence structure such as particle pairs.

Common usage are phrases (nouns, adjectives, verbs) that are common to the previous sentence. The feature of the common usage to be input to the SVM is 1 if the causal candidate sentence has that usage as a common usage, and 0 if it does not.

For the extraction of nouns, adjectives, and verbs in the sentences, we used the part-of-speech information from the MeCab analysis results.

Table 1 below shows an overview of the features.

TABLE I. OVERVIEW OF FEATURES

Feature name	Overview
Pairs of particles	All pairs of particles (excluding redundancies), with the particle in the phrase containing the clue expression (the core phrase) as the front particle and the particle in the phrase pertaining to the phrase to which the core phrase is applied (the base phrase) as the back particle.
Clue expressions in a sentence	Clue expressions contained in the target sentence
Morphological uni-gram	A unigram obtained by decomposing candidate sentences containing causal relations with a morphological analyzer.
Morphological bi-gram	A bigram obtained by decomposing candidate sentences containing causal relations with a morphological analyzer.
Inter-sentence similarity	The sentence immediately before the sentence containing the clue expression is S_i .the sentence containing the clue expression is S_j .There is a high possibility that S_i and S_j have a common word in the cohesion by use. Therefore, the value expressed in the following equation is the inter-document similarity. $$sim\left(T(S_i), T(S_j)\right) = \frac{|T(S_i) \cap T(S_j)|}{\sqrt{|T(S_i)T(S_j)|}}$$ $T(S_i)$: Word sets of nouns, verbs, and adjectives in sentence S_i.
Common usage	As with inter-sentence similarity, it is a feature related to the usage (noun, adjective, verb) that is common to the previous sentence. Unlike inter-sentence similarity, it extracts the usage itself that is common to the previous sentence.

E. Input to SVM

For each candidate causal sentence extracted based on the clue expressions, the features can be obtained.

For the input to the SVM, the features were arranged without overlap, except for the inter-sentence similarity, and a vector was created with 1 if the feature obtained from each candidate causal sentence was included and 0 if it was not.

IV. EXPERIMENTS

The experimental procedure based on the proposed method is shown.

A. Data Set

The data used in this study are Nissan's annual securities reports for fiscal years 2000, 2004, 2005, 2007~2019, Honda's annual securities reports for fiscal years 2007~2019, and Toyota's securities reports for fiscal years 2003~2020.

B. Settings

The candidate causal sentences were extracted using the clue expressions shown in Figure 6. The extracted text was extracted line by line and saved in csv format as sentence1 for the sentence immediately before the clue expression if it was included, and sentence2 for the sentence including the cue expression. After that, we manually assigned labels to the saved csv files: 1 if there was a causal relationship between sentence1 and sentence2, and 0 if there was not.

As a result of the extraction, we were able to extract 3879 sentences from the target securities reports as candidate sentences for causal relationships spanning two sentences, of which 989 sentences were assigned positive labels and 2890 sentences were assigned negative labels. The features shown in the proposed method 3.4 were used as the features of each sentence.

We used train_testsplit to split the training data and test data, and test size was set to 0.3. To avoid data bias, we used the stratify parameter. In addition, cross validation was used to reduce overtraining. The kernel of the SVM model was linear, and the value of the regularization C was 10.

V. RESULTS

The results of an experiment conducted based on the experimental method are shown below.

A. Results of existing methods

In order to make comparisons, we also conducted experiments without the addition of the features we devised in this study, and experiments with only one of the two new features.The results are shown below in tabular form.

Table 2 shows the result of existing methods.

TABLE II. RESULTS OF EXISTING METHODS

Count	Accuracy	Precision	Recall	F-measure
1	0.925	0.860	0.843	0.851
2	0.925	0.855	0.851	0.852
3	0.931	0.865	0.867	0.865
4	0.914	0.832	0.833	0.832
5	0.921	0.849	0.842	0.845
Ave	0.923	0.852	0.847	0.849

B. Results of the proposal method

Table 3 below shows the result of our proposal methods.

TABLE III. RESULTS OF THE PROPOSAL METHODS

Count	Accuracy	Precision	Recall	F-measure
1	0.934	0.875	0.864	0.869
2	0.941	0.888	0.880	0.883
3	0.931	0.862	0.872	0.867
4	0.931	0.868	0.864	0.866
5	0.926	0.855	0.856	0.856
Ave	0.933	0.87	0.867	0.868

C. Results of the proposal methods (using only common Ftausage as new feature)

TABLE IV. RESULTS OF THE PROPOSAL METHODS (USING ONLY COMMON USAGE AS NEW FEATURE)

Count	Accuracy	Precision	Recall	F-measure
1	0.932	0.871	0.862	0.866
2	0.931	0.864	0.869	0.866
3	0.925	0.850	0.858	0.853
4	0.926	0.846	0.869	0.857
5	0.931	0.866	0.866	0.866
Ave	0.929	0.859	0.865	0.862

D. Results of the proposal methods (using only inter-sentence similarity as new feature)

TABLE V. RESULTS OF THE PROPOSAL METHODS (USING ONLY INTER-SENTENCE SIMILARITY AS NEW FEATURE)

Count	Accuracy	Precision	Recall	F-measure
1	0.926	0.858	0.854	0.855
2	0.927	0.864	0.846	0.855
3	0.928	0.859	0.861	0.860
4	0.931	0.873	0.855	0.863
5	0.930	0.868	0.858	0.862
Ave	0.928	0.864	0.855	0.859

The tables from II to V shows the results of five-times classifications and their average values by the proposed method.

 ISBN: 978-1-61208-874-7

VI. DISCUSSIONS

In this section, we discuss the experimental results and show the effectiveness and problems of newly devised features.

A. Comparison with conventional methods

As can be seen from Table 2 and 3, Accuracy increased by 1%, and Precision increased by 1.8%, Recall increased by 2%, and F-score increased by 1.9% for the proposed method. In addition, as can be seen from Table 2 and Table 4, Accuracy increased by 0.6%, Precision increased by 0.7%, Recall increased by 1.8%, and the F-measure increased by 1.3% for the proposed method when only the features of common words were added to the existing method. Furthermore, as can be seen from Table 2 and Table 5, Accuracy increased by 0.5%, Precision increased by 1.2%, Recall increased by 0.8%, and the F-measure increased by 1% for the method that only added the feature of inter-sentence similarity to the existing method.

In the above comparisons, the proposed method for extracting causal relationships across multiple sentences has certain results, as it produces better results than the conventional method. However, as shown in Table 2 and Table 3, the proposed method is not always better than the conventional method in the third and fifth comparisons, respectively. The proposed method must be improved by increasing the amount of target data, dividing the training data, and appropriately adjusting each parameter of SVM.

B. Inter-sentential similarity and its validity as a feature of common speech and its problems

The comparison between the proposed method and the existing method with the addition of inter-sentence similarity and common usage as features is confirmed that all the evaluation indices increased even when each of them was added by itself. Therefore, we believe that these two features are effective as features for extracting causal relations across two sentences.

As a result, we were able to obtain a better evaluation value compared to the existing methods, and we were able to prove the effectiveness of the inter-sentence similarity and common terms.

Future work includes increasing the number of target data, re-examining the cue expressions used to extract causal candidate sentences spanning two sentences, adjusting the data partitioning method and each parameter of machine learning, and examining the extraction method when a causal sentence exists before the previous sentence.

The increase in the evaluation index value with the addition of these new features was between 0.5 and 1.0, which does not necessarily mean that the effectiveness of this study on other companies' securities reports data is guaranteed, since the learning and evaluation was conducted using a limited amount of securities report data. As a countermeasure to this problem, an increase in the number of securities reports can be considered.

In addition, when we look at the average of the absolute value of the difference from the mean for each evaluation value of the proposed method, we can see that the Accuracy is 0.4, Precision is 0.9, Recall is 0.7, and the F-score is 0.6. This is thought to be due to the influence of the data used for training, but to have a stable two-sentence causal relationship extraction method, it is necessary to reduce this blur by using appropriate training data that reflects the characteristics of all data from among all data.

VII. CONCLUSION

In this study, we used clue expressions that can be used as clues for causal sentences spanning two sentences selected independently from the ”業績の概要”(Summary of Business Results), “対処すべき課題”(Issues to Be Addressed), and “業績等のリスク”(Business Risks) where text data can be obtained from the securities reports of Nissan, Honda, and Toyota, especially for the contents related to business results. We then created a model for extracting causal relations spanning two sentences using SVM with the features of particle pairs, included clue expressions, morphological unigrams, and morphological bigrams used in existing research, plus inter-sentence similarity and common usage.

In future research, we aim to extract sentences containing causal relationships from a wide range of text data, not limited to securities reports. For this purpose, it is necessary to consider the extraction of candidate sentences other than clue expressions. In addition, we aim to extract the causal part and the resultant part from the extracted sentences.

REFERENCES

[1] Fumihiro Sato, Hiroaki Sakuma, and Shunya Kodera, “Extraciton of Causal Knowledge from Annual Securities Report”,The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 4pp. (2018)

[2] H. Sakaji and M. Sigeru(Toyohashi University of Technology), “Extraction of Causal Knowledge by Using Text Mining”, 第 54 回自動制御連合講演会, 4pp. (2011)

[3] Yuji Yamamoto, Shigeru Masuyama, and Hiroyuki Sakai, “小説自動要約のための隣接文間の結束判定手法”, 言語処理学会年次大会発表論文集 (Proceedings of the Annual Meeting of the Association for Natural Language Processing), pp.1083-1086 (2006)

 ISBN: 978-1-61208-874-7