Extraction of Causal Relationships across Multiple Sentences from Securities Reports

Takerou Aniya
Ibaraki University
Ibaraki, Japan
Email: 21nm704f@vc.ibaraki.ac.jp

Minoru Sasaki
Ibaraki University
Ibaraki, Japan
Email: minoru.Sasaki.01@vc.ibaraki.ac.jp

About the presenter

- Takerou Aniya

Major in Computer and Information Sciences

Graduated from Science and Engineering of Ibaraki University in 2021

Outline

- Motivation
- Background
- Previous Studies
- How to extract more information
- Related Methods
- Flow of extracting causal expressions
- Features used as input to the SVM
- Experiments
- Results of the experiments
- Conclusion

Motivation

- Annual securities reports, which summarize a company's business performance and financial information, are an extremely useful source of information for investors.
- However, it is difficult to read and obtain information from all the securities reports issued by many companies.
- To solve this problem, we have developed a discriminative model that automatically extracts useful information through natural language processing.

Background

- Information affecting business performance contained in securities reports is represented by statements that show cause and effect. Therefore, it is important to extract them well. However, unlike English, Japanese is not divided into word units, so it is necessary to devise a way to train discriminant models.

Previous Studies

- In the study by Sato et al[1]. constructed a discriminant model of causal sentences for single sentences using text data contained in the annual securities reports of the companies that make up the TOPIX(Tokyo stock price index) 1000 from 2008 to 2016.

- In the study by Sakaji et al.[2], four features were extracted from candidate causal sentences: particle pairs, cue expressions in sentences, morphological unigrams, and morphological bigrams, and a discriminant model was constructed using SVM(support vector machine).

How to extract more information

- The study by Sato et al. [1]aimed to extract information from securities reports, but it targeted only single cause-and-effect sentences, and the cause-and-effect relationships spanning multiple sentences were excluded from the target data.
- The study by Sakaji et al.[2] includes up to two sentences in the sentence structure pattern, but the target data is the Nikkei Shimbun newspaper, and the feature extraction is based on the features that can be extracted from a single sentence.
- In some cases, cause-and-effect relationships are expressed across two or more sentences. We considered extracting causal expressions that span two sentences as a target for extracting more information from securities reports.

Related Methods

- Morphological Analysis

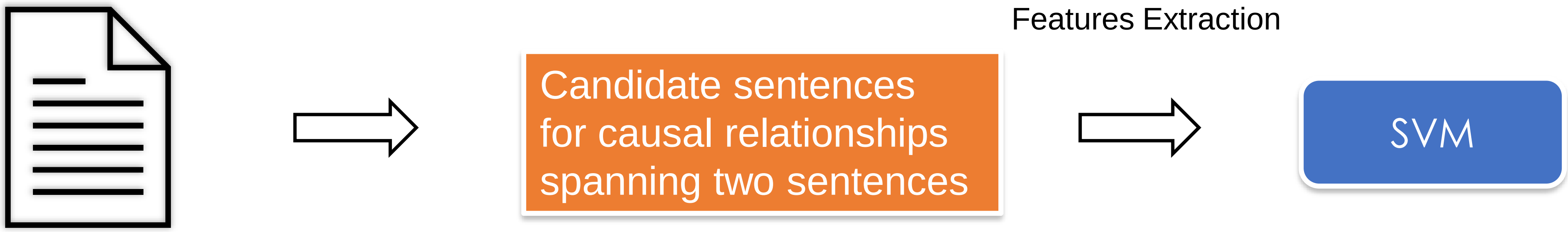
For machine learning input, a sentence is divided into words, and a vector is used based on the divided words. Since Japanese is not divided into words like English, it needs to be divided down to the word level, and this is done by a morphological analyzer. In this study, we used Mecab which is a Japanese morphological analyzer.

- Syntax Analysis

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

Flow of extracting causal expressions

Securities Report



Features Extraction

Candidate sentences for causal relationships spanning two sentences

SVM

Extraction of sentences
containing keywords
indicating causal relationships
as candidate sentences
for causal relationships.

Labeling.

From the candidate sentences,
we obtain particle pairs, clue expressions, morphological uni-grams,
morphological bi-grams, inter sentence similarity,
and common usage as input to SVM.

Features used as input to the SVM 1/5

- 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.

Features used as input to the SVM 2/5

- Examples of clue expressions

このようななか(in this situation)，この結果、(as a result)，したがって(therefore)，　主な(mainly)

Since sentences containing such expressions often contain the cause part, the effect part, or both of the causal expressions, sentences containing these expressions are extracted as candidate causal sentences.

However, this string is not necessarily a causal expression, so the discriminant model is used to classify it in the end.

Features used as input to the SVM 3/5

- Morphological unigram and bigram

A unigram obtained by decomposing candidate sentences containing causal relations with a morphological analyzer.

A bigram obtained by decomposing candidate sentences containing causal relations with a morphological analyzer.

Features used as input to the SVM 4/5

- 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-sentence similarity.

$$sim\left(T(S_i), T(S_j)\right) = \frac{|T(S_i) \cap T(S_j)|}{\sqrt{|T(S_i) T(S_j)|}}$$

This predisposition was based on the work of Yamamoto et al.[3]

Features used as input to the SVM 5/5

- 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.

Experiments

- 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.

Results of the experiments

- Results of existing methods (Not using Inter-sentence similarity and common usage)

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

Results of the experiments

- Results of proposal methods (Using Inter-sentence similarity and common usage)

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

Conclusion

- In the case of two sentences, the experimental results are better than those of the conventional method in terms of Accuracy, Precision, Recall, and F-measure.
- The effectiveness of the newly devised features of inter-sentence similarity and common usage as features was confirmed.

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)