

ロイターニュース記事からの因果関係抽出の試み

An Experiment of Causal Knowledge Extraction from Reuter News

坂地 泰紀^{*1}

Hiroki Sakaji

ベネット ジェイソン^{*2}

Jason Bennett

宮尾 祐介^{*3}

Yusuke Miyao

和泉 潔^{*1}

Kiyoshi Izumi

^{*1}東京大学

The University of Tokyo

^{*2}三井住友アセットマネジメント

Sumitomo Mitsui Asset Management Company, Limited

^{*3}国立情報学研究所

National Institute of Informatics

In this paper, we try to develop a method for extracting causal expressions from news articles. First, we investigated clue expressions that are included in news articles. Next, we focused on specific clue expressions and developed the method to extract causal expressions.

1. はじめに

近年、人工知能分野の手法や技術を、金融市場における様々な場面に応用することが期待されており、例えば、膨大な金融情報を分析して投資判断を支援する技術が注目されている。さらに、最近では証券市場における個人投資家の比重が増大しており、個人投資家に対して投資判断の支援を行う技術の必要性が高まっている。

投資家にとって、企業の業績に関する情報は、投資判断を行ううえで重要であるが、企業の業績だけでなく、その業績要因に含まれる因果関係が重要である。例えば、原因「猛暑」、結果「冷房需要の盛り上がり」といった因果関係を投資家に提示することで、「猛暑」の場合には、「冷房需要」が高まる可能性があることを個人投資家が知ることができるというメリットがある。そして、その原因「猛暑」に対する結果「冷房需要の盛り上がり」の因果関係から、猛暑の年には、冷房に関する事業を行っている企業の業績が好調に推移することが期待できる。

そこで我々は、ロイターニュース記事に含まれる因果関係を抽出するために、その調査をまず行った。その後、構文解析結果を用いた抽出手法の開発を試みた。

2. ロイターニュース記事中の因果関係の調査

我々は、ロイターニュース英文記事中の因果関係を表す手がかりとなる表現（手がかり表現）を調査するため、ランダムに 20 記事を選び、その中に含まれる因果関係と手がかり表現を調査した。ここで我々が対象とする因果関係の定義としては、The penn discourse treebank(PTDB)[Prasad 07]における“Cause”に該当する。手がかり表現と因果関係を調査した結果、381 文中 30 個の因果関係を発見することができた。加えて、表 1 に示す手がかり表現を得ることができた。

表 1 より、一般的に因果関係を示す手がかり表現である「because」や「as」などは見つけることができた。しかしながら、「since」や「from」等の他の手がかり表現は調査中に含まれていなかった。調査の結果、大きく分けて「will keep」や「cause」などの動詞である手がかり表現と、「because」や「because of」などのその他の手がかり表現があることが分かった。ここで、先行研究である Mirza et al.[Mirza 14b]の手法によると、彼女らは手がかり表現として **affect verbs**, **link verbs**, **causative verbs** と **SIGNAL** を用いている。CSIGNAL は、前置詞

連絡先: 坂地 泰紀, 東京大学大学院工学系研究科,
sakaji@sys.t.u-tokyo.ac.jp

表 1: 調査で得られた手がかり表現一覧

That's why	1	affected	1
after	1	will keep	1
as	3	because	8
because of	2	will pressure	1
buoyed by	1	due to	1
fuelled by	1	helped by	1
meant that	1	partly because of	1
partly on	1	prompted by	1
was aided by	1	will also dampen	1
will check	1	will entice	1

(because of, as a result など), 接続詞 (because, since, so that), 副詞 (so, therefore, thus) や特殊な表現 (the reason why, the result is, that is why) から構成される。このうち, affect verbs, link verbs, causative verbs が動詞である手がかり表現に該当し, SIGNAL がその他である手がかり表現に該当する。すなわち, 定義の細かさは異なるが, 我々が調査した結果は, Mirza et al.[Mirza 14b] が定義した手がかり表現と同様の分類となったことを意味する。

さらに, 調査を進めたところ, Mirza et al. や Khoo et al. などの手法でも用いられている手がかり表現「because」において, 因果関係を含まない例を見つけることができた。

”They won’t let North Korea get away with it **because** others might try the same thing,” said Redeker at BNP Paribas.

この文では, might が含まれているため, 意味として仮定を表しており, 原因・結果を表す因果関係ではない。

上記のように, 「because」が含まれていれば必ず原因・理由の因果関係を示すわけではないということが分かった。つまり, because などの手がかり表現が含まれている場合において, 因果関係を意味する文か否かを判定する手法が必要となってくる。判定手法の試みに関しては, 4. 節にて述べる。

3. 因果関係抽出の試み

我々は, まず, because と will + 動詞である手がかり表現について, 原因・結果を抽出できる手法を開発した。日本語における因果関係抽出手法としては, 我々が開発した手法 [Sakaji 08] が

存在する。それに対して、英語文における因果関係抽出手法としては、Chang et al.[Chang 06]の手法や Khoo et al.[Khoo 00]の手法、Mirza et al.[Mirza 14a]の手法が存在する。上記で挙げた手法のほとんどでは、構文解析の結果を利用して因果関係を抽出している。本研究においても、構文解析の結果を利用する。

本研究では、構文解析器として Stanford CoreNLP^{*1}を利用する。例えば、以下の文を構文解析器にかけると、図1のような結果が得られる。

It may augur well for the local economy **because** the Philippines is heavily dependent on remittances from overseas Filipino workers.

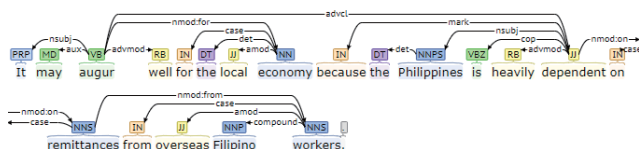


図 1: 構文解析した結果

上記例にタグ付けしたところ、以下のような結果が得られた。

<effect>It may augur well for the local economy</effect> <clue>because</clue> <cause>the Philippines is heavily dependent on remittances from overseas Filipino workers.</cause>

また、will + 動詞からなる手がかり表現の例を以下に示す。

<cause>Fears of war in Iraq and a stand-off with North Korea about nuclear weapons</cause> <clue>will keep</clue> <effect>the dollar weak in the first half of the year.</effect>

上記結果から、手がかり表現 because では、「結果表現, 手がかり表現, 原因表現」の順に出現し、手がかり表現 will + 動詞では、「原因表現, 手がかり表現, 結果表現」の順に出現することが分かった。日本語においても、手がかり表現によって、原因表現と結果表現の出現する順番が異なることは、我々が過去に行った調査で分かっている。英語文においても、同様に、手がかり表現によって原因表現と結果表現の出現する位置が異なることが分かった。

我々は、日本語においては、5つの構文パターンで因果関係を抽出する手法を開発した。英語文においても、Khoo et al.[Khoo 00]の手法が示すように、構文パターンが存在すると思われる。今後は、手がかり表現の種類によって、構文パターンは異なるのか、または、同様なのかを調査し、適切な抽出手法を開発することを試みる。

4. 因果関係判定手法の試み

前節の述べたように因果関係の有無を判定する手法の必要があったため、その手法の開発を試みた。前節で取得した手が

かり表現を含む文をランダムに 4,000 文抽出し、因果関係が含まれるか否かのタグを付与した。その結果、4,000 文中 1,599 文に因果関係が含まれていた。我々は、調査のために、タグ付けしたデータのうち 3,000 文を学習データに、1,000 文を評価データにして、いくつかの機械学習手法を試した。その結果を表2に示す。結果より、機械学習によっての分類性能の差はほとんどなかった。

表 2: 機械学習により判定した結果

	適合率	再現率	F 値
SVM	0.737	0.906	0.813
RandomForest	0.730	0.977	0.836
BiLSTM	0.741	0.950	0.832

5. まとめ

本研究では、ロイターニュース英文記事に含まれる因果関係を調査し、その抽出を試みた。調査した結果、Mirza et al.[Mirza 14b] が用いていた手がかり表現と類似した手がかり表現を得ることができた。構文解析手法を用いて、いくつかの手がかり表現を対象に、因果関係抽出手法を開発することはできたが、他の手がかり表現から抽出を行う手法はまだ開発できていない。今後の課題としては、全ての手がかり表現を利用して因果関係を抽出する手法を開発し、その性能評価を行うことが考えられる。

参考文献

- [Chang 06] Chang, D.-S. and Choi, K.-S.: Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities, *Information Processing and Management*, Vol. 42, No. 3, pp. 662–678 (2006)
- [Khoo 00] Khoo, C. S., Chan, S., and Niu, Y.: Extracting Causal Knowledge from a Medical Database Using Graphical Patterns, in *Proceedings of the 38th ACL*, pp. 336–343 (2000)
- [Mirza 14a] Mirza, P.: Extracting temporal and causal relations between events, in *Proceedings of the ACL 2014 Student Research Workshop*, pp. 10–17 (2014)
- [Mirza 14b] Mirza, P. and Tonelli, S.: An analysis of causality between events and its relation to temporal information, in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 2097–2106 (2014)
- [Prasad 07] Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., and Webber, B. L.: The penn discourse treebank 2.0 annotation manual (2007)
- [Sakaji 08] Sakaji, H., Sekine, S., and Masuyama, S.: Extracting Causal Knowledge Using Clue Phrases and Syntactic Patterns, in *7th International Conference on Practical Aspects of Knowledge Management (PAKM)*, pp. 111–122 (2008)

*1 <https://stanfordnlp.github.io/CoreNLP/>