成蹊大学理工学研究報告
J. Fac. Sci.Tech., Seikei Univ.
Vol.51 No.2 (2014)pp.23-28

新聞記事からの因果関係の抽出

坂地　泰紀*1，酒井　浩之*2，増山　繁*3

An Extraction Method of Causal Knowledge from Newspaper Corpus

Hiroki SAKAJI*1, Hiroyuki SAKAI*2, Shigeru MASUYAMA*3

ABSTRACT：This paper proposes a method that extracts causal knowledge from news paper articles via clue expressions. Our method decides whether a sentence includes causal knowledge or not when the method extracts it. Therefore, our method can extract causal knowledge accurately. Furthermore, the advantage of our decision method is to extract causal knowledge manually without dictionaries and patterns.

Keywords：Textmining, Information Extraction, Causal Knowledge

(Received September 19, 2014)

１．はじめに

現在，ウェブページや新聞記事を含む大規模な機械可読文書が入手可能になっている．多くの機械可読な文書の中には，実アプリケーションに役立つ様々な情報があり，テキストマイニング技術を用いることで獲得することが可能である．そのような情報の一つに因果関係がある．因果関係は，QAシステム[1]や因果ネットワーク構築[2]などで用いられることが期待されている．しかしながら，そのような知識を人手で抽出するためには，非常に高いコストと時間がかかる．さらに，上記の用途に用いるためには，誤った情報ではなく，正確な情報が必要である．この問題を解決するために，文書から因果関係を自動的に抽出することを試みた研究がいくつかある[3]〜[6].

本論文では，因果関係を抽出するうえで重要な手がかりとなる表現(手がかり表現と定義する)を利用して，新聞記事から自動的に抽出する手法を提案する．文献[7]に準拠し，因果関係は，出来事(結果)とその理由(原因)の組から構成されるとするが，本論文では，1 文中，または，隣り合う 2 文中に直接表現されている表層的なものに限定する．本論文では，次節で説明する予備実験で，1 文中に出現する因果関係を示す手がかり表現だけを対象として後述する判定手法を適用する．例えば，「サブプライムローンの危機により、世界不況が起こった」という文の場合，「世界不況が起こった」は結果表現，「サブプライムローンの危機」は原因表現，「により」は手がかり表現となる．
これらの結果と原因は，手がかり表現「により」によって明確に示されている．

また，手がかり表現には，因果関係以外の意味を持つものがある．例えば，「あなたの**ため**に、花を買った.」という文中の「ため」は，原因・結果ではなく，目的の意味を表している．このような場合に対応するために，半教師在り学習を用いたフィルタリング手法を適用した．手がかり表現を用いて因果関係を表す表現を高精度に抽出するアルゴリズムを作成し，それにフィルタリング手法を適用し，その評価を行った．

２．因果関係の抽出

本節では，因果関係を表す表現の抽出方法について述べる．因果関係を表す表現を抽出するにあたり，まず，文中に出現する原因・結果と手がかり表現の位置関係に着目し，調査を行った．ここで，原因・結果を，それぞれ，原因表現と結果表現と本論文では定義する．調査には，坂地ら[8]の方法を用いて，1990 年から 2005 年の日

*1：成蹊大学理工学部助教（hiroki_sakaji@st.seikei.ac.jp）
*2：成蹊大学理工学部准教授
*3：豊橋技術科学大学教授

経新聞から獲得した景気動向記事を用いた．獲得した景気動向記事 10,027 記事から無作為に 300 記事を選択し，それを対象として調査を行なった．なお，パターンの抽出自体は景気動向記事であるが，評価実験は日経新聞の分野によらない記事を用いた．

その結果，原因表現については，約 97%の場合，手がかり表現と同一文内で，かつ，手がかり表現の直前に出現することが分かった．稀ではあるが，約 3%の原因表現が直前の文にも出現することが明らかになった．それに対して，結果表現は手がかり表現が含まれる文と，その直前の文内の様々な場所に出現することが分かった．以上の結果を表現の出現位置についてまとめると，手がかり表現と原因・結果表現の出現位置は大きく分類して 5 通りであることが分かった．その 5 通りをPattern AからDとし，図 1 に示す．本手法は，この 5 通りのPatternから因果関係を獲得するアルゴリズムを用いて，因果関係を抽出する．具体的なアルゴリズムは，Sakaji et al.[9]を参考にされたい．

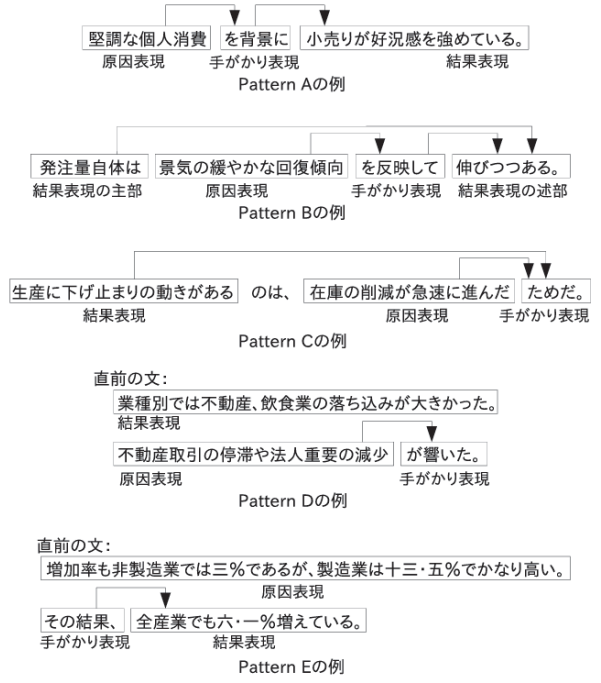

図 1　各Patternの例

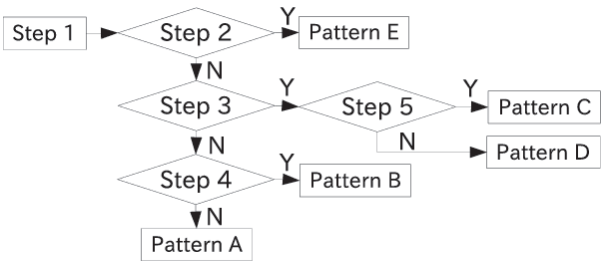

図 2　Pattern識別の概要

2．1　適切な表現形式の識別

対象文が与えられたときに，上記に示したPatternのうち，どのPatternを適用するかを識別する方法を説明する．ここで，手がかり表現が含まれる最後尾の文節を手がかり表現の**核文節**，核文節の係り先の文節を**基点文節**と定義する．構文Pattern識別の手続き(*Identification of patterns*)を以下に示す(図 2 を参照．).

[*Identification of patterns*]

Step 1:　手がかり表現を含む文を探す.

Step 2:　手がかり表現が文頭に出現する場合，Pattern Eを適用した後，Step 6 を実行する．そうでなければ，Step 3 を実行する.

Step 3:　手がかり表現に「.」が含まれている，もしくは，手がかり表現の後に「.」があるなら，Step 5 を実行する．そうでなければ，Step 4 を実行する.

Step 4:　基点文節が動詞句であり，かつ，基点文節が係り先である文節中に係り助詞，もしくは，格助詞を含むものがあれば，Pattern Bを適用する．そうでなければ，Pattern Aを適用する．Step 6 を実行する.

Step 5:　核文節に係っている文節に係り助詞が含まれている場合，Pattern Cを適用する.そうでなければ，Pattern Dを適用する.

Step 6:　手続きを終了する.　□

3．フィルタリング手法

手がかり表現が因果関係以外の意味を持つ場合があり，それを取り除くためにフィルタリング手法を用いる．フィルタリング手法は，文に因果関係が含まれているか否かを判定する手法である．ルールを用いてフィルタリングすると，因果関係を含むか否かを判定する際に用いる特徴(素性)の数が多く，ルールを作成するにも数が多すぎるという問題がある．そのため，本研究では，機械学習手法を用いた．フィルタリング手法で用いる素性として，表 1 にある素性を採用した.

表 1　素性の一覧

構文的な素性	助詞のペア
意味的な素性	拡張言語オントロジー
それ以外の素性	手がかり表現の直前形態素の品詞
	手がかり表現
	形態素ユニグラム
	形態素バイグラム

我々は，因果関係を含むか否かの判定のため，構文的な素性，意味的な素性を用いる．構文的な素性を用いることにより，日本語文において因果関係を表すためによく用いられる表現を利用するという狙いがある．例えば，「半導体**の**需要回復**を受けて**半導体メーカーが設備投資**を**増やしている．」という文に含まれる助詞と手がかり表現の並び「～の～を受けて～を～」が因果関係を表している可能性が高い．そこで，構文解析を用いて手がかり表現に関係のある助詞だけを素性として獲得する．また，意味的な素性として拡張言語オントロジー[10]を用いることにより，因果関係を示す語彙の関係を利用するという狙いがある．各素性の抽出に関しては，[11]を参照されたい．

3．1　タグなしデータからの追加学習データの獲得

フィルタリング手法は，タグなしデータから追加学習データを自動的に獲得することで，学習データを増やし，精度の向上を図る．学習データを作成するために文中に因果関係が存在するか否かを人手で判断するのは，時間やコストがかかるという問題がある．そこで，すでにタグがつけられた学習データを用いて，タグなしデータから追加学習データを自動的に獲得する．学習データの詳細については，次節に記述してある．その概要を図３に示す．

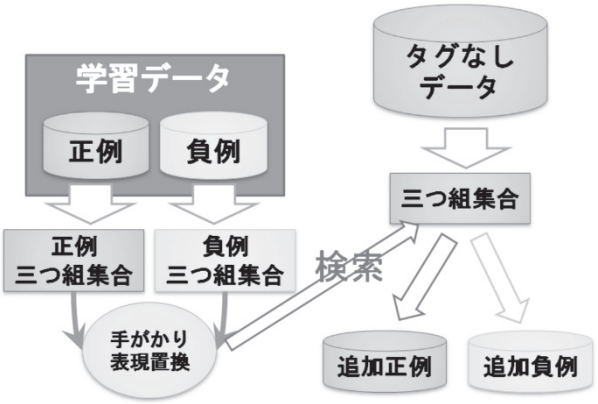

図３　追加学習データの取得

追加学習データを獲得するにあたり，我々は手がかり表現が持つ意味に着目した．そのため，本手法は手がかり表現がもつ意味が因果関係であるか否かの判定であるとも考えられる．また，手がかり表現には，因果関係以外の意味を持つ多義性のものもある．このことを利用すると，他の手がかり表現に置換した文がコーパス中に存在すれば，その文は因果関係を含む可能性が高い．例えば，文「円高により、日本経済が悪化した．」という文に含まれる手がかり表現を，「のため、」に置換した文「円高のため、日本経済が悪化した．」は因果関係を持つ．

それに対して，因果関係を持たない文では，手がかり表現とその前後に因果関係でないことを示す特徴がある．例えば，「記者会見で、不快感を示した．」という文であれば，「記者会見」と「を示した．」が特徴となる．上記の特徴を持った他の文「記者会見で、歓迎する意向を示した．」は因果関係を持っていない．負例の追加学習データを獲得する際には，上記の特徴を利用する．

追加学習データを獲得する手続き*Extracting additional learning data*[11]を以下に示す．また，アルゴリズム中の三つ組については，[11]を参照されたい．

[*Extracting addtional learning data*]

Step 1: *Acquiring ternary set*[11]により，正例から三つ組集合S，負例から三つ組集合F，タグなしデータから三つ組集合Tを抽出する．

Step 2:　Sに含まれる三つ組と，その手がかり表現部分を他の手がかり表現に置換したものの集合をPとする．Fに含まれる三つ組と，その手がかり表現部分を他の手がかり表現に置換したものの集合をNとする．

Step 3:　$AP = P \cap T, AF = N \cap T$

Step 4:　APを正例の追加学習データとして獲得する．AFを負例の追加学習データとして獲得する．　□

４．評価実験

本手法の評価実験を行い，その性能を評価する．実験には，表２に示す手がかり表現を用いた．実験データには，1990 年から 2005 年の日経新聞記事から，調査に用いた300記事と異なる100記事を抽出し，これを用いた．実験データは，景気動向記事のみから獲得したものではなく，日経新聞記事全体からランダムに獲得したことを注意されたい．また，フィルタリング手法の学習データには，1995 年から 2005 年の日経新聞記事から，調査に用いた 300 記事と実験データと異なる，ランダムに抽出した手がかり表現を含む 1,000 文を用いた．

表２　手がかり表現一覧

を背景に	を背景に、	を受け、	ため、	に伴う
に伴い、	を反映して	で、	をきっかけに	により、
に支えられて	によって	を反映し、	が響き、	ためで、
を受けて	から、	により	が響いた．	ため」
が影響した．	による．	ためで	ためだ．	を受けて、
に伴い	ため．	が響く	が響いている	が響いている．
による	このため、	このため	その結果、	この結果、

学習データ，評価データ共に人手で因果関係を示すタグを付与した．その結果，学習データ 1,000 文のうち 387 文が因果関係を含んでいた．実験データ 100 記事(1376 文)には，87 の因果関係を表す表現が含まれていた．そのうち，表 2 で表されている因果関係を表す表現の数が 72，それ以外の手がかり表現で表されている因果関係を表す表現の数が 11，手がかり表現を伴わない因果関係を表す表現の数が 4 であった．

形態素解析器としてはMecab(注 1)を用い，係り受け解析器としてはCabocha[12]を用いた．学習器にはSVMLight(注 2)を用いた．評価は再現率(Recall)と適合率(Precision)，および，その調和平均であるF値(F-Measure)で行なった．F値とは，再現率と適合率の調和平均である．

4．1　実験結果

本手法の結果を表 3 に示す．また，フィルタリング手法において，追加学習データ 1023 文を自動的に増やすことができた．

表 3　結果一覧

	適合率	再現率	F値
本手法	0.34	0.71	0.46
本手法(フィルタリングあり)	0.68	0.59	0.63

表 3 より，フィルタリング手法を用いることで，本手法の適合率を 0.34 から 0.68 に向上させることができた．しかしながら，再現率においては，0.71 から 0.59 に低下してしまった．ただし，F値に着目すると，0.46 から 0.63 に向上しているため，フィルタリング手法は有効であると考えられる．

本手法が獲得した 181 個の因果関係を表す表現のうち，正解であったものは 62 個であった．つまり，本来獲得すべきである 72 個のうち，62 個のみ獲得し，残りの 10 個は獲得できなかった．また，実験データにおいて，手がかり表現が因果関係以外の意味を示している場合が多い傾向にあったため，フィルタリングを用いない本手法の適合率が 0.34 になったと考えられる．

5．エラー解析

本手法のエラー解析を行った．本手法によって獲得されたものの中で，因果関係を持っていないものをFPとし，文に因果関係があるにもかかわらず，違う部分を獲得してしまったものをFNとする．FPとFNのそれぞれの数を表 4 に示す．

表 4　エラー解析結果

	FN	FP	FN + FP
本手法	3	117	120
本手法(フィルタリングあり)	1	23	24

表 4 より，フィルタリング手法を用いることでFPを 117 から 23 と，大幅に減らすことができた．

6．関連研究

我々の目的と同様に，因果関係の自動抽出を試みた研究は数多く行われている．乾らは接続標識「ため」を含む複文から因果関係に関する知識を獲得し，それらを 4 種類の因果関係(*cause*，*effect*，*precond*，*means*)に自動的に分類する手法を提案している[13]．接続標識「ため」は因果関係を抽出するうえで重要な手がかりである．しかしながら，彼らの研究では，「ため」のみを手がかり表現として用いているため，それ以外の手がかり表現でしか表していない因果関係を抽出することができない．また，佐藤らは重文，複文のみを対象として，核フレームを用いて因果関係を抽出している[14]．

Khooらは人手で作成したパタンを用いて，新聞記事や医療データベースから因果関係を抽出する手法を提案している[3],[4]が，結果表現と原因表現が同じ文に含まれている必要がある．これらの研究は，初期の研究として非常に重要であるが，因果関係を抽出する対象が限定されているため，抽出結果も限定的となる．それに対して，本手法では用いている手がかり表現は 35 個と豊富であり，重文，複文や文をまたがった対象からも因果関係を抽出することができるため，数多くの抽出結果が得られることが期待できる．

7．むすび

本研究では，手がかり表現を用いて，タグの付いていない新聞記事から因果関係(結果と原因の組)を抽出する手法を提案した．本手法は，因果関係を表す表現を抽出する際に，文が因果関係を含んでいるか否かを判定し，因果関係を含んでいると判定された文から因果関係を表す表現を抽出することで，精度を向上させた．

今後の課題として，手がかり表現を伴わない因果関係や，実験で用いなかった手がかり表現で表されている因果関係を獲得できるような手法が求められる．

(注 1)：http://mecab.sourceforge.net/
(注 2)：http://svmlight.joachims.org/

参考文献

[1] R. Higashinaka and H. Isozaki: “Corpus-based question answering for why-questions", in Proceeding of IJCNLP, pp. 418-425 (2008).

[2] 石井，馬，吉川：“因果関係ネットワークの増分的な構築について”，情報処理学会創立 50 周年記念（第 72 回）全国大会, pp.239-240 (2010).

[3] C. S. Khoo, J. Korn_lt, R. N. Oddy and S. H. Myaeng: “Automatic extraction of cause-effect information from news-paper text without knowledge-based inferencing", Literary and Linguistic Computing, 13, 4, pp. 177-186 (1998).

[4] C. S. Khoo, S. Chan and Y. Niu: “Extracting causal knowledge from a medical database using graphical patterns", Proceedings of the 38th ACL, pp. 336-343 (2000).

[5] R. Girju: “Automatic detection of causal relations for question answering", In ACL Workshop on Multilingual Summarization and Question Answering, pp. 76-83 (2003).

[6] D.-S. Chang and K.-S. Choi: “Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities", Information Processing and Management, 42, 3, pp. 662-678 (2006).

[7] 庵：“新しい日本語学入門”，スリーエーネットワーク(2001).

[8] H. Sakaji, H. Sakai and S. Masuyama: “Automatic extraction of basis expressions that indicate economic trends", Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), pp. 977-984 (2008).

[9] H. Sakaji, S. Sekine and S. Masuyama: “Extracting causal knowledge using clue phrases and syntactic patterns", 7th International Conference on Practical Aspects of Knowledge Management (PAKM), pp. 111-122 (2008).

[10] 小林，増山，関根：“Wikipedia と汎用シソーラスを用いた汎用オントロジー構築手法”, 電子情報通信学会論文誌 D, J93-D, 12, pp. 2597-2609 (2010).

[11] 坂地，増山：“新聞記事からの因果関係を含む文の抽出手法”，電子情報通信学会論文誌 D, J94-D, 8 (2011).

[12] 工藤，松本：“チャンキングの段階適用による日本語係り受け解析”，情報処理学会論文誌, 43, 6, pp. 1834-1842 (2002).

[13] 乾，乾，松本：“接続標識「ため」に基づく文章集合からの因果関係知識の自動獲得”, 情報処理学会論文誌, 45, 3, pp. 919-933 (2004).

[14] 佐藤，笠原，松澤：“テキスト上の表層的因果知識の獲得とその応用”，信学技報(TL98-23), pp. 27-34 (1999).