論　文

新聞記事からの因果関係を含む文の抽出手法

坂地　泰紀[†a)]　　増山　繁[†b)]

A Method for Extracting Sentences Including Causal Relations from Newspaper Articles

Hiroki SAKAJI[†a)] and Shigeru MASUYAMA[†b)]

あらまし　本論文では，新聞記事から因果関係を含む文を自動的に抽出する手法を提案する．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であり，その中には実アプリケーションに役立つ様々な情報が存在し，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り，本研究では因果関係の存在を示す手掛りとなる表現に基づいた因果関係を含む文の抽出を行った．その結果，人手により作られた辞書やパターンを用いず，自動的に因果関係を含む文を抽出することができた．本手法は，素性として構文的な素性と，意味的な素性を用いた．また，追加学習データを自動的に獲得することができる．その結果，性能が向上し，F 値 0.797 を達成した．

キーワード　因果関係，半教師あり学習，文解析

1. まえがき

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], [3] などで用いられることが期待されている．ただし，因果関係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．この問題を解決するために，手掛り表現を用いて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 [4]〜[6].

しかしながら，手掛り表現を用いて自動的に因果関係を抽出する場合，因果関係をもたない表現を獲得してしまうことがある．これは，手掛り表現が因果関係以外の意味で使われているためである．例えば，手掛り表現「から」は因果関係以外に，出発する位置を表す意味があり，「京都から大阪」などには因果関係は含まれない．このような因果関係ではない表現を抽出しないようにフィルタリングする手法がいくつか存在する [7]〜[9].

我々は，これらの因果関係をもたない場合を取り除くため，因果関係が含まれている文を自動的に抽出する手法を提案する．後述のように，本論文では，対象とする因果関係を同一文内に原因と結果が含まれている場合に限定している．そのため，手掛り表現を含む文において，それが因果関係を示しているか否かを判別するということは，その文が因果関係を含むか否かの判定と同じである．また，本論文における手掛り表現は，同一文中に原因を表す表現とその結果を表す表現が含まれていることを示唆する表現と定義する．

本手法は名詞間や動詞句間が因果関係を含んでいるか否かの判定ではなく，文に因果関係が含まれているか否かの判定を行っている．これは，因果関係を含む文を抽出すれば，動詞句で表されている因果関係，若しくは，名詞句で表されている因果関係のどちらの抽出タスクにおいても本手法をフィルタリング手法として適用することができるからである．本手法は半教師あり学習を用いて，因果関係が含まれている文を抽出する．また，本論文で扱う因果関係は原因若しくは，理由と結果を示し，手掛り表現を伴って出現するもの

† 豊橋技術科学大学，豊橋市
Toyohashi University of Technology, 1–1 Hibarigaoka, Tempaku-cho, Toyohashi-shi, 441–8580 Japan
a) E-mail: sakaji@la.cs.tut.ac.jp
b) E-mail: masuyama@tut.jp

に限定する [10]．例えば，「この食べ物は腐っていたため、食べなかった」という文の場合，表現「食べなかった」は「帰結」，「結果」のどちらにも判断できる．つまり，表現「この食べ物は腐っていた」は「理由」，「原因」のどちらにも判断でき，区別することが難しい．新聞記事中には，このような文が多数存在するため，本研究では原因と理由を区別しなかった．

2. 手掛り表現

本研究では，手掛り表現を用いて因果関係を含む文を獲得する．例えば，「日本市場では消費者などの抵抗感**から、**遺伝子非組換え品に限定していない一般大豆のニーズが減退している。」という文では，「から、」が手掛り表現となっている．

しかしながら，手掛り表現は因果関係ではない意味を表す場合がある．例えば，「同社は全国の延べ約十六万人の会員**から、**約九十六億円を集めており、法規制後のねずみ講としては過去最大規模という。」という文では，「から、」は出所を示す意味であり，因果関係は表していない．日本語において，手掛り表現を用いて因果関係を獲得する際には，この問題を伴う場合がある．そこで，文中の手掛り表現が因果関係を示しているか否かを判別する必要がある．本章で後述するように，本論文では対象とする因果関係を同一文内に原因と結果が含まれている場合に限定しているので，文中の手掛り表現が因果関係を示しているか否かを判別するということは，その文が因果関係を含むか否かの判定と同じである．そのため，本研究では手掛り表現が因果関係を示すか否かに着目した．

因果関係の中には手掛り表現を伴わないものもある．例えば，「高松市でも午後七時十五分ごろ送電線に落雷があり、市内の約九万八千世帯で一、二分停電した。」という文は，手掛り表現を伴わないが因果関係が存在する．事前調査として 1995 年から 2005 年の日経新聞記事から 100 記事をランダムに抽出し，その中に含まれる手掛り表現によって示されている因果関係と，手掛り表現を伴わない因果関係の数を調べた．その結果，手掛り表現によって示されている因果関係の数が 56，手掛り表現を伴わない因果関係の数が 22 であった．そのため，本研究では数が多い，手掛り表現によって示されている因果関係を扱う．また，2 文にまたがって存在する因果関係もあるが，数が少ない [4] ため，本研究では同一文内に原因と結果が含まれている場合に対象を限定した．

3. 因果関係を含む文の抽出手法

本論文で提案する手法は因果関係を含む文を抽出する手法である．因果関係が存在する文と，存在しない文の例を表 1 に示す．例中の太文字は手掛り表現である．

因果関係を含む文をルールを用いて抽出するとすると，因果関係を含む文を獲得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．また，手掛り表現を含む文のみを対象としているため，本手法は手掛り表現が因果関係の意味をもつか否かを判定する手法であると考えることができる．

3.1 素性の抽出

因果関係を含む文を抽出するために，文から特徴（素性）を抽出する．本研究では，表 2 の素性を抽出する．

我々は，因果関係を含むか否かの判定のため，構文

表 1　因果関係を含む文と，含まない文の例
Table 1　Examples of sentences that have causal relations and examples of sentences that do not have causal relations.

因果関係を含む文
- こうした経済指標やユーロ現金導入**に伴う**消費動向の微妙な変化**を背景に、**米同時テロ以降に広がった景気悲観論は急速に後退している。
- サリドマイドやスモン被害**を受け、**同省は一九八〇年に医薬品の副作用被害を救済する制度を創設した。

因果関係を含まない文
- 長野県信用組合（長野市、丸山彰一理事長）は五日**から、**法人向けにインターネットバンキングの取り扱いを始める。
- 大きさは幅が約三十二・八センチ、奥行き約三十・六センチ、高さ約四十八・一センチ**で、**重さは約三・七キロ。

表 2　素性の一覧
Table 2　List of features.

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
- 手掛り表現の直前形態素の品詞
- 文に含まれる手掛り表現
- 形態素ユニグラム
- 形態素バイグラム

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掛り表現の並び「〜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掛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後述する拡張言語オントロジーを用いることにより，因果関係を示す語彙の関係（表 2 を参照）を利用するという狙いがある．

助詞のペア

構文的な素性である助詞のペアの取得には，坂本ら [11] の手法を適用する．文が因果関係をもっている場合も，因果関係を含まない場合（並列関係など）にも，助詞に特徴が現れる．例えば，因果関係ではなく並列関係である場合「通信機器**が**同二〇%増と高い伸びで、コンピューター**も**好調だった。」では，助詞「が」と「も」を伴う文節の名詞句が並列であることを特徴付けている．しかしながら，特徴を表している助詞のみを獲得しなければ素性として役に立たない．そこで，構文情報を用いて助詞を獲得することで，手掛り表現に依存している助詞を素性として獲得する．本研究では，手掛り表現が因果関係を表す意味をもっているという考えに基づいているため，上記のように手掛り表現に依存しているものが重要となってくる．

また，素性抽出と追加学習データ抽出に用いる用語を以下に定義する [4]．

核文節　手掛り表現を含む最後尾の文節

基点文節　核文節の係り先となる文節

まず，核文節に係っている文節を探し，その文節に含まれている助詞を前部助詞として獲得する．次に，基点文節に係っている文節を探し，その文節に含まれている助詞を後部助詞として獲得する．そして，前部助詞と後部助詞の全ての組合せを助詞のペアとして抽出する．詳細な助詞のペア抽出は以下の手続き *Extraction of pairs of particles* に従う．

[*Extraction of pairs of particles*]

Step 1：文を文節ごとに区切る．

Step 2：文節を先頭から走査する．

Step 3：文節が核文節に係る場合

- その文節に含まれる助詞を前部助詞リストに追加する．

Step 4：文節が基点文節に係る場合

- その文節が核文節以前に位置する場合はスキップする．
- その文節に含まれる助詞を後部助詞リストに追加する．

Step 5：前部助詞が取得できなかった場合

- 手掛り表現より前に出現する表現のうち，核文節に一番近い助詞を前部助詞リストに追加する．
- 手掛り表現より前に出現する表現に助詞が存在しない場合は前部助詞リストに null を追加する．

Step 6：後部助詞が取得できなかった場合

- 手掛り表現より後に出現する表現のうち，基点文節に一番近い助詞を後部助詞リストに追加する．
- 手掛り表現より後に出現する表現に助詞が存在しない場合は後部助詞リストに null を追加する．

Step 7：前部助詞と後部助詞の全ての組合せ（助詞のペア）を素性とする．　□

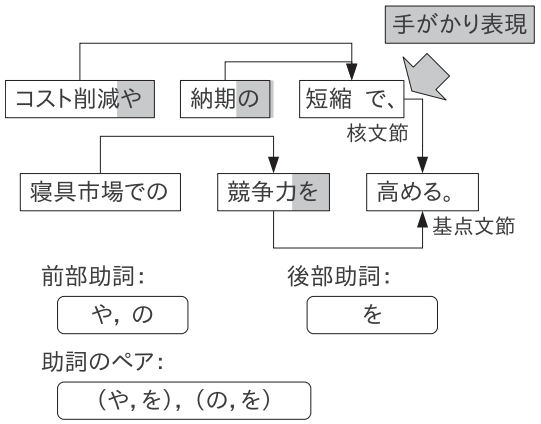

図 1　助詞ペア取得の例
Fig. 1　An example of extracting pairs of particles.

上記手続きによる処理の例を図 1 に示す．図 1 では，手掛り表現「で、」を含む文節「短縮で、」が核文節に，核文節が係っている文節「高める。」が基点文節である．まず，核文節に係っている文節「コスト削減や」，「納期の」から前部助詞として，それぞれ「や」，「の」を獲得する．その後，基点文節に係っている文節「競争力を」から後部助詞として「を」を獲得する．そして，前部助詞と後部助詞の全組合せである「や、を」と「の、を」を素性助詞のペアとして獲得する．

拡張言語オントロジー

本研究では，小林ら [12] が作成した言語オントロ

ジー（シソーラス）を拡張言語オントロジーと定義し，これを用いる．小林らは Wikipedia から抽出した語彙を既存の言語オントロジーにマッチングすることで既存の言語オントロジーを拡張しているため，語彙数が多い．そのため，様々な種類の語彙を素性として網羅することができると考え，本研究ではこの拡張言語オントロジー中の語彙を素性として採用した．本実験では，日本語語彙大系 [13] から作成された拡張言語オントロジーを用いる．

本手法では，拡張言語オントロジーの上から 6 階層目の意味カテゴリーを素性として用いる．日本語語彙大系における同階層の意味カテゴリーは，同程度の抽象度になっている．予備実験として，使用した拡張言語オントロジーが基づいている日本語語彙大系の 6 層目すべての語に対し，それぞれ上の層，下の層の語と比較した結果，それぞれ約 92%が適切であった．そこで，前部語彙，後部語彙ともに 6 階層目の語を採用した．例えば，「あんかけスパゲッティ」という語をオントロジーの上位にたどっていくと，6 階層目は「食料」という意味カテゴリーであり，5 階層目は「人工物」である．ここで，5 階層目の「人工物」としてしまうと，その子である「建造物」に下位にある語と「あんかけスパゲッティ」が同じ扱いになってしまう．更に，6 階層目の意味カテゴリー数 256 に比べ，7 階層目の意味カテゴリー数は 536 と多く，本手法では意味カテゴリーの組合せを素性に用いているため，7 階層目を用いると素性数が多くなってしまうという問題もある．

まず，手続き *Extraction of pairs of particles* と同様に核文節に係っている文節を探す．その文節に拡張言語オントロジーに含まれる語があれば，拡張言語オントロジーの上から 6 階層目の意味カテゴリーを前部語彙として獲得する．次に，基点文節に係っている文節を探す．前部語彙と同様に，文節に拡張言語オントロジーに含まれる語があれば，上から 6 階層目の意味カテゴリーを後部語彙として獲得する．前部語彙，後部語彙，それぞれ獲得できなかった場合は，“null” とする．そして，前部語彙と後部語彙の各組合せを素性として抽出する．

素性の抽出例を図 2 に示す．図 2 では，文「台風の影響で天草五橋が通行止めになった。」が係り受け解析され，係り受け情報をもった文節に分割されている．核文節「影響で、」に係る文節「台風の」に含まれる語「台風」が拡張言語オントロジーに含まれているため，「台風」の上位にあたる意味カテゴリー「気象・天象」が前部語彙として獲得されている．基点文節「なった。」に係る文節「天草五橋が」に含まれる語「天草五橋」が拡張言語オントロジーに含まれているため，「天草五橋」の上位にあたる意味カテゴリー「交通路」が後部語彙として獲得されている．その結果，素性として（気象・天象，交通路）が抽出されている．

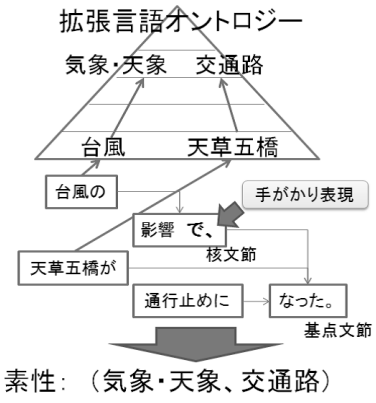

図 2　拡張言語オントロジー素性の取得例
Fig. 2　An example of extracting expanded linguistic ontology features.

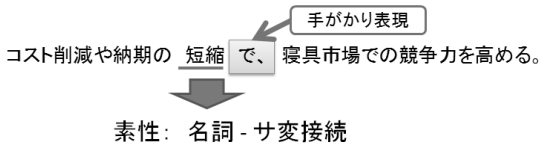

図 3　手掛り表現の直前形態素の取得例
Fig. 3　An example of extracting a morpheme that appears before a clue expression.

手掛り表現の直前形態素の品詞

本手法では，手掛り表現の直前形態素を素性に用いる．例えば，手掛り表現「ため、」の直前に助詞「の」がある場合と，直前に動詞がある場合では，因果関係を表す確率が異なることから，これを反映するために直前の品詞を素性とした．図 3 に直前形態素の取得例を示す．図 3 では，素性として「名詞–サ変接続」が取得されている．

3.2　タグなしデータからの追加学習データの獲得

本手法では，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，既にタ

グが付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，**4. 1** に記述してある．その概要を図 4 に示す．

追加学習データを獲得するにあたり，我々は手掛り表現がもつ意味に着目した．本研究では，手掛り表現を含む文を対象として，実験を行っている．そのため，本手法は手掛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，手掛り表現には，因果関係以外の意味をもつ多義性のものもある．このことを利用すると，他の手掛り表現に置換した文がコーパス中に存在すれば，**2.** で述べたように 2 文以上にまたがる因果関係が少ないことに注意すると，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掛り表現を，「のため、」に置換した文「円高のため、日本経済が悪化した。」は因果関係をもつ．

それに対して，因果関係をもたない文では，手掛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．上記の特徴をもった他の文「記者会見で、歓迎する意向を示した。」は因果関係をもっていない．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* を以下に示す．

[*Extracting addtional learning data*]

Step 1：後述する *Acquiring ternary set* 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，その手掛り表現部分を他の手掛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，その手掛り表現部分を他の手掛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． □

以下の手続き「*Acquiring ternary set*」により，追加学習データを獲得するために用いる表現対と手掛り表現の組を抽出する．抽出する組には，正例（因果関係）と負例を獲得するために必要な特徴が含まれている．

[*Acquiring ternary set*]

Step 1：手掛り表現を含む文を係り受け解析し，先頭の文節から走査する．

Step 2：手掛り表現の直前の形態素が動詞や助動詞である場合は，直前の形態素からさかのぼり，格助詞までを前部表現として獲得する．そうでない場合は，直前の形態素を前部表現として獲得する．

Step 3：基点文節が動詞句や形容詞句である場合，基点文節の最後尾の形態素からさかのぼり，格助詞までを後部表現として獲得する．基点文節が名詞句である場合は，基点文節の名詞部分を後部表現として獲得する．

Step 4：前部表現，後部表現と手掛り表現を三つ組として抽出する． □

以上の手続きにより，三つ組を抽出する．図 5 に

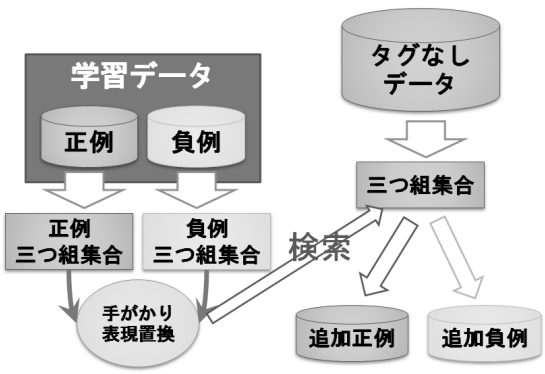

図 4　追加学習データの取得
Fig. 4　An example of extracting additional learning data.

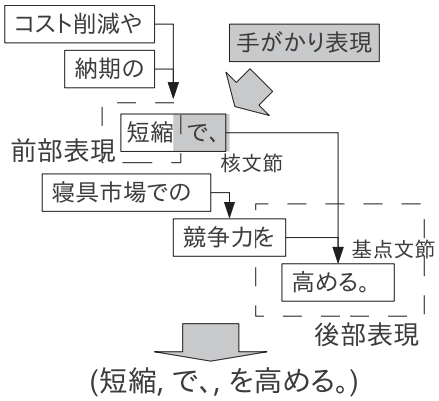

図 5　追加学習データを取得するための表現対の抽出例
Fig. 5　An example of extracting a pair of expressions for extracting additional learning data.

表現対の抽出例を示す．この例では，前部表現として「短縮」，後部表現として「を高める。」が獲得され，結果として三つ組（短縮，で、，を高める。）が抽出されている．

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．

4.1 実験方法

学習データには 1995 年から 2005 年の日経新聞記事からランダムに抽出した手掛り表現を含む 1,000 文を用い，評価データにはランダムに抽出した手掛り表現を含み，かつ，学習データとは異なる 1,000 文を用いた．5 人の評価者に人手で因果関係を含むか否かを示すタグを付与してもらい，3 人以上が因果関係ありと判断したものを正例とし，そうでないものを負例とした．評価者は工学研究科に属し，自然言語処理の研究に従事する修士課程学生である．以下の条件に従ってタグ付けを行った．

- 因果関係は，原因若しくは理由と，その結果で構成されるものに限定する．

細かな条件は出さずに上記の条件だけでタグ付けを行ってもらった．

その結果，学習データ 1,000 文のうち 379 文が，評価データ 1,000 文のうち 365 文が因果関係ありと判定された．また，学習データと評価データ中の因果関係ありと判断された文における，手掛り表現ごとの文数をそれぞれ表 4 と表 5 に示す．ただし，1 文内に複数の手掛り表現が存在する場合があるため，文の合計数は多くなっている．各評価者間の κ 値を表 3 に示す．表 3 より，各 κ 値の平均が 0.65 であったため，各評価者間の一致度合は高く，評価者の間で因果関係に対しての判断に大きな差がないことが分かった．

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 [14] を用いた．学習器には SVM^{Light} (注2) を用いた．カーネルは線形を用いた．手掛り表現には，[4] で獲得された手掛り表現の中から 1 文中で因果関係を示す手掛り表現だけを取り出したものを用いる．取り出した結果，21 種類の手掛り表現を得ることができた．21 種類の手掛り表現を，表 6 に示す．

比較手法として，乾ら [15] と Chang ら [7] の手法を用いた．乾らの手法は，SVM によって因果関係を四つの意味（*cause*，*effect*，*means* and *precond*）に分類する手法である．本論文で対象としている原因（理由）と結果を示す因果関係が乾らの *cause* に相当すると考え，比較手法として用いた．乾らの手法は，表 7 の素性と，その他様々な条件や素性に基づいて分類を行っている．乾らの手法では，しきい値を設けて評価データからいくつかデータを取り除いているが，本タスクではそのような方法は用いていないため，この部分は実装しなかった．また，乾らの素性 3 については論文 [15] 中に詳細な記述がなかったことから，実装することができなかった．そのため，表 7 と素性 1，素性 2 を用いたものを乾らの手法とし，本実験で用いた．（乾らの素性 1，2，3 については乾らの論文を参照されたい [15]）

Chang らの手法は，名詞と彼らの作成したオントロ

表 3　評価者間の κ 値
Table 3　κ statistics between each evaluator.

評価者 A & 評価者 B	0.66	評価者 A & 評価者 C	0.65
評価者 A & 評価者 D	0.78	評価者 A & 評価者 E	0.60
評価者 B & 評価者 C	0.61	評価者 B & 評価者 D	0.68
評価者 B & 評価者 E	0.65	評価者 C & 評価者 D	0.63
評価者 C & 評価者 E	0.54	評価者 D & 評価者 E	0.71
		平均	0.65

表 4　学習データ内における手掛り表現ごとの文数
Table 4　The number of sentences including each clue phrases in learning data.

を背景に：15	を背景に、：3	を受け、：21	ため、：129
に伴う：16	に伴い、：8	を反映して：1	で、：113
をきっかけに：1	により、：8	に支えられて：0	によって：9
を反映し、：0	が響き、：4	ためで、：3	を受けて：13
から、：39	により：18	ためで：6	を受けて、：3
に伴い：16			

表 5　評価データ内における手掛り表現ごとの文数
Table 5　The number of sentences including each clue phrases in test data.

を背景に：22	を背景に、：8	を受け、：29	ため、：120
に伴う：18	に伴い、：9	を反映して：2	で、：92
をきっかけに：0	により、：8	に支えられて：0	によって：12
を反映し、：1	が響き、：2	ためで、：3	を受けて：17
から、：28	により：21	ためで：4	を受けて、：4
に伴い：15			

表 6　手掛り表現
Table 6　A list of clues.

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　ためで　を受けて、　に伴い

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

表 7 乾らが用いた素性一覧
Table 7 Features of Inui et al.

単語，EDR（動詞の概念），ALT-J/E（動詞のエントリ）， 日本語語彙大系（動詞の用語意味属性）， 助詞（ガ格の有無，ヲ格の有無），要素（ガ格要素，ヲ格要素）， テンス（節末がタ形であるかル型であるか）， アスペクト（節末に「· · · ている」を含むか否か）， ヴォイス（節末に「(· · · さ）れる」を含むか否か）， ヴォイス（節末に「(· · · さ）せる」を含むか否か）， 可能（節末に「· · · できる」を含むか否か）， 否定（節末に「(· · · し）ない」を含むか否か）

ジーを素性としてナイーブベイズ法により名詞句間に因果関係が存在するか否かを判定する手法である．本タスクに適用させるため，以下の式 (1) と式 (2) のように設定した．ここでは，c_j の j は 0 か 1 の値をとり，c_1 は因果関係が含まれる文を c_0 は因果関係を含まない文を表す．

$$c^* = \arg\max_{c_j} P(c_j|t_i) = \arg\max_{c_j} \frac{P(c_j)P(t_i|c_j)}{P(t_i)} \tag{1}$$

$$P(t_i|c_j) = P(CP_{t_i}|c_j)P(SP_{t_i}|c_j)\prod_{k=1}^{|t_i|} P(LP_{t_i k}|c_j) \tag{2}$$

ここで，t_i は文，CP_{t_i} は文 t_i に含まれる手掛り表現，SP_{t_i} は文 t_i に含まれるオントロジーの要素，LP は文 t_i に含まれる形態素とした．本実験では，彼らの作成したオントロジーの代わりに，拡張言語オントロジーを用いた．

本手法と乾らの手法，Chang らの手法との違いは以下のとおりである．表 7 より，乾らは EDR や日本語語彙大系の用言の体系を用いているが，本手法で用いた小林らの拡張言語オントロジーは日本語語彙大系の一般名詞意味体系に基づいて作成されている．その他の辞書についても，乾らは動詞に着目しているが，それに対して，我々は以下のような理由で名詞に着目している．学習データにおいて，原因，若しくは，その結果の少なくとも一方が名詞句である因果関係を含む文の数を数えたところ，因果関係を含む 379 文のうち 154 文であった．つまり，用言を素性として用いると，因果関係を含む 379 文のうち 154 文に関しては，原因，若しくは，その結果から素性を獲得することができない．それに対して，この 379 文中に，原因，若しくは，その結果の少なくとも一方が動詞のみで構成される因果関係を含む文は存在しなかった．そのため，名詞を用いると 379 文に含まれる全ての原因とその結果から素性を獲得することができる．以上の理由により，我々は名詞に着目した素性を用いた．また，本手法では素性として構文的な素性を用いているが，乾らや Chang らは用いていない．更に，本手法では自動的に追加学習データを獲得している点が乾らの手法と大きく異なっている．

表 8 結 果 一 覧
Table 8 Result of methods.

	適合率	再現率	F 値
本手法	0.802	0.753	0.777
本手法 + 追加	0.807	0.756	0.781
本手法 + 追加（均一）	0.773	0.822	0.797
乾らの手法	0.780	0.649	0.708
乾らの手法 + 追加	0.819	0.619	0.705
乾らの手法 + 追加（均一）	0.717	0.729	0.723
Chang らの手法	0.729	0.611	0.665
Chang らの手法 + 追加	0.749	0.449	0.562
Chang らの手法 + 追加（均一）	0.639	0.753	0.692

評価方法は以下の式 (3)，式 (4)，式 (5) のとおりである．

$$\text{適合率} = \frac{|P \cap G|}{|P|} \tag{3}$$

$$\text{再現率} = \frac{|P \cap G|}{|G|} \tag{4}$$

$$F\text{ 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} \tag{5}$$

ただし，

P：因果関係を含むと判定された文の集合

G：評価データ中の因果関係をもつ文の集合

4.2 評価結果と考察

学習データ 1,000 文を用いて，1995 年から 2005 年までの日経新聞記事中の文で学習データ，評価データに含まれない文から追加学習データを抽出したところ，正例として 556 文，負例として 8,197 文を抽出することができた．正例として抽出された 556 文を評価したところ，554 文が因果関係を含んでおり，適合率約 99%という結果になった．また，正例に数を合わせて抽出した負例 556 文を評価したところ，545 文が因果関係を含んでおらず，適合率約 98%という結果になった．本手法を用いれば，追加正例，追加負例ともに高い適合率でタグなしデータから抽出することができることが分かった．

本手法と比較手法の実験結果を表 8 に示す．また，表 9 に正しく因果関係を含むと判定された文をいくつか示す．表 8 において，追加は追加学習データを

表 9　正しく判定された因果関係を含む文
Table 9　Sentences that have causal relations.

八月からは豪州産牛肉にセーフガードがかかる可能性もあることから、米国産牛肉の輸入再開を求める声が改めて高まりそうだ。
大型液晶テレビの開発加速などを背景に韓国や台湾のフィルターメーカーからの需要が拡大しており、生産を強化する。
営業時間の延長に伴う人件費の増加で売上高販管費比率は○・八ポイント上昇。
客の多くがロッカーとカードの暗証番号を同じにしていたため、預金引き出しの被害に遭った。
ディーゼル車の排ガス規制開始で、更新時期を迎える前に使用できなくなる車両が大量に出た。
一株当たりの最終損益が市場予想を下回ったため、失望売りが出た。

学習用データに加えたものであり，追加（均一）は追加で獲得した正例 556 文に併せて，追加する負例の数を 556 文にしたものである．追加する負例は獲得した 8,197 文からランダムに抽出したものである．

表 8 より，「本手法 + 追加（均一）」が最も良い F 値となった．これにより，本手法の有用性を示すことができたと考えられる．ただ，適合率に関しては「乾らの手法 + 追加」が最も良い結果となった．しかしながら，「本手法 + 追加」と比べ，精度の差が 0.12 のみである．全体的な傾向として，追加学習データ（均一）を追加することにより，F 値が向上している．

また，追加学習データの正例数と負例数を同一にすることにより，全体的に再現率が向上している．これは，学習をするときに，獲得した負例の方が正例より 7,641 文も多いため，過学習になってしまい，適合率は向上するが再現率が下がってしまったと考えられる．

全ての手法において，追加学習データを用いることで性能が向上している．この結果より，文が因果関係を含むか否かを判定する際には，本手法により獲得した追加学習データを用いることで，性能の向上を図ることができる．

各素性の精度向上への寄与率を調べるために，表 8 にある「本手法」において，各素性を一つずつ除いた場合の精度・再現率を求めた．その結果を表 10 に示す．表 10 と表 8 より，形態素バイグラムの寄与が一番高いことが分かった．手掛り表現や拡張言語オントロジーでは，これらの素性を含めない場合の F 値が含める場合に比べて，それぞれ極めて小幅だが向上した．しかしながら，エラー数を調べたところ，「本手法」では評価データ 1,000 文のうち 158 件あったのに対し，手掛り表現を含めない場合のエラーが 156 件，拡張言語オントロジーを含めない場合のエラーが 155 件であり，誤差の範囲であることが分かった．

また，更に各素性の寄与率を調べるために，各評価者がタグ付けしたデータそれぞれで実験を行い，そのときの性能を調べた．その結果を表 11 に示す．表 11 は，全ての素性を使った場合の F 値から，それぞれ各素性を含めない場合の F 値を引いた値を表示している．表 11 では，形態素ユニグラムはユニグラム，形態素バイグラムはバイグラムと，拡張言語オントロジーはオントロジーと，それぞれ省略して記述している．各 A〜B は評価者 A〜B を意味している．

表 10　各素性を含めない場合の適合率・再現率
Table 10　Results of methods that do not include each features.

	適合率	再現率
形態素ユニグラム	0.795	0.753
形態素バイグラム	0.793	0.726
直前の品詞	0.791	0.745
手掛り表現	0.803	0.759
助詞のペア	0.801	0.751
拡張言語オントロジー	0.804	0.762

表 11　各評価者がタグ付けしたデータにおける結果
Table 11　Each annotator's results.

	A	B	C	D	E
ユニグラム	−0.007	−0.001	0.002	0.009	0.009
バイグラム	0.03	0.02	0.022	0.009	0.001
直前の品詞	0.013	0.006	0.004	0.003	−0.002
手掛り表現	−0.005	0.017	0.007	−0.019	−0.009
助詞のペア	−0.003	0.001	0.004	0.001	−0.004
オントロジー	0.003	0.003	0.009	−0.01	−0.006

表 11 より，全ての評価者のデータにおいて，形態素バイグラムが精度に寄与していることが分かった．また，評価者 5 人中 3 人以上において，形態素ユニグラム，直前の品詞，助詞のペアと拡張言語オントロジーの精度への寄与が確認された．しかしながら，形態素バイグラム以外の素性に関して，大きな寄与が確認されなかったため，それらの素性の使い方を考える必要がある．

5. エラー解析

各本手法の False Positive と False Negative の数を表 12 に示す．表 12 より，エラー数が最も少なかったのは，「本手法 + 追加（均一）」であった．また，「本手法 + 追加（均一）」の場合のみ，FP の数が FN の数より多くなっている．

表 12　False Positive と False Negative の数
Table 12　Numbers of false positives and false negatives.

	FP	FN	合計
本手法	68	90	158
本手法 + 追加	66	89	155
本手法 + 追加（均一）	88	65	153

5.1 False Positive

以下に，False Positive をいくつか示す．

> 自動車・家電用部品事業では車載電装品の販売が堅調で、洗濯機用電子制御基板も好調に推移。

上記の例では，並列を示す素性助詞のペア「(が，も)」が含まれていたが，「堅調」や「好調」といった学習データ中の正例に多く現れた語が含まれていたため，False Positive となってしまった．

> 石原は連合福島の支援を受けて大票田の福島市や伊達郡を重点に巻き返す。

上記の例では，手掛り表現「を受けて」の直前の形態素「支援」が正の重みが大きい素性である名詞–サ変接続であったため，正への重みが大きくなり，False Positive になってしまっていた．直前形態素の素性と手掛り表現の共起を素性にすることにより，手掛り表現ごとに因果関係をもっている場合の直前形態素を学習できるのではないかと考えている．

5.2 False Negative

以下に，False Negative をいくつか示す．

> 日銀は金融機関への資金払い戻しとなる資金吸収オペの期日を大量に設定するなどで、当座預金残高を約五兆円に維持した。

上記の例では，正に重みが大きい直前形態素の素性「助詞–副助詞」が含まれていたが，False Negative となっていた．これは，負に重みが大きいオントロジー素性「年月日::null」や，負の重みが付与されている形態素「資金」「預金」「金融」などが含まれていたためである．この問題を解決するためには，学習データを更に増やして，学習するときにこれらの形態素に負の重みが付かないようにする必要がある．

表 13　本手法における手掛り表現ごとの精度・再現率
Table 13　Precisions and recalls of each clue expressions by our method.

	精度	再現率		精度	再現率
を背景に	0.96	1.0	を背景に、	1.0	1.0
を受け、	0.72	0.97	ため、	0.90	0.99
に伴う	0.75	1.0	に伴い、	1.0	1.0
を反映して	1.0	1.0	で、	0.76	0.42
をきっかけに	N/A	N/A	により、	0.78	0.88
に支えられて	N/A	N/A	によって	0.53	0.67
を反映し、	1.0	1.0	が響き、	1.0	1.0
ためで、	1.0	1.0	を受けて	0.55	0.94
から、	0.58	0.75	により	0.66	0.90
ためで	1.0	1.0	を受けて、	1.0	1.0
に伴い	0.83	1.0			

> 四国四県都を結ぶ高速道路の全通や愛媛県南西部（南予）地域の大洲市までの松山自動車道の延伸効果で、南予地域への観光客が大幅に伸びた。

上記の例では，正の重みをもつ形態素「効果」や形態素バイグラム「効果–で」が含まれていたが，負に重みをもつ助詞のペア「の–へ」が含まれていたため，False Negative となってしまっていた．

5.3 手掛り表現ごとの精度・再現率

「本手法 + 追加（均一）」を用いたときの，手掛り表現ごとの精度・再現率を表 13 に示す．

表 13 より，精度・再現率ともに 1.0 となっている手掛り表現がいくつか存在した．これらの手掛り表現を含む文は全て因果関係を含んでいた．もし，精度良く因果関係を含む文を抽出したいのであれば，これらの手掛り表現のみを用いればよいと考えられる．しかしながら，これらの手掛り表現が含まれている文は 25 文しかなく，数多くの因果関係を網羅することができないという問題がある．

また，乾ら，Chang らの手法における手掛り表現ごとの精度・再現率を表 14 と表 15 に示す．表 13，表 14，表 15 より，本手法における手掛り表現「で、」の精度が乾らや Chang らの手法より高くなっている．また，本手法における手掛り表現「を受け、」の精度が，乾らや Chang らの手法に比べ，低くなっているが，再現率に関しては非常に高くなっている．これらの手掛り表現に関しては，F 値で評価したとき，本手法が乾らや Chang らの手法に比べ優位であることが分かった．

表 14　乾らの手法における手掛り表現ごとの精度・再現率
Table 14　Precisions and recalls of each clue expressions by Inui's method.

	精度	再現率		精度	再現率
を背景に	0.95	0.86	を背景に、	1.0	1.0
を受け、	0.88	0.48	ため、	0.91	0.96
に伴う	0.83	0.83	に伴い、	1.0	0.67
を反映して	1.0	1.0	で、	0.48	0.41
をきっかけに	N/A	N/A	により、	0.86	0.75
に支えられて	N/A	N/A	によって	0.41	0.58
を反映し、	1.0	1.0	が響き、	1.0	1.0
ためで、	1.0	1.0	を受けて	0.54	0.76
から、	0.56	0.82	により	0.64	0.76
ためで	1.0	1.0	を受けて、	1.0	0.75
に伴い	0.85	0.73			

表 15　Chang らの手法における手掛り表現ごとの精度・再現率
Table 15　Precisions and recalls of each clue expressions by Chang's method.

	精度	再現率		精度	再現率
を背景に	1.0	1.0	を背景に、	1.0	1.0
を受け、	0.94	0.52	ため、	0.92	0.81
に伴う	0.8	0.89	に伴い、	1.0	0.78
を反映して	1.0	1.0	で、	0.36	0.64
をきっかけに	N/A	N/A	により、	0.83	0.63
に支えられて	N/A	N/A	によって	0.43	0.5
を反映し、	1.0	1.0	が響き、	1.0	0.5
ためで、	1.0	1.0	を受けて	0.54	0.76
から、	0.65	0.71	により	0.62	0.76
ためで	1.0	1.0	を受けて、	1.0	0.5
に伴い	0.87	0.87			

6. 関連研究

Inui ら [8] は二つの事象間に因果関係が存在するか否かを自動的に判断するモデルの提案をしている．しかしながら，Inui らの手法の実験においては，手掛り表現「ので」と「ため」の前後に出現する動詞句に含まれる動詞を対象として実験を行っている．そのため，手掛り表現「ので」「ため」以外の手掛り表現に適用できるか否かは示されていない．また，Inui らの実験で得られた精度より，本実験で得られた精度の方が良い結果となった．これは，Inui らの手法は，動詞間に因果関係が含まれている否かを判定するものとなっており，情報として動詞しか用いていない．それに対して，本研究では文全体の情報（構文的な素性など）を利用できる．このことから，本研究よりも Inui らの研究の方が難しい問題設定を用いている．以上のことから，本研究で得られた精度が Inui らの研究で得られた精度より，高くなったと考えている．

Bethard らは Syntactic 素性と Semantic 素性を用いて，動詞対に対して因果関係があるか否かの判定を行う手法を提案している [9]．彼らが用いている Semantic 素性には WordNet を利用している [16]．

上記で述べた研究では，名詞句や動詞句に限って因果関係を抽出している．しかしながら，因果関係を構成する原因・結果表現は名詞句や動詞句のみとは限らない．本手法では，手掛り表現を対象に因果関係を表す意味かどうかを判定しているため，動詞句でも名詞句でも判定することが可能である．また，自動的に学習データを増やすことを本手法の新規性として挙げることができる．学習データを増やすことにより，F 値が向上することが確認できた．更に，本手法だけではなく，いくつかの既存手法に適用することでも F 値が向上することが確認でき，汎用性もある．

半教師あり学習を用いた手法として，Blum et al. [17] や Yarowsky [18] などの研究がある．Blum et al. は，少数のラベル付けされたデータと，二つの学習器を用いてラベルの付与されていないデータにラベルを付与する手法である Co-Training を提案している．Yarowsky は，言語がもつ特徴（連語や談話は一つだけの意味をもつ）と，ラベルの付与されていないデータを用いて語の曖昧性解消を行う手法を提案している．それに対して本手法の特徴は，二つの学習器や言語がもつ特徴に基づいてラベルが付与されていないデータを扱うのではなく，手掛り表現がもつ因果関係という意味に基づいて追加学習データを獲得する点にある．

7. むすび

手掛り表現を含む文が因果関係を含むか否かを判定する手法を提案した．素性には意味的な素性として拡張言語オントロジー，構文的な素性として助詞のペアを用いた．また，学習データを自動的に増やし，半教師あり学習を実行することにより，性能が向上した．評価実験の結果，「本手法 + 追加（均一）」のときに F 値 0.797 を達成した．

今後の課題として，手掛り表現と各素性を関連づけて学習することにより，より性能が向上することが期待できる．今回は，作成された拡張言語オントロジーをそのまま用いたため，意味カテゴリーの分類がされておらず，これを利用することができなかった．作られた拡張言語オントロジーの意味カテゴリーをグループ化することで，より効果的に拡張言語オントロジーを素性として利用できると考えている．そのため，拡

張言語オントロジーの意味カテゴリーの分類を行う必要がある．

また，本手法をフィルタリングに用いることで，既存の因果関係抽出手法の精度向上に役立てることができる．本手法は，動詞句や名詞句に限らず用いることができるため，様々な手法に適用することができ，更に自動的に学習データを増やすことができるため，少ない学習データだけでも利用可能である．

謝辞 ニューヨーク大学関根聡研究准教授には，本研究に関して有益なご助言を頂いた．本研究は文部科学省グローバル COE プログラム「インテリジェントセンシングのフロンティア」，日本学術振興会科研費（C）22500129，人工知能研究振興財団，電気通信普及財団の支援を受けた．

文　献

[1] R. Higashinaka and H. Isozaki, “Corpus-based question answering for why-questions,” Proc. IJCNLP, pp.418–425, 2008.

[2] 石井裕志，馬　強，吉川正俊，“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] 青野壮志，太田　学，“要因検索による因果関係ネットワークの構築と因果知識の獲得，” Forum on Data Engineering and Information Management, 2010.

[4] 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.

[5] R. Girju, “Automatic detection of causal relations for question answering,” ACL Workshop on Multilingual Summarization and Question Answering, pp.76–83, 2003.

[6] C.S. Khoo, J. Kornfilt, R.N. Oddy, and S.H. Myaeng, “Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing,” Literary and Linguistic Computing, vol.13, no.4, pp.177–186, 1998.

[7] D.-S. Chang and K.-S. Choi, “Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities,” Inf. Process. Manage., vol.42, no.3, pp.662–678, 2006.

[8] T. Inui, H. Takamura, and M. Okumura, “Latent variable models for causal knowledge acquisition,” 8th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2007), pp.85–96, 2007.

[9] S. Bthard and J.H. Martin, “Learning semantic links from a corpus of parallel temporal and causal relations,” Proc. ACL-08, pp.177–180, 2008.

[10] 庵　功雄，新しい日本語学入門，スリーエーネットワーク，2001.

[11] 坂本大祐，坂地泰紀，酒井浩之，増山　繁，“企業業績要因文における因果関係の有無判定手法の提案，” 言語処理学会第 15 回年次大会，pp.925–928, 2009.

[12] 小林暁雄，増山　繁，関根　聡，“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，” 信学論（D），vol.J93-D, no.12, pp.2597–2609, Dec. 2010.

[13] 池原　悟，宮崎正弘，白井　諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林　良彦，日本語語彙大系，岩波書店，1997.

[14] 工藤　拓，松本裕治，“チャンキングの段階適用による日本語係り受け解析，” 情処学論，vol.43, no.6, pp.1834–1842, 2002.

[15] 乾　孝司，乾　健太郎，松本裕治，“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，” 情処学論，vol.45, no.3, pp.919–933, 2004.

[16] C. Fellbaum, WordNet, An Electronic Lexical Database, The MIT Press, 1998.

[17] A. Blum and T. Mitchell, “Combining labeled and unlabeled data with co-training,” COLT: Proc. Workshop on Computational Learning Theory, pp.92–100, 1998.

[18] D. Yarowsky, “Unsupervised word sense disambiguation rivaling supervised methods,” Proc. 33rd Annual Meeting of the Association for Computational Linguistics, pp.189–196, 1995.

（平成 22 年 11 月 18 日受付，23 年 3 月 21 日再受付）

坂地　泰紀

2009 豊橋技術科学大学大学院工学研究科修士課程知識情報工学専攻了．現在，同大学院工学研究科博士後期課程電子・情報工学専攻在学中．自然言語処理の研究に従事．

増山　繁　（正員）

1977 京大・工・数理卒．1982 同大大学院博士後期課程単位取得退学．1983 同修了（工博）．1982 日本学術振興会奨励研究員．1984 京都大学工学部数理工学科助手．1989 豊橋技術科学大学知識情報工学系講師．1990 同助教授．1997 同教授．2010 同大大学院工学研究科情報・知能工学専攻教授．2005 同大学インテリジェントセンシングシステムリサーチセンター教授併任．2010 同大学人間・ロボット共生リサーチセンター教授併任．アルゴリズム工学，特に，並列アルゴリズム等，及び，自然言語処理，特に，テキスト自動要約等の研究に従事．言語処理学会，情報処理学会等会員．