人工知能学会研究会資料
SIG-FIN-011-09

企業業績発表記事からの因果関係抽出

Extraction of Causal Knowledge from Articles Concerning Business Performance of Companies

坂地泰紀 [1*]　酒井浩之 [1]　増山繁 [2]
Hiroki Sakaji[1]　Hiroyuki Sakai[1]　Shigeru Masuyama[2]

[1] 成蹊大学 理工学部 情報科学科
[1] Department of Computer and Information Science, Faculty of Science and Technology
[2] 豊橋技術科学大学 大学院 工学研究科 情報・知能工学専攻
[2] Department of Computer Science and Engineering, Toyohashi University of Technology

Abstract: This paper proposes a method that extracts causal knowledge from Japanese financial articles concerning business performance of companies via clue expressions. Our method decides whether a sentence includes causal knowledge or not when the method extracts it. For example, a sentence fragment "World economy recession due to the subprime loan crisis ..." contains causal knowledge in which "World economy recession" is an effect phrase and "the subprime loan crisis" is its cause phrase. These relations are found by clue phrases, such as "ため (*tame*: because)" and "により (*niyori*: due to)". We found that some specific syntactic patterns are useful to improve accuracy of extracting causal knowledge. Therefore, our method can extract causal knowledge accurately.

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の支援を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おこな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だけでなく，その業績要因が重要である．なぜなら，業績拡大の要因が，その企業の主力事業が好調であることであったならば株価への影響は大きいが，株式売却益の計上などの特別利益の計上が要因であるならば株価への影響は軽微であるからである．しかしながら，証券市場の上場企業数は約 3,500 社と多いうえに，近年では年に 4 回，決算発表がある．さらに，大幅な業績の修正を行う場合にも業績修正発表を行う必要があるため，人手によって全ての企業の業績要因を取得するのには多大な労力を要する．そのため，酒井らは，経済新聞記事から企業の業績発表記事を抽出し，その中から業績要因（例えば,「主力の半導体製造装置の受注が好調」）を抽出する手法を提案した [Sakai 08].

*連絡先：成蹊大学理工学部情報科学科
〒 180-8633 東京都武蔵野市吉祥寺北町 3-3-1
E-mail: hiroki_sakaji@st.seikei.ac.jp

しかしながら，酒井らの手法では,「暖冬により暖房用燃料の販売が低調だった。」という業績要因を含む文に出現する原因「暖冬」，結果「暖房用燃料の販売が低調だった。」という因果関係を示す表現を抽出することができない．そのため，業績要因に加えて，原因と結果の対として因果関係を経済新聞から抽出することで，個人投資家に対して多くの重要な情報を提供できるようになる．例えば，原因「猛暑」，結果「冷房需要の盛り上がり」などの因果関係を提示することで,「猛暑」の場合には,「冷房需要」が高まる可能性があるという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．過去の業績要因記事から因果関係を大量に入手しておくことにより，上記のような対応が可能となる．そこで，本研究では，業績発表記事から因果関係を抽出する手法の開発を行う．

本研究では，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，業績発表記事から因果関係を自動的に抽出する手法を提案する．文献 [庵 12] 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．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

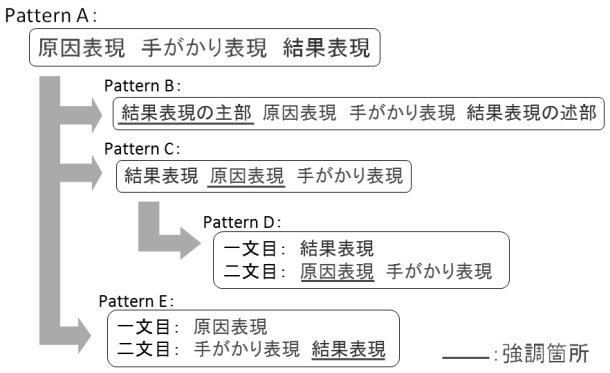

図 1: 各 Pattern の関連図

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．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．我々は，経済新聞記事を調査することにより，手がかり表現と原因・結果表現の出現位置を 5 通りに分類した [Sakaji 08]．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．具体的な各 Pattern に対応する抽出アルゴリズムは，文献 [Sakaji 08] を参考にされたい．

図 1 において，我々は Pattern A は基本型であると考えた．Pattern B は，基本型から結果の主部が強調のため文頭へ移動したものである．Pattern C は，結果を強調するため基本型を倒置したものである．Pattern D と E は一文にすると長くなるので，原因と結果を 2 文に分割したのものである．Pattern A を分割したものが，Pattern E であり，Pattern C を分割したものが Pattern D となっている．また，Pattern D と E では，それぞれ，手がかり表現を含む文が強調されるようになっている．

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．)．

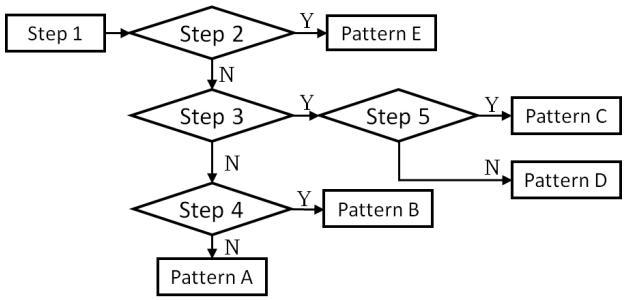

図 2: Pattern 識別の概要

[*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． □

例えば,「暖冬により暖房用燃料の販売が低調だった。」という文の場合は，まず Step 1 で手がかり表現「により」でこの文を発見することができる．次に，Step 2 で手がかり表現が文頭に存在しないため，Step 3 へ行く．Step 3 では，手がかり表現に句点が含まれていないので，Step 4 へ行く．最後に，この文の基点文節は,「低調だった。」という動詞句であるが，基点文節に係っている文節の中に係り助詞，もしくは，格助詞を含む文節が存在しないため，Pattern A が適用される．

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 (SVM) 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
- 文に含まれる手がかり表現
- 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 [小林 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．次節では，拡張言語オントロジーの素性への適用について述べる．それ以外の素性の抽出に関しては，[坂地 11] を参照されたい．

3.1　拡張言語オントロジー

本研究では，小林ら [小林 10] 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 [池原 97] 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，後部語彙共に 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辿っていくと，6 階層目は「食料」という意味カテゴリであり，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．さらに，6 階層目の意味カテゴリ数 256 に比べ，7 階層目の意味カテゴリ数は 536 と多く，本手法では意味カテゴリの組み合わ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，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み合わせを素性として抽出する．

素性の抽出例を図 3 に示す．図 3 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，係り受け情報を持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「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「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．

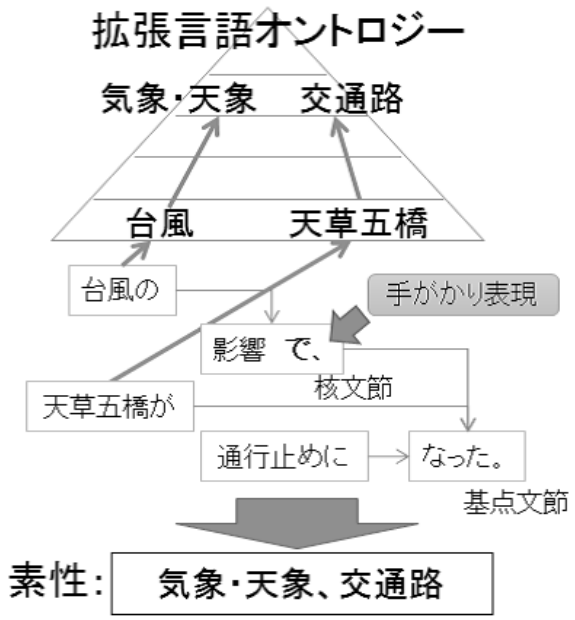

図 3: 拡張言語オントロジー素性の取得例

3.2 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 4 に示す．

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．

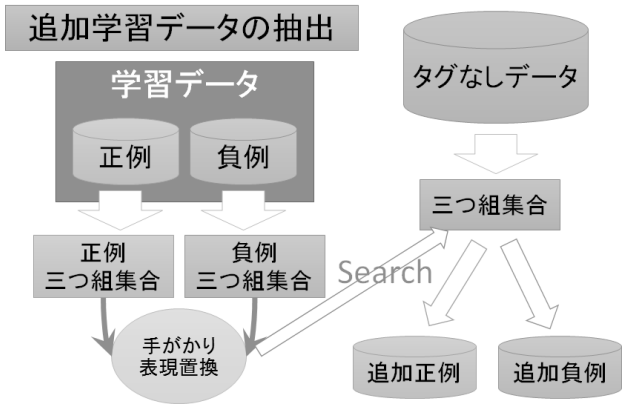

図 4: 追加学習データの取得

追加学習データを獲得する手続き *Extracting additional learning data*[坂地 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，[坂地 11] を参照されたい．

[*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: $P \cap T$ を正例の追加学習データとして獲得する．$N \cap T$ を負例の追加学習データとして獲得する． □

4 評価実験

1990 年から 2008 年の日経新聞記事集合から 71,070 個の業績発表記事を取得し，その記事集合から，手がかり表現を用いて因果関係の抽出を行った．形態素解析器としては Mecab[1] を用い，係り受け解析器としては Cabocha[工藤 02] を用いた．学習器には SVM^{Light}[2] を用い，カーネルは線形を用いた．手がかり表現には，[Sakaji 08] で獲得された手がかり表現を用いた．実験に用いた手がかり表現を，表 2 に示す．

[1] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[2] http://svmlight.joachims.org/

表 2: 手がかり表現

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

4.1 評価結果

71,070 個の業績要因記事から因果関係を抽出したところ，109,050 個の因果関係を抽出した．抽出した 109,050 個の因果関係からランダムに 200 個の因果関係を選び，精度を評価したところ，83%であった．さらに，200 個の因果関係において，パターン別の抽出数と精度を算出した．その結果を，表 3 に示す．また，抽出した因果関係の例を表 4 に示す．

表 3: パターン別の抽出数と精度

	A	B	C	D	E
抽出数 (全)	79,966	12,031	1,105	13,303	2,645
抽出数 (200)	142	21	3	28	6
精度	0.85	0.76	0.33	0.79	1.00

4.2 考察

本手法は，精度 83%と高い値を達成することができた．これは，業績要因記事がある種定型的に記述されていることに起因すると考えられる．本手法は，Pattern A から E という手がかり表現と原因・結果表現の位置により，抽出方法を変更するというものであるので，文体の影響を受けやすい．そのため，定型的な文書には適用しやすいという特徴があることから，高い精度となったと考える．

係り受け解析ミスによる抽出ミスがいくつか見られた.「店舗増床、改装に伴う償却負担で経常利益は同三三・三%減の三億八千百万円。」という文から，原因表現「改装」，結果表現「焼却」という因果関係が抽出されていたが，原因表現としては「店舗増床、改装」を抽出すべきであった．これは，係り受け解析器が並列関係をうまく解析できなかったため，起こったエラーである．一方,「原油や古紙、木材チップの価格高騰を背景に原材料コストが膨らむ」という文から，原因表現「原油や古紙、木材チップの価格高騰」，結果表現「原材料コストが膨らむ」と正しく抽出できている例もある．

エラーの解析を進めると，Pattern A から E を作成するときには，隣り合った 2 文間にしか因果関係を見つけることができなかったが，企業の業績発表記事では間に 1 文はさんだ，1 文飛ばしの形で因果関係が存在することが分かった．例えば,「セキテクノトロンの二〇〇二年三月期の経常損益は二億二千万円程度の赤字（前期は三億七千百万円の黒字）になる見通し。従来予想は一億四千万円の黒字。半導体市況の低迷で主力の半導体製造装置の販売不振が響く。」という文では，1 文目に結果表現が存在し，3 文目に原因表現が存在する．上記の例のように，2 文目が 1 文目の補足説明となっているというものが業績要因記事特有に存在するものと考える．企業業績要因記事において，3 文に渡って因果関係が存在する場合，エラー解析した中では，2 文目は「従来予想〜」と始まるものだけであったため，この場合のみ抽出方法を変更することで対応可能である．

表 3 より，Pattern C の精度が 0.33 と他の Pattern と比べ，低い値となっている．しかしながら，Pattern C の抽出数が 3 と少なく，正しく評価できていない可能性があるため，今後の課題として評価データ数を増す必要がある．

5 関連研究

Khoo らは人手で作成したパターンを用いて，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 [Khoo 98, Khoo 00] 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の種類が 35 種類と豊富であり，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．

Chang らは手がかり表現と語の組の出現確率を用いて，2 つの名詞句間の因果関係を抽出する手法を提案している [Chang 06]. また，Girju は手がかり表現に基づいて自動的に WordNet[Fellbaum 98] に含まれる名詞句間の因果関係の検出と抽出を行う手法を提案している [Girju 03]. 彼らの研究は名詞句の組を因果関係の対象としているため，他の表現間の因果関係を抽出することができないが，本手法では名詞句だけでなく動詞句や文をも対象としている．

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

表 4: 抽出された因果関係の例

抽出対象文	ガラスバルブなどの増産効果で、営業利益は三四％増となった。
Pattern	A
手がかり表現	で、
原因表現	ガラスバルブなどの増産効果
結果表現	営業利益は三四％増となった
抽出対象文	主力業態のロイヤルホストでは、サービスの徹底や新規メニューの導入 により、これまで低下傾向にあった客単価が上昇に転じた。
Pattern	B
手がかり表現	により、
原因表現	サービスの徹底や新規メニューの導入
結果表現の主部	主力業態のロイヤルホストでは
結果表現の述部	これまで低下傾向にあった客単価が上昇に転じた
抽出対象文	最終益が前の期に比べて大幅に伸びたのは、栃木県佐野市にある土地を売却した結果、 法人税や住民税などの支払いが減少したため。
Pattern	C
手がかり表現	ため。
原因表現	木県佐野市にある土地を売却した結果、法人税や住民税などの支払いが減少した
結果表現	最終益が前の期に比べて大幅に伸びた
抽出対象文	従来の水平型から大容量の「垂直磁気記録方式」への移行に苦戦し、 生産工場の本格立ち上げが遅れているため。
Pattern	D
手がかり表現	ため。
原因表現	従来の水平型から大容量の「垂直磁気記録方式」への移行に苦戦し、 生産工場の本格立ち上げが遅れているため。
結果表現	利益率の低いペンタックスの新規連結が主因だが、 ハードディスク駆動装置（ＨＤＤ）用ガラスディスクの採算悪化も足かせとなる。
抽出対象文	このため、年間配当を二円減らして五円にする。
Pattern	E
手がかり表現	このため、
原因表現	東海染は個人消費の低迷で売り上げが減少、九四年三月期は五億円の営業赤字になる見通し。
結果表現	年間配当を二円減らして五円にする。

う手法を提案している [Bthard 08]．彼らが用いている Semantic 素性には WordNet[Fellbaum 98] を利用している．

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，手がか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．

6 むすび

本研究では，経済新聞記事における企業の業績発表記事を対象としたテキストマイニングの一環として，業績発表記事から因果関係の抽出を行った．手がかり表現と 5 つのパターンを用いた抽出手法で，精度 83%と高い値を示した．今後の課題として，抽出した因果関係を活用するために，原因・結果表現にどのような種類の語が多く含まれるかなどの傾向分析を行う必要があると考える．

参考文献

[Bthard 08] Bthard, S. and H.Martin, J.: Learning Semantic Links from a Corpus of Parallel Temporal and Causal Relations, in *in Proceedings of ACL-08*, pp. 177–180 (2008)

[工藤 02] 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, 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842 (2002)

[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)

[Fellbaum 98] Fellbaum, C.: *WordNet: An Electronic Lexical Database*, The MIT Press (1998)

[Girju 03] Girju, R.: Automatic detection of causal relations for Question Answering, in *In ACL Workshop on Multilingual Summarization and Question Answering*, pp. 76–83 (2003)

[Khoo 98] Khoo, C. S., Kornfilt, J., Oddy, R. N., and Myaeng, S. H.: 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)

[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)

[Sakai 08] Sakai, H. and Masuyama, S.: Cause Information Extraction from Financial Articles Concerning Business Performance, IEICE Trans, *IEICE Trans. Information and Systems*, Vol. E91-D, No. 4, pp. 959–968 (2008)

[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)

[庵 12] 庵 功雄：新しい日本語学入門（第 2 版）, スリーエーネットワーク (2012)

[坂地 11] 坂地 泰紀, 増山 繁：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, Vol. J94-D, No. 8, pp. 1496–1506, (2011)

[小林 10] 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, 電子情報通信学会論文誌 D, Vol. J93-D, No. 12, pp. 2597–2609 (2010)

[池原 97] 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)