ニュース記事の内容と構造特徴を考慮した因果関係マイニング

野島　裕輔† 　馬　　強†† 　吉川　正俊††

† 京都大学工学部情報学科　〒 606–8501 京都市左京区吉田本町
†† 京都大学大学院情報学研究科　〒 606–8501 京都市左京区吉田本町
E-mail: †nojima@db.soc.i.kyoto-u.ac.jp, ††{qiang,yoshikawa}@i.kyoto-u.ac.jp

あらまし　本稿では，ニュース記事から因果関係を抽出する手法を提案する．ニュース記事における因果関係の記述には，結果事象の記述が省略されたものが存在する．このような記述からは既存の手法では因果関係の抽出を行うことができない．そこで，本研究では，ニュース記事の構造的特徴を利用した因果関係抽出手法を提案する．提案手法では，ニュース記事のタイトルと因果関係の関係，ニュース記事の主題と結果事象の一致性，そして原因記述同士の類似性に着目し，関連記事集合を一つの単位として，因果関係を抽出する．また，提案手法に関する評価について論じる．

キーワード　因果関係，テキストマイニング，因果関係ネットワーク

Causal Relation Mining by Utilizing Contents and Structural Features of News Articles

Yusuke NOJIMA†, Qiang MA††, and Masatoshi YOSHIKAWA††

† School of Informatics and Mathematical Science, Faculty of Engineering, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto 606–8501 Japan
†† Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto 606–8501 Japan
E-mail: †nojima@db.soc.i.kyoto-u.ac.jp, ††{qiang,yoshikawa}@i.kyoto-u.ac.jp

1. 序　　論

事象と事象との間の因果関係を理解することは，意思決定を行う際において，非常に重要である．我々は，新聞やテレビなどで報じられるニュースから，因果関係に関する知識を得ている．しかし，事象の中には，多くの因果関係が複雑に絡み合って発生しているものも存在し，このような関係を理解することは必ずしも容易ではない．我々は，事象間の因果関係を，ネットワークの形に整理し，ユーザに提示して，ユーザの意思決定を支援する技術について研究を行っている．

因果関係は原因事象と結果事象の組で表現できる．原因事象と結果事象は，それぞれ因果関係の原因と結果となる事象である．また，本研究では，原因事象と結果事象を自然言語で表現したものを原因記述と結果記述と呼ぶ．

従来から因果関係の抽出や組織化に関する研究は数多く行われている [1]〜[5]．これらの研究では，言語的な特徴として手がかり表現を用いて因果関係を発見して抽出する手法を多く提案している．しかしながら，ニュース記事には，因果関係に関する記述の一部が欠落している場合があり，このような場合，従来手法は因果関係の抽出に不十分である．例として，次のニュース記事を考える．

> **ガソリン価格、５カ月ぶり下げ止まり**
> 前週まで 19 週連続で値下がりしていたガソリンの店頭価格が約 5 カ月ぶりに下げ止まった。（中略）原油価格が投機マネーの流入などで反発したことが背景だ。

このニュース記事において,「原油価格が投機マネーの流入などで反発したことが背景だ。」は，因果関係の原因記述である．この因果関係の結果である，ガソリンの店頭価格が下げ止まったという事象は，冒頭で説明されているため，最後の文では繰り返されていない．

多くの応用では，因果関係の出現回数は，より有用な因果関係を発見するのに重要である．例えば，因果関係を可視化する際において，大量の因果関係を全てユーザに提示するのではなく，重要度を因果関係の出現回数に基いて計算し，重要度が高い因果関係だけをユーザに提示するという応用が考えられる．しかしながら，従来手法は，因果関係を網羅的に抽出することよりも，因果関係を正確に抽出することを重視しており，抽出

される因果関係の網羅性に問題があるため，正確に因果関係の出現回数を計算することができない．そこで，本研究では，ニュース記事の構造的特徴を考慮して，因果関係を抽出する手法を提案する．

ニュース記事は雑多な文の集合体ではなく，そのニュース記事が取り上げる主要な事象 (主題と呼ぶことにする) に対する説明を綴ったものであるため，あるニュース記事に現れる因果関係は，主題と関係したものであることが多い．さらに，ニュース記事の主題は，そのニュース記事のタイトルに記述されている場合が多いことも観測された．例えば，先のニュース記事の例では,「ガソリン価格、５カ月ぶり下げ止まり」が主題の記述になっている．これにより，本文で結果記述が省略されている場合でも，そのニュース記事のタイトルから，主題を結果事象して抽出することで，因果関係を抽出することができる．

ニュース記事の関連記事集合，すなわち主題を共有するニュース記事の集合において，同じ関連記事集合内に存在する因果関係は，同じ因果関係を表現している可能性が高い．特に，同じ関連記事集合内に存在する原因記述は，同じ原因事象を表現している可能性が高い．提案手法では，この性質を利用し，原因記述の抽出時に，原因記述の類似性を利用し，手掛り表現だけでは原因記述か否かがはっきりしない記述に対して，他の類似する記述の原因記述のスコアを伝播させることによって，識別を行う．

提案手法は，与えられた関連記事集合を，タイトルと本文に分け，タイトルから主題を，本文から原因事象をそれぞれ抽出し，原因事象から主題への因果関係を出力として返す．

本稿では，第 2 節で，因果関係抽出における先行研究を紹介し，提案手法との差異について議論する．第 3 節で，因果関係の抽出アルゴリズムと，そのアルゴリズム内で使用する識別器の構築法を議論する．第 4 節で，事象の同一性と同一関連記事集合内の原因記述の類似性に関する予備実験の結果を示した後，第 3 節で議論した原因記述の識別器の性能と，因果関係の抽出アルゴリズムの性能を評価する．第 5 節では，提案手法と実験結果についてまとめ，今後の課題を述べる．

2. 関 連 研 究

因果関係抽出手法の主流は，手がかり標識，手がかり表現などと呼ばれる，因果関係の存在を示す表現や，文の接続関係を利用することで，因果関係を含む文を抽出し，その文から因果関係を抽出する手法である [1]〜[4].

佐藤ら [1] の手法では，まず，文書を「手がかり標識」を用いて単文に分解する．各単文が一つの事象を表していると考え，文末の形態素を使って次の単文との因果関係の有無を判定する．さらに,「必ず」「たぶん」といった形態素から，事象間の結びつきの強さを調べる．事象は，単文から抽出された重要単語で表現される．手がかり標識は予め与えられるものとしている．

乾ら [2] の手法では，必然性の高い因果関係を明示的に表現する接続標識「ため」に着目し，複文形式を持つ文から因果関係を抽出している．さらに，抽出した因果関係を cause, precond, effect, means の 4 つに分類する方法を述べている．

坂地ら [3] の手法では，まず，文書集合から手がかり表現の候補にスコアを付けることによって，手がかり表現を獲得する．そして，文書集合から手がかり表現を含む文を抽出する．抽出された文が 4 つの構文パターンのうちどれに相当するのかを判定し，それぞれのパターン毎に決められたルールにより，根拠表現と結果表現を抽出する．

酒井ら [4] の手法では，手がかり表現とは原因表現によって頻繁に修飾される表現であると定義し，業績に関する記事の集合から統計的に手がかり表現と原因表現を抽出する.「が好調」「が不振」を最初の手がかり表現とし，これらの句を頻繁に修飾する表現 (頻出表現) を抽出する．さらに，頻出表現によって頻繁に修飾されている句を抽出することにより，手がかり表現を拡張する．このように，手がかり表現と頻出表現を交互に拡張していくことによって多数の手がかり表現を取得し，手がかり表現と頻出表現を使うことで原因記述を抽出する．

これらの手法では，序論で述べたような不完全記述は考慮せず，基本的に一つの文を単位に因果関係を抽出する(注1)．このため，因果関係の原因と結果の片方が省略された場合，因果関係の抽出を行うことができない．提案手法では，ニュース記事が持つ構造的特徴を考慮することで，この問題に対処する．

また，青野ら [5] は，Web 検索を利用して，与えられた事象に関する因果関係を表す文を自動的に取得する手法を提案している．ある事象 R と，手がかり表現を一つを入力とし，Web 検索によって自動的に手がかり表現と R の要因となる事象の集合を抽出する．まず，R と与えられた手がかり表現で Web 検索を行い，R の要因となる事象の初期集合 F を得る．R と F で Web 検索を行い，R とその要因の記述の間に存在する表現のうち，条件を満たすものを手がかり表現として取得する．R と手がかり表現を組み合わせた検索式で Web 検索を行い，検索結果から因果関係を抽出する．

この手法は，抽出したい因果関係に含まれる事象を入力として与える必要があるため，データセット内のあらゆる因果関係を取得するという目的には使用することができない点と，データセット内の因果関係を網羅的に抽出することは考慮していない点が本研究と異なる．

また，既存の手法と異なり，提案手法では手掛り表現をナイーブベイズモデルを使用して確率的に扱っている．

3. 因果関係抽出

3.1 概　　要

ニュース記事の集合から因果関係の集合を抽出する手法について述べる．

序論で述べたように，ニュース記事に出現する因果関係の記述には不完全記述の問題が存在する．因果関係の記述には (a) 原因及び結果の両方の記述が存在する (b) 原因のみが記述される (c) 結果のみが記述される という 3 つのパターンがあり，(b),(c) の場合，1 文を単位とした既存の手法では，因果関係の

(注1)：例外として，坂地らの手法では，手がかり表現が「。」で終わる場合，手がかり表現を含む文の一つ前の文から結果記述を抽出する．

抽出を行えない．本研究では (b) に対処することを考える．(c) については，本研究の提案手法を拡張して適応可能である．

我々は，次の 2 つの仮説を立てた．

結果・主題一致仮説　因果関係の結果事象は，その因果関係が含まれるニュース記事の主題と一致するという仮説．

原因記述類似仮説　同じ関連記事集合内に存在する原因記述同士は類似度が高いという仮説．

これらの仮説は，予備実験によって妥当性を検証した (4.1 節及び 4.2 節参照)．

(b) の場合，結果・主題一致仮説により，主題を結果事象として因果関係を抽出することができる．

因果関係抽出は，図 1 に示すように，関連記事集合をタイトルと本文に分け，タイトルから主題を，本文から原因事象をそれぞれ抽出することで行う．

タイトルは，短い文の中に主題を記述するのに必要なキーワードが十分に含まれていることが多く，序論で述べたキーワードの欠落の問題が少ない．

本文から原因記述を抽出するため，ナイーブベイズモデルを用いて記述の末尾の表現をモデル化し，機械学習によって識別器を構築する．識別器は，与えられた記述が原因記述である確率を計算し，初期スコアとし，類似度の高い記述に対して初期スコアを伝播させ，最終的なスコアを得る．

3.2 節で識別器の学習及びそれを使った原因記述の識別について議論する．3.3 節で原因記述の抽出について詳しく議論する．

3.2　原因記述の識別

因果関係抽出を行うためには，ある記述 d が与えられたとき，d が原因記述であるかどうかを識別する必要がある．そこで，予め原因記述を人手で識別しておいたコーパスを与え，手掛り語や文末の単語の確率分布などを学習し，これらを用いて原因記述の識別を行う．

まず，原因記述の末尾にどのような単語が含まれるのかを学習する．末尾に注目するのは，原因記述の末尾は，因果関係の手掛りとなる単語を含んでいる場合が多いからである．次の 5 つの原因記述を例として考える．

- 番組の合間に流すスポット広告が堅調だったことから、
- 世界的な金融緩和を背景にした
- 朝鮮半島の緊張やアイルランドへの金融支援に関する懸念を背景に海外で株安が進んだことを受け、

それぞれ,「から」「背景」「受け」が因果関係の存在を示す手掛り語である．実際に経済ニュースの記事を調査したところ,「背景」という単語が記述の末尾にある場合，その記述は 97% 以上の確率で原因記述だった．しかし,「から」という単語は,出発点を表す用法があるため,「から」があるからといって確実に原因記述であるとは言いきれない．実際，末尾に「から」を含む記述の 42%程度は非原因記述だった．

そこで，学習ステップでは，各単語 w に対して，原因記述において，w が記述の末尾に存在する確率 $p(w|C)$ と，非原因記述において，w が記述の末尾に存在する確率 $p(w|\neg C)$ を学習する．

識別ステップでは，学習ステップで得たパラメータを使って，与えられた記述の末尾の原因記述らしさを 0 以上 1 以下のスコアで評価する．

さらに，関連記事集合内の類似する記述にスコアを伝播させ，最終的なスコアを得る．スコアの伝播を行うのは，一つの関連記事集合に含まれる原因記述は，互いに類似しやすい傾向にあるからである．

最終的なスコアが閾値を超えていれば，その記述は原因記述であると判定する．

ここで，以下の議論で使用する用語及び記号の定義を行う．**記述**とは，1 個以上の単語の列である．$\mathrm{tail}(d, L)$ を記述 d の末尾 L 単語から成る記述とする．ただし，d の長さが L 未満の場合，$\mathrm{tail}(d, L) = d$ とする．$\mathrm{tf}(d, w)$ を記述 d に単語 w が出現した回数であるとする．$\mathrm{df}(D, w)$ を記述集合 D 内の記述であって単語 w を含むものの数とする．W を全ての単語の集合とする．

3.2.1　学習ステップ

D_C を原因であるとラベル付けされた記述の集合とする．$D_{\neg C}$ を原因であるとラベル付けされていない記述の集合とする．D_C と $D_{\neg C}$ は与えられるものとする．

ここで，l を，記述が原因記述か非原因記述かを表すラベルとする．すなわち，$l \in \{C, \neg C\}$ である．記述の末尾 L 単語のみを抽出した記述集合を次のように定義する．

$$D'_l := \left\{ \mathrm{tail}(d, L) \;\middle|\; d \in D_l \right\} \tag{1}$$

まず，手掛り語として有用な単語を抽出する．全ての単語に対して，**手掛りスコア**を計算する．手掛りスコアは，w が原因記述に含まれる割合であり，次式で与えられる．

$$\text{clue-score}(w) := \frac{\mathrm{df}(D'_C, w) + 1}{\mathrm{df}(D'_C, w) + \mathrm{df}(D'_{\neg C}, w) + \frac{|D_C| + |D_{\neg C}|}{|D_C|}} \tag{2}$$

分母の $(|D_C| + |D_{\neg C}|)/|D_C|$ の項と，分子の 1 の項は，df の値が小さい時に手掛りスコアが極端に大きくなることを防止するために存在する．

手掛りスコアの大きい単語を**手掛り語**と呼ぶことにする．

W を手掛りスコアでソートし，上位 I 単語を取り出して作った単語集合を，**本文の手掛り語集合**と呼び W_I で表す．W_I に含まれない単語を，その単語の品詞に置き換えることによって，素性の次元の削減を行う．

$$\mathrm{rep}(w) := \begin{cases} w & (w \in W_I) \\ \mathrm{pos}(w) & (\text{otherwise}) \end{cases} \tag{3}$$

$$D''_l := \left\{ \{\mathrm{rep}(w) | w \in d\} \;\middle|\; d \in D'_l \right\} \tag{4}$$

ただし，$\mathrm{pos}(w)$ は単語 w の品詞である．

ある記述に出現する単語の出現確率は，その記述が原因記述であるかどうかという情報が与えられれば，他の単語の出現確率と独立であり，多項分布に従うと仮定する．

$$p_l := p(l), \quad p_{wl} := p(w|l) \tag{5}$$

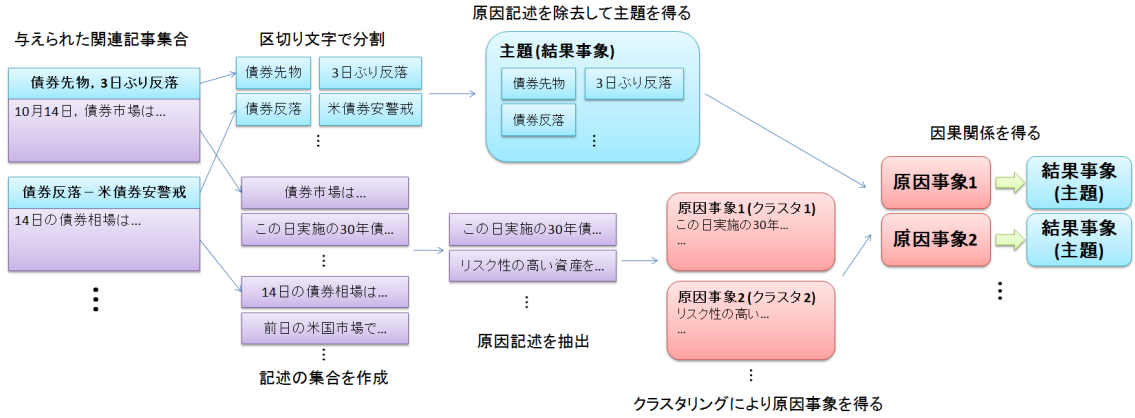

図 1　因果関係抽出の流れ

と置くと，ラベルが与えられた条件下での記述の出現確率は次のように書ける．

$$p(d|l) = \frac{\left(\sum_{w\in W} \mathrm{tf}(d,w)\right)!}{\prod_{w\in W} \mathrm{tf}(d,w)!} \prod_{w\in W} p(w|l)^{\mathrm{tf}(d,w)} \tag{6}$$

さらに，単語の出現確率と，記述が原因記述である確率は，α をパラメータとするディリクレ分布に従うと仮定する．

$$p(p_l;\alpha) \propto p_C^{\alpha-1} p_{\neg C}^{\alpha-1}, \quad p(p_{wl};\alpha) \propto \prod_{w\in W} p_{wl}^{\alpha-1} \tag{7}$$

このとき，パラメータの事後確率を最大にするようにパラメータを選ぶと，次のようになる．

$$p_{wl} = \frac{\sum_{d\in D_l''} \mathrm{tf}(d,w) + \alpha - 1}{\sum_{d\in D_C''\cup D_{\neg C}''} \mathrm{tf}(d,w) + |W|(\alpha-1)} \tag{8}$$

$$p_l = \frac{|D_l''| + \alpha - 1}{|D_C''| + |D_{\neg C}''| + 2(\alpha-1)} \tag{9}$$

以上で求めた，手掛り語集合及び p_l, p_{wl} をデータベースに格納して学習は終わる．

3. 2. 2　識別ステップ

R を一つの関連記事集合とする．R は，1 個以上の記述の集合として表現されているとする．

R が与えられたとき，R 内の各記述 $d \in R$ が原因記述であるかどうかを識別する．

まず，記述の末尾に注目してスコアを付ける．

$$\text{p-score}(d) := \frac{q_C(d)}{q_C(d) + q_{\neg C}(d)} \tag{10}$$

ただし，$l \in \{C, \neg C\}$ として

$$\log q_l(d) := \log p_l + \sum_{w\in W} \mathrm{tf}(d',w) \log p_{wl}(w) \tag{11}$$

$$d' := \{\mathrm{rep}(w) | w \in \mathrm{tail}(d, L)\} \tag{12}$$

学習ステップで得たパラメータを使って，d が原因記述である確率を計算したものが p-score(d) である．

$$\begin{aligned} p(l|d) &\propto p(d,l) = p(l)p(d|l) \\ &\propto p(l) \prod_{w\in W} p(w|l)^{\mathrm{tf}(d,w)} = q_l(d) \end{aligned} \tag{13}$$

であることと，

$$p(C|d) + p(\neg C|d) = 1 \tag{14}$$

であることに注意されたい．

「日米の金融政策の動向を見極めたいとして模様眺めムードが広がり、」のように，記述の末尾だけでは判定が難しい記述を分類するため，2 つの記述が共に原因記述であることと，2 つの記述の類似度が高いことの間に相関があることを利用する．p-score(d) に，同じ関連記事集合内の他の記述 d' のスコア p-score(d') を，d と d' の類似度で重み付けして足し込むことによって，スコアを伝播させ，類似している記述が原因記述ならばスコアが大きくなるようにする．

記述 d と d' の類似度を次ようにコサイン類似度により計算する．

$$\mathrm{sim}(d,d') := \frac{\sum_{w\in W} \mathrm{tfidf}(d,w)\,\mathrm{tfidf}(d',w)}{\sqrt{\sum_{w\in W} \mathrm{tfidf}(d,w)^2}\sqrt{\sum_{w\in W} \mathrm{tfidf}(d',w)^2}} \tag{15}$$

ただし $\mathrm{tfidf}(d,w)$ は，単語 w の記述 d における重要度であり，

$$\mathrm{tfidf}(d,w) := \mathrm{tf}(d,w) \log \frac{|\mathcal{D}|}{\mathrm{df}(\mathcal{D},w)} \tag{16}$$

で定義される．$\mathcal{D}$ は記述の集合であり，この集合はできるだけ大きな集合を使うのが望ましい．$\mathcal{D}$ と学習で用いた記述の集合が一致する必要はない．

s-score(d) は，p-score(d) と，伝播スコアの重み付きの和で定義される．

$$\begin{aligned} \text{s-score}(d) &:= (1-\mu)\,\text{p-score}(d) \\ &+ \mu \frac{\sum_{d'\in R\setminus\{d\}} \mathrm{sim}(d,d')\,\text{p-score}(d')}{\sqrt{\sum_{d'\in R\setminus\{d\}} \mathrm{sim}(d,d')^2}\sqrt{\sum_{d'\in R\setminus\{d\}} \text{p-score}(d')^2}} \end{aligned} \tag{17}$$

ただし，μ は，伝播の強さを表す係数であり，p-score(d) のエントロピーで決まる．

$$\begin{aligned} \mu := {} & \text{p-score}(d) \log_2 \text{p-score}(d) \\ & + (1 - \text{p-score}(d)) \log_2 (1 - \text{p-score}(d)) \end{aligned} \tag{18}$$

s-score の 2 番目の項は，類似度と p-score のコサイン類似度と見ることができる．すなわち，自分以外の記述が原因記述である確率と，自分と自分以外の記述の類似度が近ければ近いほど，その記述は，原因記述らしいと判定される．また，p-score の項と伝播項との混合係数として p-score(d) のエントロピーを使用するのは，p-score(d) による識別が困難である記述，すなわち，p-score によって d が原因記述である確率が 0.5 付近であると判定された記述に対しては，伝播項の重みを大きくし，そうでない記述は伝播項の重みを小さくするためである．

s-score(d) $> \theta$ のとき，d を原因記述と識別する．θ はパラメータである．我々の実験 (4.3 節) では，θ の値は 0.3 から 0.5 程度に設定すれば良い識別結果を得られることを示している．

3.3 因果関係の抽出

関連記事集合 R が与えられたとき，R から主題と原因事象を抽出したい．

まずは主題の抽出について考える．主題はタイトルから抽出する．多くの場合，ニュース記事のタイトルは，主題の記述のみから成る場合と，主題の記述と主な原因の記述の 2 つから成る場合の 2 通りである．よって，タイトルから主題を抽出するためには，タイトルに含まれる原因記述を排除する必要がある．次の例を考える．下線部が原因記述である．

- 債券は続落、30 年入札や株高を警戒
- ＮＹ債券、長期債続落 3 年債入札を嫌気

タイトルは,「、」や「ー」，空白などによって複数の部分に分かれており，いくつかの部分は原因記述である可能性がある．ある部分が原因記述であるかどうかを判定するために，末尾 1 単語が手掛り語集合に含まれているかどうかを調べる．手掛り語集合に含まれていれば，その部分は原因記述であり，そうでなければ非原因記述である．

手掛り語集合は，3.2 節の手掛り語集合と同様に，手掛りスコアを計算することによって求める．本文の手掛り語集合は，ニュース記事の本文に含まれる記述から手掛りスコアを計算した．タイトルの場合は，タイトルに含まれる記述から手掛りスコアを計算し，スコアの上位 I 単語を手掛り語として使用する．こうして得られた手掛り語の集合を**タイトルの手掛り語集合**と呼び，$\widetilde{W_I}$ で表す．

タイトルから原因記述の部分を取り除いて残った部分を R の主題とする．主題の抽出のアルゴリズムは次のようになる．

（1） R 内のニュース記事の全タイトルを一―‐-＝∶・、…／()「」[]【】及び空白で区切る．

（2） 1 で区切られた各部分の末尾の 1 単語がタイトルの手掛り語集合 $\widetilde{W_{10}}$ に含まれているかどうかを調べ，含まれている場合はその部分を削除する．

（3） 残った部分全体を R の主題とする．

次に，与えられた R 内の本文から，以下の手順で記述の集合を作成する．

（1） まず，本文を句読点で区切って，記述の集合を作る．

（2） 1 で作った集合に含まれる各記述に対して，本文の手掛り語集合 W_{10} に含まれる単語を含んでいるか調べる．含んでいる場合，その記述の先頭からその手掛り語までの文字列を新たな記述として記述の集合に追加する．ただし，句読点の 2 単語前以降は，手掛り語があっても無視する．

（3） 2 で作った記述の集合から単語数が 4 以下の記述を削除する．

こうして作られた記述の集合から，3.2 節の識別器を使って，原因記述の集合を取り出す．

次に，抽出された原因記述の集合を群平均法を使ってクラスタリングすることで，原因事象の集合を得る．クラスタリングのアルゴリズムは次のようになる．

（1） 原因記述の個数を n とする．各原因記述を，各単語の tf-idf 値を要素として持つベクトル $v_1, \ldots v_n$ で表す．

（2） 初期状態におけるクラスタ集合は，原因記述のベクトルを一つずつ含む n 個のクラスタの集合である．各クラスタ k_j は，tf-idf 値を要素として持つベクトル w_j を持つ．初期状態では $w_j = v_j$ である．

（3） クラスタの数が 1 以下であれば，現在のクラスタの集合を出力して終了する．

（4） 全てのクラスタの対の中で，最も類似度の高いクラスタの対を求める．クラスタ k_i と k_j の類似度は，w_i と w_j のコサイン類似度 $w_i \cdot w_j/(\|w_i\|\|w_j\|)$ で定義される．

（5） 4 で求めたクラスタ対の類似度がある閾値より小さい場合，現在のクラスタの集合を出力して終了する．閾値は予備実験によって決める．

（6） 4 で求めたクラスタ対 (k_i, k_j) に含まれるクラスタをマージし，クラスタ k_i 及び k_j を削除する．マージされて作られた新たなクラスタ k_l が持つベクトル w_l は $w_l := (w_i + w_j)/2$ で定義される．

（7） 3 に戻る

クラスタリングによって得られたクラスタ一つ一つが，原因事象に対応する．

こうして得られた原因事象と主題の組を R から得られた因果関係として抽出する．

表 1　実験で用いた記事集合

記事数	文数	単語数	分野	取得元	期間
349	2969	89504	ビジネス	Google News	2010/10/14〜2010/11/4

表 2　予備実験 1 の結果

因果関係数	仮説を満たす因果関係の数	仮説の成立率
605	496	0.82

4. 実　　験

この節では，事象の同一性と同一関連記事集合内の原因記述の類似性に関する予備実験の結果を示した後，第 3 節で議論した原因記述の識別器の性能と，因果関係の抽出アルゴリズムの性能を評価する．

4.1 予備実験 1: 結果・主題一致仮説

「因果関係の結果事象は，その因果関係が含まれるニュース記事の主題と一致する」という仮説について検証した．

表 3　予備実験 2 の結果

組み合わせ	平均コサイン類似度
原因記述 – 原因記述	0.45
原因記述 – 非原因記述	0.29
非原因記述 – 非原因記述	0.27
全ての記述 – 全ての記述	0.28

表 4　識別器の実験で用いた記述集合

原因記述の数	非原因記述の数	関連記事集合の数
456	3680	21

検証に使ったのは表 1 の記事集合である．この記事集合から因果関係を人手で抽出し，仮説が成立しているかどうかを調べた．結果は表 2 の通りである．この結果から，経済ニュースの場合，仮説が 82% という高確率で成立していることがわかる．

4.2　予備実験 2: 原因記述類似仮説

「同じ関連記事集合内に存在する原因記述同士は類似度が高い」という仮説について，4.1 節と同じ記事集合で検証した．

関連記事集合 R に対して，以下の手順で平均コサイン類似度を求めた．まず，R 内の原因記述を人手で抽出し，原因記述の集合を作る．次に，R 内の記事から原因記述の部分を削除し，残った部分を句読点で区切って非原因記述の集合を作る．さらに，全ての記述のペアに対して，tf-idf によるコサイン類似度を計算する．最後に，コサイン類似度を平均する．

原因記述が全体で 605 個抽出され，非原因記述が全体で 5589 個抽出された．

すべての関連記事集合に対してこの計算を行い，結果を平均すると表 3 のようになった．原因記述と原因記述の類似度は，それ以外のどの類似度よりも 50%以上高く，確かに同じ関連記事集合内の原因記述同士は類似度が高いことが分かる．

4.3　因果関係の識別実験

原因記述識別器の適合率および再現率を評価した．

この小節における適合率及び再現率を次のように定義する．pp をシステムが原因記述と判定し，実際に原因記述であった記述の数であるとする．pn をシステムが原因記述と判定し，実際には非原因記述であった記述の数であるとする．np をシステムが非原因記述と判定し，実際には原因記述であった記述の数であるとする．このとき，適合率 $=$ pp $/$(pp $+$ pn) であり，再現率 $=$ pp $/$(pp $+$ np) である．

まず，入力データとして，関連記事集合から原因記述の集合を人手で抽出した．さらに，原因記述を削除した記事集合を句読点で区切り，非原因記述の集合を作成した．その結果，表 4 に示す記述の集合を得た．

このデータセットに対し，10 分割交差検定を行い，性能を評価した．$\theta = 0.1, 0.2, \ldots, 0.9$ としたときの結果を表 5 及び図 2 に示す．ただし，学習時のパラメータは $L = 3, \alpha = 2, I = 200$ である．

s-score において，θ の値が 0.3 から 0.5 までのとき，適合率が 0.7 を上回っており，かつ再現率も 0.4 を超えている．よって，θ の値はこの範囲の値に設定するのが良いと考えられる．

表 5　識別器の実験結果

	p-score による識別		s-score による識別	
θ	適合率	再現率	適合率	再現率
0.1	0.41	0.75	0.38	0.68
0.2	0.54	0.69	0.55	0.57
0.3	0.63	0.65	0.72	0.51
0.4	0.67	0.59	0.78	0.46
0.5	0.68	0.51	0.84	0.41
0.6	0.74	0.46	0.86	0.38
0.7	0.79	0.45	0.87	0.29
0.8	0.82	0.40	0.88	0.20
0.9	0.87	0.30	0.91	0.13

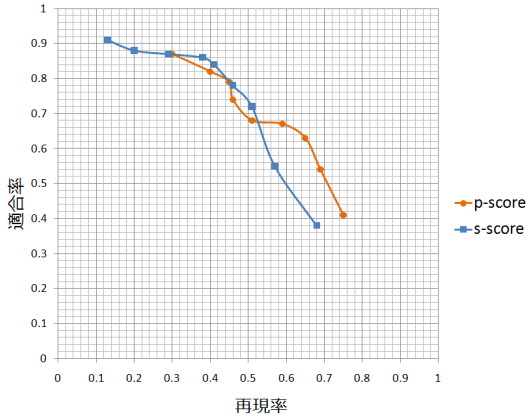

図 2　識別器の実験結果

表 6　因果関係の抽出実験に用いた記事集合

関連記事集合の数	記事数	異なる因果関係の数
6	52	31

表 7　因果関係の抽出実験における適合率と再現率

適合率	再現率
0.74	0.65

表 8　因果関係の抽出実験におけるクラスタリング精度の評価

正しい	過剰に分割	過剰に結合	不明
14 (45 %)	2 (6 %)	4 (13 %)	11 (34 %)

適合率が 0.7 から 0.9 の範囲では p-score よりも s-score の方が平均的に再現率が高く，スコアの伝播により，この範囲では，識別器の性能を向上させることができると分かった．

4.4　因果関係の抽出実験

提案手法の因果関係の抽出性能を実験により評価した．さらに，結果記述が省略された場合における，因果関係の抽出性能を評価した．また，ケーススタディによって，提案手法における事象の表現について考察した．

4.4.1　因果関係抽出の評価

因果関係の抽出の適合率と再現率を調べるため，表 6 に示す記事集合に対して，因果関係の抽出を行い，抽出された因果関係の正誤を人手で調査した．結果は表 7 の通りである．評価は以下の手順に従って行った．

（1）　各関連記事集合 R に対して，以下の 2 から 4 を行う．

表 9　因果関係の種類ごとの個数及び割合

原因と結果の両方が存在	原因のみが存在	合計
531 (80.8 %)	126 (19.2 %)	657 (100 %)

（2） R に対して，提案手法を適用して因果関係を抽出する．

（3） 抽出された因果関係に対して，その因果関係が正しいかどうかを人手で判定する．正しいと判定された因果関係の数を pp とし，そうでない因果関係の数を pn とする．

（4） システムにより抽出されなかった因果関係を R 内から人手で抽出する．この手順により抽出された因果関係の個数を np とする．

（5） 適合率 Pr と再現率 Rc を以下のように計算する．

$$\mathrm{Pr} = \sum \mathrm{pp} / (\sum \mathrm{pp} + \sum \mathrm{pn}), \quad \mathrm{Rc} = \sum \mathrm{pp} / (\sum \mathrm{pp} + \sum \mathrm{np})$$

ただし，記述集合 C から記述集合 E への因果関係が正しいとは，以下の条件が全て成立するときを言う．

- C に含まれる記述から，ある原因事象 c が読み取れる．
- E に含まれる記述から，ある結果事象 e が読み取れる．
- R 内に c から e への因果関係を示す表現が存在する．

抽出には，スコアとして s-score を使用し，$\theta = 0.4$ とした識別器を使用した．この識別器の記述識別の適合率は 0.78 であり，再現率は 0.46 である．因果関係抽出の再現率が，識別器の記述識別の再現率よりも高いのは，関連記事集合内に同じ因果関係を表す記述が複数存在していたからであると考えられる．

さらに，抽出された因果関係の原因事象のクラスタに対して，クラスタリングが適切であったかどうかを評価した．結果は表 8 の通りである．ただし，評価は以下の手順で行なった．

（1） 各クラスタ k に対して次の 2 から 5 を行う．

（2） k から原因事象が読み取れない場合，k を「不明」にカウントする．

（3） k から複数の原因事象が読み取れる場合，k を「過剰に結合」にカウントする．

（4） k から読み取れる原因事象が，k とは異なるクラスタ k' からも読み取れる場合，k を「過剰に分割」にカウントする．

（5） 2 から 4 のいずれの条件も満たさなかった場合，k を「正しい」にカウントする．

実験結果から，クラスタが過剰に結合される傾向があることが分かった．現在，クラスタリングにおいて，単語と単語の類似性は考慮していないため，本来は類似した内容を表す記述同士の類似度が小さくなることがある．これに対処するために，結合を行う類似度の閾値を 0.1 という比較的小さな値に設定している．この閾値の小ささが，過剰な結合を招いたものと考えられる．単語と単語の類似度をシソーラスなどを用いて計算し，これに基いて記述と記述の類似度を計算することで，この問題に対処できると考えられる．

4.4.2　結果記述の省略時における因果関係抽出の評価

結果の記述が省略され，原因のみが記述されているような場合に，提案手法がどれだけ因果関係を抽出可能か評価した．

まず，結果の記述が省略される場合がどれだけあるかを調べるため，表 1 の記事集合から，原因と結果が共に記述されている因果関係記述と，原因のみが記述されている因果関係記述を人手で抽出した．その結果，表 9 に示すように，結果記述が存在しない因果関係記述が約 2 割存在した．

表 10　結果の記述が省略されている因果関係の抽出

	p-score を使用		s-score を使用	
θ	再現率	結果＝主題であった割合	再現率	結果＝主題であった割合
0.1	0.53	0.91	0.51	0.91
0.2	0.47	0.90	0.39	0.88
0.3	0.41	0.88	0.32	0.90
0.4	0.36	0.87	0.24	0.93
0.5	0.34	0.82	0.19	0.92
0.6	0.24	0.88	0.19	0.92
0.7	0.20	0.86	0.17	0.91
0.8	0.17	0.91	0.17	0.91
0.9	0.15	0.90	0.14	0.89

- ＮＹ金、史上最高値を更新 １３８０ドル台に
- ＮＹ金、最高値更新＝１３８８．１０ドル
- ＮＹ金、最高値更新 一時 1370 ドル台に上昇
- ＮＹ金、続伸 12 月物は 1354.4 ドルで終了、米金融緩和観測で
- ＮＹ金、史上最高値更新

図 3　因果関係の抽出に用いた関連記事集合のタイトルの一部 (下線部は原因記述と判定された部分)

さらに，結果記述が省略されている原因記述の識別の再現率と，システムが原因記述であると判定した記述に対応する結果事象が，実際にその記事の主題と一致した割合を調べた．その結果を表 10 に示す．

この結果から分かるように，結果記述が省略された場合，原因記述から主題への因果関係を抽出することで高い確率で正しい因果関係を抽出できる．

この実験における再現率は，全ての記述を対象にした場合 (表 5) よりも低い．これは，「〜の理由は，…である」「〜の背景は，…だ」のような形の原因記述を正しく識別できていないからであると考えられる．

4.4.3　ケーススタディ

関連記事集合から因果関係を抽出する実験を行った．また，図 3 は，関連記事集合に含まれるタイトルの一部である．タイトルの中で原因記述と判定された部分には下線が引いてある．下線の引いてない部分全体が，この関連記事集合の主題である．

この関連記事集合に対して因果関係の抽出アルゴリズムを適用したところ，原因事象の集合として図 4 に示す 4 つのクラスタが得られた．

クラスタ 4 は，「外国為替市場」とその省略された言い方である「外為市場」という 2 つの表現を含んでいる．このように，クラスタリングにより「外国為替市場」「外為市場」という表記の揺れを，原因事象の表現の中に取り込むことが出来た．また，クラスタ 3 の 3 つの記述のうち一つだけが「米連邦準備制度理事会 (FRB)」という追加金融緩和を行う主体の表現を含んでいる．このように，クラスタリングにより，一つ一つの記述だけを見ると不足していたキーワードをクラスタリングによ

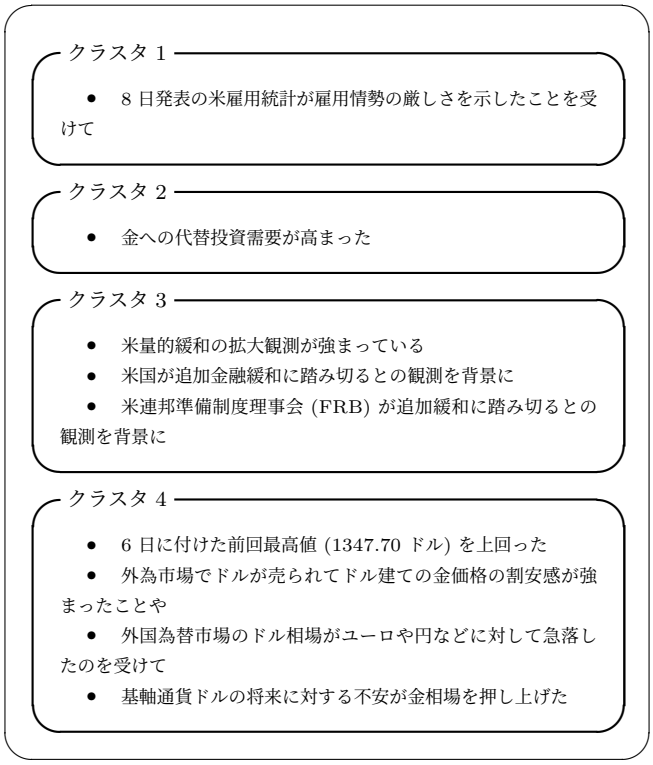

図 4　原因記述のクラスタリングの結果

り補うことが出来た．

また，クラスタ 3,4 は，クラスタ 1 と比べて多くの記述を含んでいる．このため，クラスタ 3,4 はクラスタ 1 と比べてより注目されており，より重要な原因事象を表現していると考えることができる．

この関連記事集合からは，クラスタリングにより，この 3 つの原因事象が抽出されるべきであるが，今回の結果では，4 つのクラスタが生じてしまった．クラスタ 2 とクラスタ 4 は本来は一つのクラスタであるべきである．関連記事集合内には「ドルの代替資産とされる金が買われやすい状態が続くとの思惑が」という記述が存在し，これは今回の原因記述抽出では抽出できなかった．この記述は，クラスタ 2 とクラスタ 4 を間を繋ぐ役割をする記述である．なぜなら「代替」「金」というクラスタ 2 において tf-idf の高い単語を持ち,「ドル」というクラスタ 4 において tf-idf の高い単語を持つからである．よって，より高い再現率を持つ識別器を構築することで，クラスタリングの精度を向上させることができると考えられる．

5. 結　　論

主題と結果事象の一致性や，原因記述同士の類似性といったニュース記事が持つ構造特徴を利用して，記事集合から因果関係を抽出する手法を提案した．提案手法では，与えられた関連記事集合のタイトルから主題を，本文から原因事象をそれぞれ抽出する．主題の抽出は，タイトルから原因記述と判断される部分を削除することで行う．原因事象の抽出は，本文から記述の集合を作成し，各記述に対して，機械学習によって得た識別器を使って原因記述であるか否かの判定を行う．原因記述であると判定された記述を群平均法でクラスタリングすることにより，原因事象の集合を得る．得られた原因事象から主題への因果関係を結果として返す．

予備実験により，主題と結果事象の一致性や，原因記述同士の類似性を確かめた．関連記事集合内の原因記述の類似性を利用した s-score は，それを利用しない p-score と比べ，適合率が 7 割から 9 割のとき，平均的に p-score よりも再現率が高かった．結果記述が省略された場合でも，提案手法は，原因記述を 3 割から 4 割程度の再現率で抽出でき，さらに抽出された原因記述の約 9 割は対応する結果事象と主題が一致していた．抽出された因果関係の具体例に対して，提案手法により，表記の揺れや，キーワードの省略に対応可能であることを確認した．また，原因事象のクラスタに含まれる記述の種類や量に基づいて，重要な因果関係を抽出可能であると分かった．複数の関連記事集合から因果関係を抽出を行い，74%の適合率と 65%の再現率が得られた．

本研究では，因果関係の不完全記述について，原因のみが記述され，結果が省略される場合のみを考えたが，原因が省略され，結果のみが書かれる場合もありうる．省略される記述は，多くの場合，主題であるため，これは主題を原因とする因果関係であることが予想される．将来的にはこの予想を検証し，原因が省略される場合において因果関係を抽出する手法を考えていきたい．さらに，原因記述，結果記述が両方ある場合，いずれかのみ存在する場合を判定することにも取り組んでいきたい．

また，本研究では，実験対象のニュース記事として経済ニュースのみを用いたが，経済ニュース以外のニュース記事に対して実験を行い，提案手法が他の分野のニュース記事に対して有効であるかどうかを調べていきたい．

謝　　辞

本研究の一部は、科研費（20700084 と 20300042）の助成を受けたものである．

文　　献

[1] 佐藤岳文, 堀田昌英. Web マイニングを用いた因果ネットワークの自動構築手法の開発. 社会技術研究論文集, Vol. 4, pp. 66–74, 2006.

[2] 乾孝司, 乾健太郎, 松本裕治. 接続標識「ため」に基づく文書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, Vol. 45, No. 3, pp. 919–933, 2004.

[3] 坂地泰紀, 竹内康介, 増山繁, 関根聡. 構文パターンを用いた因果関係の抽出. 言語処理学会第 14 回年次大会論文集, pp. 1147–1147, 2008.

[4] Hiroyuki Sakai and Shigeru Masuyama. Cause information extraction from financial articles concerning business performance. *IEICE Transactions*, Vol. 91-D, No. 4, pp. 959–968, 2008.

[5] 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. *DEIM Forum 2010 B9-1*, 2010.

[6] Hiroshi Ishii, Qiang Ma, and Masatoshi Yoshikawa. An incremental method for causal network construction. In *Proc. of WAIM*, pp. 495–506, 2010.

[7] 高村大也. 言語処理のための機械学習入門 (自然言語処理シリーズ). コロナ社, 7 2010.

[8] 北研二, 辻井潤一. 言語と計算 (4) 確率的言語モデル. 東京大学出版会, 11 1999.