

因果関係ネットワークの構築によるニュースの理解支援

石井 裕志[†] 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

^{††} 京都大学情報学研究科社会情報学専攻 〒 606-8501 京都市左京区吉田本町

E-mail: [†]ishii@db.soc.i.kyoto-u.ac.jp, ^{††}{qiang,yoshikawa}@i.kyoto-u.ac.jp

あらまし ニュースの理解支援を行うために、本論文では TEC モデル (Topic-Event Causal relation model) と、モデルを利用した因果関係ネットワークの構築手法を提案する。TEC モデルでは、因果関係の原因と結果をそれぞれ始点と終点ノードとしたエッジラベル付き有向グラフで表現する。また、ノードはニュース記事の因果関係を示す文節から得られる語、記事のタイトルから得られるトピックを表す語とそれぞれの語の重要度によって構成し、枝は重要度を保持する。我々はまず、手がかり表現を利用して記事から因果関係を抽出し、ネットワークとして表現する。さらに、語の頻度を利用したノードの類似度計算によるグラフの結合と、枝の重要度を利用したグラフの簡略化を行いながらネットワークを構築していく。また、本論文ではノード抽出の評価実験を行い、提案手法の有効性を確認した。キーワード 因果関係、ネットワーク構築、リダクション、ニュース

Causal Network Construction for News Understanding Support

Hiroshi ISHII[†], Qiang MA^{††}, and Masatoshi YOSHIKAWA^{††}

[†] School of Informatics and Mathematical Science, Faculty of Engineering Kyoto University

Honmachi, Yoshida, Sakyo-ku, Kyoto, 606-8501

^{††} Department of Social Informatics, Kyoto University

Honmachi, Yoshida, Sakyo-ku, Kyoto, 606-8501

E-mail: [†]ishii@db.soc.i.kyoto-u.ac.jp, ^{††}{qiang,yoshikawa}@i.kyoto-u.ac.jp

Abstract To support understanding of news, we propose a novel model, named the TEC model (Topic-Event Causal relation model), and its construction method. Within the TEC model, causal relations are represented by an edge-labeled directed graph. A source vertex expresses a cause of an event, and a destination vertex expresses the result of that event. Each node contains event keywords, topic keywords, and an importance score for each keyword. The label of edge is the importance score of that causal relation. To construct a causal network, we extract causal relations from articles based on 'clue phrases' and merge similar event nodes, and reduce the causal network based on the importance of each causal relation. In this paper, we carried out a preliminary experiment to validate the proposed method.

Key words causal relation, network construction, reduction, news

1. 緒 論

テレビや新聞、Web 上などのニュースで報道される出来事を理解する時に、その出来事の起こる背景知識が重要となる場合がある。特に、出来事の動きが速い場合や、出来事に対して複雑に事象が絡み合っているような場合である。例えば、昨今話題になっている「金融危機」では「サブプライム問題」や「リーマンブラザーズ破綻」など、たくさんの事象が複雑に背景として存在し、また「金融危機」から「AIG 救済」や「景気後退」などの多くの事象が起こっている。このようなニュースは背景

知識を持たずに見聞きしても、深く理解できなかったり正しく理解できなかったりする。しかし、ユーザーが一つ一つ記事を検索して背景知識を得るには非常に労力が必要であり、また背景知識の見落としなども発生する。そこで、ユーザーに対してニュースの背景知識を提供し、理解支援を行うシステムが必要である。

ニュースの背景知識を提示する上で、ニュースがなぜ起こったかという因果関係が大きな手掛かりとなる。そこで本研究では、ニュースに含まれる因果関係を利用したニュースの理解支援システムを提案する。従来、ニュース記事などから因果関係

を抽出する手法が研究開発 [1] ~ [6] されてきた。しかし、抽出処理が記事単位で行われているために、得られる因果関係は断片的なものが多く、ニュースの背景知識として大きな意味を持つものは少ない。因果関係をニュースの背景知識として利用するには、一つ一つの事象に対して、その事象を取りまく因果関係全体やそれが起こった大元の原因、波及して起こった結果を提示することが必要である。そのために本研究では、Web 上のニュース記事から因果関係を抽出し、それらを結合して因果関係ネットワークとして逐次構築し、因果関係のつながりを提示することでユーザーのニュースの理解を支援する。

本研究では、類似事象を意味している事象ノード同士をマージ (結合) し、因果関係のつながりを表現できるネットワークを構築する。また、重複している因果関係や重要度の低い因果関係を削除して、因果関係ネットワークのリダクション (簡略化) を行う。一般に、因果関係ネットワークは原因の事象を始点事象ノード、結果の事象を終点事象ノードとした有向グラフで因果関係を表現するが、マージやリダクションを行うために、我々はニュースの背景理解のための因果関係ネットワークのモデルとして TEC モデル (Topic-Event Causal relation model) を提案した。ノードのマージを行うにはノード同士の類似度を計測する必要があるため、事象ノードを元の記事で重要と思われる語 (キーワード) の集合で表現する。しかし、文書集合から因果関係ネットワークを構築する手法 [4] ~ [6] では、因果関係が記述されている文のみからキーワードを抽出するため、キーワードの類似度による類似事象ノードの判断が精度よく行えなかった。そこで TEC モデルでは、以下の 2 種類のキーワードと各キーワードの重要度をノードに保持し、キーワード事象ノードの類似度計算に利用する。

- ・ 事象キーワード：因果関係に含まれる語
- ・ トピックキーワード：記事のタイトルに含まれる語

因果関係のリダクションでは、因果関係を重要度で区別する必要がある。そのため TEC モデルでは、枝の重要度をラベルとして因果関係の枝に保持する。

本研究で提案する因果関係の構築システムでは、まず、記事集合から「手がかり表現 (「を背景に」や「ため、」など)」を基に因果関係を抽出し、事象ノードのマージ、リダクションを行って因果関係ネットワークを構築する。さらに、一日単位で追加された新しい記事からも同様に因果関係ネットワークを構築し、これまでに構築した因果関係ネットワークと合わせて、マージ、リダクションを行い、因果関係ネットワークを更新する。

以下、2 節で関連研究について説明し、3 節で TEC モデルについて、4 節で TEC モデルにおける因果関係ネットワークの構築について、5 節でシステムのプロトタイプについて、6 節で事象ノード抽出の実験について述べる。最後に 7 節で研究のまとめと今後について述べる。

2. 関連研究

文章からの因果関係の抽出手法、因果関係ネットワークの構築方法の関連研究について説明する。

2.1 因果関係抽出手法

文章から因果関係を自動抽出する手法として、乾らの接続標識「ため」を用いる手法 [3]、格フレームを用いる手法 [4] ~ [6]、坂地らの手がかり表現による因果関係の構文パターンを用いる手法 [1], [2] などが提案されている。乾ら [3] の手法は、必然性の高い因果関係が成り立つ接続詞として「ため」に注目し、「ため」が使われやすい複文のみから因果関係を抽出している。

佐藤・笠原ら [5], [6] の手法では、文が 1 文中に複数の格フレームを持ち、さらに接続関係のある場合に因果関係を抽出し、そこからキーワードを抽出する。佐藤・堀田 [4] は、Web 文章から「手がかり標識」によって因果関係があると判断される文について、佐藤・笠原らの方法を元に原因文節 (単文) と結果文節 (単文) から重要と思われる単語を抽出する。以上の手法 [3] ~ [6] は記事からの因果関係文の抽出を複文・重文のみで行えなかったのに対し、坂地らの手法 [1], [2] では、複文・重文に限らず「手がかり表現」(佐藤・堀田らの論文では「手がかり標識」に相当) を含む文全体を対象にして因果関係文節を取得できる。本論文では坂地らの手法を利用している。

2.2 因果関係ネットワークの構築

上記で述べた因果関係の抽出法を使用して因果関係ネットワークを構築する手法として、佐藤・笠原ら [5], [6] や佐藤・堀田 [4] の手法が提案されている。

佐藤・笠原らの手法は、人間が常識的に持っている因果関係知識をデータベース化することを目的として、一般の文章から因果関係ネットワークを構築する。佐藤・堀田の手法は、Web 上の文章から因果関係ネットワークを構築する。両手法ともに、文章の因果関係を含む文節から得られる重要単語を事象ノードのキーワードとしている。単語が重要かどうかの判断は係受け解析と格フレームを用いる自然言語処理で行い、単語は重要 (事象データに含める) か冗長 (事象データに含めない) の 2 分である。2 つの手法のいずれも、因果関係を表現している文節しかキーワードの対象にならないことと、冗長とされた単語は全く考慮されないため、原因・結果の一つの事象ノードは数語の単語からのみ構成される。この状態ではキーワードが不足しており、

- ・ 事象データが何を意味しているのかわからない
- ・ 事象データから実際に起きた事象を特定できない

ということが生じる。例えば、佐藤・堀田 [4] の論文では、キーワードとして「慣例」のみを持つノードが取得された例が紹介されている。「慣例」だけでは、元の文が何について述べているのかわからず、また他に「慣例」のみを持つノードが取得されると、2 ノード間の類似度は高くなるが、この 2 つのノードが類似した内容を意味しているとは言えない。これを解決するために、本研究ではノードの持つキーワードを拡張する。

佐藤・堀田の手法はノードのマージは行っておらず、ノード間の類似度の高さをノード間の距離の近さとして表現している。しかし、類似事象を示すノードが多い場合には因果関係のつながりを理解しづらい。本研究では、類似度計算により類似ノードをマージすることで因果関係のつながりを表現する。なお、佐藤・笠原らの手法は、ノード間の言葉の揺らぎに対してシソー

ラスを用い同概念の事象を一つにまとめて、キーワードが完全一致した場合に類似ノードと判定している．これは、常識としての因果関係を対象にネットワークを構築することを研究の目的としているからだと考えられる．我々のニュースの理解支援では、一つの事象それぞれを区別する必要があるため、ノード間の言葉の揺らぎに対してシソーラスは用いず、キーワードの完全一致ではなく類似度の計算を行うことで対処する．

佐藤・堀田、佐藤・笠原らの両手法ともに生成したネットワークを整理していないが、ネットワークが複雑であるとユーザーが理解できない．本研究では、枝の重要度を用いた因果関係のリダクションによりネットワークの整理を逐次行いながら、ネットワークの構築を行う．

3. TEC モデル

本研究では、ニュースの背景理解を支援するため、TEC モデル (Topic-Event Causal relation model) を提案する．‘T’ は事象ノードが保持するキーワードの「トピックキーワード (topic keyword)」、‘E’ は事象ノードが保持するキーワード「事象キーワード (event keyword)」の頭文字に由来している．まず、因果関係とは、2つの事柄に対して一方が原因で他方が結果であるような関係があることをいう．TEC モデルでは、原因部、結果部それぞれに事象ノードを作成して、始点を原因の事象ノード、終点を結果の事象ノードとする．また、枝は因果関係の重要度を表すラベル (枝の重要度と呼ぶ) を保持し、因果関係ネットワークをエッジラベル付き有向グラフで表現する．エッジラベル付き有向グラフはグラフ G のノード集合 V 、辺 (枝) 集合 E 、辺のラベル h に対して、写像 f が式 (1) を満たすように存在し、 $G := (V, E, f, h)$ で表わされる．

$$f: E \rightarrow V \times V \quad (1)$$

TEC モデルでの因果関係の表現例を図 1 に示す．

例)「金融危機」が原因となり、結果として「自動車の減産」が起こったことを示すグラフ

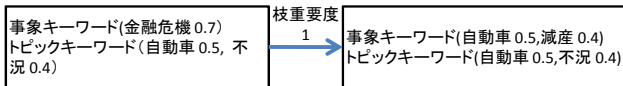


図 1 TEC モデル

グラフの各枝 e は、始点ノード (原因ノード) v_{s_e} 、終点ノード (結果ノード) v_{t_e} 、枝の重要度 h_e に対し、式 (2) で表す．

$$e := (v_{s_e}, v_{t_e}, h_e) \quad (2)$$

枝の重要度は、因果関係の頻度によって決定する．因果関係の因果関係が抽出された時の枝の重要度 h の初期値は 1 であり、頻度が高い場合には値が大きくなる．枝の重要度の計算方法は 4.3 節で述べる．

TEC モデルでは各事象ノードに、因果関係を表すキーワード集合と記事のトピックを表すキーワード集合およびこれらのキーワードの重要度を保持する．事象ノード v は、キーワード k とその重要度 w_k のペアの集合であり、式 (3) のように表す．

$$v := \{(k, w_k) + \} \quad (3)$$

キーワード k は、事象キーワード k_e とトピックキーワード k_t の 2 種類ある．つまり、

$$v := \{(k_e, w_{k_e}) +, (k_t, w_{k_t}) + \} \quad (4)$$

である．ただし「+」は 1 個以上存在していることを表している．次に、 k_e, k_t について述べる．キーワードの重要度の計算方法は、4.1.3 節で述べる．

(1) 事象キーワード k_e

事象キーワードは取得した因果関係を表現するキーワードである．因果関係を含むとして取得した文節から抽出できる語を事象キーワードとする．

(2) トピックキーワード k_t ノードが持つ因果関係のトピックを表すキーワードである．元の記事のタイトルに含まれる語をトピックキーワードとする．

4. TEC モデルによる因果関係ネットワークの構築

TEC モデルでの因果関係ネットワークの構築手法について述べる．記事から TEC モデルに合うように因果関係を抽出し、そこから類似ノードの結合や不要因果関係の削除を行い、因果関係ネットワークとして構築する．因果関係の抽出では、手がかかり表現を用いて文章から因果関係を抽出し、事象ノードと因果関係の枝を作成する．ノードのマージでは、ノードの類似度を利用して類似事象を示すノードを結合し、ネットワークのリダクションでは、枝の重要度を利用してネットワークを簡略化する．

4.1 因果関係の抽出

ニュース記事から因果関係を抽出し、事象ノードを作成するまでの手順を以下に示す．

(1) 記事からの因果関係文節の抽出

(2) 事象ノード、枝の作成

(a) 因果関係の事象キーワードの抽出

(b) 記事のトピックキーワードの抽出

(3) キーワードの重要度計算

記事から因果関係を抽出し、事象ノードを作成する例を図 2 に示す．

4.1.1 記事からの因果関係文節の抽出

本研究では、文章からの因果関係抽出に坂地ら [1] の手法を用いた．まず「を背景に」や「ため、」などの因果関係の存在する文を示す「手がかかり表現」を含む文の記事から抽出し、因果関係を含む文と判断する．抽出された文を cabocha [7] によって係受け解析し、次の 4 種の構文パターンに分類する．

Pattern A: 結果表現の主部と述部が存在する場合

「<原因表現>のため、<結果表現主部>が<結果表現述部>した。」

Pattern B: 手がかかり表現の直後に結果表現が存在する場合

「<原因表現>のため、<結果表現>した。」

Pattern C: 結果表現が抽出文の前文である場合

「<結果表現>した、<原因表現>のためだ。」

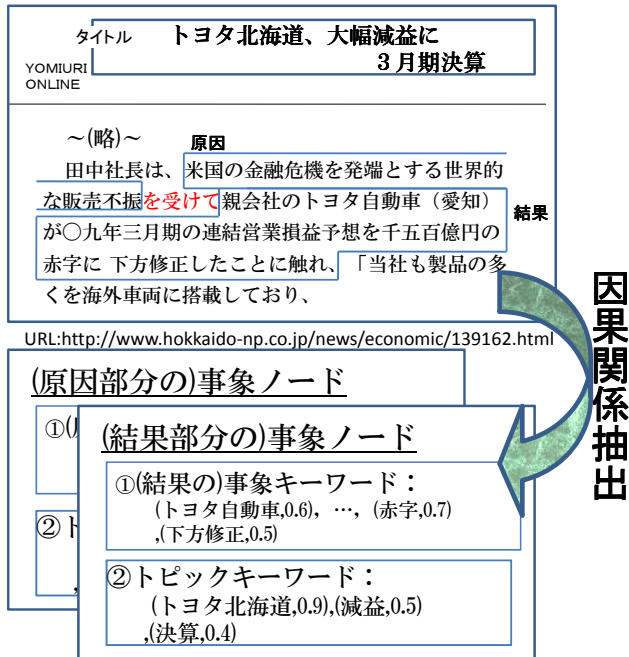


図 2 因果関係抽出

手掛かり表現
「を背景に」、「を挙げる」、「ため」、「に伴う」、「に加え」、「ためだ」、「により」、「を受けて」、「の効果」、「の影響も」、「の影響が」、「によって」、「から」

表 1 手掛かり表現

Pattern D: 結果表現が根拠表現より前に出現する場合

構文パターン別の文法的特徴により、元の文章から根拠表現文節と結果表現文節を抽出する。また、主な手掛かり表現を表 1 に示す。

図 2 の記事の例では、手がかり表現「を受けて」を検索することで因果関係を含む文が抽出され、係受け解析によって Pattern A として判定され、原因表現「米国の金融危機を発端とする世界的な販売不振」と結果表現「親会社のトヨタ自動車（愛知）が〇九年三月期の連結営業損益予想を千五百億円の赤字に下方修正したことに触れ」が抽出される。

4.1.2 事象ノードの作成

得られた因果関係のそれぞれについて原因の事象ノードと結果の事象ノード、それを結ぶ枝を作成する。この時、枝の重要度は 1 とする。そして、事象ノードごとにキーワードを抽出し挿入していく。(以降、ここで因果関係の文節を抽出した元の記事のことを注目記事と呼ぶ。)

(a) 事象キーワード k_e の抽出

事象ノードの事象キーワード k_e となる語を抽出する。因果関係文節の抽出で抽出された文節を茶筌 [8] を用いて形態素解析し、名詞と未知語を抽出する。抽出された語を事象キーワード k_e とする。

(b) トピックキーワード k_t の抽出

事象ノードのトピックキーワード k_t となる語を抽出する

まず、注目記事のタイトルの部分を茶筌 [8] を用いて形態素解析し、名詞と未知語を抽出する。抽出された語をトピックキーワード k_t とする。

図 2 の事象ノードの例では、抽出された結果文節から結果の事象キーワード「トヨタ自動車」などが抽出される。また、記事タイトルからトピックキーワード「トヨタ北海道」や「減益」が抽出されている。なお、キーワードとペアになっている数値はキーワードの重要度を示している。

4.1.3 キーワードの重要度計算

抽出した事象キーワード、トピックキーワードの各単語について重要度を求める。重要度は、マージの類似度計算時にそれぞれの語の重みとして使用され、また、重要度の低い語は不要な語としてノードから削除される。記事中に多く出現する単語や、タイトルに含まれる単語は重要だと考え、重要度を高くする。重要度 $weight$ は式 (5) によって計算する。

$$weight(k) = ntf(k) + ttf(k) \quad (5)$$

ただし、 $ntf(k)$ は語 k の本文中での正規化された頻出度であり、式 (6) で計算する。

$$ntf(k) = \frac{freq(k)}{\sqrt{\sum_{t \in T} freq(t)^2}} \quad (6)$$

ここで、 $freq(k)$ は単語 k の注目記事中の頻出度を表し、ここでは記事中の k の出現回数である。 T は注目記事中から抽出された語集合である。

$ttf(k)$ は語 k のタイトル中での正規化された頻出度であり、式 (7) で計算する。

$$ttf(k) = \frac{freq'(k)}{\sqrt{\sum_{t \in T'} freq'(t)^2}} \quad (7)$$

ここで、 $freq'(k)$ は注目記事タイトル中の k の出現回数、 T' は注目記事タイトル中の語集合である。

4.2 ノードのマージ

マージでは類似した事象ノードと事象ノードの結合を行う。

4.1 節で一つつの因果関係の取得について述べたが、このマージによって複数の因果関係を結びつけ、因果関係の連鎖を表現することができる。図 3 に示すマージの例では、事象キーワードが「自動車、減産」の 2 ノードが、類似ノードとしてマージされ、因果関係の連鎖が取得できる。

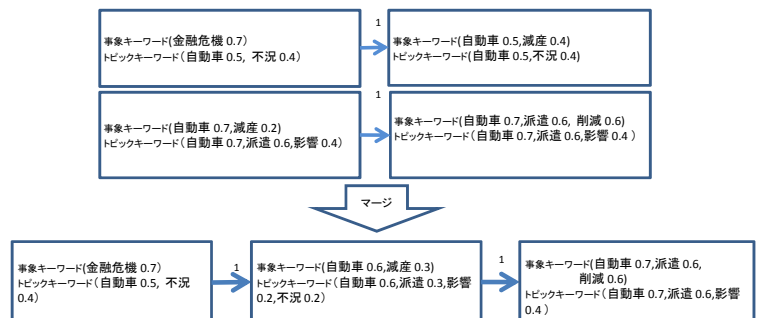


図 3 マージ

4.2.1 類似度計算

マージは、直接は枝で結ばれていない相異なるノード間の類似度を計算し、類似度が閾値以上の時にノードのマージを実行する。ノード間の類似度計算は、キーワードをベクトルとするベクトル空間モデルでノードを表し、2ノード間のコサイン相関値を計算する。ノード v_a とノード v_b の類似度 $sim(v_a, v_b)$ は式 (8) で表わされる。

$$sim(v_a, v_b) = \frac{w_{a1} \cdot w_{b1} + w_{a2} \cdot w_{b2} + \dots + w_{an} \cdot w_{bn}}{\sqrt{w_{a1}^2 + \dots + w_{an}^2} \cdot \sqrt{w_{b1}^2 + \dots + w_{bn}^2}} \quad (8)$$

ただし、 w_a, w_b は 4.1.3 節で述べた v_a, v_b のキーワード重要度 *weight* である。ベクトルの次元は v_a, v_b のキーワード集合の和集合であり、片方のノードにのみに含まれるキーワードは、含まれていない側のノードでの重要度 w を 0 とする。

4.2.2 マージ

ノード v_a, v_b において、類似度 $sim(v_a, v_b)$ が閾値以上の時にマージを実行し、 v_{a+b} を作成、 v_a, v_b を削除する。 v_{a+b} のキーワードは次のように構成する。

$$K_{e_{a+b}} = K_{e_a} \cup K_{e_b} \quad (9)$$

$$K_{t_{a+b}} = K_{t_a} \cup K_{t_b} \quad (10)$$

ここで K_e, K_t はキーワードの集合を示している。つまり、 v_{a+b} のキーワードは事象キーワード集合、トピックキーワード集合ごとに、 v_a, v_b の各キーワード集合の和集合を取ったものである。キーワード重要度 *weight* は式 (11) で求める。

$$weight_{a+b}(t) = \frac{weight_a(t) * (i+1) + weight_b(t) * (j+1)}{i+j+2} \quad (11)$$

ただし、 i はノード v_a のマージを行ってきた回数、 j はノード v_b のマージを行ってきた回数である。式 (11) では、各キーワードについて元々記事から取得された時の重要度を平均した値が、マージ後の重要度となる。

4.3 ネットワークのリダクション

リダクションでは重複する枝をまとめ、まとめた枝の重要度 (因果関係の重要度を表す) を求める。これを利用して重要度の低い因果関係の削除やユーザーへ提示時する因果関係の判断を行う。枝が重複しているとは、ノードのマージの結果、式 (12) のように2つのノード間に同じ方向の枝が2つ以上存在していることをいう。

$$\left. \begin{aligned} e_{k_1} &= (v_1, v_2, h_{e_{k_1}}) \\ e_{k_2} &= (v_1, v_2, h_{e_{k_2}}) \\ &\vdots \\ e_{k_n} &= (v_1, v_2, h_{e_{k_n}}) \end{aligned} \right\} \quad (12)$$

リダクションでは、まず各ノード間で重複した枝が存在するかどうかを調べる。存在する場合には、重複する枝それぞれの重

要度の和を重要度とした枝を作成し、元の枝を削除する。式 (12) の場合は枝をまとめた枝 $e_{k_1+k_2+\dots+k_n}$ を式 (13) のように作成し、枝 $e_{k_1}, e_{k_2}, \dots, e_{k_n}$ を削除する。図 4 に枝がまとめられる例を示す。

$$e_{k_1+k_2+\dots+k_n} = (v_1, v_2, h_{e_{k_1}} + h_{e_{k_2}} + \dots + h_{e_{k_n}}) \quad (13)$$

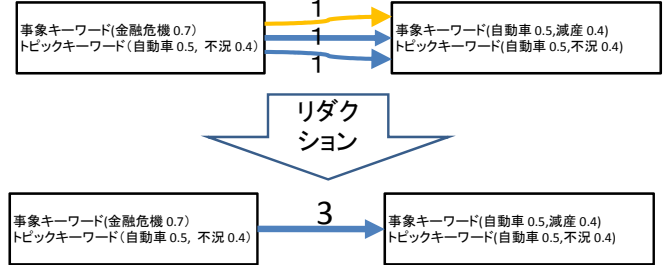


図 4 リダクション

4.4 ネットワークのインクリメンタル構築

ニュースは毎日新しいものが報道されるため、因果関係ネットワークを更新する必要がある。本研究では、追加された記事で構築する因果関係ネットワークとそれまでに構築したネットワークを逐次構築することでネットワークの更新をする。その様子を図 5 に示す。図 5(a) では、一日単位で新たに収集された記事から因果関係の抽出し ((1),(2))、抽出された因果関係からマージ・リダクションを行い一日分の因果関係ネットワークを構築 ((3),(4)) し、そして、一日分のネットワークとそれまでに構築したネットワークをマージ (図 5(b))、リダクション (図 5(c)) を行う。つまり、図 5 の (a),(b),(c) を (a) → (b) → (c) → (a) → ... と一日単位で繰り返し行うことで、ネットワークを更新する。

5. プロトタイプシステム

現在、有用な因果関係ネットワークが構築が実際に可能であるかを評価するため、3 節、4 節の手法に基づいた因果関係ネットワークを構築・提示するためのプロトタイプシステムを実装中である。システムの構成とその実装について述べる。

5.1 システムの構成

システムの構成を図 6 に示し、システムの各モジュールについて説明する。

(1) ニュース収集モジュール

ニュース収集コンポーネントにより Google News [9] から新聞サイトの記事の更新情報を取得し、各新聞サイトから新規記事を取得する。

(2) ノード作成モジュール

収集された記事から、3 節、4.1 節の手法に基づいてノードの作成を行う。

(3) マージモジュール

4.2 節の手法に基づいて、因果関係ネットワーク DB に格納された相異なる 2 つのノードの組み合わせそれぞれに対し類似度計算を行い、閾値を超えたものについてマージを行う。

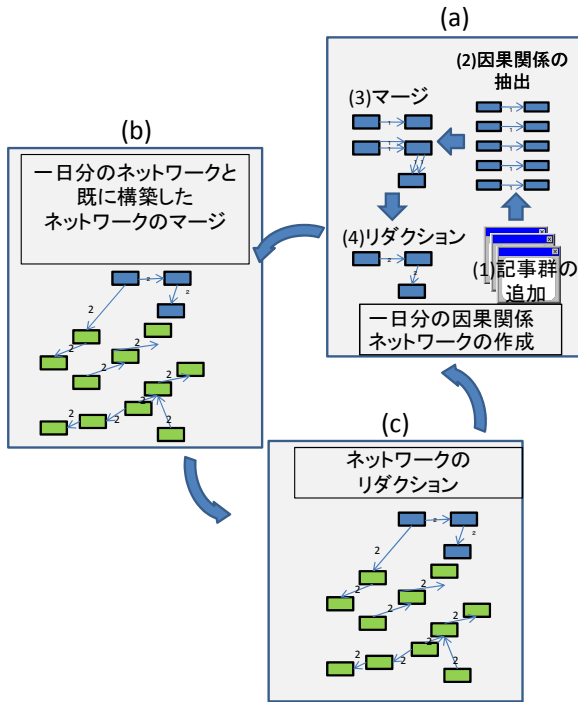


図 5 ネットワークのインクリメンタル構築

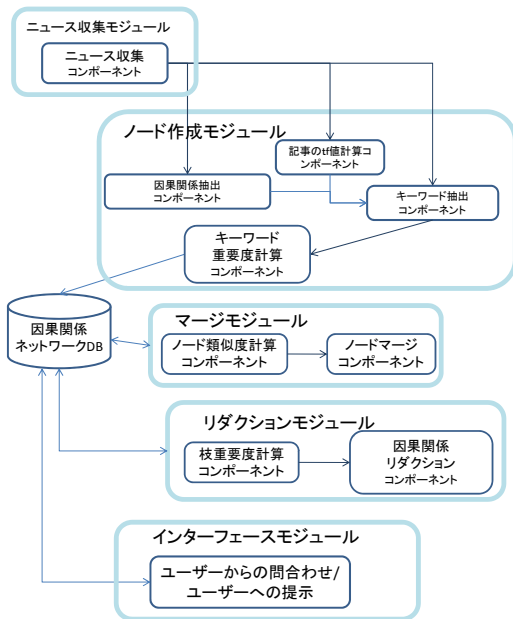


図 6 システム構成図

(4) リダクションモジュール

4.3 節の手法に基づいて、重複した枝の統合と枝の重要度の計算、不要因果関係の削除を行う。

(5) インターフェースモジュール

ユーザーからの問い合わせに応答し、ユーザーの理解支援をするグラフの提示を行う。

システムが因果関係を抽出し、そこから事象ノードをマージ、リダクションする処理の流れを図 7 を示す。図 7 は上から順に、因果関係の抽出のみを行ったネットワーク、マージ後のネットワーク、リダクション後のネットワークとなっている。また、

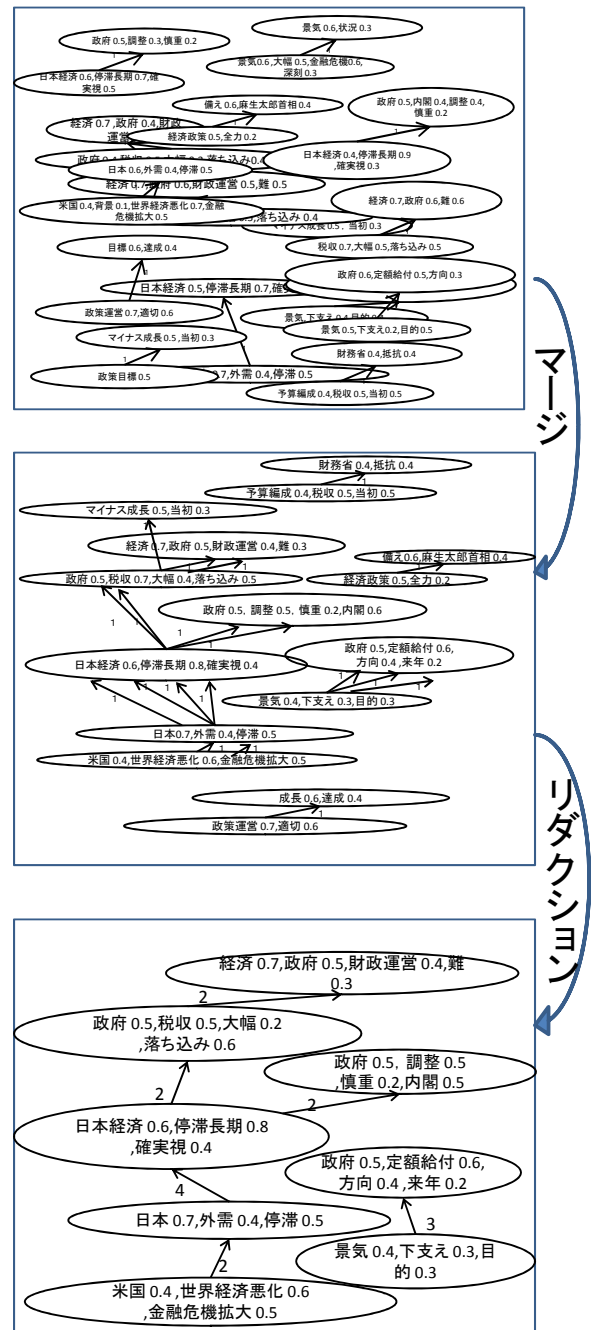


図 7 システムの処理の流れ

ユーザーインターフェースは、例えば「日本、経済」というキーワードでユーザーが問い合わせると、システムが「日本、経済」に関するノードとその周辺ネットワークを検索し、図 7 のリダクション後のようなネットワークが表示される。

5.2 実装

現在、ノード作成モジュールとノード類似度計算コンポーネントについて実装済みである。残りのシステムについては今後、順次実装を行っていく。現在のアプリケーションの実行例を図 8 に示す。

6. 実験

提案手法に基づいて取得できる事象ノードの内容が、元の記

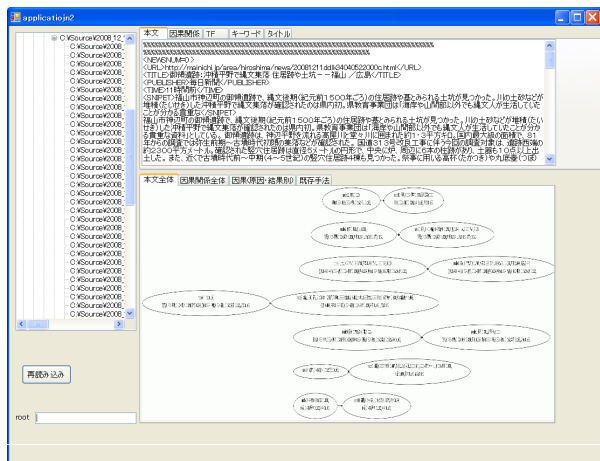


図 8 アプリケーション

事中での意味を表現しているかについて実験を行った。

6.1 実験準備

実験に際して、因果関係を抽出する時の手がかり表現は、表 1 の手がかり表現を使用した。実験に使用した記事群は、2008 年 12 月 15 日の Google News の経済カテゴリから取得した「石油価格下落に対する石油減産の決定」に関する記事 (9 記事) である。各記事のタイトルと URL を表 2 に示す。

サウジ、総会を前に供給を 8 %削減 = O P E C 議長 http://jp.reuters.com/article/topNews/idJPJAPAN-35446920081215
O P E C : 大規模な追加減産も 1 7 日に臨時総会 http://mainichi.jp/select/biz/news/20081216k0000m020092000c.html
O P E C が大幅減産へ、17 日に臨時総会 http://www.nikkei.co.jp/news/kaigai/20081215AT2M1503815122008.html
ロシア、O P E C 加盟に意欲—原油価格下落に警戒感 http://news.searchchina.ne.jp/disp.cgi?y=2008&d=1215&f=business_1215_045.shtml
原油価格急落阻止へ大幅減産か O P E C、1 7 日に総会 http://www.47news.jp/CN/200812/CN2008121501000408.html
O P E C、大幅減産協議へ = 1 7 日にアルジェリアで総会 http://www.jiji.com/jc/c?g=int_30&k=2008121300225
金融危機と原油相場 http://www.gci-klug.jp/crudeoil/2008/12/15/004146.php
石油市場は約 1 億バレル供給過剰 = O P E C 事務局長 http://jp.reuters.com/article/worldNews/idJPJAPAN-35444820081215
原油週間展望 : N Y 原油、O P E C 減産なら 5 0 ドル回復へ http://news.searchchina.ne.jp/disp.cgi?y=2008&d=1215&f=business_1215_007.shtml

表 2 実験に使用した記事

6.2 実験結果と考察

実験に用いた 9 記事からは 24 個の因果関係、つまり、24 本の枝と 48 個の事象ノードにより構成された因果関係ネットワーク (マージ、リダクションを行っていない) が抽出された。その

うち 12 個の事象ノードは手がかり表現が因果関係以外の意味で用いられた文から抽出されていた。事象ノードの表現の精度を検証するため、それら 12 個の事象ノードを除いた 36 個の事象ノードの分析結果を表 3 に示す。

	事象ノード数	割合 (%)
正しく事象が取得できる	19	50.0
事象を表現できていないノード	8	21.1
自然言語解析の失敗	11	28.9
計	38	100

表 3 事象ノードの精度

ノードから元の記事中での因果関係を想起できる成功例と想起できない失敗例を図 9 に示す。図 9(a) の例では、原因事象ノードの事象キーワードは「低迷、精製」であった。これだけでは何に対して精製と述べられているのか分からない。つまり、因果関係を含む文節のみからキーワードを取得した既存手法では元の文での意味を想起できないが、TEC モデルでは記事のタイトルから「原油」を取得できたことで元の文での意味を想起できる。キーワードを拡張したことにより、既存の手法よりも事象ノードが元の因果関係の意味を想起しやすくなったことが確認できた。また、図 9(b) の結果事象ノードは元の意味を想起できなかった。しかし、元の文での結果表現「相場を下押し」の「下押し」が事象キーワードとして取得できていれば、元の意味を想起できた。「下押し」は名詞でないために抽出できなかったが、事象ノードの内容をより正確にするために、名詞以外の単語や否定語もキーワードとして取得する必要があると考えられる。

7. 結 論

本研究ではニュース記事を対象とした因果関係ネットワークを構築するため、因果関係に含まれる語とタイトルに含まれる語から構成される事象ノードと枝の重要度ラベルを定義した TEC モデルを提案した。また、TEC モデルを用いたノードの類似度によるノードのマージと、枝の重要度による因果関係のリダクションを行うことでネットワークを構築する手法を提案した。事象ノードの抽出実験を行い、事象ノードのキーワードを拡張したことによって、事象キーワードのみを事象ノードのキーワードとする既存の手法と比べて、事象ノードから因果関係の意味を想起しやすくなったことが確かめられた。

今後は、プロトタイプシステムを完成させ、提案手法を用いたネットワークの構築を行い、マージやリダクションの精度を検証する。また、古いノードの重みを小さくすることや、古い因果関係の重要度を低くするなど、因果関係の時間推移による変化を考慮したい。さらに、キーワードの上位概念や下位概念などの体系情報などを考慮し、より有用なネットワークを作成していきたい。

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。また、本研究の一部は、NICT 委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」によるものである。

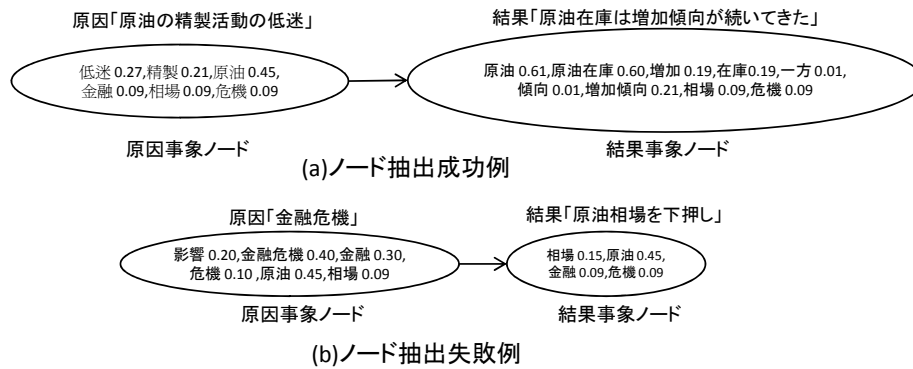


図 9 ノード抽出の成功例と失敗例

文 献

- [1] 佐藤浩史, 笠原要, 松澤和光: テキスト上の表層的因果知識の獲得とその応用, 電子情報通信学会技術研究報告. TL, 思考と言語, Vol. 98, pp. 27–32 (1999).
- [2] 佐藤浩史, 笠原要, 松澤和光: 表層的因果知識ベースによる事象推移予測方式, 情報処理学会 全国大会講演論文集, Vol. 第 56 回平成 10 年前期, No. 2, pp. 251–252 (1998).
- [3] 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, pp. 66–74 (2006).
- [4] 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 (自然言語), 情報処理学会論文誌, Vol. 45, No. 3, pp. 919–933 (2004).
- [5] 坂地泰紀, 竹内康介, 増山繁根根聡: 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. 1144–1147 (2008).
- [6] Sakaji, H., Sekine, S. and Masuyama, S.: Extracting Causal Knowledge Using Clue Phrases and Syntactic Patterns, *Proc. of the 7th International Conference on Practical Aspects of Knowledge Management (PAKM2008)*, pp. 111–122 (2008).
- [7] 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, Vol. 43, No. 6, pp. 1834–1842 (2002).
- [8] 茶筌: <http://chasen-legacy.sourceforge.jp/>.
- [9] GoogleNews: <http://news.google.co.jp/>.