大規模テキストコーパスを用いた因果関係の自動抽出手法

高橋 拓誠　目良 和也　黒澤 義明　竹澤 寿幸
広島市立大学大学院 情報科学研究科
takahashi@ls.info.hiroshima-cu.ac.jp
{mera, kurosawa, takezawa}@hiroshima-cu.ac.jp

1 はじめに

日常的に行われる会話や行動決定に至るまでの推論には，因果関係知識に基づいた判断が重要である．因果関係とは，一般に原因と結果を表す事象間の関係性のことを指しており，こうした常識的な知識は知的エージェントの推論にも応用される．しかし，因果関係知識を人手で構築するには膨大なコストがかかってしまうという問題点がある．そのため，近年では新聞記事や Web ページをクロールしたテキストコーパスから因果関係を自動収集する研究が行われている．従来では，手がかり表現を用いて因果関係を抽出する手法や，構文パターンを用いる手法が提案されてきた．また，因果関係になる事象対は共起しやすいという特性 [1] も報告されており，既存の多くの手法でこの特性が用いられている．

そこで本稿では，手がかり表現と共起情報の両方を用いて，高い精度を保ちつつ大規模な量に対して因果関係となる事象対を抽出する手法を提案する．

2 提案手法

2.1 概要

本稿では，大規模文書コーパスから文章中のサ変名詞 X, Y の間に因果関係が存在するかどうかを自動判定し，因果関係となる名詞対を自動抽出する手法を提案する．本研究では手がかり表現の前後の節に着目することで，共起するサ変名詞対を獲得し因果関係の有無を判定する．因果関係を表す名詞対は様々な手がかり表現で頻出すると考えられることから，本研究では，このような名詞対の抽出に大規模コーパス ClueWeb09[2] を用いる．

提案手法の処理の流れを図 1 に示す．図中の例において，「勉強すれば合格する」では，因果関係を表す名詞対 $< X, Y >$ は $<$ 勉強, 合格 $>$ のように表す．はじめに，手がかり表現を用いて，文書コーパスからサ変名詞を含む文を抽出する．続いて，抽出した文に対して手がかり表現の前後の組み合わせでサ変名詞対を獲得する．獲得したサ変名詞対について，手がかり表現ごとの共起頻度情報を計算し特徴ベクトルの生成を行う．最後に，生成した特徴ベクトルに基づいて，因果関係の有無の自動判定を行う．

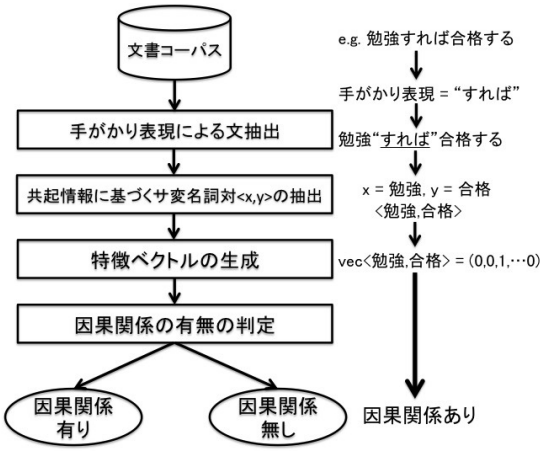

図 1: 提案手法の処理の流れ

2.2 手がかり表現に基づく候補文の抽出

ある文章中に含まれている因果関係を抽出する場合，因果関係を示す手がかり表現を用いて抽出する手法 [3] が一般的な手法の一つとして挙げられる．2.1 節の例では手がかり表現 “ すれば ” を用いることで，「勉強」と「合格」という因果関係を抽出することができる．そのため，因果関係を表す特徴的な手がかり語に着目することで，コーパス中の文章から因果関係となりやすい文の抽出を行う．

本手法で用いた手がかり表現を表 1 に示す．24 個の手がかり表現のうち 18 個 [1] は，津川ら [3] の研究より引用した．残り 6 個 [2] の手がかり表現は，引用した 18

個の手がかり表現では抽出できない例もあると考えたため予備実験の結果に基づいて新しく追加した．

本稿では，以下の手がかり表現を含む文すべてをClueWeb09 のデータの一部から抽出した．

表 1: 文抽出に用いた手がかり語の一覧

を受けて [1]，をきっかけに [1]，を反映し [1]，のため [1]，の背景に [1]，の原因は [1]，によって [1]，に伴う [1]，に支えられて [1]，が理由で [1]，を受け [1]，を背景に [1]，の結果 [1]，の原因として [1]，の影響で [1]，により [1]，に伴い [1]，が影響した [1]，するために [2]，を行うと [2]，による [2]，しないと [2]，すれば [2]，が原因の [2]

2.3 共起情報に基づく $< X, Y >$ の抽出

2.2 節の手がかり表現によりコーパス中から抽出した文書から，因果関係の候補となる名詞対 $< X, Y >$ の獲得を行う．名詞対 $< X, Y >$ 抽出までの流れを以下に示す．なお，2.2 節の手順により抽出した文書集合を D，D 中の文章 i を $sentence_i$ とする．

STEP1: 文書集合 D より，$sentence_i (i = 1, ..., |D|)$ を参照する．

STEP2: $sentence_i$ をコーパスから抽出する際に用いた手がかり表現によって 2 文に分割する．このとき，手がかり表現の直前にあった文を fs，手がかり表現の直後にあった文を bs とする．

STEP3: fs，bs に含まれるサ変名詞をすべて抽出する．ここで，fs から抽出したサ変名詞の系列を $List_{fn} = \{fn_1, fn_2, ..., fn_n\}$，$bs$ から抽出したサ変名詞の系列を $List_{bn} = \{bn_1, bn_2, ..., bn_m\}$ とする．

STEP4: $List_{fn}$ の要素と $List_{bn}$ の要素をすべて組み合わせたサ変名詞対をリストに格納する．($\{< fn_1, bn_1 >, < fn_1, bn_2 >, ..., < fn_n, bn_m >\}$)

STEP5: D 中のすべての文章 i に対して，STEP1〜STEP4 の処理を行う．

2.4 因果関係の有無の判定

2.4.1 手がかり表現と共起情報を用いた特徴ベクトル生成

2.3 節の手順により獲得したサ変名詞対について，因果関係の有無を判定する．因果関係の有無を判定するためには，サ変名詞対を特徴量として表現する必要があるためベクトル表現に変換する．あるサ変名詞対 $< x, y >$ の特徴ベクトルは，24 次元ベクトル $\mathbf{vec}_{x,y} = (score_{cw_1}, score_{cw_2}, ..., score_{cw_{24}})$ で表現される．

$score_{cw_j}$ の計算式を式 (1) に示す．なお，$cw_j (j = 1, ...24)$ は表 1 の手がかり表現，$score_{cw_j}$ は手がかり表現 cw_j に対する $< x, y >$ のスコアである．

$$score_{cw_j} = \log_2\{TF(x,y) * IDF(x) * IDF(y) * PMI(x,y)\} \quad (1)$$

$TF(x, y)$ は文書集合 D における x と y の共起頻度である．また，$IDF(x), IDF(y)$ は文書集合 D における x と y の逆出現頻度である．$PMI(x, y)$ は文書集合 D における x と y の自己相互情報量である．PMI は x と y の共起のしやすさを表しており，式 (2) において $PMI(x, y) > 0$ であれば共起しやすい，$PMI(x, y) < 0$ であれば共起しにくいことを表す．$PMI(x, y) = 0$ であれば各単語が単独で出現する確率と共起する確率が等しいことを表す．

$$PMI(x,y) = log_2 \frac{P(x,y)}{P(x)P(y)} \quad (2)$$

2.4.2 因果関係の有無の判定のための閾値選択

本稿では，因果関係の有無を自動判定するために，機械学習 SVM を用いる手法と閾値を用いる手法の 2 つを提案する．どちらの手法においても特徴ベクトル $\mathbf{vec}_{x,y}$ を用いるが，本節では閾値を用いて判定する方法について説明を行う．

はじめに，獲得したサ変名詞対から表 2 の x に対して $score_{cw_j}$ の値が上位 5 件の $< x, y >$ を解析用データとしてサンプリングした．なお，1 つの x に対してこの操作を $cw_j (j = 1, ..., 24)$ ごとに行い，合計 999 件解析用データとして収集した．

解析用データセットは，ある x, y の間に因果関係があるかどうかを人手でアノテーションしたものである．アノテータは自然言語処理を専門としている大学生 3

表 2: 解析用データのサンプリングに用いた x

練習，飲酒，感染，麻酔，研究，検索，病気，取引，紛争，予約，運動，質問，試験，喫煙，死亡，怪我

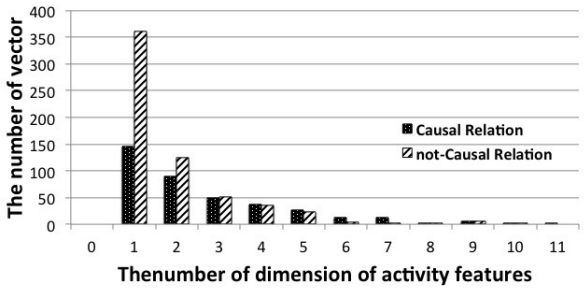

図 2: 活性化している次元数と特徴ベクトルの数の分布

名 (男性，22〜24 歳) に依頼し，2 人以上が因果関係があると判定したデータを正解とした．

解析用データについて，以下の 2 つの数値に対して自動判定のための閾値設定を行う．

1. $\mathbf{vec}_{x,y}$ の活性化している次元の数
2. $\mathbf{vec}_{x,y}$ の大きさ

活性化している次元の数による閾値設定

$\mathbf{vec}_{x,y}$ において，活性化している次元数が多いほど，様々な手がかり表現で頻出する名詞対ということを意味しており，因果関係が成り立つ可能性が高いといえる．活性化している次元とは，24 次元ベクトル $\mathbf{vec}_{x,y}$ において，要素として 0 以外の値をもつ次元のことを指す．

解析用データセット 999 件の特徴ベクトルの活性化している次元数とその特徴ベクトルの数の分布を図 2 に示す．図 2 より，活性化している次元数が 1 のみの時は因果関係のない名詞対が多いことが分かる．一方で，活性化している次元数が 6 以上になるとほとんど因果関係のある名詞対しかないことが確認できる．

以上より，活性化している次元数 = 6 で閾値を設定した．

特徴ベクトルの大きさによる閾値設定

$\mathbf{vec}_{x,y}$ について，大きいベクトルであるほど式 (1) の値が高い，あるいは多くの次元で活性化しているということから，大きいベクトルであるほど因果関係が

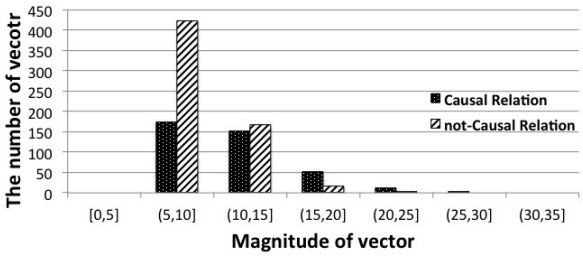

図 3: ベクトルの大きさと特徴ベクトルの数の分布

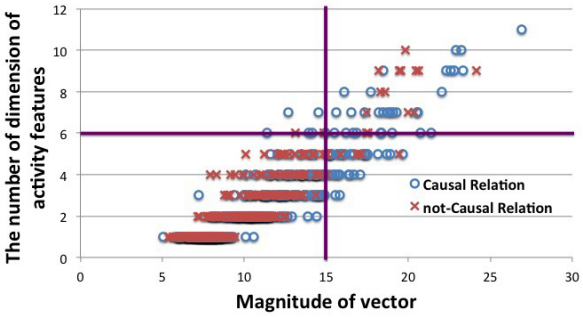

図 4: 活性化している次元数と特徴ベクトルの大きさによる解析用データの分布

成り立つ可能性が高いといえる．特徴ベクトルの大きさは式 (3) により計算した．

$$Mag = \sqrt{\sum_{j=1}^{24} score_{cw_j}{}^2} \tag{3}$$

特徴ベクトルの大きさとその特徴ベクトルの数の分布を図 3 に示す．図 3 より，特徴ベクトルの大きさが 15 以下の場合，因果関係のない名詞対が多いことが分かる．一方で，大きさが 15 より大きくなるとほとんど因果関係のある名詞対しかないことが確認できる．

以上より，特徴ベクトルの大きさ = 15 で閾値を設定した．

さらに，活性化している次元数と特徴ベクトルの大きさの 2 次元空間でプロットした散布図を図 4 に示す．

3 実験

2.4.2 節で用いたデータとは別の評価用データ 800 件について，因果関係の有無の自動判定について性能評価を行う．評価用データは解析用データと同様に，ClueWeb09[2] から 24 個の手がかり表現に基づき抽出した文書集合 D から生成したサ変名詞対である．ClueWeb09 は Web 上のデータをクロールして収集したコーパスであり，今回 24 個の手がかり語を全体の

表 3: 各手法により抽出された因果関係の例

	抽出された因果関係
SVM	< 乾燥, 肌荒れ >, < 努力, 実現 >, < 手術, 矯正 >, < 運動, 発汗 >, < 喫煙, 病気 >
閾値法	< 攻撃, 防御 >, < 乾燥, 化粧 >, < 努力, 成功 >, < 手術, 治療 >, < 運動, ダイエット >

一部に適用することで 807279 文を抽出した．ここで得られた名詞対の中から，解析用データと同様の手法で 800 件を評価用データとしてサンプリングした．評価用データとする名詞対 $< x, y >$ を収集するために用いた x を表 4 に示す．

さらに，評価用データも解析用データと同様にアノテーションをあらかじめ行った．評価用データの内訳は，因果関係あり:400 件，因果関係なし:400 件となっている．

表 4: 評価用データのサンプリングに用いた x

攻撃，借金，乾燥，努力，手術，倒産，転倒，戦争，混雑，告発，ミス，混乱，捜査，破産，紹介，推薦，点滴，評価

提案手法の性能評価をするために以下の 3 つの分析手法について比較を行う．

ランダム法: 評価用データに対してランダムに因果関係の有無を判定する．

SVM: 機械学習 SVM(SupportVectorMachine) を用いて，因果関係のある名詞対の自動分類を行う．学習に用いる素性は 2.4.1 節で生成した 24 次元ベクトルを用いた．評価は Leave-One-Out 手法によって行う．

閾値法: 2.4.2 節により決定した 2 つの閾値 (活性化している次元 = 6，特徴ベクトルの大きさ = 15) を用いる．

各手法によって因果関係として抽出された名詞対の数，精度，再現率，F 値を表 5 に示す．

実験結果より，SVM を用いた手法ではランダムで因果関係の判定をした場合と比較して精度，再現率ともに向上していることが確認できる．一方で，閾値を用いた手法では他のどの手法よりも高い精度で因果関係を抽出することができた．しかし，再現率はどの手法よりも低い値を示した．

閾値を用いた手法は高い精度を得ることができたが，再現率は最も低いという結果を示した．原因として，図 4 より活性化している次元数が低いかつ特徴ベクトルが小さい領域にも，因果関係があるデータが多く存在していることが挙げられる．このため，再現率が低くなったと考えられる．しかし，今回のように大規模データを用いることで，多少の再現率の低下を許容することができる．

表 5: 評価用データを自動分類した各手法の性能比較

	抽出数	精度	再現率	F 値
ランダム法	200	0.50	0.50	0.50
SVM	221	0.70	0.55	0.62
閾値法	52	0.85	0.13	0.23

各手法により抽出された因果関係の例を表 3 に示す．

4 おわりに

本研究では，大規模文書コーパス ClueWeb09 を用いて，サ変名詞間の因果関係の自動抽出を行った．判定手法として，特徴ベクトルから機械学習を用いる手法と閾値を用いる手法を提案した．

閾値を用いた手法では，かなり高精度で因果関係の抽出を行うことができた．しかし，機械学習を用いた手法と比較して，再現率が著しく低い結果を示していた．今回提案した手法では，活性化している次元数が閾値以上であれば因果関係になると判定したため，活性化している次元数が低ければ因果関係として抽出されない．

今後の課題として，再現率の向上のために，因果関係をもつ名詞対同士で共起しやすい次元を設定することが必要である．特定の次元が共起しやすいことが確認されれば，図 4 の散布図において左下の大部分のデータに対しても因果関係があるデータの抽出を行うことができ，再現率の向上が期待できる．

参考文献

[1] 乾孝司, 奥村学. 文書内に現れる因果関係の出現特性調査. 情報処理学会研究報告, SLP-056, pp. 81–86, 2005.

[2] The clueweb09 dataset. `http://www.lemurproject.org/clueweb09.php/` (最終アクセス日: 2016 年 1 月 7 日).

[3] 津川敦朗, 新妻弘崇, 太田学. 交絡事象の発見による因果関係ネットワークの拡張. In *DEIM Forum 2015 E2-3*, 2015.