B3IM2020

修士論文

事象間の因果関係の獲得と一般化に関する研究

佐藤 貴大

2015 年 3 月 25 日

東北大学 大学院
情報科学研究科 システム情報科学専攻

本論文は東北大学 大学院情報科学研究科 システム情報科学専攻に
修士 (情報科学) 授与の要件として提出した修士論文である。

佐藤 貴大

審査委員：

乾 健太郎 教授　　（主指導教員）
篠原 歩 教授
大町 真一郎 教授
岡崎 直観 准教授　（副指導教員）

事象間の因果関係の獲得と一般化に関する研究*

佐藤 貴大

内容梗概

事象の間にある因果関係をとらえることは自然言語を理解する上で非常に重要な問題である。

因果関係の知識は主に自然言語処理における推論や関係抽出、質問応答システムといったタスクに応用される。事象を対象とした因果関係の獲得の手法として統語パターンを利用するものと共起頻度をもととするものが存在するが、多くの研究において因果関係は単語レベルで獲得されるため汎用性が低いという問題点が存在した。

本研究では共起頻度をもととした統計的手法に基づきクラスタにより集約された事象間の因果関係を獲得することで一般化された因果関係獲得の手法を提案する。実験により、最短距離法により作成されたクラスタによる集約を用いることで知識の獲得精度が向上することが分かった。

キーワード

自然言語処理、知識獲得、情報抽出、因果関係、事象

*東北大学 大学院情報科学研究科 システム情報科学専攻 修士論文, B3IM2020, 2015 年 3 月 25 日.

目 次

図 目 次

表 目 次

1　はじめに

本節では本研究の背景と、その目的とするところについて述べる。
また、本論文の構成についても述べる。

1.1　研究の背景

近年、インターネットの発達に伴い利用可能な文書データが人間の可読量を大きく超えて膨大している。このようなデータを効率的に扱うためには計算機による自然言語の解析が必要となる。

計算機を用いて文書データを計算する上で、文書中で生じた「事象」をとらえ、その結果何が起きるのかという「因果関係」をとらえることが重要となる。

その重要性のため、これまでに因果関係の知識を獲得するためのさまざまな試みがなされてきた。因果関係の種類として名詞間の因果関係、事象間の因果関係が挙げられるが、その重要性に反して事象間の因果関係知識は未だ十分であるとはいえない。

本研究では事象を動詞とその項となる名詞の組みとしてとらえ、文書中に出現する 2 つの事象の間の因果関係知識の獲得を目指す。

1.2　目的

本研究は以下のような知識を一般化して獲得することを目的とする。

$$Kill\ (A,someone) \Rightarrow Arrest\ (police,A)$$
$$Murder\ (A,someone) \Rightarrow Arrest\ (policeman,A)$$

上記はいずれも A が誰かを殺したなら警察は A を逮捕するということを意味する。

本論文において $Kill\ (A,someone)$ のように動詞とその項となる名詞からなる組みのことを**事象**と呼ぶ。

また、2 組の事象の間に原因と結果の関係が存在するとき、それらの事象の間には**因果関係**があるとする。

本論文では事象のクラスタによる集約を行い、集約された事象間の因果関係をとらえることで一般化された因果関係知識の獲得を行う。

1.3 本論文の構成

本論文の構成は以下の通りである。

本章では研究の背景と、その目的とするところについて述べた。第 2 章では因果関係に関する関連研究を紹介し、本研究の位置づけを述べる。第 3 章では提案手法について述べ、入力と各処理手順、使用するツールについての説明を行う。第 4 章では提案手法を評価するための手順と評価に用いたデータセットについて説明し、その結果を示す。第 5 章では評価実験により得られた結果をもとに提案手法の分析を行い、問題点と解決策について述べる。第 6 章では本論文のまとめを行う。

2 関連研究

因果関係知識は文書の談話的つながりをとらえる上で非常に重要である。このため因果関係獲得のための多くの研究がなされてきた。因果関係知識は関係をとらえる対象により、名詞間の因果関係知識と事象間の因果関係知識に分けることが出来る。

- 名詞間の因果関係の例

$$Mosquito \Rightarrow Malaria$$

$$Earthquake \Rightarrow Tidal\ Wave$$

- 事象間の因果関係の例

$$Kill\ (X,someone) \Rightarrow Arrest\ (police,X)$$

$$Offer(A) \Rightarrow Refuse(A)$$

また、獲得の方法としては大きく以下のように分類出来る。

- 統語的パターン

  cause や *because* のような因果関係に関わる統語的なパターンを用いて因果関係を獲得する。

  知識の獲得精度は高いがその規模が小さくなる傾向がある。

- 共起頻度

  因果関係にある単語は文書中で連続して出現し易いことを利用して単語間の連続性を計算し因果関係を計算する。

  大規模な知識を獲得出来るが獲得精度が低くなる傾向がある。

以下に関連する研究の紹介を行う。

2.1 統語パターンによる名詞間因果関係

Girju[1] の研究では *cause* や *generate* のような因果関係を記述する統語パターンにもとづき因果関係の獲得を行った。例えば、

Mosquitoes cause Malaria.

のような文から *Mosquito* と *Malaria* の間の因果関係が抽出される。

2.2 共起頻度による名詞間因果関係

Sun ら *[2]* の研究では検索エンジンのクエリログを対象として単語間の共起頻度をもとに名詞間の因果関係を獲得した。例えば、*Earthquake* と $\widehat{Meltdown}$ の共起頻度がそれぞれ単独で出現する場合に比べて大きい場合、*Earthquake* と $\widehat{Meltdown}$ の間の因果関係が抽出される。

2.3 統語パターンによる事象間因果関係

Blanco ら *[3]* は*"because"*などの事象間のの因果関係を表すパターンを用いて因果関係を獲得した。例えば、

The police arrested him because he killed someone.

のような文から *Arrest (police,him)* と *Kill (he,someone)* の間の因果関係が抽出される。

2.4 共起頻度による事象間因果関係

Beamer ら *[4]* は一文中の動詞間の共起頻度を計算することで事象間の因果関係を獲得した。例えば、*offer* と $\widehat{refuse}$ の共起頻度がそれぞれ単独で出現する場合に比べて大きい場合、*Offer(X)* と $\widehat{Refuse(Y)}$ の間の因果関係が抽出される。*Do* ら *[5]* の研究では文書中の任意の事象中の動詞と名詞を対象に共起頻度を計算することで因果関係を獲得した。

本研究では *Do* らの手法における因果関係らしさの計算式である、*Causal-Effect Association*（*CEA*）を用いる。

2.4.1 Causal-Effect Association

事象 e を構成する動詞を p、名詞の集合を A とする。事象 $e_i : p^i(A^i)$ と事象 $e_j : p^j(A^j)$ の間の因果関係を

1. 動詞 p^i と動詞 p^j の間の関係

2. 動詞 p^i と項 A^j、動詞 p^j と項 A^i の間の関係

3. 項 A^i と項 A^j の間の関係

それぞれを計算することで事象間の因果関係らしさを評価する。計算には（*1*）式を用いる。

$$CEA(e_i, e_j) = s_{pp}(e_i, e_j) + s_{pa}(e_i, e_j) + s_{aa}(e_i, e_j) \tag{1}$$

s_{pp} では動詞間の関係、s_{pa} では動詞と名詞の間の関係、s_{aa} では名詞間の関係を計算する。

図 *1* の左上に s_{pp}、左下に s_{pa}、右上に s_{aa} をそれぞれ示す。

2.4.2 動詞間の関係

(1) 式における動詞間の関係を計算する項 s_{pp} についての説明を行う。s_{pp} は以下のように定義される。

$$\begin{aligned} s_{pp} = PMI(p^i, p^j) \times max(u^i, u^j) \\ \times IDF(p^i, p^j) \times Dist(p^i, p^j) \end{aligned} \tag{2}$$

動詞 p^i と動詞 p^j の間の関係は互いの共起のし易さとして $PMI(p^i, p^j)$、*(be,be)* のような多くの文書に出現するペアへのフィルタリングとして $IDF(p^i, p^j)$、文章

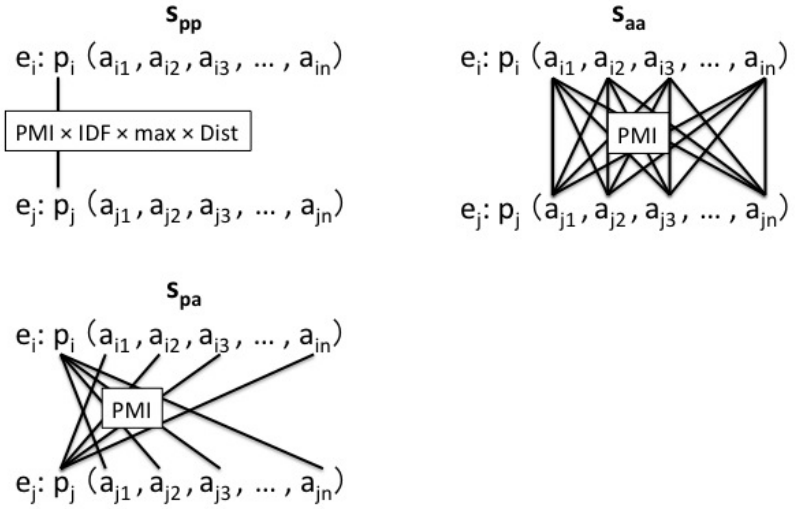


図 1: CEA

単位での距離の近さを評価するために $DIST(p^i, p^j)$、動詞 $p^i(p^j)$ と共起する動詞の内で動詞 $p^j(p^i)$ が占める割合を評価するために $max(u^i, u^j)$ を計算する。以下にそれぞれの項を説明する。

- $PMI(p^i, p^j)$

  動詞間の共起のし易さを図るために自己情報量（PMI）を用いる。

  *Suppes[10]*は事象 e_i が出現する際に単独で出現する頻度よりも事象 e_j を伴って出現する頻度が高い場合、事象 e_i と事象 e_j の間には因果関係にある可能性が高いとした。すなわち $P(e_j|e_i) > P(e_j)$ の時に因果関係になり易い。$P(e_j|e_i) = P(e_i, e_j)/P(e_i)$ であることから上式は $P(e_i, e_j)/P(e_i)P(e_j) > 1$ と書き直せる。

  動詞間の PMI は次式の用に表現され、$P(p^i, p^j)/P(p^i)P(p^i) > 1$ の時、すな

わち事象 e_i と事象 e_j が因果関係になり易いとき、$PMI(p^i, p^j) > 0$ となる。

$$PMI(p^i, p^j) = \log \frac{P(p^i, p^j)}{P(p^i)P(p^j)}$$

- $IDF(p^i, p^j)$

  動詞 p^i と動詞 p^j が多くの文書で出現する場合、例えば $(p^i, p_j) = (be, be)$ のようなペアの場合、因果関係を考える上での重要性は低いことが多い。このため逆文書頻度 *(IDF)* を用いることで動詞ペアの重要性を計算する。

$$IDF(p^i, p^j) = idf(p^i) \times idf(p^j) \times idf(p^i, p^j)$$

  ただし $idf(p) = \log \frac{D}{1+N}$、$D$ は全文書数、N は動詞の出現した文書数とする。

- $max(u^i, u^j)$

  u^i, u^j はそれぞれ以下で定義される。$max(u^i, u^j)$ は u^i と u^j の内より大きい値を返す。

$$u^i = \frac{P(p^i, p^j)}{max_k[P(p^k, p^j)] - P(p^i, p^j) + \varepsilon}$$

$$u^j = \frac{P(p^i, p^j)}{max_k[P(p^i, p^k)] - P(p^i, p^j) + \varepsilon}$$

  u^i,u^j はいま考えている動詞の組みが互いに最も良く共起する時に最大となる。

- $Dist(p^i, p^j)$

  2 つの事象が文単位でどれだけはなれているかを測る。

  事象が離れていればいるほど低い値を返す。

  CEA では文書中の事象について事象間の因果関係を評価する。このため、考える *2* つの事象の距離を計算する必要がある。

  事象間の距離をその事象を含む文の番号の差で表現し、$DIST(p^i, p^j)$ は以下の式で定義される。

$$u^i = -\log \frac{|sent(p^i) - sent(p^j)| + 1}{2 \times ws}$$

  ただし *sent(p)* は事象のセンテンス番号、ws はウィンドウサイズとする。

  本研究では、$ws = 3$ とした。すなわち事象 e^i に対し前後 *2* 文中に存在する事象との関係を考える。

2.4.3 動詞と名詞間の関係

(1) 式における動詞と名詞間の関係を計算する項 s_{pa} についての説明を行う。

$$s_{pa} = \frac{1}{|A_{e_j}|} \sum_{a \in A_{e_j}} PMI(p^i, a) + \frac{1}{|A_{e_i}|} \sum_{a \in A_{e_i}} PMI(p^j, a) \tag{3}$$

動詞と全名詞の間の *PMI* を計算し、平均を求める。

2.4.4 名詞間の関係

(1) 式における名詞間の関係を計算する項 s_{aa} についての説明を行う。

$$s_{aa} = \frac{1}{|A_{e_i}||A_{e_j}|} \sum_{a \in A_{e_i}} \sum_{a' \in A_{e_j}} PMI(a, a') \tag{4}$$

事象間の全名詞についてそれぞれ *PMI* を計算し、平均を求める。

3　クラスタ化による因果関係知識の獲得

本研究は文書中に現れる因果関係にある 2つの事象を認識し、事象間の一般化された因果関係知識を得ることを目的とする本章では知識獲得のための提案手法について、それを行うために必要となるツールとともに説明する。

3.1　システム概要

文書を入力として受け取り、システムは以下のように処理を行う。

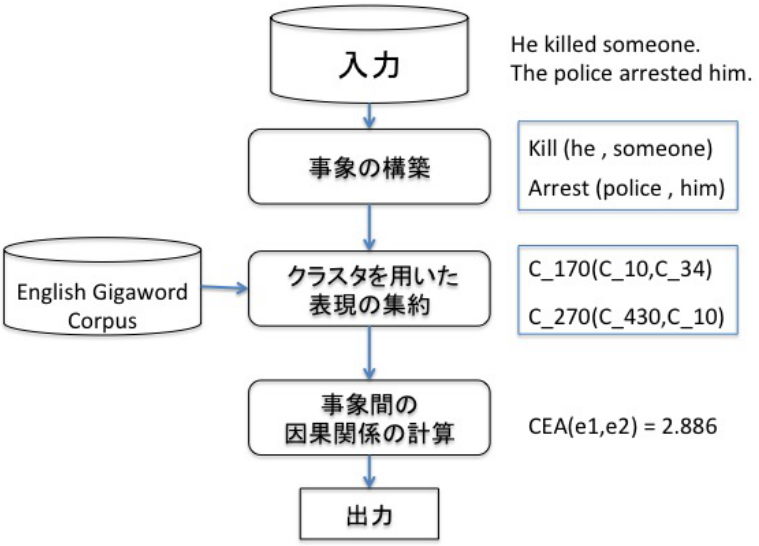


図 2: システム概要

1. 事象の構築

   文書に対し品詞タグ付け、レンマ化、構文解析を行う。動詞に対して構文解析結果をもとに項を取り出し事象を構築する。文書の解析は *Stanford CoreNLP*

を用いて行われる。

2. クラスタを用いた表現の集約

   事象の集約による因果関係知識の一般化を目的としてクラスタ化を行うクラスタは *Word2Vec* により作成され、事象中の動詞と名詞はクラスタ *ID* によって置換される。

3. 事象間の因果関係の計算

   動詞と名詞がクラスタ化された事象の集合から任意の *2* つの事象について頻度を主とした統計情報により因果関係を計算する

これらの処理により、文書中の *2* つの事象と、その事象対が因果関係にあるかどうかを計算した結果を出力とする。

3.2 事象の構築

本研究では、*1.1* 節で述べたように、事象を動詞とその項となる名詞の組みとして扱う。このため、与えられた文書に対して品詞タグ付けにより動詞と名詞を認識すること、構文解析により動詞の項を認識することが必要となる。

また、頻度を主とした統計情報を扱うために、レンマ化を必要とする。

事象の構築は以下の手順で行われる

1. 文書の解析文書に対して品詞タグ付け、構文解析、レンマ化を行う。

2. 動詞の検出品詞タグをもとに動詞を検出する。

3. 項の選択検出された動詞毎に構文解析をもとにして項を選択する。

   項は動詞に対して *nsubj*（名詞主語）、*nsubjpass*（受動節の名詞主語）、*dobj*（直接目的語）、*iobj*（間接目的語）、*agent(*受動態に対し *by* で係る補語*)* により係っている名詞を全て選択する。

図 *3* に事象の構築手順を例示する。

文書の解析の各処理は *Stanford CoreNLP* を用いて行う。

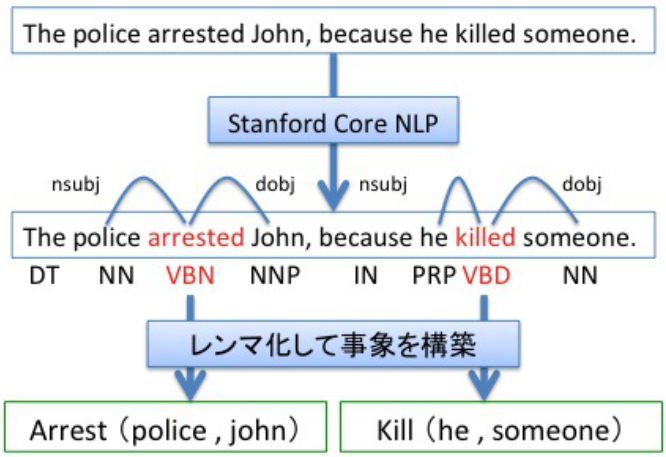


図 3: 事象の構築例

3.2.1 Stanford CoreNLP

Stanford CoreNLP は *Stanford Natural Language Processing Group* によって開発されている英語テキスト用の自然言語開発ツールである *[6]*。

本研究では事象の構築とその後の頻度を主とした統計情報を扱うため、入力された全文書に対して以下の解析を行う。

品詞タグ付け

文中の単語に対して品詞をラベル付けする。タグは *Penn TreeBank* タグセット *[7]* をもとに付けられる。今回利用した動詞と名詞に関するタグの一覧を表 *1* に示す。

表 1: Penn TreeBank タグセット：動詞と名詞の例

品詞タグ	説明	事象中での扱い
NN	普通名詞（単数形）	名詞
NNS	普通名詞（複数形）	
NNP	固有名詞（単数形）	
NNPS	固有名詞（複数形）	
PRP	代名詞	
PRP$	所有確代名詞	
VB	動詞（基本形）	動詞
VBD	動詞（過去形）	
VBG	動詞（現在分詞・動名詞形）	
VBN	動詞（過去分詞形）	
VBP	動詞（現在形：非三人称単数形）	
VBZ	動詞（現在形：三人称単数形）	

構文解析 付与された品詞タグをもととして単語間の依存構造を決定する。例として、以下のような文を解析した結果を図 *4* に示す。

The police arrested John, because he killed someone.

図 *4* では、動詞 *arrested* に対し *police* が *nsubj*（名詞主語）、*John* が *dobj*（直接目的語）として係る。

レンマ化 単語に対してレンマ化（見出し語化）を行う。

例として *police* と *polices* のレンマは共に *police* となる。

動詞の基本形/ 活用形や名詞の単数形/ 複数形などを同一の単語として扱うことが出来るようになるため、頻度を主とした統計情報を扱う上で有用となる。

3.3 クラスタを用いた事象の集約

構築した事象を集約するためにクラスタの作成を行う。単語のベクトルを構築し、単語間の *cos* 距離を求めることでクラスタリングを行う。

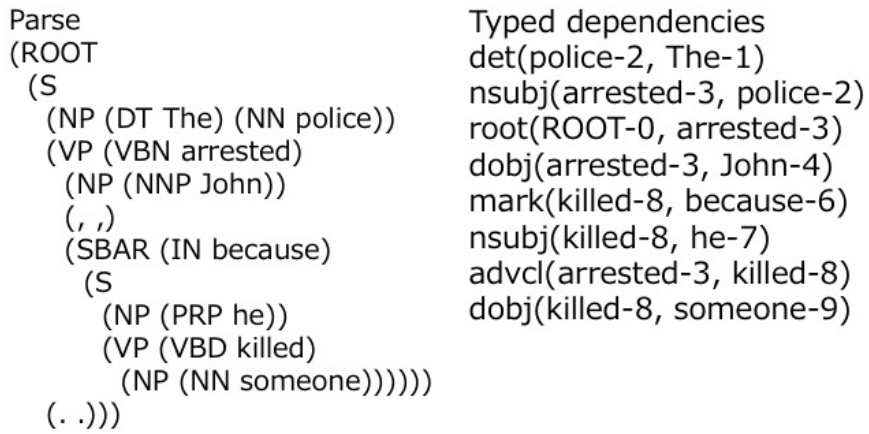


図 4: 構文解析

3.3.1 ベクトルの構築

単語の意味をとらえたベクトルの構築のためのツールとして、*word2vec* を用いた。*[9]*

word2vec では「文脈の似た単語は近い意味を持つ」という前提をもととして単語を *200* の要素の組み合せで表現する。本研究では入力コーパスとして *English Gigaword Fifth Edition* を用いた。

3.3.2 English Gigaword Fifth Edition

English Gigaword Corpus はペンシルバニア大学の *Linguistic Data Consortium (LDC)* によって作成されている、英語の新聞記事が蓄積されたコーパスである。*[8]* 今回はその第５版のうち、*NewYorkTimes* からの *20*万文書を入力として *word2vec*

による単語のベクトル化を行った。

3.3.3 単語のクラスタリング

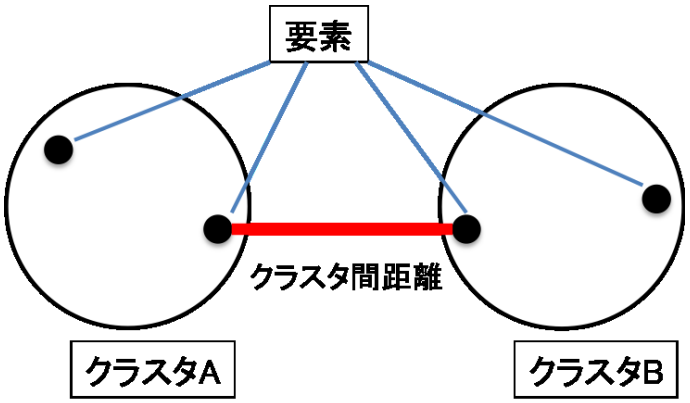
代表的なクラスタリング手法として以下のような手法が存在する。

- 階層型クラスタリング
  - 最短距離法（図 *5.*（*a*））
    最も近い要素間の距離をクラスタの距離とする。
    1 つのクラスタに順に要素が *1* つずつ吸収されてしまう鎖効果が生じることがある。
  - 最長距離法（図 *5.*（*b*））
    最も遠い要素間の距離をクラスタの距離とする。
    要素が多いクラスタほど他のクラスタとの距離が離れてしまう拡散現象が生じることがある。
  - 群平均法
    要素間の距離の平均をクラスタの距離とする。
    鎖効果や拡散現象を起こさないが計算量が多い
  - *Ward* 法
    クラスタ P と Q に対してクラスタ間の距離を以下で定義する

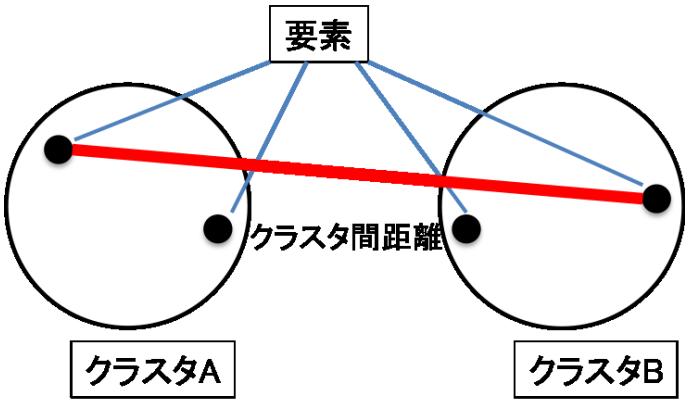
$$E(P \cup Q) - E(P) - E(Q)$$

    ただし *E(A)* は A の全要素から重心までの距離の *2* 乗和。分類感度は高いが計算量が多い。

- 非階層型クラスタリング
  - *k-means* 法（図 *6* ）
    クラスタ数 k を与える。



(a)最短距離法



(b)最長距離法

図 5: 階層型クラスタリング手法

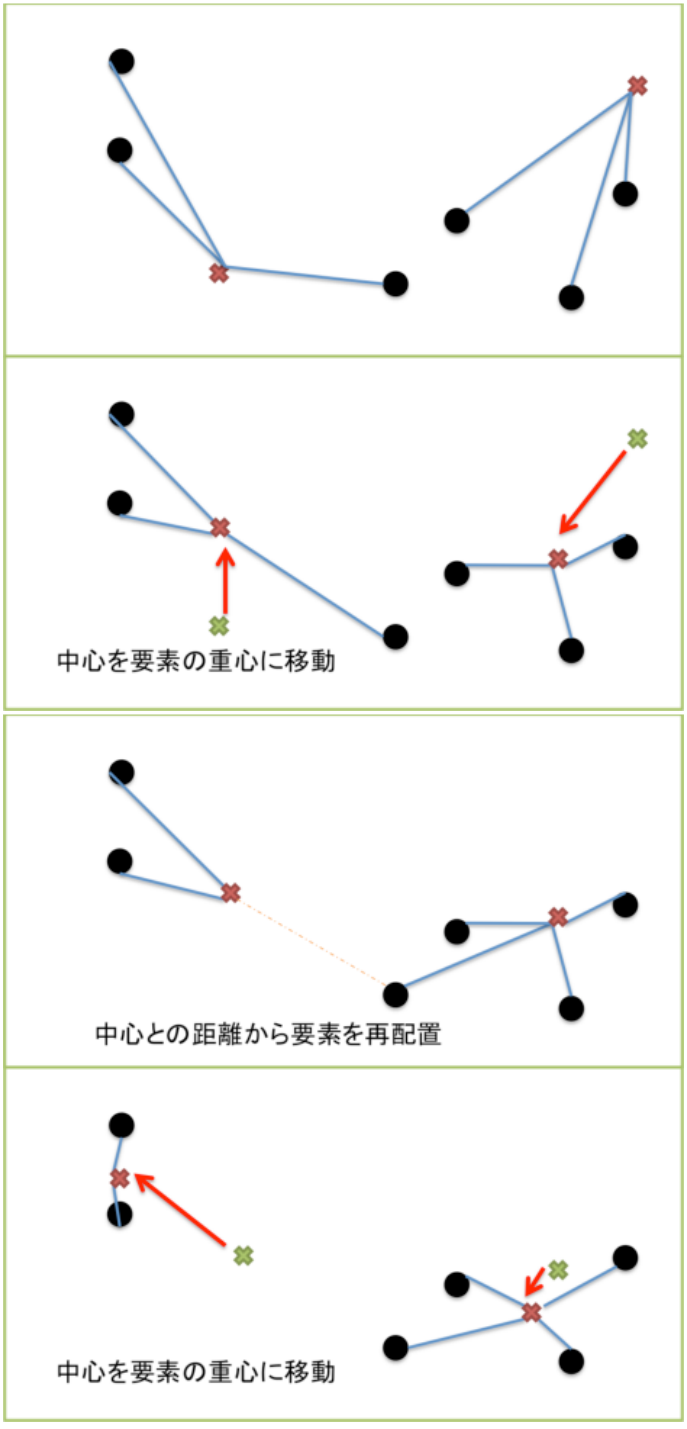


図 6: k-means 法

k 個の要素をランダムに選択しクラスタの中心とする。以下の処理を収束するまで繰りかえす。

1. 全要素を最も近いクラスタに割り当てる
2. クラスタのの中心を現在のクラスタの重心に更新

本研究では階層型クラスタリング手法として最短距離法、最長距離法、非階層型クラスタリング手法として *k-means* 法を用いてクラスタを作成した。単語間の距離として、*word2vec* により学習された単語ベクトルの *cos* 距離を用いた。表 *2* に最短距離法を用いて作成したクラスタの一部、表 *3* に最長距離法を用いて作成したクラスタ、表 *4* に *k-means* 法を用いて作成したクラスタをそれぞれ例示する。

それぞれのクラスタの特徴として、最短距離法を用いたものは多くの単語が集約されるが意味的に遠い単語も同一クラスタに集約されることがある。最長距離法を用いたものはクラスタ内には意味的に近い単語が集約されるが近い単語が別々のクラスタにまとまることがある。*k-means* 法を用いたクラスタは始めにランダムに割り当てたクラスタの中心によって作成されるクラスタが毎回変わることがある。

3.3.4 事象の集約

事象を入力として、クラスタ化の処理は以下の手順で行う。

1. 事象を構成する動詞と名詞それぞれに対し、同一の表現が含まれるクラスタが存在するかを確認する。
2. 同一の表現が含まれるクラスタが見つかった場合、単語をクラスタの *ID* で置き換える
3. 含まれるクラスタが存在しない場合（例えば固有名詞等）、*1* 単語で *1* クラスタとしてあつかう。

表 2: 最短距離法により作成したクラスタ例（一部）

クラスタ ID	token	(意味)
1639	anthropology	(人類学)
	astronomy	(天文学)
	biologist	(生物学者)
	census	(国勢調査)
	conclusion	(結論)
	ecosystems	(生態系)
	enzymes	(酵素)
	experiment	(試す)
	extinction	(絶滅)
	findings	(発見)
	gene	(遺伝子)
	geneticist	(遺伝学者)
	graduate	(卒業)
	inference	(推論)
	laboratory	(研究室)
	microbiology	(微生物学)
	mutation	(突然変異)
	neurologist	(神経科医)
	ornithologists	(鳥類学者)
	physicist	(物理学者)
	poll	(世論調査)
	prof	(教授)
	report	(レポート)
	researcher	(研究者)
	scientific	(科学)
	sociology	(社会学)
	statistical	(統計)
	study	(勉強)
	survey	(調査)
	teach	(教える)
	zoologist	(動物学)

表 3: 最長距離法により作成したクラスタ例

クラスタ ID	token	(意味)
1109	crawl	(はう)
	roam	(歩き回る)
	stroll	(散歩する)
	walk	(歩く)
	walking	(歩く)
	wander	(ぶらつく)
1935	poll	(世論調査)
	research	(研究)
	researcher	(研究者)
	study	(勉強する)
	survey	(調査)
10246	ambivalence	(ためらい)
	bitterness	(反感)
	despair	(失望)
	hopelessness	(絶望的)
	misery	(惨め)
	vindictiveness	(執念深い)

表 4: k-means 法により作成したクラスタ例

クラスタ ID	token	(意味)
213	care	(治す)
	checkup	(検査)
	clinic	(クリニック)
	dentist	(歯医者)
	disease	(病気にかからせる)
	doctor	(医師)
	health	(健康状態)
	hospice	(ホスピス)
266	agitation	(動揺)
	alienation	(仲違い)
	annoyance	(困りごと)
	bemusement	(困惑)
	chivy	(しつこく悩ます)
	despair	(失望)
	desperation	(自暴自棄)
365	abscond	(誘拐する)
	assault	(暴行)
	behead	(打ち首にする)
	kill	(殺す)
	kidnap	(誘拐する)
	murder	(殺害する)
	suicide	(自殺)

3.4　事象間の因果性の計算

統計情報にもとづき事象間の因果関係を計算する。

クラスタを用いて集約がなされた事象間に対して（*1*）式をもとに因果関係らしさを評価する。

（*1*）式は動詞、名詞からなる *2* つの事象に対して適用されるものであるが、集約により置き換えられたクラスタ *ID* を *1* つの単語とみなして計算を行うことで同様の評価が可能となる。

4　実験

本章では、提案手法の評価実験とその結果について述べる。

式の *CEA* を計算するにあたり、実験ではその統計情報を得るため *ClueWeb12* データセットを用いた。

ClueWeb12 はウェブからクロールされ収集された大規模データセットである。

データセットの約 *7* 億文書の内、今回は約 *2000* 万文書を対象として統計情報を求めた。

4.1　評価用・開発用データセット

評価用、開発用のデータセットとして、*Do* らにより公開されている因果関係が人手でアノテートされたデータセットを用いた。

データセットは評価用に *20* 文書、開発用に *5* 文書に対してそれぞれ人手によるアノテートがなされ、文書中の因果関係にあると思われる *2* つの単語に対して関係のタグが付与される。

評価用データセットでは *492* のタグ、開発用データセットでは *92* のタグが含まれ、その一致率は *0.67* である。

4.2　実験設定

評価用データセットをもとに因果関係を評価する。

評価指標として *Precision*、*Recall* を用いる。クラスタ化による知識獲得精度の変化を調査するため、*CEA* を用いた計算により獲得され得る関係を正解データの全数とした。すなわち、システムにより得られる関係を全て因果関係とみなした時 *Recall* が *100%* となるようにした。

$$
\begin{aligned}
Precision &= \frac{\text{システムから得られた正しい因果関係}}{\text{システムから得られた因果関係}} \\
Recall &= \frac{\text{システムから得られた正しい因果関係}}{\text{手法により獲得され得る因果関係}}
\end{aligned}
$$

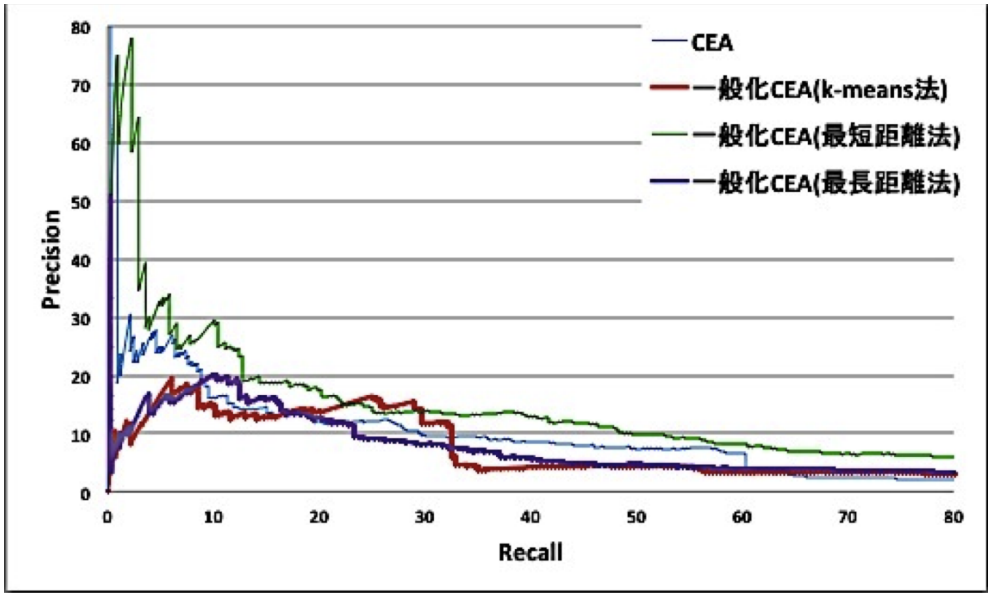


図 7: クラスタリング手法の比較

4.3 クラスタリング手法による比較

クラスタ化を用いた因果関係の知識獲得において有効なクラスタリング手法の評価のために評価実験を行う。

階層型クラスタリング手法として最短距離法と最長距離法、非階層型クラスタリング手法として *k-means* 法によるクラスタ化を用いた *CEA* 及び、クラスタ化を用いない *CEA* を対象として実験を行った。階層型クラスタリング手法では最短距離法ではしきい値としてクラスタ間の *cos* 距離 *0.6*、最長距離法ではしきい値としてクラスタ間の *cos* 距離 *0.7* として与えた。また、非階層型クラスタリングではクラスタ数を *500* として与えた。

図 *7*は *CEA* に対して *k-means* 法、最短距離法 *(nn* 法*)*、最長距離法 *(fn* 法*)* によるクラスタ化を行った場合の *Precision* と *Recall* のグラフである。*PR* 曲線の比較から、最短距離法を用いたクラスタ化ではクラスタ化を行わない *CEA* と比較し

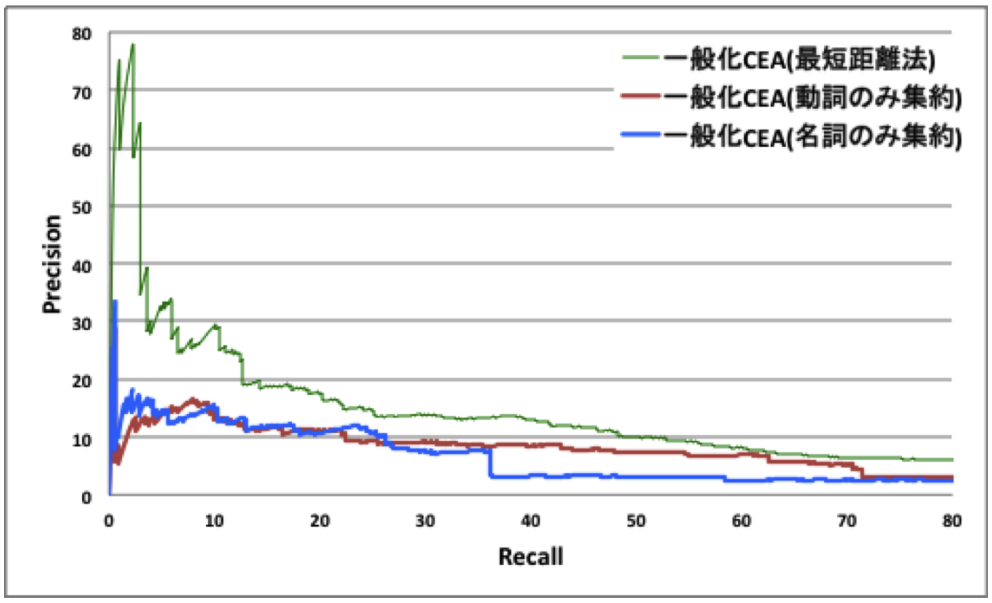


図 8: 動詞・名詞に対するクラスタ化

獲得精度が上昇しているのに対し、最長距離法を用いたクラスタ化では逆に獲得精度が下がっていることが分かる。

次に最短距離法を用いたクラスタ化をもとに、動詞・名詞それぞれに対するクラスタ化の有効性を評価する

4.4 動詞・名詞に対するクラスタ化

最短距離法によるクラスタ化に対して、動詞のみのクラスタ化、名詞のみのクラスタ化をそれぞれ比較することでクラスタ化が動詞・名詞それぞれに対する有効性を評価する。

図 *8* は動詞のみのクラスタ化を行った場合と名詞のみのクラスタ化を行った場合の *Precision* と *Recall* のグラフである。

グラフから、動詞のみのクラスタ化を行った場合と、名詞のみのクラスタ化を行った場合の孰れにおいても動詞・名詞共にクラスタ化した結果よりも獲得精度が下がることが分かる。名詞、動詞の一方をクラスタ化しないことで獲得精度が

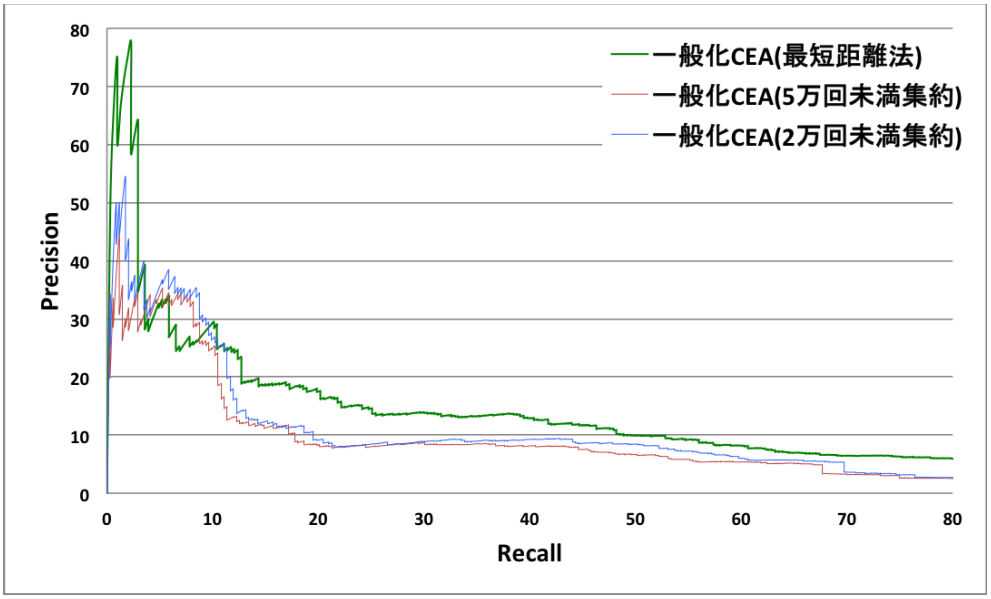


図 9: 低頻度単語に対するクラスタ化

下がることから、クラスタ化は因果関係知識獲得において動詞・名詞のどちらに対しても有効に働いていることが分かる。

次に、低頻度の単語のみに対するクラスタ化を実験する。

4.5 低頻度に対するクラスタ化

コーパス中に高頻度で出現する単語をクラスタ化の対象外とすることで、低頻度の単語のみの集約を実験する。

図 *9* はコーパス中に *2* 万回以上出現する単語と *5* 万回以上出現する単語をそれぞれクラスタ化の対象から外した場合の *Precision* と *Recall* のグラフである。

グラフから、高頻度の単語のクラスタ化を行わなかった場合には獲得の精度が落ちてしまうことが分かる。

5　考察

実験の結果についてそれぞれ原因の分析と考察を行う。

5.1　クラスタリング手法に対する分析

5.1.1　最長距離法による精度の低下

実験の結果、最長距離法を用いたクラスタ化ではクラスタ化を行う前よりも精度が下がっていた。これは、*1*つ*1*つのクラスタに集約される単語数が非常に少ないためと考えられる。最長距離法ではクラスタ間の距離はそれぞれのクラスタに属する最も遠い要素の距離で与えられる。このため、しきい値を*cos*距離 *0.6*として与えたとき、クラスタ間の単語対のどれか*1*組でも *0.6*以上はなれているときクラスタは別々に分割されることとなる。

作成されたクラスタはそれぞれ数単語ずつのみの集約がなされていた。このため、本来期待していた低頻度の単語間の集約はほとんどなされず、逆に高頻度の単語が集約されることで獲得の精度が下がる結果となった。

5.1.2　最短距離法におけるエラー分析

獲得された知識の上位*50*件を対象として分析を行った。

- クラスタ化により獲得出来なくなった知識

  上位*50*件においてクラスタ化を行わない*CEA*により獲得されていた正しい因果関係知識*13*件の内、クラスタ化により獲得がより困難となった関係が*7*件存在した。このうち、クラスタ化により高頻度の単語が過度に集約されてしまったエラーが*71%*、学習データに含まれずクラスタ化出来なかったことによるエラーが*29%*存在した。

- クラスタ化により獲得されるようになった誤った知識

  クラスタ化により獲得された誤った因果関係知識*34*件のうち、クラスタ化を行う前には得られなかった誤った知識が*16*件存在した。このうち、クラ

表 5: 最短距離法により過度に高頻度単語が集約されたクラスタ

クラスタ ID	token
13	able
	can
	do
	essential
	help
	intended
	must
	necessary
	need
	needed
	required
	sufficient
	try
	want

スタ化により高頻度の単語が過度に集約されてしまったエラーが *19%*、クラスタリングのエラーが *89%*存在した。

最短距離法ではクラスタ間の距離はそれぞれのクラスタに属する最も近い要素の距離で与えられる。表 *5*に最短距離法により過度な集約が生じたクラスタの例を示す。

5.2 低頻度に対するクラスタ化の分析

高頻度の単語へのクラスタ化を行わなかった場合、逆に精度が低くなる結果となった。

評価用データにおいて因果関係を持つとされる *350*種類の単語を調査した結果、コーパス中で *5*万回以上出現する頻出単語が *163*種類、*2*万回以上出現する単語

は *232* 種類存在していた。このため、真に計算すべき対象の約半数がクラスタ化による集約がなされない結果となってしまっていた。

高頻度の単語全てを対象外とするのではなく、高頻度の単語の一部については例えば変化形の単語のみ集約するといったように単語毎にルールベースにより処理を行うことや、高頻度の単語のみ最長距離法によるより近い意味の単語のみのクラスタリングのように別のクラスタリング手法により作成されたクラスタで集約を行うといった処理により集約する対象を適切に選択することが求められる。

6　まとめ

最短距離法、最長距離法、k-$means$法によりそれぞれ作成されたクラスタにより事象の集約を行い、共起頻度により統計的に因果関係を計算することで一般化された因果関係知識を獲得する手法を提案した。

最短距離法を用いて作成されたクラスタによる集約を行ったものが最も因果関係の獲得精度を向上させた。

動詞・名詞それぞれにのみ集約を行った結果よりどちらも集約を行った場合の方が精度が高いことから、クラスタによる集約は動詞・名詞ともに有効であることが分かった。

獲得された知識に対する分析から、コーパス中での高頻度単語に対する過度な集約が問題となっていた。しかし単純に高頻度の単語の集約を行わない場合、逆に獲得精度が下がってしまう結果となった。これは評価データ中の因果関係にあるとされる単語の頻度が大きいものが多かったためと考えられる。

今後の展望としては、クラスタリングのエラーの割合が高いことから、クラスタリングの精度向上に伴い、システムの精度の向上が見込まれる。また、高頻度単語に対するクラスタ化による集約を適切に行う必要があると考えられる。

謝辞

本研究を進めるにあたって、多くの方々のご協力を頂きました。心よりの感謝を申し上げます。

主指導教員である乾健太郎教授には、研究会において多くのコメントや助言を頂き、また、研究を行う上での心構えを教えて頂きました。

副指導教員である岡崎直観准教授には研究の方向性を始めとして多くの御指導を頂き、また、お忙しい中たくさんのコメントや助言を頂きました。

また、研究室の先輩や同期、後輩の皆様方には研究のことのみならず、日頃から多くの助けを頂きました。充実した研究生活を送ることが出来たのは皆様方のおかげであると考えております。

研究室で得た知識や多くの刺激をもとに今後の人生に役立てていく所存です。

参考文献

[1] Girju, Roxana "Automatic detection of causal relations for question answering", Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12 76–83 2003 Association for Computational Linguistics

[2] Sun, Yizhou, et al. "Causal relation of queries from temporal logs." Proceedings of the 16th international conference on World Wide Web. ACM, 2007.

[3] Blanco, Eduardo, Nuria Castell, and Dan I. Moldovan. "Causal Relation Extraction." LREC. 2008.

[4] Beamer, Brandon, and Roxana Girju. "Using a bigram event model to predict causal potential." Computational Linguistics and Intelligent Text Processing (2009): 430-441.

[5] Do, Quang Xuan, Yee Seng Chan, and Dan Roth. "Minimally supervised event causality identification." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011.

[6] Manning, Christopher D., et al. "The Stanford CoreNLP natural language processing toolkit." Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2014.

[7] Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. "Building a large annotated corpus of English: The Penn Treebank." Computational linguistics 19.2 (1993): 313-330.

[8] Parker, Robert, et al. "English gigaword fifth edition, june." Linguistic Data Consortium, LDC2011T07 (2011).

[9] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in Neural Information Processing Systems (2013): 3111-3119.

[10] Patrick Suppes. 1970. A Probabilistic Theory of Causality. Amsterdam: North-Holland Publishing Company.

発表文献一覧

国内会議・研究会論文

- 佐藤貴大, 岡崎直観, 乾健太郎. ウェブ文書の構造を利用した場所名・住所ペアの獲得. 人工知能学会第 *27*回全国大会, *3E3-5, June 2013.*