

# 因果関係知識獲得のための隠れ変数モデル<sup>1</sup>

乾 孝司 †

tinui@lr.pi.titech.ac.jp

† 日本学術振興会 特別研究員

高村大也 ‡

takamura@pi.titech.ac.jp

‡ 東京工業大学 精密工学研究所

奥村 学 ‡

oku@pi.titech.ac.jp

## 1 はじめに

因果関係に関する知識は、質問応答システムや対話システムなど、幅広い自然言語アプリケーションにとって重要な知識のひとつである。近年では、大規模な文書集合から自動的に因果関係知識を獲得する手法が既に幾つか提案されている。

文書内では、因果関係にある出来事対は共起して出現しやすく、特に、日本語では、係り受け関係を持ちやすい [16]。因果関係知識の自動獲得に関する既存研究には幾つかのアプローチがあるが、その多くは、明示的あるいは暗黙的に、因果関係にある出来事対は文書内で共起しやすいという出現傾向を利用している [13, 4, 2, 15]。

本研究では、出来事対の共起情報に加え、少数の教師信号（因果関係にあるか否かが既にわかっている出来事対）を利用して、任意の出来事対が因果関係にあるか否かを判定する確率的因果モデルを構築、提案する。提案モデルには、次の特徴がある：

- 半教師あり学習によるモデル構築、
- 手がかり語句への非依存性、
- 隠れ変数の導入による疎データ問題への対応。

因果関係に関する大規模な教師信号を作成することは負荷が高く現実的でないため、教師信号と共起出来事対の情報を混合して利用する。教師の有無という観点から見れば、共起出来事対は教師なしデータであり、提案モデルは半教師あり学習アルゴリズムにより構築される。次に、ここでいう共起出来事対は因果関係の手がかり語句（「ため」や「ので」など）の有無を問わず収集されることを仮定する。このことは、従来モデル（e.g., [7]）と違い、提案モデルは手がかり語句の情報に依存せずに構築されることを意味する。結果、因果関係知識の獲得における高い被覆率が確保できる。また、一般には、モデル構築に利用可能なデータが疎になり、十分に信頼できる出来事対の共起情報を得ることは困難である（ゼロ頻度問題）。提案モデルでは、原因、結果となる出来事に対する隠れ変数を導入することで、この問題を回避し、訓練データに出現しない出来事対についても因果関係にあるか否かの判定を可能にする。

本稿の構成を示す。まず、2 節で関連研究について述べ、3 節で提案モデルを説明する。その後、4 節で評価実験について述べ、5 節で本稿をまとめる。

## 2 関連研究

因果関係に関する知識、あるいは、言語表現を文書集合から自動獲得する研究について述べる。

Girju et al. [4] は、手がかり語句を利用して、英語文書から因果関係を表す名詞句対を獲得した。Sanchez-Graillet et al. [11] は、英語名詞句間の因果性を判定するベイジアンネットワークを構築している。

Torisawa [13] は、動詞並列句に注目した、手がかり語句に依存しない手法を提案しているが、そこでは教師信号は利用されない。Chang et al. [2] や山田ら [15] は、ナイーブベイズ分類器（NB）に教師なしデータを取り込む方法 [10] を用いて、因果関係知識を自動獲得した。ここで、因果関係をモデル化するには、出来事間の依存関係を捉えることが重要であると考えられる。しかし、彼らは分類器として NB を用いているために、出来事対の間の依存関係は陽にモデル化されておらず、改良の余地がある。本研究では、出来事対の間の依存関係を陽にモデル化する。なお、Torisawa [14] や Inui [7] では、因果関係の種類を下位分類したり、一部の下位分類のみに注目しているが、本研究では因果関係の下位分類は考慮しない。

## 3 確率的因果モデル

原因側、結果側の出来事を  $x$  および  $y$ 、それらの対を  $\langle x, y \rangle$ 、出来事対が因果関係にあるか否かを表す因果関係クラスを  $c \in \{0, 1\}$  で表し、教師信号となる共起出来事対を  $\langle x, y, c \rangle$  で表す。我々の目標は、教師信号の集合  $D^l = \{\langle x, y, c \rangle_1, \dots, \langle x, y, c \rangle_{|D^l|}\}$  と共起出来事対の集合  $D^u = \{\langle x, y \rangle_1, \dots, \langle x, y \rangle_{|D^u|}\}$  からなるデータ  $D = D^l \cup D^u$  を利用して、任意の出来事対  $\langle x, y \rangle$  が因果関係にある ( $c = 1$ ) か否 ( $c = 0$ ) かを判定する確率モデルを構築することである。

### 3.1 モデル構造

図 1 を参照しながら提案モデルを説明する。図 1 下段の 3 つのモデルが本稿で新たに提案するモデルである。

まず、共起データのための確率モデルを拡張する。図 1 の (a) および (b) は、Hofmann et al. [6] が提案し

<sup>1</sup>Latent Variable Models for Causal Knowledge Acquisition

$$P(x, y, z^x, z^y, c) = P(x|z^x)P(y|z^y)P(z^x|c)P(z^y|z^x, c)P(c) \quad (1)$$

$$Q(\theta) = \sum_{x_i} \sum_{y_j} \sum_{c_m} N_{x_i y_j c_m} \sum_{z_k^x} \sum_{z_l^y} \bar{P}(z_k^x, z_l^y | x_i, y_j, c_m) \log P(x_i | z_k^x) P(y_j | z_l^y) P(z_k^x | c_m) P(z_l^y | z_k^x, c_m) P(c_m) \quad (2)$$

$$+ \sum_{x_i} \sum_{y_j} N_{x_i y_j} \sum_{z_k^x} \sum_{z_l^y} \sum_{c_m} \bar{P}(z_k^x, z_l^y, c_m | x_i, y_j) \log P(x_i | z_k^x) P(y_j | z_l^y) P(z_k^x | c_m) P(z_l^y | z_k^x, c_m) P(c_m)$$

$$\bar{P}(z_k^x, z_l^y | x_i, y_j, c_m) = \frac{P(x_i | z_k^x) P(y_j | z_l^y) P(z_k^x | c_m) P(z_l^y | z_k^x, c_m) P(c_m)}{\sum_{z_k^x} \sum_{z_l^y} P(x_i | z_k^x) P(y_j | z_l^y) P(z_k^x | c_m) P(z_l^y | z_k^x, c_m) P(c_m)} \quad (3)$$

$$\bar{P}(z_k^x, z_l^y, c_m | x_i, y_j) = \frac{P(x_i | z_k^x) P(y_j | z_l^y) P(z_k^x | c_m) P(z_l^y | z_k^x, c_m) P(c_m)}{\sum_{z_k^x} \sum_{z_l^y} \sum_{c_m} P(x_i | z_k^x) P(y_j | z_l^y) P(z_k^x | c_m) P(z_l^y | z_k^x, c_m) P(c_m)} \quad (4)$$

$$P(c_m | x_i, y_j) = \frac{\sum_{z_k^x} \sum_{z_l^y} P(x_i | z_k^x) P(y_j | z_l^y) P(z_k^x | c_m) P(z_l^y | z_k^x, c_m) P(c_m)}{\sum_{z_k^x} \sum_{z_l^y} \sum_{c_m} P(x_i | z_k^x) P(y_j | z_l^y) P(z_k^x | c_m) P(z_l^y | z_k^x, c_m) P(c_m)} \quad (5)$$

図 2: クラス付きプロダクトモデルの (1) モデル構造, (2) Q 関数, (3)(4) E ステップでの計算式と (5) クラス判定式

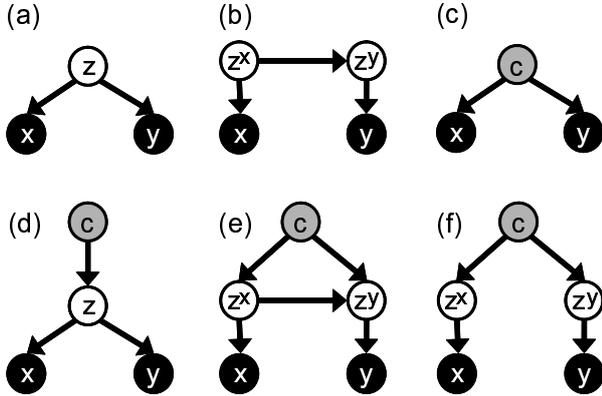


図 1: 提案モデルのグラフ表現:  $x, y$  は観測変数 (原因側出来事, 結果側出来事),  $c$  は部分観測変数 (因果関係クラス),  $z, z^x, z^y$  は隠れ変数.

た共起データのための確率モデルである. (a) をアスペクトモデル (Asp), (b) をプロダクトモデル (Pro) と呼ぶ. モデル内の  $z$  および  $z^x, z^y$  は隠れ変数である. 隠れ変数の各状態は  $x, y$  に関する意味クラスタを形成していると解釈でき, 意味クラスタが与えるスムージング効果により疎データ問題が解消される. アスペクトモデルでは隠れ変数の数は 1 つであるが, プロダクトモデルでは 2 つあり, 2 つの隠れ変数間 (原因側の意味クラスタから結果側の意味クラスタ) の依存関係が陽に表される.

(a) と (b) はどちらも出来事対  $\langle x, y \rangle$  のための確率モデルである. 各モデルに対し, 教師信号の情報を考慮するためにクラス変数 (因果関係クラス)  $c$  を加えたモデルが (d) および (e) である. ここでは, (d) をクラス付きアスペクトモデル (Class-Asp)[12], (e) をクラス付きプロダクトモデル (Class-Pro) と呼ぶ. すべての出来

事対に対して因果関係クラスのラベル値を得ることは現実的に不可能であるので,  $x$  と  $y$  が観測変数である一方で,  $c$  は部分観測変数となる. 図からわかるように, 因果関係クラスの情報は, 隠れ変数を介して, 原因側, 結果側の各出来事の生成確率に影響を与える.

続いて, 図 1 の (f) は, (e) から隠れ変数間の依存関係を無視した構造をしており, (d) と (e) の中間の複雑さを持っている. (c) は因果関係クラスが与えられたという条件で出来事  $x$  と  $y$  の独立性を仮定したモデルとなっており, いわゆるナイーブベイズモデル (NB)[9] の形をしている. (f) は (c) に隠れ変数を加えた形をしていることから, 本稿では (f) を潜在ナイーブベイズモデル (Latent NB: LNB)[8] と呼ぶ.

以上, 提案モデルを含む 6 つの確率モデルを紹介した. 4 節では, 評価実験を通して, これらの確率モデルの因果性判定課題に対する有効性を検証する.

### 3.2 推定法

確率モデルの推定方法について述べる. ここでは紙面の制約から, 最も複雑さの高いクラス付きプロダクトモデルのみを記述するが, 他のモデルも同様である.

図 1(e) のクラス付きプロダクトモデルにおいて, 出来事の出現を表す確率変数  $x$  および  $y$ , 因果関係クラスを表す確率変数  $c$ , それに隠れ変数  $z^x, z^y$  を加えた 5 つの変数の同時確率は図 2 の式 (1) になる. 入手可能なデータは, 部分観測変数  $c$ , および隠れ変数  $z^x, z^y$  を含む不完全データであるため, モデル推定には Expectation-Maximization (EM) アルゴリズム [1] を適用する. EM アルゴリズムで用いられる Q 関数, すなわち, 隠れ変数の事後確率に関する対数尤度の条件つき期待値は図 2 の式 (2) で表される. ここで,  $\theta$  はパラメータの集合,  $N_{x_i y_j c_m}$  は訓練データ中での教師信号  $\langle x_i, y_j, c_m \rangle$  の出

現頻度,  $N_{x_i y_j}$  は訓練データ中での因果関係クラスが不明な共起出来事対  $\langle x_i, y_j \rangle$  の出現頻度をそれぞれ表す. また,  $\bar{P}(\cdot)$  は古いパラメータを用いて計算された確率値であることを表す. この時, E(Expectation) ステップは隠れ変数の事後確率の計算に帰着される (式 (3), (4)). また, M(Maximization) ステップは  $Q$  関数の各パラメータに対する最大化問題となり, ラグランジュの未定乗数法を用いることによって解ける.

4 節で述べる評価実験では, 実際には, EM アルゴリズムの変種である tempered EM[5] を採用する. tempered EM では, 更新途中の確率値に対して重み  $\beta$  を付与することができる. これにより, 更新途中の確率値を過信することによる推定の失敗を回避しやすくなる. また, 教師信号  $\langle x, y, c \rangle$  の量に比べ, 教師なし事例, すなわち, 因果関係クラスが不明な出来事対  $\langle x, y \rangle$  が大量に存在する場合は EM アルゴリズムがうまく機能しないことがある. そこで, 教師なし事例に対して重み  $\lambda$  ( $0 \leq \lambda \leq 1$ ) を設け (e.g., Nigam et al. [10]), 確率値の更新過程において, 教師信号に対する教師なし事例の影響を抑えてモデルを推定する.

### 3.3 判定法

構築したそれぞれのモデルから  $x$  や  $y$ ,  $c$  に関する幾つかの確率が得られる. これらの確率値を利用して出来事対の因果性を判定する方法について述べる.

まず, 因果関係クラス  $c$  を含む, (c) ナイーブベイズモデル, (d) クラス付きアスペクトモデル, (e) クラス付きプロダクトモデル, (f) 潜在ナイーブベイズモデルでは, 出来事対  $\langle x, y \rangle$  が与えられたとき, 次の式 (6) で因果関係クラス  $c$  の状態を決定することで, 因果性を判定する:

$$\hat{c} = \operatorname{argmax}_{c_m} P(c_m | x, y). \quad (6)$$

モデルによって  $P(c_m | x, y)$  の具体的な計算式は異なる. 例えば, クラス付きプロダクトモデルの場合では, 図 2 の式 (5) になる.

提案モデルとの比較モデルとなる, (a) アスペクトモデルと (b) プロダクトモデルは因果関係クラスを含まない. そこで, これら 2 つのモデルに関しては, 次の方法で判定する. 出来事対  $\langle x, y \rangle$  が与えられたとき, まず, 共起尺度として頻りに利用される式 (7) の PMI (pointwise mutual information) [3] を求める:

$$PMI(x, y) = \log \frac{P(x, y)}{P(x)P(y)}. \quad (7)$$

そして, 判定には, PMI に対する閾値  $\sigma$  を予め定めておき,  $\sigma$  以上の PMI 値をもつ  $\langle x, y \rangle$  の間には因果関係があり,  $\sigma$  より低い値をもつ  $\langle x, y \rangle$  の間には因果関係がないと判定した.

## 4 評価実験

### 4.1 設定

訓練データについて説明する. 我々は, 以前, 新聞記事を対象として, 文書内に出現する因果関係の出現傾向を調査した [16]. まず, この調査過程で得られたデータの中で, 因果関係をもつ出来事対集合と因果関係をもたない出来事対集合のそれぞれから, 動詞を述部にもつ出来事対を 200 件ずつ抽出した. そして, この抽出データから教師信号を生成し訓練に用いた.

今回の実験では, 出来事 (の表現) を動詞一語で近似し, 因果モデルを構築した. 例えば,  $\langle$  目の異状を訴える, 病院へ運ぶ  $\rangle$  という出来事対が手元にある場合は, そこから  $\langle$  訴える, 運ぶ  $\rangle$  という動詞対を生成してモデルの訓練に用いた. この決定は文書中に含まれる因果関係の観察から得た次の仮定に基づく:

共通の述部をもつ複数の出来事対の集合の中で, 述部の対が係り受け関係にある出来事対は因果関係の有無に関して同様の傾向をもつ.

すなわち, 係り受け関係にある出来事対に注目した場合, それらの出来事対の因果性は述部の情報から決定される傾向があると仮定する. 例えば, 次の 2 文中のそれぞれの出来事対は共通の述部をもち係り受け関係にある. これらの出来事対の間には共に因果関係があると考えられ, 上記の仮定を満たしている.

- 目の異状を訴え, 病院へ運ばれた
- 背中の痛みを訴えたので, 救急車で運んだ

当然, 仮定を満たさない例外も存在するが, この近似によりモデルの複雑さを軽減できる利点がある.

教師なしデータには, 教師信号が得られた文書とは異なる新聞記事から, 係り受け関係にある共起動詞対を抽出して用いた. なお, 得られた動詞対の大部分は, 抽出元の文書内において手がかり語句を有していなかった. 教師なしデータの量の違いによる因果性判定の精度変化を調査するために, 量の異なる 3 セットの教師なしデータ (100 件, 1000 件, 10000 件) を用意した.

評価は, 教師信号に関する 5 分割交差検定で行った. ただし, どの分割データに対しても教師なしデータは共通である. 評価尺度には, 以下の式で得られる再現率 ( $R$ ) と適合率 ( $P$ ) をもとに,  $F$  値 ( $2RP/(R+P)$ ) を算出した.

$$R = |A \cap B| / |A| \quad P = |A \cap B| / |B|$$

$A$  = 因果関係をもつ出来事対の集合

$B$  = 因果関係があると判定された出来事対の集合

### 4.2 結果

まず, 因果関係クラスをもつ 4 つのモデル (c) NB, (d) Class-Asp, (e) Class-Pro, (f) LNB を比較する.

実験の結果を表 1 に示す. 教師なしデータの量ごとに, 最も高い  $F$  値を太字にした. 各モデルは, 隠れ変

表 1: 評価実験の結果 ( $F$  値)

	(c) NB	(d) Class-Asp	(e) Class-Pro	(f) LNB
0	.298	.319	<b>.583</b>	.533
教師なし 100	.328	.588	<b>.610</b>	.569
データの量 1,000	.459	<b>.644</b>	.641	.595
10,000	.623	.677	<b>.678</b>	.631

数の状態数に加え, 3.2 節で述べた  $\beta$  と  $\lambda$  というハイパーパラメータをもつ. 表 1 は, 幾通りかのパラメータ値の組合せを試行した中で最良であった結果を示している.

隠れ変数をもたない (c) NB に比べ, 提案モデルである (d) Class-Asp, (e) Class-Pro, (f) LNB では良好な結果を得た. 提案モデルの中では, 顕著な差はないが, 総合的には (e)Class-Pro が良いと言える. また, 表 1 から, いずれのモデルにおいても教師なしデータの量を増やすことで  $F$  値も上昇しており, ここから半教師あり学習の効果が確認できる.

続いて, クラス変数 (因果関係クラス) を導入した効果を調査するために, 因果関係クラスをもたない (a) Asp, (b) Pro と提案モデル (d) Class-Asp, (e) Class-Pro を比較する. まず, (a) Asp と (b) Pro では 3.3 節で述べた判定方法において,  $PMI$  に対する閾値  $\sigma$  を変化させながら  $F$  値を観測し, 最も高い  $F$  値を求めた. 一方の (d) Class-Asp と (e) Class-Pro でも同様に, まず, 判定式 (6) で得られた  $\hat{c}$  に対応する  $P(\hat{c}|x, y)$  に対して閾値  $v$  を設定し,  $P(\hat{c}|x, y) \leq v$  となる出来事対については判定を保留する. そして, 閾値  $v$  を変化させながら  $F$  値を観測し, 最も高い  $F$  値を求めた. 結果を表 2 に示す. 表 2 は, 10000 件の教師なしデータを用いて獲得された推定モデルによる結果である.

表 2: クラス変数の効果 ( $F$  値の最大値)

(a) Asp	.638	(b) Pro	.641
(d) Class-Asp	.730	(e) Class-Pro	.829

表 2 から, (d) Class-Asp, (e) Class-Pro は (a) Asp および (b) Pro のどちらよりも高い  $F$  値を得た. この結果から, クラス変数を導入することの有効性が確認できる.

## 5 おわりに

本稿では出来事対の因果性を判定するために, 共起モデルを拡張した幾つかの確率モデルを提案した. 評価実験を通して, クラス変数, 隠れ変数を導入した確率モデルが因果性判定課題に有効であることを示した. 謝辞

本研究は, 科学研究費補助金学術創成研究 (課題番号:13NP0301) の支援を受けて行われた.

## 参考文献

- [1] Nan M. Laird Arthur P. Dempster and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, Vol. 34, pp. 1–38, 1977.
- [2] Du-Seong Chang and Key-Sun Choi. Causal relation extraction using cue phrase and lexical pair probabilities. In *Proc. of the 1st IJCNLP*, pp. 61–70, 2004.
- [3] Kenneth W. Church and Patrick Hanks. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th. Annual Meeting of the Association for Computational Linguistics*, pp. 76–83, 1989.
- [4] Roxana Girju. Automatic detection of causal relations for question answering. In *Proc. of ACL 2003, Workshop on Multilingual Summarization and Question Answering - Machine Learning and Beyond*, 2003.
- [5] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning Journal*, Vol. 42(1), pp. 177–196, 2001.
- [6] Thomas Hofmann and Jan Puzicha. Statistical models for co-occurrence data. Technical Report AIM-1625, Artificial Intelligence Laboratory, MIT, 1998.
- [7] Takashi Inui. *Acquiring Causal Knowledge from Text Using Connective Markers*. PhD thesis, Graduate School of Information Science, Nara Institute of Science and Technology, 2004.
- [8] Nevin L.Zhang, Thomas D. Nielsen, and Finn V. Jensen. Latent variable discovery in classification models. *Artificial Intelligence in Medicine*, Vol. 30, No. 3, pp. 283–299, 2004.
- [9] Tom Mitchell. *Machine Learning*. McGraw-Hill, 1997.
- [10] Kamal Nigam, Andrew K. McCallum, Sebastian Thrun, and Tom M. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, Vol. 39, No. 2/3, pp. 103–134, 2000.
- [11] Olivia Sanchez-Graillet and Massimo Poesio. Acquiring bayesian networks from text. In *Proc. of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, pp. 955–958, 2004.
- [12] Hiroya Takamura, Takashi Inui, and Manabu Okumura. Latent variable models for semantic orientations of phrases. In *Proc. of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, 2006. (To appear).
- [13] Kentaro Torisawa. An unsupervised learning method for commonsensical inference rules on events. In *Proc. of the Second CoLogNet-ElsNET Symposium*, pp. 146–153, 2003.
- [14] Kentaro Torisawa. Automatic acquisition of expressions representing preparation and utilization of an object. In *Proc. of the Recent Advances in Natural Language Processing*, pp. 556–560, 2005.
- [15] 山田一郎, 小早川健, 三浦菊佳, 住吉英樹, 八木伸行, 崔杞鮮. クローズドキャプションを対象とした因果関係知識抽出の検討. In *FIT*, 2005.
- [16] 乾孝司, 奥村学. 文書内に現れる因果関係の出現特性調査. *計量国語学*, Vol. 25(3), pp. 123–144, 2005.