テキストから獲得可能な因果関係知識の類別およびその自動獲得の試み -接続助詞「ため」を含む文を中心に-

乾孝司　乾健太郎　松本裕治
奈良先端科学技術大学院大学 情報科学研究科
{takash-i,inui,matsu}@is.aist-nara.ac.jp

1 はじめに

因果関係という概念は，古代ギリシャ哲学の時代から今日にいたるまで心理学，哲学をはじめ多くの学問領域で研究対象になってきた．人工知能の領域でも，初期の頃から知能の源として常識・因果知識に関心が向けられ，それらの知識を蓄積した知識ベースを利用した推論や言語理解研究が活発になされてきた [11, 2]．例えば，プラン認識に基づく談話理解過程では，行為の前提条件や効果に関する知識を用いる．図 1 (a) はそのような知識の例であり，洗濯物を干せばその結果としてどのような状況になるか (効果)，洗濯を干すにはどのような状況である必要があるか (前提条件) という因果関係によって構成されている．深い言語理解の実現にこうした因果関係知識が必要であることは広く認識されている．

しかし，CYC[6] や OpenMind[12] に見られるように常識的知識を人手でどこまで書き尽せるかは依然として未知数である．また，それらの知識を自動的に獲得する方法論についても目覚しい進展があるとは言い難い．これに対し近年では，WWW に代表される電子化テキストを知識源と見なし，そこから (半) 自動的に因果知識・世界知識を獲得する試みがいくつか報告され始めている [10, 5, 1]．

以上の背景から我々も，大量の電子化文書集合から知識を自動的に獲得・蓄積し，深い言語理解に利用することを目指している．我々の当面の目標は，

> 図 **1(a)** を構成している図 **1(b)** のような，幾つかの関係種類をもつ **2** 項関係の知識を文書集合から自動獲得する

ことである [4]．

2 どのような情報源からどのような知識を獲得するか

2.1 因果関係の種類

獲得対象とする因果関係の種類としては，因果関係に立つ 2 つの事態がそれぞれ意志的な行為 (例えば「洗濯物を干す」という行為) であるか，非意志的事態 (例えば「洗濯物が乾く」という状態) であるかという基準に基づいて分類する．多くの談話理解研究が指摘しているように，言語理解において行為と状態の区別は重要である．行為には主体の意図がその背後にあり，そうした意図を理解することが談話理解の主要なゴール

(a)

干す (Agent, 洗濯物, 屋外)	
precond:	天気 (晴れ)
effect:	乾く (洗濯物)

(b) 【構成要素】
effect (洗濯物を干す, 洗濯物が乾く)
precond (天気が晴れ, 洗濯物を干す)

図 1: プランオペレータの例とその構成要素

の一つとされてきた．また，2 つの事態がともに有意志行為であるとき，それらの行為の主体が同一かどうかの区別も必要である．

ここで，時間的に先行する一方の事態を $e1$，もう一方の事態を $e2$ と表し，有意志行為を a，非意志的事態を s で表すとすると，我々が獲得を目指す因果関係は表 1 に示す 5 種類に分類できる．

2.2 知識獲得の情報源

(2) に列挙しているのは，例文 (1) から獲得が期待できる知識の例である．

(1) a. 風が強いため，洗濯物がはやく乾く．
b. 乾燥機にかけたため，洗濯物がはやく乾く．
c. 湿度が高かったが，洗濯物が乾いた．
d. 晴れていれば，洗濯物がよく乾く．

(2) a. *cause* (風が強い, 洗濯物がはやく乾く)
b. *effect* (乾燥機にかける, 洗濯物がはやく乾く)
c. *cause* (湿度が高い, 洗濯物が乾かない)
d. *cause* (晴れる, 洗濯物がよく乾く)

このように因果関係知識はいろいろな接続標識をもつ文から獲得できる．このような接続標識の中から，本稿では，相対的に出現頻度が高いこと (表 2 参照)，典型的に因果関係を表すことの 2 つの理由から接続助詞「ため」に対象を絞って議論していく．さらに「ため」の出現頻度を周辺形態素ごとに見ると (表 3 参照)，用言 (主に動詞) を伴って「ため」が出現する頻度が高いことがわかる．また，事態は基本的に節 (用言と格要素) で表現されることを考え合わせ「ため」の中でも「節-ため-節」となる複文 (以下，タメ複文) を取り扱うことにした．

ただ，一口にタメ複文と言っても，(2a) や (2b) のように，そこから獲得できる因果関係は，*cause* の関係の

表 1: 因果関係の種類

“言語テストの例” の列に書かれている表現は “関係名” の判断用のテスト表現である．あるテキストが “言語テストの例” の列に書かれているテスト表現に言い換え可能であるならば，そのテキストからはテスト表現と同一行に記載されている “関係名” の因果関係が獲得できる．

関係名 ($e1$,$e2$)	主体同一性	関係の意味	言語テストの例
cause (s,s)		$e1$ は $e2$ の原因である．	$e1$ が起こった結果として $e2$ が起こった．
effect (a,s)		$e2$ は $e1$ の効果である．	$e1$ をした結果 $e2$ が起こる．
precond (s,a)		$e1$ は $e2$ の前提条件である．	$e1$ が成り立っている状況ではしばしば $e2$ する．
means (a,a)	+	$e1$ は $e2$ の手段である．	主体 X が $e2$ を達成する手段として，$e1$ した．
enable(a,a)	−	$e2$ は $e1$ の応答である．	主体 X1 が $e1$ すれば，主体 X2 が $e2$ できる．

表 2: 接続助詞の出現頻度

日本経済新聞 [15]1990 年の 1 年分から次の 3 条件を満たす文の出現頻度を求めた．(1) 接続助詞の直前に現れる自立語が動詞である，(2) 接続助詞を含む文節の係り文節が文末文節である，(3) 係り文節の主辞形態素が動詞である．

が	7,450	たら	336
と	5,179	から	325
(れ) ば	3,362	なら	134
ため (に)	3,252	のに	109
ながら	862	φ、	18,886
ので	514	て、	4,088

表 3: 接続助詞「ため」の出現傾向

日本経済新聞 1990 年の 1 年分のデータ．“用言” は動詞，形容詞，形容動詞が主辞となる文節，“ための” は「ため」に「の」が後接して名詞に係る文節を表す．

		係り先文節			
		用言	連体化「の」	名詞	その他
「ため」	用言	36,670	4,505	4,896	5,934
を	ための	9,787	4,744	2,298	209
含	連体詞	478	284	36	186
む	名詞	901	11	110	171
文節	その他	35	0	4	174

時もあれば *effect* の関係の時もあるため，これらを正しく分類する必要がある．従って，我々の目標は，タメ複文に関する限り，1 節で述べた目標の部分目標として，

> タメ複文集合中の個々の文を表 **1** の因果関係の種類ごとに自動分類する

という分類問題を解くことである．

3 タメ複文に現れる因果関係

例文 (2a)，(2b) で見たように，タメ複文からは幾つかの異なる因果関係を獲得することができる．そこで，表 1 の 5 種類の因果関係がタメ複文からどの程度の頻度割合で獲得できるのかを調査した．

調査には，表 2 で求めたタメ複文から 1,000 事例を選び，形態素解析などの前処理による解析誤りや疑問文を省いた 994 事例を用いた (以下，この事例集合を $\mathcal{S}_{tame}$ で参照する)．

まず，節が意志性を持つか否かによってタメ複文を 4 つの文集合のクラス (A ~ D) に分割した．表 4 の “頻度” の列に頻度の内分けを示す．次に，個々のクラスに対して，そこから獲得できる因果関係の種類と頻度分布を調べた．各関係の存在の有無を判断する際には，表 1 の “言語テストの例” に示したような表現への言い換え可能性に従って判断した．

結果として，A ~ D の 4 クラスのいずれにおいても，表 4 の “最頻出関係” の列に示した関係が最も多く存在

表 4: タメ複文に現れる因果関係

”従意”は従属節の意志性，”主意”は主節の意志性を表す．

クラス	従意	主意	頻度	最頻出関係とその頻度	
A	−	−	229	*cause* (従節, 主節)	220 (0.96)
B	+	−	161	*effect* (従節, 主節)	139 (0.93)
C	−	+	225	*precond* (従節, 主節)	202 (0.90)
D	+	+	379	*means* (主節, 従節)	323 (0.85)

しており，いずれも全体の 85%以上を占めていた．各クラスについて，最頻出関係の例文を (3) に示す．

(3) A1. 十月は気温が比較的高めに推移したため、衣料品の売り上げが伸び悩んだ。《*cause*》
A2. 間伐、下枝刈り、下草刈りなどの手入れが行き届かないため流木被害などが起きやすい。《*cause*》
B1. バングラデシュではマングローブを破壊したために大水害に見舞われた。《*effect*》
B2. 石油会社も先物のドル買い予約を入れたため、円は軟調に推移した。《*effect*》
C1. 国内とアジアで段ボールの需要が急増しているため生産拠点の拡充に踏み切る。《*precond*》
C2. その後、配管室の白煙が増加したため、手動で炉を停止している。《*precond*》
D1. 街並みを美しくするため、十月末から歩道に花の苗を設置している。《*means*》
D2. 北京から福建省への切符を買うため、駅の発券所に並んだ。《*means*》

表 4 からもわかるように，クラスと最頻出関係が一対一に対応しているわけでは必ずしもない．最頻出でなかった例文を (4) に示しておく．

(4) A1. 企業の成長が見込めるようになったため、組合の設立機運が高まっている。(《*cause*》でない)
B1. しかも、たくさん収容するため、日本のホールは扇形になっている。(《*effect*》でない)
B2. 大量の受注をさばくため、一部で中途採用の動きが見られる。(《*effect*》でない)
C1. 微妙な古楽器が舞台の温度になれるためにオーケストラは、開演までに十分余も調整した。(《*precond*》でない)
D1. 和菓子と洋菓子両方を取り扱っていたため、その製造ノウハウを活用した。(《*means*》でない)

以上をまとめると，タメ複文に関する限り，5 種類の因果関係のうち，*enable* を除く *cause*，*effect*，*precond*，*means* の 4 つの関係について，従属節，主節の意志性がわかりさえすれば，85%以上の精度で因果関係が分類できる．以下ではこれら 4 つの関係の自動分類の可能性ついて調べていく．

4 意志性の自動推定

これまでの議論は意志性が既知であるという前提の基でおこなってきたが，実際には，従属節，主節の各々の意志性を推定する必要がある．

4.1 動詞と意志性

まず，意志性の特徴を探るために，$\mathcal{S}_{tame}$ の文中で同じ動詞を主辞とする節に対して，動詞以外の文脈が節の意志性に及ぼす影響を調査した．

$\mathcal{S}_{tame}$ 中の 720 の異なり動詞のうち，2 度以上出現した動詞は，異なりで 297 あり，75%に相当する 227 動詞は意志性が均一であった．しかし，残りの 70 動詞 (のべ 561 動詞) では意志性が文脈によって変化していた．

(5) a. 生産能力を 拡大する《意志性+》ため設備投資する。
b. 管理費が 横ばいにとどまったため、営業利益が 拡大した《意志性 −》。

4.2 SVM による意志性推定

以上の結果を踏まえ，節の意志性 $(+/-)$ がどの程度推定できるかを，機械学習アルゴリズムである Support Vector Machine[13] を用いて実験的に検証した．

4.2.1 実験条件

[素性] 先ほどの 70 動詞の分析から得られた，意志性の変動要因を基にして，表 5 の 6 種の素性を利用した．動詞クラスについて，辞書エントリの存在しない動詞については新たに情報を追加した．

[対象データ] 事例集合 $\mathcal{S}_{tame}$ の 994 文．各節の意志性の分布は，従属節が ($+:539$ / $-:455$)，主節が ($+:603$ / $-:391$) であり大きな偏りはない．

例文 (5) のように出現位置の影響が強いことから，今回は従属節，主節のそれぞれについて分類器を作成し，5 分割の交差検定を行なった．

4.2.2 結果

推定精度は，従属節が 94.8% (945/994)，主節が 95.6% (950/994) ，全体で 95.3%(1895/1988) であった．分類事例を見てみると，例文 (5) で示した動詞「拡大する」は 13 事例中 5 例が+，8 例が − であったが，すべて正しく意志性を推定できていることが確認できた．$\mathcal{S}_{tame}$ 中の動詞で，意志性が一意であったものは辞書的知識として意志性属性を整備し，文脈に依存するものについては事例数が多かった値を採用することとしてベースの推定精度を概算すると $1863/1988 = 93.7\%$ となる．この数値との比較から，SVM による結果は意志性辞書を構築するコストを省き，かつ，文脈による変動にもある程度対応できていることがわかる．

さらに SVM が出力する判別関数値の絶対値 (分離超平面からの距離) が推定の信頼度を表していると解釈し，信頼度に閾値 α を設け，α 以上の信頼度をもつ事例のみに対して判断結果を出力することを考える．すると，α を変化させることにより，図 2 の被覆率-精度曲線[1] が得られる．

この結果は次節で述べる因果関係の分類の際，10%の事例を犠牲にすれば，97% 超の推定精度を有する意志性の属性値が利用可能であることを示している．

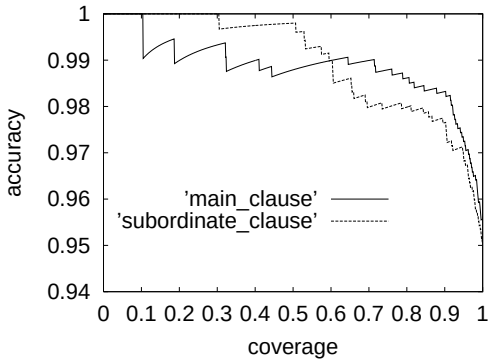

図 2: 被覆率-精度曲線 (節の意志性)

5 因果関係の自動分類

前節より，ある程度の精度で意志性を決定できる見通しを得た．そこで次に，推定した意志性の値を利用することで，タメ複文に含まれる因果関係がどの程度自動的に分類できるかを調査した．なお，分類器には 4 節と同様，SVM を用いた．

5.1 実験条件

まず，判別するクラスとして *cause*，*effect*，*precond*，*means* の 4 つと残りの関係をまとめたクラスの計 5 クラスを設定した．素性には，4.2 節で用いた素性の他に，意志性 (被覆率 90%点での推定値)，主体一致性 (人手で付与[2]) を利用した．また，SVM は 2 値分類器のため，One vs. Rest 法を適用し，多値分類に応用した．ただし，この手法により複数の分類器から得られる判別関数値が正となる場合は判別関数値が最大の分類器の結果を優先する．これらの実験条件のもと，$\mathcal{S}_{tame}$ の 994 文を用い，5 分割の交差検定を行なった．

5.2 結果

One vs. Rest 法により得られた各分類器の判別関数値のうち，最大値を s_1，2 番目に大きい値を s_2 としたとき，$\beta_1 s_1 + \beta_2(s_1 - s_2)$ の値を判別の信頼度とする．そして，この信頼度について 4.2.2 節と同様の処理を行なうことにより図 3 の再現率[3]-精度曲線を得た．ただし，図は $\beta_1 = 1$，$\beta_2 = 1$ での結果である．

まず，すべての関係で，再現率を落せばより高い精度が望めることが確認できる．表 4 に示した出現割合を比較対象とすると，それほど分類精度が好ましくない *effect* では，再現率を 60%まで落すことで表 4 の数値 (93%) を上回り，すべての関係概念で 98%以上の精度を得るには，再現率を 30%まで落とせばよいことがわかる．

なお，誤り原因としては，まず，用言クラスの選択が最適でないことが挙げられる．また，反例事例として頻出する，未来事象に対する備えの関係では「来年」などの明示的な時間標識をもたない場合 (例えば (3d)) への対応が課題となる．

[1] 被覆率 $= \frac{\text{出力された事例数}}{\text{全事例数}}$，
精度 $= \frac{\text{正しく該当クラスが出力された事例数}}{\text{出力された事例数}}$．

[2] 今回は人手によって情報を付与したが，複文の主体が一致するか否かは既に [9] の手法で自動判定が可能である．

[3] 再現率 $= \frac{\text{正しく該当クラスが出力された事例数}}{\text{該当クラスの事例数}}$．

表 5: 意志性の推定に利用した素性

動詞クラス 1	EDR 電子化辞書 [16] . 概念辞書の上位層に位置する概念を参照することで，対象動詞の上位概念が “移動” か “行為”(1)，上位概念が “状態” か “変化” か “現象”(2)，前 2 つどちらにも該当する (3)，該当なし (4) の 4 つのクラス値を割り当てた．
動詞クラス 2	ALT-J/E 翻訳システム用辞書 [3][14] に含まれる動詞に関する (該当/未該当) フラグから以下の 11 エントリ．“状態動詞”，“継続動詞”，“瞬間動詞”，“自動詞”，“他動詞”，“補助動詞”，“可能動詞”，“自発動詞”，“使役動詞”，“受身”，“受身 (被害)”．
動詞クラス 3	日本語語彙大系 [14] に掲載されている用言意味属性．
主体意志性	主体が人や組織などのように意志を持ち得るか否か (人手による付与)．
ガ格/ヲ格	格要素が存在するか否か．ある場合は要素を [14] に含まれる情報で抽象化した概念．
モダリティ	テンス (ル形/タ形)，アスペクト (ている)，態 (受動，使役)，可能 (できる)，否定 (ない) の有無．

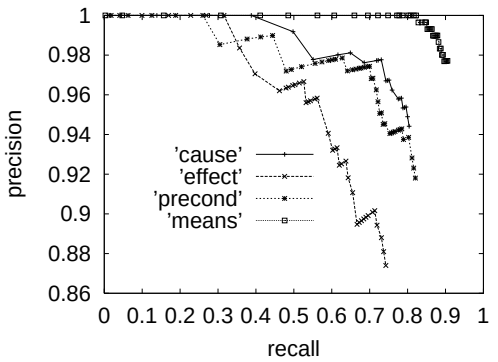

図 3: 再現率-精度曲線 (因果関係別)

5.3 獲得できる知識量の見通し

今回取り扱った事例は，表 3 の 1 行 1 列要素 (“用言-用言”) であるが，処理の観点から見れば，表 3 の “名詞” の大多数は「名詞+だ」形式であり，本稿と同様の議論ができる．1 行 1 列要素と，主節，従属節のいづれかに “名詞” を含む各要素 (1 行 3 列，4 行 1 列，4 行 3 列) をあわせた要素の合計は 42,577 事例である．因果の関係概念を 98%以上の精度で得るには，図 3 より再現率を 30%まで落とせばよいので，新聞記事データ 1 年分からは，概算で $42,577 \times 0.3 = 12,773$ 件の知識が獲得可能と見積ることができる．

6 関連研究

Girju ら [1] や佐藤ら [10] も我々と同様，因果関係をあらわす言語標識を利用することで，テキストから因果知識を獲得することを試みている．しかし，彼らの方法では，本稿で述べたような因果関係の分類は考慮されていない．

因果関係の分類は，修辞構造解析における修辞関係の同定 (例えば [7]) と同じ問題であるように見えるかも知れない．しかし，我々の考える因果関係は，モーダル情報や情報構造などが捨象された，修辞関係より抽象度の高い関係である．このことを次の例文を用いて説明する．

(6) a. 洗濯物を干せば洗濯物が乾く．
b. 洗濯物を乾かすために洗濯物を干した．
*c. 洗濯物を干したが洗濯物が乾いた．

修辞関係を考えると，(6a) は仮定，(6b) は目的と一般的には解釈される．ここで，これらの修辞関係がどちらも一貫していると感じられるその理由は *effect*(洗濯物を干す, 洗濯物が乾く) という我々がもつ因果知識と整合するからであると考えられる．このことは，同じ事態に対して例文 (1c) のような逆接の修辞関係を考えると一貫性がなくなることからも理解できる．Marcu[8] は “because” などの言語標識を利用して修辞関係を同定しているが，上で述べたように，我々の考える因果関係は修辞とは異なる層の関係であることから，本研究に修辞解析の成果を直接利用できるわけではない．

謝辞 NTT コミュニケーション科学基礎研究所より日本語語彙大系, ALT-J/E 翻訳システム用辞書を利用させて頂きました．深く感謝いたします．

参考文献

[1] R. Girju and D. Moldovan. Mining answers for causation questions. In *Proc. the AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases*, 2002.
[2] J. R. Hobbs, M. Stickel, D. Appelt, and P. Martion. Interpretation as abduction. *Artificial Intelligence*, Vol. 63, pp. 69–142, 1993.
[3] S. Ikehara, S. Shirai, A. Yokoo, and H. Nakaiwa. Toward an MT system without pre-editing – effects of new methods in **ALT-J/E**–. In *Third Machine Translation Summit: MT Summit III*, pp. 101–106, Washington DC, 1991.
[4] 乾孝司, 乾健太郎, 松本裕治. 接続助詞「ため」を含む複文から因果関係知識を獲得する. 情報処理学会自然言語処理研究会 (2002-NL-150), pp. 171–178, 2002.
[5] 黒橋禎夫, 酒井康行. 辞書とコーパスからの世界知識の自動抽出. 「知識発見のための自然言語処理」シンポジウム, 1999.
[6] D. Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Comunications of the ACM*, Vol. 38, No. 11, 1995.
[7] D Marcu. The rhetrical parsing of natural language texts. In *The Proc. of ACL97/EACL97*, pp. 96–103, 1997.
[8] D. Marcu. An unsupervised approach to recognizing discourse relations. In *Proc. of the ACL conference*, 2002.
[9] 中岩浩巳, 池原悟. 語用論的・意味論的制約を用いた日本語ゼロ代名詞の文内照応解析. 自然言語処理, Vol. 3, No. 4, pp. 49–65, 1996.
[10] 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. 信学技報 (TL98-23), 1998.
[11] R. Schank and R. Abelson. *Scripts Plans Goals and Understanding*. Lawrence Erlbaum Associates, 1977.
[12] D. G. Stork. Character and document research in the open mind initiative. In *Proc. of Int. Conf. oon Document Analysis and Recognition (ICDAR99)*, pp. 1–12, 1999.
[13] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.
[14] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦. 日本語語彙大系. 岩波書店, 1997.
[15] 日本経済新聞社. 日本経済新聞 cd-rom 版.
[16] 日本電子化辞書研究所. EDR 電子化辞書仕様説明書, 1995.