因果関係タグ付きコーパスの構築と分析 [1]

乾 孝司 †
† 東京工業大学 精密工学研究所
†COE21「大規模知識資源の体系化と活用基盤構築」
tinui@lr.pi.titech.ac.jp

奥村 学 ‡
‡ 東京工業大学 精密工学研究所
oku@pi.titech.ac.jp

1 はじめに

因果関係に関する知識は，質疑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，大規模なテキスト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（例えば，文献 [2, 6, 3]）．しかしながら，知識獲得の方法論に関する研究開発が進展するその一方で，テキスト中に含まれる因果関係の出現特性に関しては，これまでにまとまった知見が得られているとは言い難く，未だ未知な点が多い．

本研究では，このような背景を受け，一定量のテキスト集合に対して因果関係情報を注釈付け，その出現特性を調査することを目標としている．注釈付けの対象には，例文 (1a) のような有標形式に加え，(1b) のような無標形式，(1c) のように出来事が名詞句として表現される場合も含んでおり，表現形式に対する制約は原則として設けていない．

(1) a. 大雨が降ったため、川が増水した。
b. 大雨が降り、川が増水した。
c. 大雨で川が増水した。

本稿では，因果関係情報の注釈付けの過程について述べた後，現在までに注釈付けが完了しているデータを用いて，手がかり標識の有無，出来事表現の統語カテゴリ，出来事表現の文内位置の３つの観点から因果関係の出現傾向を調査したので，その結果を報告する．

2 注釈付ける因果関係情報

2.1 因果関係タグ

因果関係情報を注釈付けるために *head*（主辞要素），*mod*（修飾要素），*causal_rel*（因果関係）の３種類の基本タグを用いる．例文 (2a) へのタグ付与結果は図 1 のようになる．ここでは (2b) と (2c) の２つの出来事間に因果関係があるとする．e_1 は因果関係の前件（原因側）出来事，e_2 は後件（結果側）出来事を示す．

(2) a. そして、遠方からの観光客がGWに入って増える。
b. e_1 = GWに入る
c. e_2 = 遠方からの観光客が増える

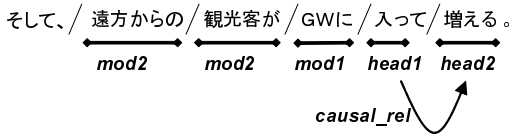

図 1: タグ付与の例

本研究では，出来事は 1 つの主辞要素に 0 個以上の修飾要素が付随して構成されると仮定する．そこで，まず，出来事の主辞となる要素（主に述語）に *head* タグを付与し（図中の下線部がタグの範囲），出来事を構成する要素が *head* タグの箇所以外にもある場合は，それぞれの箇所に *mod* タグを付与することで出来事を注釈付ける．図中の各タグの接尾数字は，それぞれの出来事（e_1 と e_2）の接尾数字に対応している．２つの出来事へタグを付与した後，それぞれの *head* タグの間にリンク情報として *causal_rel* タグを付与することで，当該の出来事間に因果関係があることを注釈付ける．もし，手がかり標識が存在する場合は，該当個所に *marker* タグを付与する[2]．

2.2 注釈付けの基準

ある２つの出来事の間に因果関係があるか否かの判断は，次のような言語テンプレートを用いた言語テストに基づく．

> 『e_1』（という）状態になれば、それに伴い、≪頻度を表す副詞≫『e_2』（という）状態になる。
>
> ≪頻度を表す副詞≫ := しばしば | 大抵 | 常に

まず，e_1 と e_2 の候補となる２つの出来事表現をテキストから抽出し，それらを言語テンプレートの各スロット（テンプレート中の鉤括弧）に代入する．この時，完成したテンプレート文の文意が適格であれば，それらの間には因果関係があると判断し，2.1 節の記述に従ってタグを付与する．

各スロットに代入される出来事表現について，その主辞要素が活用していれば，基本形に戻した状態でスロットに代入する（ex. GWに入っ→GWに入る）．また，出来事表現の主辞要素が名詞句として表現されている場合は，次のいずれかの書き換え操作を施した結果をスロットに代入する．

[1] Creating an Annotated Corpus for the Analysis of Causal Relations

[2] 実際には，３つ以上の複数の出来事が関連していることを注釈付ける *relevant* タグなど，その他の付加情報も併せて注釈付けているが，本稿では触れない．

書き換え操作

- np → np + する（ex. 停電→停電する）
- np → np + が起こる（ex. 地震→地震が起こる）
- np → np + になる（ex. 大雨→大雨になる）
- 動詞の名詞化 → 動詞の基本形（ex. 疲れ→疲れる）

今回の作業では，18 通りの言語テンプレートを用意し，それらの中でいずれかのテンプレート文が適格と判断できれば，テンプレートに代入した e_1 と e_2 の間に因果関係があると判断する．ただし，実際には，18 通りの言語テンプレートであらゆるすべての因果関係を包含できているわけではない．そこで，さらなる言語テンプレートの追加，洗練のためのデータを収集することを目的とし，言語テンプレートでは判断できない事例については作業者の純粋な主観判断に従って因果関係情報を注釈付けることを許可した[3]．

言語テンプレート中の≪頻度を表す副詞≫に挙げた３つの副詞は，必然性の強い２つの出来事と共に現れた時のみ語用論的に適格な文として認識される性質をもつ．この性質を利用し，必然性の強さに関する属性を *causal_rel* タグに付加することを考える．つまり，18 通りの言語テンプレートのそれぞれについて，≪頻度を表す副詞≫部分を３つの副詞のいずれかで置き換えたテンプレートを用いて因果関係があると判断される場合には「蓋然」の属性値を付加する．一方，≪頻度を表す副詞≫を含む言語テンプレートでは因果関係があると判断できないが，≪頻度を表す副詞≫を削除した言語テンプレートでは因果関係があると判断される場合には「偶然」の属性値を付加する．周知の通り，因果関係という概念自体が，現在でも哲学を含めた学問領域における議論の対象となっており，多様な事象間関係がその範疇下におかれる．最終的に獲得される因果関係知識を工学的な応用システムに適用することを念頭においた場合，注釈付けされた因果関係事例の中で,「蓋然」の強さをもつ事例は特に有用性が高いと期待している．

2.3 注釈付けの対象と優先規則

因果関係の出現特性に関する網羅的な調査分析を実現するため，手がかり標識を伴って表現される事例に加え，手がかり標識を伴わない事例や，出来事が名詞句として表現される事例も注釈付けの対象とする．

理想的には，テキスト中で表現されるあらゆる出来事のペアに対して因果関係が含まれるか否かの判断を試みるべきである．しかしながら，任意の２つの出来事は，テキスト上での出現位置が離れるに従い，因果関係が含まれる確率は減少すると考えられる．そこで今回は，隣接する２文以内に２つの出来事の主辞要素が含まれる場合のみ因果関係の判断を行う[4]．

また，実際のテキスト中には，e_1 に対して複数の e_2 が存在したり，複数の因果関係が複雑に絡み合うケースも存在する．しかしながら，今回は，タグ付与作業にかかる作業者への負荷を軽減することを優先し，ひとつの e_1 にはやはりひとつの e_2 のみを割り当てる．注釈付ける出来事ペアを選択する際は，下記の優先規則を順に適用して選択した．

出来事ペアの優先規則

1. テキスト中で近い位置で表現されている出来事ペアを優先する[5]．
2. 「蓋然」関係となる出来事ペアを優先する．

3 対象テキスト

毎日新聞 1995 年版 [7] から抽出した記事を採用した．因果関係の情報を注釈付けるには，記事の意味内容を十分に理解する必要がある．そこで，事故や事件の報道など，内容の理解が比較的容易な社会面の記事を採用した．また，予備調査において，記事の長さに比例して，注目個所とその前文脈との関連性が強くなり，注釈付け作業が難しくなる傾向が観察された．この結果を踏まえ，社会面の中でも記事あたり 10 文以内で構成されている比較的短い記事を 750 記事（3912 文）抽出し，注釈付けの対象とした．

4 注釈付け作業の流れ

作業者への訓練期間[6]の後，約１ヶ月をかけ，上記の 750 記事に注釈付けを行った．以下に作業の流れを示す．

1. タグ付与インタフェースを介し，記事単位でテキストが作業者に提示される．記事内の文はあらかじめ文節単位に分割されている．また，接続助詞，動詞などの因果関係の発見に寄与しそうな語句があらかじめ作業画面上でハイライト表示され，作業者へ注意を促す．各作業者は，因果関係を含む箇所を探しながら作業画面上の記事を読み，因果関係を発見次第，キーボードとマウスを使った簡単な操作で該当個所にタグを付与する．
2. 各作業者は，一定数の記事（今回は 30 記事）に対して注釈付け作業を完了した時点で，タグが付与された文字列部分のみを記事から抽出し，付与されたタグの範囲や種類を確認する．誤りがあれば，この時点で修正を施す．修正個所が無くなった時点で新たな記事群（30 記事）への注釈付けに進む．

作業者は言語学の素養をもつ２名に筆者の１人を加えた計３名（以下，A～C）である．注釈付けの訓練期間中は作業者間での相談を許可していたが，本作業中は相談をもたず，互いに独立に注釈付けを行った．

すべての作業者の作業が完了した後，e_1 と e_2 がペアになっていないなど，明らかに注釈付けが不適切であると思われる事例について，機械的に修正，削除処理

[3] 主観判断に基づいて得られたデータは 5 節，6 節で述べる評価からは除外している．

[4] この制約は *head* タグを付与する場合のみ適用する．照応や省略表現によって，出来事の主辞要素とは異なる前文脈の文に主辞への修飾要素が現れる場合は，文の位置に関わらず *mod* タグを付与する．

[5] 各出来事の位置は主辞要素を含む文節の位置で代表させる．

[6] 初期に設計したタグセットとタグ付与基準を用いた予備作業の期間を含めれば２ヶ月～３ヶ月．

表 1: タグが付与された事例（“転落”，“負う” を含む例）

mod1	*head1*	*mod2*	*head2*
中学校舎から	転落する		死亡
6 階から	転落する		意識不明
川に	転落		助け上げ
	転落	胸などを	打つ
	殴る	けがを	負う
軽乗用車と	衝突	打撲傷を	負う
郵便物が	爆発する	重傷を	負う
顔や 手などに 火傷を	負う		重傷
重傷を	負う		休職する

表 2: 付与された因果関係数

A	2014 (2.7)	1224/766/24
B	1587 (2.1)	1094/492/ 1
C	1048 (1.4)	603/431/14

を施し，その結果得られたデータに対して，次節以降で評価ならびに考察を行う．一般には，メタ作業者による人手の修正過程を設けることがあるが，現在の言語テストに基づく因果関係の判断基準では，メタ作業者の主観が入る余地が残されているため，人手による修正過程は設けなかった．

5 注釈付けの結果

5.1 総数

作業者ごとに付与された因果関係の総数を表 2 左に示す．括弧内は記事あたりの平均因果関係数である．また，タグが付与された事例を表 1 に例示する．表 2 左を見ると，同一の基準を用いていたにもかかわらず，付与された総数は A と C で半数近い差があることがわかる．この結果から，因果関係の認識が作業者の主観に強く依存することが確認できる．

表 2 右は必然性の強さの内訳である．“/” 記号の左から,「蓋然」,「偶然」および属性値が未付与であった数を示す．必然性の強さに関する作業者間の比率はほぼ一致しており，おおよそ 6 割の事例が「蓋然」の強さであると判断された．

5.2 一致度

次に作業者間のタグの一致度を求める．作業者間で e_1 と e_2 の *head* タグが共に共通の文節内にある場合を「判断が一致している」と見なし，一致数を調査した．

表 3 左の “1” は該当する作業者がタグを付与したことを示し，“0” はタグを付与していないことを示す．例えば，第 4 行目（“110”）は，A と B のみが共通した文節内に *head* タグを付与し，C はタグを付与していない事例が 567 件あったことを表している．

必然性の強さの違いを区別せずに作業差間の一致度を求めた場合（図中の “混合”），2 人以上で一致した事例が計 1605 件（$= 567 + 167 + 182 + 689$），3 人全員で一致した事例が 689 件であった．また，作業者ごとに見た場合，それぞれ，A: 40%，B: 62%，C: 80% の割合で他のいずれかの作業者と一致した箇所にタグを付与していた．

表 3: 作業者間の一致数

A	B	C	混合	蓋然	偶然
1	0	0	1222	632	535
0	1	0	602	487	255
0	0	1	218	134	207
1	1	0	567	230	90
1	0	1	167	92	77
0	1	1	182	107	83
1	1	1	689	270	64

必然性の強さを区別して一致度を求めた場合[7],「蓋然」に比べて「偶然」の関係の一致度が顕著に下がることがわかる．この結果から,「蓋然」の強さをもつ事例中には,「偶然」の関係に比べて客観性と信頼性をもつ事例が集中していると期待できる．

6 考察

以下では,「蓋然」の強さをもち，かつ，2 人以上の作業者から因果関係があると判断された 699 件（$= 230 + 92 + 107 + 270$）を対象にして考察する．

6.1 手がかり標識の有無

今回，注釈付け作業時に，因果関係タグを付与する手がかりとなった語句に *marker* タグを合わせて付与した．ここでは，*marker* タグの有無の割合を調査した．結果を表 4 に示す．

表 4: 手がかり標識の有無

あり	219
なし	480

手がかり標識を一切伴わなずに因果関係をもつ 2 つの出来事が表現されることに関しては従来からも言及があるが（例えば，文献 [1]），今回，このことを定量的に確認した．取り扱うデータへの依存性をもちろん考慮しなければならないが，今回の場合，手がかり標識が存在している事例は約 3 割に満たない．これより，高い被覆率で因果関係知識を獲得するには，鳥澤の手法 [6] のように，標識を伴わない場合をうまく考慮することが望まれる．

6.2 出来事表現の統語カテゴリ

因果関係をもつ 2 つの出来事のそれぞれについて，それらが動詞句（vp）と名詞句（np）のどちらの形式で表現される傾向にあるかを調査した．統語カテゴリの決定には，主辞要素の末尾形態素の品詞に注目し，動詞か形容詞であれば動詞句，名詞であれば名詞句とした．品詞情報は ChaSen[5] の解析結果を利用した．

結果を表 5 に示す．e_1，e_2 のいずれにおいても動詞句を形成する場合が過半数を占めている．しかしながら，名詞句を形成する事例も決して少なくはない．従来の因果関係知識の獲得手法では，動詞句に注目する

[7] “蓋然” と “偶然” の和が “混合” の数となっているわけではないことに注意されたい．例えば，“111” に該当する事例において，A のみが「蓋然」の強さ，B と C が「偶然」の強さとしてタグを付与していた場合，“混合” では “111” でカウントされるが，“蓋然” では “100”，“偶然” では “011” でカウントされる．

表 5: 出来事表現の統語カテゴリ

カテゴリ	該当例	e_1	e_2
vp	動詞-自立（“焼く”） 形容詞-自立（“難しい”）	365	412
np	名詞-サ変接続（“停電”） 名詞-一般（“火災”）	322	269
	その他（“うっとり”）	12	18

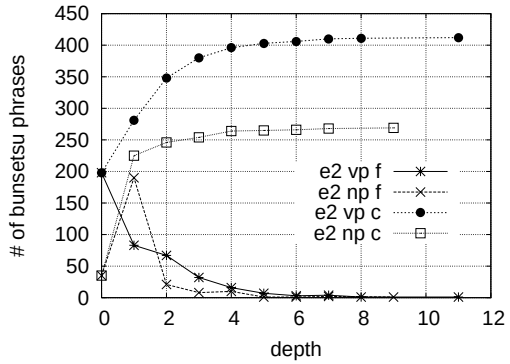

図 2: 出来事表現の出現位置（e_2）

傾向があったのに対し，今回の結果は，今後，名詞句への対応も必要であることを示唆している．

6.3 出来事表現の出現位置

次に，タグが付与された元文のそれぞれを，文末文節を根（深さ＝0）とする係り受け木と考え，出来事の主辞要素を含む文節が位置している深さを調査した．係り受け情報は CaboCha[4] の解析結果を利用した．

図 2 に e_2 の主辞位置の分布を示す．図中の “f” は，ある深さでの頻度を示し，“c” はその深さまでの累積頻度を示す．図 2 から，動詞句を形成する場合は文末に集中しており，名詞句を形成する場合は深さ＝1，すなわち，文末に係っている文節に集中していることがわかる．また，いずれの統語カテゴリにおいても，深さ＝2 までに 8 割以上の事例が集中している．この結果から，因果関係をもつ出来事（の主辞）を発見的に探査する場合，それほど深い位置まで走査をしなくとも大部分の事例が発見できることがわかる．なお，紙面の都合上グラフは割愛するが，e_1 については深さ＝3 までに 8 割以上の事例が集中していることを確認した．

6.4 出来事表現間の相対的な出現位置関係

最後に，e_1 の主辞要素と e_2 の主辞要素の相対的な出現位置関係を調査した．先程は個々の主辞の深さに注目したが，ここでは個々の主辞の深さの差に注目する．

結果を表 6 に示す．表中の “⇒” は，e_1 が e_2 よりも深い位置にあった場合を示し，“⇐” は，e_1 が e_2 よりも浅い位置にあった場合を示す．文内の “×” は，e_1 から係り受け関係を辿った際，先祖あるいは子孫の位置に e_2 がなかった場合を示す．

直感的には e_1 の主辞要素が e_2 の主辞要素に直接係っている場合が多いと予想され，実際の結果も 259 件と最も多かった．しかしながら，それ以外の位置関係にある場合も少なからず存在していることが明らかになった（“×”: 72 件，“文間”: 141 件など）．深さの差＝2 となる事例には,「濡れたシートで足を滑らせる」の「濡れる」と「滑る」のように，連体修飾関係を挟む場合が多く含まれていた．文内 “×” に該当する事例には，抽出された元文中に並列関係が存在している場合が多く，現在の二項関係という枠組みでは捉え難い事例が多く集まっていた．

表 6: 出来事表現間の相対的な出現位置関係

		e1 ⇒ e2	e1 ⇐ e2
文内	深さの差 ＝1	259	15
	＝2	152	23
	＞2	33	4
	×	72	
文間		141	

7 おわりに

本稿では，まず，一定量のテキスト集合に対して因果関係情報を注釈付けたその過程について述べた．また，注釈付けの結果得られたコーパスを用いて，テキスト中に含まれる因果関係の出現傾向を調査し，その結果を報告した．

今回の調査を通して，幾つかの形態や位置に因果関係が含まれていることが定量的に示された．今後は，これらの結果を考慮して，因果関係知識の自動獲得手法を開発する予定である．

謝辞

コーパスの作成にあたり，ランゲージウェアの衛藤純司氏，日本システムアプリケーションの植田禎子氏，十河則子氏，奈良先端科学技術大学院大学の高橋哲朗氏の諸氏から多大な協力を頂きました．諸氏の皆様に感謝いたします．

参考文献

[1] 有田節子. 因果の言語学. 月刊言語, Vol. 25, No. 5, pp. 20–23, 1996.

[2] 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.

[3] 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. 情報処理学会論文誌, Vol. 45, No. 3, pp. 919–933, 2004.

[4] 工藤拓, 松本裕治. チャンキングの段階適用による係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.

[5] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸. 日本語形態素解析システム『茶筌』 version 2.2.1 使用説明書. 奈良先端科学技術大学院大学, 2000.

[6] 鳥澤健太郎.「常識的」推論規則のコーパスからの自動抽出. 言語処理学会第 9 回年時大会, pp. 318–321, 2003.

[7] 毎日新聞社. 毎日新聞 CD-ROM 版（1995）.