SDRTに基づく因果関係認識日本語評価データ構築手法の提案

金子 貴美[1]　　戸次 大介[1,2,3]

[1] お茶の水女子大学大学院 人間文化創成科学研究科 理学専攻 情報科学コース

[2] 国立情報学研究所

[3] 独立行政法人科学技術振興機構, CREST

[1]{kaneko.kimi, bekki}@is.ocha.ac.jp

1　はじめに

高度な意味処理の実現が急務の課題と目され、その実現に向け様々な研究が行われている。そのなかに、文章に内在する因果関係認識、テキストからの因果知識抽出に関する研究 [1,2,3,5] がある。これらの研究は、注目を集めるようになってきており、進展が見られるものの、解決すべき課題も多く残されている。

乾ら [2] は、「から」「ので」などの表現を手がかりに、前件の出来事を原因、後件の出来事を結果とする因果知識を獲得している。たとえば (1a) では、前件「雨が降った」が原因、後件「水たまりができた」が結果として獲得される。

(1) a. 雨が降った ので、水溜りができた。

しかし、これらの表現には必ずしも前件が後件の原因・理由とならない用法も存在する。以下に例を示す。

(1) b. 人身事故が起きた から、電車が遅延したというわけではない。

この例文 (1b) では、前件「人身事故が起きた」ことが後件「電車が遅延した」ことの「原因」ではない。このように、因果表現が存在する一方、前件と後件の出来事に因果関係がないものがあるが、[2] などの既存研究では、因果表現があれば因果関係があるものとして知識を獲得してしまうため、このようなケースを排除するのが難しいという問題がある。

したがって、本研究では計算機に因果関係がある表現をより正確に抽出させるために必要な情報を調査し、日本語評価データとその構築手法を確立することを目標とする。本論文ではまず関連研究の分析を行い、評価データに付与すべき情報を述べた後、分節談話表示理論 (以下、“SDRT”)[4] を元にした提案手法を説明する。また、提案手法により、実際に注釈付けた文を評価し、その結果を報告・考察する。

2　関連研究

因果・時間などの関係のアノテーションについての研究、および、因果・時間関係を言語学的に分析した研究について述べる。

Bethard ら [5] は、因果関係を注釈付けた英語のデータを構築している。また、因果関係と同時に時間関係も併せてアノテーションし、因果と時間順序の関わり方を調査すると共に、作成したデータを一致率、認識精度の観点から評価した。評価において、因果関係は2種類（因果関係あり、なし)、時間関係は3種類（前節のイベントが先、前節のイベントが後、不明）と分類が粗かったため、より詳細に関係を定義して再分析する必要があると彼らは指摘している。

一方、乾ら [6] は、日本語文書における因果表現の出現割合を調査し、日本語の因果関係のタグ付きコーパスを作成している。しかし、彼らは、前述の (1b) などの一部の用法については考慮していない。また、因果は時間順序に影響を受ける関係であるが、その相互関係についての分析がなされていないという課題が残る。

日本語における、時間・因果関係を言語学的に分析した研究として田村 [7] のものがある。これによると、日本語の理由・目的表現は、時制形式から予測される出来事の時間についての予測と、実際の解釈が齟齬を起こすことがあり、因果表現における時間の前後関係は、認識視点や文の意味、意図によって決まると述べられている。また、日本語の因果構文には、主節と従属節の過去・非過去の選択が全く自由であり、絶対時制・相対時制のシステムに従わないものもある [8]（次節で例を示す)。したがって、時制は必ずしも因果関係の有無よりも先に決められるものではなく、時間関係と因果関係は同時に決める必要がある。

Asher[4] は、SDRT という、談話関係が意味に及ぼす影響を考慮した意味論を構築している。談話関係に

は因果に関する関係が含まれている。そして、上で述べてきたように、因果関係は時間順序に影響され、日本語の因果表現における時間順序は、認識視点や文の意味、意図により決まるので、因果関係の有無もまた認識視点、文意や意図によって決まってくる。したがって、因果関係と共に談話関係を注釈付けることで、文間の意味的関係を考慮することができ、因果関係を認識する上でより有用なコーパスが構築可能になる。また、文の時制、イベント発生時間、イベントを認識した時間、文の意味、文構造、因果の有無、のどれが事前手がかりとなり、どれが出力となるかは時と場合により変わり、時間・因果関係と共に談話関係も同時に判定する必要があると考えられるため、談話関係と共に時間・因果関係を付与することで、これらの情報を同時判定させる際の正解データとすることができる。

しかし、SDRT の談話関係は時間関係と分けておらず、時間関係と談話関係を同時に付与するとなるとややこしくなるという問題や分けていないために関係の種類が不必要に複雑になってしまっているという問題があり、談話関係の整理をする必要があると考えられる。以下に例を示す。

(2) 犬は庭をかけ回った。猫は炬燵でまるくなっていた。

この文は、対句であるので、SDRT の “Contrast” ラベルが付与される。一方で、1 文目の状況は 2 文目の状況と時間的に重なっているため、“Background” ラベルにも該当する。このように、SDRT では、複数の談話関係に当て嵌まってしまうケースが少なからず存在するが、これは本研究のように時間関係と分離させることで、談話関係の重複を避けることが可能である。

これらの点を踏まえ、本研究では、SDRT を談話関係、時間関係と因果関係に分離して整理し、ある程度網羅的な談話関係の理論を再構築すると共に、それに基づいてアノテーションすることとした。また、イベント発生時間とイベントを認識した時間を分けて捉えたい場合もある。以下に例を示す。

(3) 明日テストがあるので、今日は勉強することにした。

ここで、前件「明日テストがある」という事実を認識したのは、後件「今日は勉強する」の判断を行う前のはずである。前件を後件の前に認識したこととして捉えたいか、後に起こるイベントとして捉えたいかはケースバイケースであると考えられる。したがって、こうした区別を行いたい場合に対応できるよう、認識視点情報を付与することとした。

3 提案手法

文の主節と従属節、等位接続、およびその中間的なもの（日本語の連用接続など）と連続する 2 文、および隣接する文（節）のノードに対して、1 つの談話関係を付与することとし、1 つの命題につき、それぞれ 1 つの時間関係と因果関係を付与することとした。例文 (4a) への関係ラベルの付与結果は以下の (4a') のようになる。

(4) a. 風が吹いた。張り紙が剥がれ、飛んだ。
a'. [**Precedence**(π1,π3),**Explanation**(π1,π3), **CAUSE**(π1,π3)],
[**Precedence**(π2,π4), **Explanation**(π2,π4), **CAUSE**(π2,π4)]
$_{\pi 2 \pi 1}$風が吹いた。$_{\pi 4 \pi 3}$張り紙が剥がれ、飛んだ。

以下、3.1 節で今回提案する時間関係を、3.2 節で因果関係、3.3 節で談話関係を示し、3.4 節で認識視点について説明する。

3.1 時間関係

時間関係は、以下の 3 種類を用意した（表 1）。これらは 2 つの「命題に含まれるイベント[1]」間の時間的関係を表し、イベントは時間上のインターバルとして、開始時間と終了時間を持つものと仮定する。また、任意のイベントについて、(e の開始時間)≤(e の終了時間)、と仮定する。2 つの引数の順序を考慮すれば、このように定義することにより、任意の 2 つのイベントの時間的な配置は、表 1 の 3 つに限られる。

関係ラベル	説明
Precedence(A,B)	終了時間 (A) <開始時間 (B), すなわちイベント A がイベント B に先行する。
Overlap(A,B)	開始時間 (A)≤ 開始時間 (B)≤ 終了時間 (B)≤ 終了時間 (A), すなわちイベント A とイベント B は重なっている。
Subsumption(A,B)	開始時間 (A) <開始時間 (B), かつ終了時間 (A) <終了時間 (B), すなわちイベント B はイベント A に時間的に真に包含される。

表 1: 時間関係一覧

しかし、日本語の非過去形述語は、「習慣的繰り返し」を表すことがあり、この場合は「参照点より未来の出来事」を指す用法とは区別されなければならない。本論文では、「習慣的繰り返し」は、以下の例のようにそのスコープを注釈付けることで明示する。

(5) a. 退院後、{ 公園を走る }$_{repeat}$ ようにしている。
b.{ スポーツ飲料を飲んだ後、公園を走る }$_{repeat}$ ようにしている。

[1] これは一般的仮定ではないが、自然言語の文では一般的に成立すると考えて差し支えない。

3.2 因果関係

「命題に含まれるイベント」間に因果関係がある場合のみ、以下の関係を付与することにした（表 2)。表 2 の関係のみだと、既存研究 [5] と同じく 2 種類（因果関係あり、なし）の記述しかできないが、談話関係、認識視点、時間と組み合わせることによって、既存研究 [5] より細かく因果関係の有無の記述ができる。

関係ラベル	説明
CAUSE(A,B)	A のイベントと B のイベントに因果関係がある。

表 2: 因果関係一覧

3.3 談話関係

談話関係は、SDRT を元に、表 3 の 7 種を用意した。

関係ラベル	説明
Alternation(A,B)	「A か B」のように、論理の「∨」の関係と対応するもの。
Consequence(A,B)	「A ならば B」のように、論理の「→」の関係と対応するもの。
Elaboration(A,B)	B が A の詳細を説明する談話関係。 B のイベントは A のイベントの部分をなす。
Narration(A,B)	A と B を同じ状況に配置し、論理の「∧」と対応する談話関係。
Explanation(A,B)	A が B の原因・理由であることを述べる談話関係。
Contrast(A,B)	A と B を逆説的に対比する談話関係。
Commentary(A,B)	A の内容を B で要約したり、補足したりする談話関係。

表 3: 談話関係一覧

ここで挙げた談話関係は、SDRT 同様、時間関係と因果関係に制約を課すものが存在する。時間関係、因果関係と談話関係がどのように影響を及ぼし合うか、そして、本論文の時間関係と因果関係と談話関係の組み合わせが、SDRT の談話関係とどのように対応するかを表 4 に示す。尚、ここで**太字の関係**は本論文では統廃合した関係である。

SDRT	本論文	規則
Alternation(A,B)	Alternation(A,B)	-なし-
Consequence(A,B)	Consequence(A,B)	-なし-
Elaboration(A,B)	Elaboration(A,B)	∀ A,B(Elaboration(A,B) → Subsumption(A,B))
Narration(A,B)	Precedence(A,B) ∧ Narration(A,B)	なし
Background(A,B)	Subsumption(A,B) ∧ Narration(A,B)	なし
Result(A,B) **Explanation(A,B)**	Explanation(A,B) Cause (A,B)	∀ A,B(Cause(A,B) → Temp_rel(A,B)) [2]
Contrast(A,B)	Contrast(A,B)	-なし-
Commentary(A,B)	Commentary(A,B)	-なし-

表 4: SDRT と本論文の関係と、適規される規則の対応

このような時間関係に対する制約は、日本語の因果表現の「脱テンス」[3] [8] 構文において、時間順序、および前件と後件どちらが原因・理由表現であるかを判断する上で役に立つ。以下に例を示す。(5) の例は、従属節が非過去、主節が過去となっている文であり、一見主節のイベントの方が後であると判断しかねないが、“Explanation” の制約によって、そうではないことが分かる。

(5) [**Precedence**(π1,π3),**Explanation**(π1,π3), **CAUSE**(π1,π3)],
[**Precedence**(π2,π4),**Explanation**(π2,π4), **CAUSE**(π2,π4)]
π2π1昨日あんなに食べる から、
π4π3今日お腹が痛くなった んだ。

3.4 認識視点

認識視点付き命題と、認識視点なし命題を別々にマークすることで、認識レベルと事実レベルを区別する。本節でその区別の方法を述べる。

「〜のだ」は認識視点付き命題の代表的な例である。したがって、「〜」の部分と、「〜のだ」の部分を別々にマークする。また、「〜ので」「〜から」が田村のいう「認識的因果用法」で用いられている場合は、「ので」「から」の前に空の「のだ」が存在すると考えるため、「〜」と「〜（のだ）」の両方をマークする。結果的に、「〜」は二重にマークされる。 このように定義することにより、下記のような、田村 [7] のいう「根拠用法」における認識付命題同士の因果関係を書き分けることが可能となる。

(6) [**Precedence**(π3,π1), **Explanation**(π3,π1), **CAUSE**(π3,π1)],
[**Precedence**(π2,π4), **Explanation**(π2,π4), **CAUSE**(π2,π4)]
π2π1今朝何も報道されなかった ので、
π4π3昨日はめぼしい事件は起こらなかった のだ。

ここで、事実レベルの時間関係は、π3 が前、π1 が後である一方で、認識レベルでは、π1 の事実を認識した後に π4 の判断を行っていると考えられるが、π1 と π3 の関係と π2 と π4 の関係をそれぞれ分けて記述することで、事実レベル、認識レベル両方の関係性を再現できる。

3.5 提案手法の利点

ここまでで定義してきたように、認識レベルと事実レベルを区別した上で、談話関係と時間関係、因果関係を同時に付与することで、既知の因果情報のほか、文と文の意味的な影響関係、接続詞、文（節）間の時間順序などの様々な手がかりから、事実レベルと認識レベルで因果関係の有無を多角的に判断する正解データを構築できるようになった。

[2] Temp_rel(A,B) ≡
Precedence(A,B) ∨ Overlap(A,B) ∨ Subsumption(A,B)

[3] 沈 [8] によると、「脱テンス」とはテンス的な意味を失っているもの、つまり前件と後件の意味関係の論理的な面が強調されることにより、時間関係が裏に引っ込んでしまうようなもののことである。

また、談話関係と因果関係を分けて定義することで、因果表現であるが因果関係がない場合は因果関係がないものとして、因果表現ではないが因果関係がある場合は因果関係があるものとして書き分けられるようになった。

4 評価

BCCWJ のデータの一部に本手法を適用した。53 文を 1 名で 156 セグメントに分割し、2 名のアノテータがラベル付けを行った。アノテーションには、3 節で挙げた時間関係, 因果関係および談話関係のラベルを用いた。156 セグメント中、96 セグメント (38 文) は本手法の策定に用い、残りの 60 セグメント (17 文) について一致率と平均セグメント数、ラベル付けの平均時間を産出した。今回扱ったデータにおける一致率は 0.68、カッパ値は 0.80 であった。一致率は以下の式で産出した。

$$\text{一致率} = \text{ラベル一致数}/\text{全体数}$$

一致率の値から、実データによる分析と方法論の改良を行っていけば、本手法は実用性のあるものとなると考えられる。

セグメント数は全体で 60 セグメントであり、平均 3.53 セグメント/文であった。これは認識レベル、事実レベル両方を含めた数であるため、事実/認識レベルを分けなければ、平均 1.77 セグメント/文だったことになる。複雑な構造の文も存在しているにも関わらず、このように分割数が少ないのは、本手法では区切らない、節内に埋め込まれた関係節などが多いためであると考えられる。

表 5 に本研究における関係ラベルの出現頻度を示す。

関係ラベル	セグメント数		
	全体	事実レベル	認識レベル
Precedence	16	9	7
Overlap	2	1	1
Subsumption	40	19	21
total	**58**	**29**	**29**
CAUSE	9	4	5
total	**9**	**4**	**5**
Alternation	–	–	–
Consequence	2	1	1
Elaboration	2	1	1
Narration	44	22	22
Explanation	10	5	5
Contrast	–	–	–
Commentary	–	–	–
total	**58**	**29**	**29**

表 5: 本研究における関係ラベルの分布

「Narration」が高頻出、「Alternation」「Contrast」「Commentary」が 1 度も出現しないという、大きく偏った結果となった。サンプル数が少ないため、もっと多くのデータで改めて分析する必要があると考えられるが、この結果のみからでも、高頻出な関係と、低頻出な関係に二分されるであろうことが予測できる。

また、ラベル付けの平均時間は、1 セグメントあたり平均 1.46 分であった。平均セグメント数から、1 文あたり平均約 5.15 分だったことになる。平均セグメント分割時間も測定した上で改めて判断を行うべきではあるが、ラベル付けに関してのみ言えば、本手法は妥当だと考えられる。

5 まとめ

因果関係がある表現をより正確に抽出させるために必要な情報を調査し、SDRT をベースにした、因果関係認識のための日本語評価データとその構築手法を提案した。また、156 セグメント（53 文）について、時間関係、談話関係と因果関係をアノテーションし、その結果を分析・報告した。

6 謝辞

本研究の一部は、情報・システム研究機構データ中心科学リサーチコモンズ事業の助成を受けたものである。ここに謝意を表する。

参考文献

[1] 乾孝司，高村大也，奥村学：因果関係知識獲得のための隠れ変数モデル，言語処理学会第 12 回年次大会，pp. 959-962.（2006）

[2] 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.42，No.12，pp. 3160-3172.（2001）

[3] Riaz，Mehwish. and Roxana Girju：Toward a Better Understanding of Causality between Verbal Events:Extraction and Analysis of the Causal Power of Verb-Verb Associations，Proceedings of the SIGDIAL 2013 Conference，pp. 21-30.（2013）

[4] Asher，Nicholas and Alex Lascaridas：Logics of Conversation: Studies in Natural Language Processing，Cambridge University Press.（2003）

[5] Bethard，Steven and William Corvey，Sara Kilingenstein，James H. Martin：Building a Corpus of Temporal Causal Structure，LREC2008.（2008）

[6] 乾孝司，奥村学:因果関係タグ付きコーパスの構築と分析，言語処理学会第 11 回年次大会，pp. 486-489.（2005）

[7] 田村早苗：認識視点と因果:日本語理由・目的表現の研究，博士論文，京都大学.（2012）

[8] 沈矛一：複合文の接続助詞でくくる節の述語のテンスー「スルが」と「シタが」、「スルので」と「シタので」などー，語学教育研究論叢，pp. 120-122.（1984）