Vol.2013-NL-214 No.11
2013/11/15

SDRTに基づく日本語談話アノテーションの試み

金子 貴美[1,a)]　戸次 大介[1,2,3,b)]

概要：
因果関係認識は，深い意味処理の実現を目指す上で重要な問題の 1 つであり，近年，そのニーズが高まってきている．認識に必要な，因果知識を自動獲得する方法論についての研究は進展が見られる一方で，コーパス作成の方法論については言語学的観点からの改善の余地がある．したがって，本研究では，言語学における因果関係の知見を用いた，SDRT ベースのアノテーションの枠組みを提案し，実際にアノテーションを行った文を分析した結果について考察・報告する．

キーワード：アノテーション, コーパス，SDRT（分節談話表示理論）, 因果関係認識，意味処理

Discourse Annotation of Japanese based on SDRT

Abstract: In recent years, the need for recognizing causal relations has been increasing, and thus motivated techniques for the automatic acquisition of causal knowledge. However, we claim that the corpus annotations for this purpose face several methodological problems. We propose an annotation scheme based on SDRT, integrated with the recent development of causality in Japanese linguistics.

Keywords: Annotation, SDRT(Segmented Discourse Representation Theory), Recognizing causal relations, Semantic processing

1. はじめに

因果関係に関する知識は，高度な意味処理の実現を目指す上で重要な要素の 1 つである．近年では，因果関係を自動認識するシステムの必要性が高まっており，認識に必要な因果知識を自動獲得する方法論として，乾ら [1], [2],Riaz ら [3] のものなどが既に提案されている．このように知識獲得の手法についての研究は進展しているが，一方で言語学における因果関係の知見が必要十分に反映されているとは言えず，こうした観点からの改善の余地があると考えられる．

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. 今朝の新聞に何も載っていない ので，昨日は特筆すべき事件はなかったのだ．

この根拠用法の例文 (1b) では，前件「新聞に何も載っていない」ことは後件「事件がなかった」ことの「原因」ではない．このことは，テキストから因果関係を自動抽出する，というタスクを考える上で重要な区別となる．

したがって，本研究では計算機に因果関係を正しく捉え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．

[1] お茶の水女子大学大学院人間文化創成科学研究科
Ochanomizu Uniersity, Graduate School of Humanities and Sciences
[2] 国立情報学研究所
National Institute of Infomatics
[3] 独立行政法人科学技術振興機構，CREST
CREST, Japan Science and Technology Agengy
[a)] kaneko.kimi@is.ocha.ac.jp
[b)] bekki@is.ocha.ac.jp

© 2013 Information Processing Society of Japan

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，計算機に文間の因果関係の有無を推論させる際，適切なコーパスを与えて学習させると関係が捉えやすくなり，精度の高い分類器を作成できると述べ，因果関係を注釈付けた英語の評価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を行い，因果関係と時間の前後関係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 3 種類，時間関係は 2 種類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．

一方で，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た研究として，乾ら [6] のものが存在する．しかし，彼らが手がかり表現として用いている表現は限られており，「わけ」「ということ」など因果表現となりうる表現や，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もなされていないという問題がある．

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は，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時についての予測と，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（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．そのため，談話関係に広げて注釈付けを行うことで，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．しかしながら，SDRT には, 談話関係は因果に関する関係が少ないという問題と，時間関係と因果関係を分けていないため，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) 太郎は犬が好きだ．花子は猫が好きだ．

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を行うこととした．

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 文に対して，1 組のイベントにつき，それぞれ 1 つの時間関係と談話関係を付与することとした．例文 (3a) への関係ラベルの付与結果は以下の (3a') のようになる．

(3) a. 風が吹いた．剥がれた張り紙が飛んで行った．
a'. **[Precedence(π1,π2),Cause(π1,π2)]**
π1風が吹いた．π2剥がれた張り紙が飛んで行った．

以下，3.1 節で今回提案する時間関係を，3.2 節で談話関係を示す．

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 つのイベント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．

© 2013 Information Processing Society of Japan

(4) a. 退院した後，{ 公園を走る }$_{repeat}$ ようにしている．
b.{ スポーツドリンクを飲んだ後，公園を走る }$_{repeat}$ ようにしている．

3.2 談話関係

談話関係は，SDRT を元に，表 2 の 8 種類を用意した．

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)	下記 (5a) のように，B が A の詳細を説明する用法． イベント B はイベント A に包含される．
Narration(A,B)	下記 (5b) のように，トピックに繋がりがあるもの．
Cause(A,B)	(1a) のように，前件が後件の事柄の原因・理由となる用法．
Account(A,B)	(1b) のように，前件が後件の判断の根拠になる用法．
Contrast(A,B)	対句法などのように，A と B が類似した意味構造と対照的な 意味を持つもの．および，「A だが B」のような逆接の用法．
Parallel(A,B)	A と B が類似した意味構造を持つが，A と B がテーマを 共有していたり，類似の意味を持つもの．

表 2 談話関係一覧

(5) a.**[Subsumption(π1,π2),Elaboration(π1,π2)]**
$_{\pi 1}$コース料理をいただいた．$_{\pi 2}$まず，食前酒を飲んだ．
b. **[Precedence(π1,π2),Narration(π1,π2)]**
$_{\pi 1}$東京駅に行った．$_{\pi 2}$新幹線に乗った．

ここで挙げた談話関係は，SDRT 同様，時間関係に制約を課すものが存在する．時間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 3 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)	Cause(A,B)	∀ A,B(Cause(A,B) → Temp_rel(A,B)) *1
-該当なし-	Account(A,B)	-なし-
Contrast(A,B)	Contrast(A,B)	-なし-
Parallel(A,B)	Parallel(A,B)	-なし-

表 3 SDRT と本論文の関係の対応と，適用される規則一覧

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」*2[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．

(6) **[Precedence(π1,π2),Cause(π1,π2)]**
$_{\pi 1}$昨日あんなに食べるから，$_{\pi 2}$今日お腹が痛くなったんだ．

この例は，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，Cause の制約によって，そうではないことが分かる．

4. 分析と考察

本手法に基づき，試験的に 5 文を実際に注釈付けた．以下では，その 5 文を対象に考察する．

Cause の関係に該当するか，その他の談話関係に該当するかの判別が難しいケースとして，まず「〜たら」「〜れば」などの条件文の場合が存在することが分かった．以下に例を示す．

(7) a. **[Precedence(π_1,π_2),Cause(π_1,π_2)]**
$_{\pi 1}$自販機を蹴ったら，$_{\pi 2}$故障して，
$_{\pi 3}$商品が出なくなった．
a'. **[Overlap(π_2,π_3),Cause(π_2,π_3)]**
$_{\pi 1}$自販機を蹴ったら，$_{\pi 2}$故障して，
$_{\pi 3}$商品が出なくなった．
a". **[Precedence(π_1,π_3),Cause(π_1,π_3)]**
$_{\pi 1}$自販機を蹴ったら，$_{\pi 2}$故障して，
$_{\pi 3}$商品が出なくなった．

この文は条件文であるので，今回策定した談話関係の分類では，(7a) および (7a") の談話関係は，それぞれ Cause(π_1,π_2)，Cause(π_1,π_3) か Consequence(π_1,π_2)，Consequence(π_1,π_3) のどちらかに該当することになると考えられ，判断が揺れることが予想される．ゆえに，このような条件文に関しては，用法分類を行う必要がある．また，π_3 は π_2 の具体的な症状の 1 つであるので，(7a') は Cause(π_2,π_3) と Elaboration(π_2,π_3) で判断が揺れる可能性もあり，これら 2 関係についても，より具体的な線引きが要ると考えられる．

判断が難しいもう 1 つのケースとして，以下のような Contrast が出現する例も存在する．

(8) a. **[Subsumption(π_2,π_1),Contrast(π_1,π_2)]**
$_{\pi 1}$太郎はコンタクトレンズを買ったが，
$_{\pi 2}$1 日使い捨てのものだったので，
$_{\pi 3}$洗浄液を買う必要がなかった．
a'. **[Subsumption(π_2,π_3),Cause(π_2,π_3)]**
$_{\pi 1}$太郎はコンタクトレンズを買ったが，
$_{\pi 2}$1 日使い捨てのものだったので，
$_{\pi 3}$洗浄液を買う必要がなかった．
a". **[Precedence(π_1,π_3),Contrast(π_1,π_3)]**
$_{\pi 1}$太郎はコンタクトレンズを買ったが，
$_{\pi 2}$1 日使い捨てのものだったので，
$_{\pi 3}$洗浄液を買う必要がなかった．

ここでは，π_1：「太郎はコンタクトレンズを買ったが，」と，$\pi_2+\pi_3$：「1 日使い捨てのものだったので，洗浄液を買う必要がなかった．」とで Contrast の関係になっていると考えられ，意味的には，B 節と C 節を 1 つの節として捉えた上

*1 Temp_rel(A,B) ≡
Precedence(A,B) ∨ Overlap(A,B) ∨ Subsumption(A,B)
*2 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って，時間的関係の面が裏に引っ込んでしまうようなもののことである．

© 2013 Information Processing Society of Japan

でラベル付けを行ったほうが自然であるように思われる．今回は 1 つのイベントを 1 つの節とみなしてラベル付けを行ったが，上記のような例もあるため，どのような単位でラベル付けを行うか，は改めて検討すべきである．

また，以下の例のように，因果関係の有無の判断が曖昧になるケースも存在する．

(9) a. **[Precedence(π1,π2),Narration(π1,π2)]**
$_{\pi1}$昼休みになり，$_{\pi2}$次郎は食堂へ行った．
$_{\pi3}$次郎は財布を忘れて行った．
a’. **[Subsumption(π2,π3),Narration(π2,π3)]**
$_{\pi1}$昼休みになり，$_{\pi2}$次郎は食堂へ行った．
$_{\pi3}$次郎は財布を忘れて行った．
a”. **[Precedence(π1,π3),Narration(π1,π3)]**
$_{\pi1}$昼休みになり，$_{\pi2}$次郎は食堂へ行った．
$_{\pi3}$次郎は財布を忘れて行った．

(9) は，「昼休みなり，次郎は食堂に行った．その結果，財布を研究室に置いて行く羽目になった」という意味の文であるが，「財布を研究室に置いて行く羽目になった」ことの直接的原因は「食堂に行く際に財布を持たなかった」ことであって，「食堂に行った」ことそのものは直接的な原因ではない．この「食堂に行く際に財布を持たなかった」のように，暗黙の前提となっている文を書き下すことで因果関係を明確にするなど，対処法を検討する必要があると考えられる．

さらに，根拠用法の場合は，時間順序の判断が難しいケースがある．以下に例を示す．

(10) a. **[Subsumption(π1,π2), Account(π1,π2)]**
$_{\pi1}$11 月 23 日に大学の文化祭がある．
$_{\pi2}$そろそろ，本格的に文化祭の出し物の準備をしなければならない．
b. **[Precedence(π1,π2), Narration(π1,π2)]**
$_{\pi1}$11 月 23 日に大学の文化祭がある．
$_{\pi2}$文化祭後は打ち上げをするので，$_{\pi3}$その会場を予約しておかなければならない．
b’. **[Subsumption(π2,π3), Account(π2,π3)]**
$_{\pi1}$11 月 23 日に大学の文化祭がある．
$_{\pi2}$文化祭後は打ち上げをするので，$_{\pi3}$その会場を予約しておかなければならない．
b”. **[Overlap(π1,π3), Narration(π1,π3)]**
$_{\pi1}$11 月 23 日に大学の文化祭がある．
$_{\pi2}$文化祭後は打ち上げをするので，$_{\pi3}$その会場を予約しておかなければならない．

(10a) の場合，π_1：「11 月 23 日に大学の文化祭がある．」という文は π_2：「そろそろ，本格的に文化祭の出し物の準備をしなければならない．」の文よりも未来の予定について述べたものであるが，発話者が π_1 の事実を認識したのは，π_2 の発話をする前である．このように，認識視点を考慮すべきケースが存在する一方で，(10b)(10b’)(10b”) のように，文と文の認識視点を考慮すべき場合とそうでない場合が混在するケースも存在する．したがって，考慮すべきか否かの判断を計算機が行えるよう，認識視点の情報もアノテーションする必要性があると考えられる．また，Bethard ら [6] は，「全ての CAUSAL 関係は BEFORE 関係の影響下にある（つまり，原因と結果は時間の前後関係がある）と予想したが，CAUSAL 関係の 32%は BEFORE 関係と無関係であった．」と述べており，分類方法をより洗練してから再分析すべきであることを示唆していた．本研究のように時間関係を定義すること，および認識視点の情報を追加することで，Bethard ら [6] が因果関係と時間関係の関連性を言及できなかった 32%のデータに関しても，それらのデータが Precedence(A,B)，Overlap(A,B) か Subsumption(A,B) のどれかに該当することから，関連性が示せると考えられる．

5. まとめ

本論文では，SDRT を元にし，時間関係と談話関係を別々に注釈付ける手法を提案した．また，5 文についてアノテーションを行い，その結果を分析・報告した．

分析から，用法分類，因果関係の分解や認識視点の追加などによって，分類が難しいケースへ対処する必要があること確認された．今後は，この分析結果を考慮してアノテーションスキーマを改良し，因果関係の自動獲得に利用可能なコーパスを作成する予定である．

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)

© 2013 Information Processing Society of Japan