1K3-4

因果関係認識のための日本語談話関係アノテーションとその分析

Annotation with Japanese Discourse Relations and the Analysis for Causal Relation Extraction

金子 貴美[*1] 戸次 大介[*1*2*3]
Kimi Kaneko Disuke Bekki

[*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
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

[*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
National Institute of Informatics CREST, Japan Science and Technology Agency

This paper analyzes specialized Japanese data sets for causal relation extraction created by the methodology of (Kaneko 2014b) and discusses what the causality in text is actually like and what the framework for the causal relation extraction should be.

1. はじめに

兼ねてより、深い意味処理の実現のためには、因果知識が必要であることがよく知られており、テキストからの因果関係認識に関する様々な研究が行われている [1,2,3]。しかしながら、このタスクには未だ解決すべき課題が数多く残っており、実現は難しい。何故難しいのであろうか？

因果の有無は決定的には決まるものではない。以下の例のように、不確実性を伴う状況下で話題にのぼる場合や、例外がある場合が多く存在する：

(1) 目がやたら痒い ので、遂に花粉症になったのかもしれない。
(2) 風邪を引く と、熱が出る。(→ 風邪を引いても熱が出ないこともある)

また、部分的な情報から導き出されることも多い。故に確率的なモデルを使って認識させるのが妥当である。一方で、「因果」の語義に法則的な必然性という意味が含まれていることからもわかるように、因果関係の有無は無作為に決まっているわけでもない。テキスト中の因果関係を捉える場合、話者の世界観や言語表現などが手がかりとなるが、因果を示す言語表現があるからといって必ずしも因果関係があると断定できるとは限らない。これらのことが、適切なリソースの選択と適切な認識器の設計が難しくする成因となっている。適切なリソースや認識モデルの選択や構築をするために、何が因果関係で、何がそうではないのかを整理する必要がある。

したがって、本稿では、金子ら (Kaneko 2014b) のアノテーション手法により注釈付けた、因果関係認識のための日本語評価データの分析を行い、テキスト中の因果関係が実際はどうなっているのか、および、因果関係認識の枠組みはどのようなものにすべきかを議論する。

2. 関連研究

本稿の分析対象とする、因果関係認識のための日本語評価データ (Kaneko 2014b) について述べる。この研究では、Asher(Asher 2003) の談話関係の理論 SDRT を元に談話関係、因果関係を日本語に合わせて再構築し、提案手法に基づき、必要に応じて、認識レベル (イベントを認識した時間) と事実レベル (イベント発生時間) 両方の談話セグメントの対に対してアノテーションを行っている。談話関係と因果関係を分けているという点、つまり、因果を示す言語表現と因果関係の有無を別々に注釈付けようと試みている点が本稿の趣旨に合っていると考え、本データを分析に用いることにした。

金子ら (Kaneko 2014b) は、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、各文（節）ノードの命題[*1]のペアにつき、1 つの因果関係を付与することとした。また、事実レベル・認識レベルについては、すべてのセグメントについて両方のレベルにアノテーションするのではなく、モダリティ表現や接尾辞「のだ」が存在する場合のうち、事実/認識レベルの区別が必要な場合に限り、両方のレベルにアノテーションを行うことにした。たとえば、本手法における例文 (3)(4) への関係ラベルの付与結果は以下のようになる：

(3) a. 雨が降ったので、水溜りができた。
a'. [**Explanation(π2,π3), CAUSE(π1,π3)**],
$_{\pi 2 \pi 1}$雨が降った ので、$_{\pi 3}$水溜りができた。
a". 時間関係：Precedence(π1,π3), Precedence(π2,π3)

(4) a. 今朝、首が酷く痛かったので、寝違えたのだろう。
a'. **[Explanation(π2,π4), CAUSE(π3,π1)]**
$_{\pi 2 \pi 1}$今朝、首が酷く痛かった ので、
$_{\pi 4 \pi 3}$寝違えた のだろう。
a". 時間関係：Precedence(π3,π1), Precedence(π2,π4)

(3) の後件部 π3 は事実/認識レベルに分けられていないが、前件の事実レベルの節である π1 と対にすることにより事実レベルにおける π3 のふるまいを記述することができ、前件の認識レベルの節である π2 と対にすることにより認識レベルにおける π3 のふるまいを記述することができる。また、この研究では、CAUSE は事実レベルの因果関係とし、Explanation

連絡先: 金子 貴美, 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 情報科学領域 戸次研究室, 〒 112-8610 東京都文京区大塚 2-1-1,03-5978-5789, kaneko.kimi@is.ocha.ac.jp

[*1] ここでの命題とは、イベントもしくはステートを示す動詞のことで、例文 (3)(4) だと「降った」「できた」「痛かった」「寝違えた」などが該当する。

は認識レベルのノードに付与され、認識レベルの因果関係を暗にマークするが、事実・認識両方に作用する談話関係としており、CAUSE(A,B)、Explanation(A,B) はともに Precedence(A,B) という時間関係（本論文では時間関係のアノテーションは行わない）を要求する。しかしながら、上の例では、事実レベルの因果関係である CAUSE(π3,π1) と、認識レベルの因果を暗にマークする談話関係である Explanation(π2,π4) は、それぞれ別のセグメントについての関係として区別されているため、それぞれがもたらす時間関係 Precedence(π3,π1). Precedence(π2,π4) は矛盾しないようになっている。

続いて、以下に 2.1 節で (Kaneko 2014b) における因果関係を、2.2 節で談話関係について詳細に説明する。

2.1 因果関係

事実レベルにおける、「文に含まれる命題」のペアの間に因果関係がある場合のみ、以下の関係を付与する（表 1）。認識レベルにおける因果関係は、談話関係 Explanation が暗に導入する。

関係ラベル	説明
Cause(A,B)	A の命題と B の命題に因果関係がある。

表 1: 因果関係一覧

この因果関係 Cause(A, B) は、命題 B が命題 A に先行するという時間関係 Precedence(A, B) を暗に要求する。つまり、Precedence(A, B) を満たさなければ CAUSE(A, B) は付与されず、Cause(A, B) が付与されれば Precedence(A, B) が（暗に）導入されることとなる。

2.2 談話関係

談話関係は、表 2 の通りであり、イベント（event）、状態（state）の判断には宇津木ら (宇津木 2015) の分類を採用している。また、これらの談話関係が、SDRT、および (Kaneko 2014a) の談話関係とどのように対応するかを表 3 に示す。表 3 に示すように、本研究の談話関係は、(Kaneko 2014a) の時間関係と談話関係を統合したものになっている。

関係ラベル	説明
Conjunction(A, B)	A の情報に B の情報を追加する談話関係。 論理の「$\vee$」の関係と対応するもの。
Alternation(A,B)	「A か B」のように、論理の「$\vee$」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「$\rightarrow$」の関係と対応するもの。
Adversative(A,B)	A と B が順接的になっている関係。
Contrast(A,B)	A と B を逆説的に対比する談話関係。
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、B … のように、前から順番に事実を述べるもの。 イベント A と イベント B は同じ状況に配置される。
Explanation(A,B)	A が B の原因・理由であることを述べる談話関係。
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。
Addition(A, B)	「状態 A 。また、状態 B」のように状態を並列する談話関係。
Background(A, B)	B が、A の背景的状況を述べる談話関係。A はイベント、B は状態。
Parallel(A, B)	イベントを並列する談話関係。 Narration と異なり、A、B に時間的重なりがある場合に用いる。
Introduction(A, B)	イベント B が状態 A の参照点を受け継がず、 新しい参照点を導入するような談話関係。
Instance(A, B)	「A、たとえば、B」のように、A の例を B が述べる談話関係。

表 2: 談話関係一覧

*4 Temp_rel(A,B) ≡
Precedence(A,B) $\vee$ Overlap(A,B) $\vee$ Subsumption(A,B)

本論文	SDRT	(Kaneko 2014a)
Alternation(A,B)	Alternation(A,B)	Alternation(A,B)
Consequence(A,B)	Consequence(A,B)	Consequence(A,B)
Elaboration(A,B) Instance(A, B)	Elaboration(A,B)	Elaboration(A,B)
Contrast(A,B)	Contrast(A,B)	Contrast(A,B)
Commentary(A,B)	Commentary(A,B)	Commentary(A,B)
Explanation(A,B)	Result(A,B)	Explanation(A,B)
Cause(A, B)	Explanation(A,B)	CAUSE(A, B)
Narration(A,B)	Narration(A,B)	Narration(A,B) $\wedge$ Precedence (A,B)
Introduction(A, B)	Narration(A, B)	Narration(A,B) $\wedge$ Temp_rel(A,B)*4
Addition(A, B) Parallel(A,B)	Parallel(A,B)	Narration(A,B) $\wedge$ Overlap (A,B) Parallel(A,B)
Background(A,B)	Background(A,B)	Narration(A,B) $\wedge$ Subsumption (A,B)

表 3: 先行研究と本論文の関係の対応

次に、本手法における談話関係をどのように特定するかを以下に示す。

- 手順 0：「しかし」「ところが」など、明らかに逆接の接続詞から始まる場合、Constant ラベルを付与する。また、「だから」「従って」「故に」など、明らかに因果表現の接続詞から始まる場合、Explanation を付与する。「例えば」「例として」などの、例を示す表現があれば、Instance を付与する。接続詞からでは判断しきれない場合、手順 1 の判断を行う。
- 手順 1：Conjunction($\wedge$) であるか、Disjunction($\vee$)（$\equiv$ Alternation） であるか、Conditional($\rightarrow$)（$\equiv$ Consequence）であるかを判断する。Conjunction($\wedge$) に該当する場合は、手順 2 に移る。
- 手順 2：Adversative(順接) であるか、Contrast(逆接) であるかを判断する。Adversative(順接) であれば、手順 3 の判断を行う。
- 手順 3：Elaboration, Explanation, Commentary, その他のうち、どのラベルに該当するかを判断する。
  - 一方のイベントがもう一方のイベントの一部を説明している場合、Elaboration ラベルを付与。
  - 談話の意味的に、接続詞が因果関係を表現している場合（例：「つまり」）、Explanation ラベルを付与する。
  - 後に来る談話ユニットが要約や補足説明となっている場合、Comentary ラベルを付与する。接続詞「つまり」はこちらに該当することもある。
  - その他に該当する場合は、手順 4 の判断を行う。
- 手順 4： 関係 (A, B) の A, B がイベントであるか、状態であるかを判断する。
  - A も B もイベントで、（ほぼ）同時に起こっていれば、Parrallel を付与する。
  - A がイベントまたは状態、B がイベントであり、A、B の順に成り立っていれば、Narration を付与する。
  - A が状態、B がイベントであり、A と B とで時間の参照点が切り替わっていれば、Introduction を付与する。
  - A がイベント、B が状態であれば、Background を付与する。
  - A も B も状態であれば、AND ラベルを付与する。

この一連の特定手順を決定木の形で示すと、以下の図 1 のようになる。

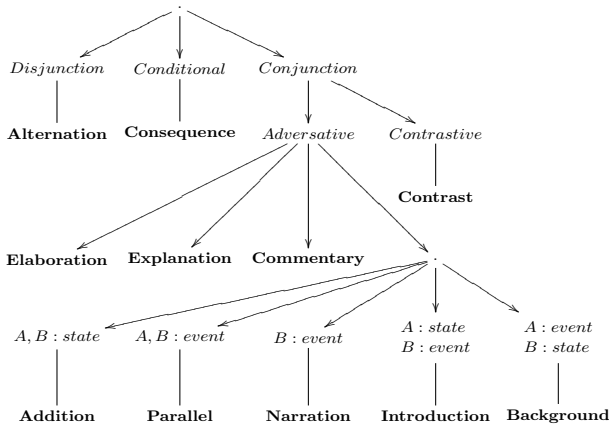

図 1: 談話関係の決定木

3. 分析と考察

2 節で述べた方法論を「現代日本語書き言葉均衡コーパス」(前川 2008) のデータの一部に適用した、因果関係認識日本語評価データ (Kaneko 2014b) に対して、エラー分析を行った。本節ではその分析結果を述べる。

1 つ目の問題として、アノテーターにみられる 2 つの対立する考え方と、その対立がもたらすアノテーションの不一致について述べる。アノテーションを行っていると、ガイドラインに指定されている手順が示す結果と、アノテーターが考えるガイドライン設計の意図が異なる場合があるが、前者にしたがうアノテーターと、後者にしたがうアノテーターが存在する。前者は、ガイドラインは客観的かつ厳密であるべきであると考えており、ガイドラインの指定と意図が異なるならば、ガイドラインが反証されたと捉える。これは、アノテーションガイドラインを一種の言語学の理論である、と捉える視点と関係が深い。後者は、ガイドラインを客観的もしくは厳密にすることは原理的に不可能であると考えており、アノテーターの主観的な判断結果を集めながら、そのような揺れのあるデータから取り出せる情報を取ろうとする。このように考えの異なるアノテーターが混在していた場合、必然的に不一致が生じることになる。

この問題の 1 つの例として、以下の文 π1 と π2 を取り上げる。このアノテーションでは、一連の特定手順により判断される談話関係と直感的に判断される談話関係が一致せず、「Narration(π1, π2)」「Background(π2, π1)」の間で揺れた。ガイドラインに従うと「忘れてくる」はイベントを指すが、このイベントは状態「忘れている」を伴うため、直感的に「状態を指す」と判断するアノテータが存在した。Narration はイベント間にのみ成立する談話関係であるため、それらのアノテータは Background を選択した。

π1. 定期券を家に忘れてきてしまった。
π2. いったん家に戻ることにした。

上記の場合のように、述語が表すイベントが起きた結果として生じる状態が存在する場合、アノテータがその述語を状態述語と取り違えるという場合は数多く見られた。こうした場合も判断が揺れないよう、イベント・状態の判定を自動的に行い、あらかじめアノテーションしておく等の処理を行う必要がある。一方で、イベントが起こった結果として、暗黙的に成り立つ状態があるという情報によって媒介される因果関係は数多く存在すると考えられるため、こうした情報を加味できるようにする意義はあると考えられる。

続いて、以下の文章を見てみる。ここでは、まず π3 と π4 の間の談話関係は、状態からイベントへの遷移であり、ガイドラインに従えば、Introduction(π3, π4) などが候補となる。

π3. パソコンの画面や本などに集中しながら、自分の入れた飲み物に手を伸ばし、飲み物にはまったく目を遣らないまま飲む、というのはだれでもやることだろう。
π4. 自分で入れたのだから、それがなんなのかは見なくてもわかる。
π5. だからたいがい、なんの問題もない。
π6. ところが、ごくごく稀に、変なことが起こる。
π7. たとえば、紅茶を入れたのに、どういうわけか、コーヒーを入れたと勘違いしてしまう。

しかし、π4 の表すイベントは、実は「飲む」というイベントに後続するものである。イベント「飲む」は、π3 のモダリティ [ダロウ] のスコープに含まれている形であるため、現在の談話関係の仕様では直接 π4 と談話関係を持つことはない。しかしながら、π4 の参照点（[ワカル] という状態が成立していることが主張されている時間的区間）は、「飲む」というイベントの直後に位置づけられるべきものである。この [ダロウ] のように、状態を表すモダリティのスコープ中にいくつかの埋め込み文が存在し、その中に現れるイベントや状態と、後続文が談話関係を持つ場合も考慮しなければならないと考えられる。

また、π5 は π3-π4 がもたらした generic な状況の帰結であるが、一方で、逆接の接続詞「ところが」に導かれる π6 は、その例外を述べている。したがって、π5 の時間軸と、π6-π7 の時間軸は異なっている。さらには、π6 の時間軸は π3-π4 の時間軸の部分でありながら、それ自体 generic なものであり、π7 はさらにその例示となっている。

このように、因果関係の有無を判断する上で、前の文（節）と後ろの文（節）が、同一の時間軸、時空間、状況で述べられているかどうかという情報が手がかりになり、談話関係がその役割を担っていると考えられる。したがって、それらの情報が談話関係から捉えられるようなラベル体系にするのが適切である。これらの情報は時間関係の判断にも役立つと予想される。

また一方で、どの時点で、イベントや状態が起こる可能性があり得て、どの時点からあり得なくなるのか（時間軸、時空間、状況が繋げられるか否か）を判断をする手がかりになるような、「完了」などの何らかの時間的な情報からのフィードバックを得ることで、談話関係や因果関係の有無の判断がなされる場合もあると推察されるため、その「何らかの時間的な情報」と談話関係や因果関係の影響関係を整理する必要がある。

ここまで述べてきた分析結果、問題点から、因果関係認識の枠組みは談話関係や因果関係、時間的関係やイベント・状態の情報などのうち、明らかになっている部分的な情報を元に、相互に推論でき、また、これらから複合的・多段的に判定できるような認識モデルを設計するのが望ましいと思われる。

4. まとめ

因果関係認識日本語評価データ (Kaneko 2014b) のガイドラインを用いて談話関係・因果関係のアノテーションを行う際に生じるアノテータ間不一致について分析を行った。特に、テキスト中の因果関係が実際はどうなっているのか、および、適切なリソースや認識モデルの選択や構築をどのように行うべきかを議論した。

参考文献

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

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

[乾 2006] 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)

[Kaneko 2014a] Kimi Kaneko, Daisuke Bekki：Building a Corpus of Temporal-Causal-Discourse Structures Based on SDRT for Extracting Causal Relations, EACL-2014 Workshop on Computational Approaches to Causality in Language, pp. 33-39. (2014)

[Kaneko 2014b] Kimi Kaneko and Daisuke Bekki：Toward a Discourse Theory for Annotating Causal Relations in Japanese, 28th Pacific Asia Conference on Language, Information and Computation, pp. 460-469. (2014)

[Riaz 2013] Mehwish Riaz 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)

[宇津木 2015] 宇津木 舞香, 稲田 和明, 金子 貴美, 戸次 大介, 乾 健太郎：「形式意味論に基づく出来事間関係認識に向けて リソース構築の展望とテンス「タ」のアノテーション」, 言語処理学会第 21 回年次大会, pp. 1036-1039. (2015)

[前川 2008] 前川喜久雄：KOTONOHA 『現代日本語書き言葉均衡コーパス』 の開発, 日本語の研究, Vol. 4, No. 1, pp. 82—95. (2008)