

日本語文のモダリティ・極性情報を捉えるために

佐尾ちとせ 江口萌 松吉俊 乾健太郎

奈良先端科学技術大学院大学 情報科学研究科

1. はじめに

一般に、文章に記述される情報は単純な命題（とそれに対する評価）のみではなく、そこには判断主体やその態度、否定表現の有無（極性）といった情報も記述される。文書集合から情報を抽出するにあたり、個々の部分に対してこれらを解析し、その解析結果に基づいて情報を整理することは重要である。しかしながら、このような解析技術はこれまで十分に研究されてきたとは言えない。我々は、日本語を対象に、文章の各部分の判断主体、態度、極性を解析する技術の開発に取り組んでいる。

日本語においては、判断主体の態度が多様な言語表現で記述されるため、その高精度な解析器を作成することは容易ではない。本論文では、この解析の基礎となる日本語モダリティタグ体系について提案するとともに、分析、学習、評価のためのコーパスを作成するにあたり、＜判断主体、モダリティ、極性＞をタグ付けする方法について説明する。

本論文は以下のように構成される。まず、2 節において、モダリティと極性の本研究での定義を述べる。次に、3 節で、先行研究について述べる。4 節において、個々の事象に対して、その判断主体、モダリティ、極性を記述する方法を説明し、続く 5 節で、提案するタグ体系に基づくコーパスとその構築時に生じた問題点について報告する。最後に、6 節で、今後の展望を述べる。

2. モダリティと極性

文には、命題に対する態度とは別に、判断主体の態度が示される。文(1)においては判断主体の〈願望〉が、文(2)においては判断主体の〈推量〉が表明されている(下線は態度の対象を指す)。

(1) あなたが行ってくるとありがたいんだけど。

(2) 明日は雨が降らないんじゃないかな。

本研究では、このような態度を対象部分の**モダリティ**と呼ぶ。日本語におけるモダリティは、主に、モダリティ表現の列によって表明される。

また、本研究では、形式的な「否定表現の有無」ではなく、実際に対象部分が肯定か否定かを表す指標として**極性**という用語を用いる。例えば、文(1)における対象部分の極性は、単純に〈肯定〉と判断するが、文(2)、文(3)における対象部分の極性は、いずれも〈否定〉と判断する。

(3) もっと早く起きられたら余裕で間に合ったのにな。

3. 先行研究

判断主体、態度、極性の解析に関する先行研究として、英語においては、Prasad らの研究[2]、Sauri らの研究[3]、TimeML に関する研究[4]などが挙げられる。

Prasad ら[2]は、Penn Discourse TreeBank (PDTB) に存在する談話関係とその項に対して、情報の発信源やその情報が事実であるかどうかを判断するための **attribution** という属性をタグ付けする枠組みを提案している。この **attribution** は、次の 4 つの要素からなる。

Source 情報の発信源もしくは判断主体を表し、文書の筆者、筆者を除く特定の他者、一般論、関係に対する値を継承のいずれかの値をとる

Type 対象を主張、信念、事実、成り行き of the どれかに分類し、対象が事実である度合いを推論するための動詞のクラス

Scopal Polarity スコープを考慮した否定の有無

Determinacy 上の 3 つの要素のいずれかが、より広い文脈で非決定的になりうるか

例として、文(4)に対する **attribution** タグを以下に示す。ここで、下線は談話関係を示唆する語列を、斜体は項 1 を、太字は項 2 を指す。

(4) John thinks *that Mary will get cured* not because she took the medication (but because she has started practising yoga).

	関係	項1	項2
Source	Ot (特定の他者)	Inh (継承)	Inh (継承)
Type	PAnt (信念)	Null	Null
Scopal Polarity	Neg (否定)	Null	Null
Determinacy	Null	Null	Null

彼らの **Type** は英語に特有の分類であり、他言語にそのまま転用することは不可能である。

Sauri ら[3]は、＜事実らしさに対する判断主体の確信度、否定の有無＞という対を事象の事実性と定義するとともに、Wiebe ら[5]が導入した「判断主体の入れ子構造」を **Source** として採用し、入力された依存木に対して、トップダウンに、各事象の事実性を **Source** ごとに解析するアルゴリズムを提案している。

TimeML[4]は、文章に存在する事象と時間表現、および、それらの間の関係を記述するための言語であり、この

体系においては、個々の事象例に対して、そのモダリティ表現を表す **modality** 属性(例えば, **modality**="SHOULD")と否定の有無を表す **polarity** 属性(例えば, **polarity**="NEG")を持つ<MAKEINSTANCE>要素が付与される。加えて、主節の事象が示唆する、従属事象の事実らしさ(**Modal**, **Factive**, **Counter-factive**, **Evidential**, **Negative evidential**, **Conditional**)が、<SLINK>要素によって表現される。TimeML では、判断主体に関する情報はタグ付けされない。

日本語においては、森田ら[1]が、事象の事実性を時間情報、極性、話者態度で捉えるモデルを構成し、機械学習に基づく事実性解析手法を提案している。彼らは、判断主体については考慮していない。

4. 判断主体、モダリティ、極性タグ体系

我々は、日本語を対象に、文章の各部分の判断主体、態度、極性を解析する技術の開発に取り組んでいる。本節では、この解析の基礎となる、我々が新たに提案する<判断主体、モダリティ、極性>のタグ体系について述べる。

4.1 タグ付与の対象

我々は、文章に存在するすべての**事象(行為、出来事、状態の総称)**をタグ付与の対象とする。具体的には、実際の文脈において事象を表現していると判断される体言と用言のすべてに対してタグを付与する。例えば、文(3)においては「起き」、「間に合っ」を、文(5)においては「購入」、「中止し」、「思う」¹をタグ付与の対象とする。

(5) 来週からこの雑誌の購入を中止しようと思う。

4.2 モダリティタグ

事象に対して、そのモダリティを付与する方法は、大きく2つに分類される。1つは、事象に関係するモダリティ表現や副詞などが固有に持っている「意味」をリスト形式で記述し、事象に対するモダリティをそれらから計算する方法である。もう1つは、関係する表現から総合的に判断し、事象のモダリティそのものを記述する方法である。英語においては、動詞の意味クラスや個々の助動詞の意味がモダリティの決定に大きく寄与し、かつ、それらが複合することは少ないことから、先行研究では、前者の方法が採用されている。

日本語には数多くのモダリティ表現が存在するため、日本語文においては、これらを複合的に用いて、同じモダリティを多様な言語形式で表現することが可能である。例えば、「降る」という事象に対する<推量>のモダリティは、以下のように、様々な言語形式で表現可能である。

「降るだろう」、「降ると思う」、「降るんじゃないかな」、「降るかもしれない」、「降る模様だ」、「降る見込みだ」これらの表現を構成するすべての要素に対して意味を記述し、複合した場合の意味を計算する規則集合を作成することは非常に困難である。このような理由により、我々は、上で述べた方法のうち、後者の方法を採用した。

事象にモダリティをタグ付けするにあたり、モダリティタグの体系を構築する必要がある。その際、情報抽出への応用を考慮して、以下の6つの分類項目を定め、論文末尾の表1に示したモダリティタグ体系を構築した。

- (a) **大まかな態度**：真偽を判断しているのか、価値を判断しているのか、そのいずれでもないのか
- (b) **モダリティ時間に対する相対的な事象時間**：事象は生じたのか、これからするのか
- (c) **事象の主体と判断主体の異同**：事象の主体が判断主体であり、判断主体自らが事象を制御することが可能であるか
- (d) **働きかけの程度**ならびに(e) **事象生起の確実性**：表現されているモダリティの程度
- (f) **仮想性**：条件付きかどうか、反実的かどうか(仮想的なら“+”とする)

表1の見出し行においては、これらを「態度」、「E/M」、「主体」、「程度」、「確実」、「仮想」と略記した。

4.3 極性タグ

2節で述べた定義に従い、事象に対して、<肯定>、<否定>、<不明>のいずれかを極性タグとして付与する。それぞれの定義は以下のとおりである。

	E/M	肯定	否定	不明
真偽	既	実現	非実現	
	未	実現の予想	非実現の予想	予想不明
価値	既	実現	非実現	実現不明
	未	実現の期待	非実現の期待	期待不明

タグ付与の例を表1の「例(極性)」の欄に示す。ここでは、上記の文字列を(肯)、(否)、(不)と略記した。

4.4 判断主体の記述

本研究では、Wiebeら[5]が導入した「判断主体の入れ子構造」を用いて、事象の判断主体(の列)を記述する。

4.5 否定・疑問の焦点の扱い

否定や疑問の焦点が、事象のモダリティや極性に影響を及ぼすことがある。例えば、文(6)～文(9)において、主節の事象「元気になる」の<モダリティ、極性>は、それぞれ、<断定、肯定>、<断定、否定>、<断定、肯定>、<推量過去、肯定>である(焦点を太字で示す)。

(6) 薬を飲んだから太郎が元気になるわけではない。(食

¹ 1つの事象を表現しているともモダリティ表現であるとも解釈できる語に対しては、判断のゆれを防ぐため、それらすべてにタグを付与することとする。

生活を改善したからだ.)

(7) 太郎はまだ治療中だから元気になったわけではない.

(8) 栄養を摂ったから太郎は元気になったのだろうか.

(9) 太郎は元気になったのだろうか.

現在, 我々は, 談話関係に焦点がある場合, 試験的に, 焦点の情報を主節の事象に記述している. 例えば, 文(6), 文(8)においては, 「原因否定(から, 飲ん)」、「原因推量(から, 摂っ)」と記述する. これらの情報によって, 焦点に関する修辞構造が陽に示される. Prasad ら[2]は, 文(4)のタグ付け結果に示されるように, 談話関係に否定の焦点がある場合を取り扱っているが, 疑問の焦点を対象としていない.

扱うべき現象の範囲 (例えば, 疑問副詞など) やそれらを体系的に記述する方法を検討していきたい.

4.6 タグ付与の具体例

文(10)に対する我々のタグを以下に示す. ここで, A, B, C は対象事象を, X はタグ付け対象外の用言を表す. また, “wr”は, 文の筆者を意味する.

(10) 恐ろしい_Xことに, 愛する_A旦那様は旦那様_Bで行く_Cつもりだったようだ.

	A	B	C
判断主体	wr	wr, 旦那様	wr, 旦那様
モダリティ	断定	意志	意志
極性	肯定	肯定	肯定

5. タグ付与コーパスの構築

5.1 コーパスの概要

我々は, 一般の Web 文書とブログを文抽出源として, 前節のタグ体系に基づくコーパスを構築している. 現時点における, これらのコーパスの規模とモダリティタグの分布を表 2 に示す. 判断主体は筆者かそれ以外かのみの記述にとどまっており, 「雨」のような, 事象性を持つ一般名詞はまだ作業対象としていない.

表 2: タグ付与コーパスの規模とモダリティタグの分布

文抽出源	Web 文書	ブログ記事
対象事象	1,753 (100%)	7,458 (100%)
断定	1,225 (70%)	5,186 (70%)
断定-反実	1 (0%)	1 (0%)
推量	99 (6%)	495 (7%)
推量-過去	6 (1%)	114 (1%)
推量-仮想	1 (0%)	59 (1%)
意志 or 予定	157 (9%)	417 (6%)
当為	0 (0%)	28 (0%)
働きかけ	90 (5%)	235 (3%)
欲求	21 (1%)	259 (3%)
問いかけ	23 (1%)	51 (1%)
仮定	130 (7%)	613 (8%)

5.2 タグ付与の問題点

タグ付与作業で生じた主要な問題点を以下に列挙する.

- **条件:** 「きのう甘いみかんを食べた」(断定)に対して, 「甘いみかんが食べたい」の「甘い」はみかんに対する陳述とはなっておらず, 名詞を限定する用法と捉え, 事象とは見なしていない. しかしながら, 「甘いみかんはよく売れる」における「甘い」は限定的用法であるものの, 総称的な仮定表現となっている. このように用言の限定的な用法が文の仮定条件を構成しているかどうかを判別することは難しいため, タグ付けの指針について, 十分な検討が必要である.
- **接続助詞の解釈:** 「1本飲むのに苦労する」における「飲む」のモダリティは, (断定) か (仮定) か
- **可能動詞の意志性の判断:** 「これ以上もう飲めません」の「飲め」のモダリティは, (断定) か (意志) か

6. おわりに

今後は, 4.5 節で述べた焦点, 総称名詞による条件, 判断が揺れやすいタグの適切な取り扱い等を検討し, 簡潔で整合性のあるタグ体系となるよう改良を加えていきたい. なお, 詳細なタグ付与指針と「現代日本語書き言葉均衡コーパス」²内の「Yahoo!知恵袋」に対するタグ付与結果を <http://cl.naist.jp/nltools/modality/>にて公開する予定である.

謝辞

本研究は, (独) 情報通信研究機構の委託研究「電気通信サービスにおける情報信頼性検証技術に関する研究開発」の一環として実施した. また, 本研究の一部は次の研究費の支援を受けている: 科研費若手研究(スタートアップ)「類義述語句同定のための語彙的知識の体系化と集積」(課題番号: 20800029, 代表: 松吉俊), 科研費特定領域研究「情報爆発時代に向けた新しいIT 基盤技術の研究」の公募研究「経験マイニング: Web 文書からの個人の経験の抽出と分類」(課題番号: 19024057, 代表: 乾健太郎).

参考文献

- [1] 森田啓, 佐尾ちとせ, 松吉俊, 松本裕治, 乾健太郎: テキスト情報の事実性解析, FIT2008 講演論文集第2 分冊, pp.259-260 (2008).
- [2] R. Prasad, N. Dinesh, A. Lee, A. Joshi and B. Webber: Annotating Attribution in the Penn Discourse TreeBank, Proceedings of the COLING/ACL Workshop on Sentiment and Subjectivity in Text, pp.31-38 (2006).
- [3] R. Saurí, and J. Pustejovsky: Determining Modality and Factuality for Text Entailment, Proceedings of ICSC2007, pp. 509-516 (2007).
- [4] R. Saurí, J. Littman, B. Knippen, R. Gaizauskas, A. Setzer, and J. Pustejovsky: TimeML Annotation Guidelines Version 1.2.1, http://www.timeml.org/site/publications/timeMLdocs/annguide_1.2.1.pdf (2006).
- [5] J. Wiebe, T. Wilson, and C. Cardie: Annotating Expressions of Opinions and Emotions in Language, *Language Resources and Evaluation*, 39 issue 2-3, pp. 165-210 (2005).

² <http://www.kokken.go.jp/kotonoha/>

表 1: 提案するモダリティタグ体系 (例において、対象とする事象に下線を引く)

態度	E/M	主体	程度	確実	仮想	例 (極性)	タグ
真偽判断	既			高	—	きのう学校を <u>休んだ</u> (肯) 去年 <u>スキーをしなかった</u> (否)	断定
					+	タベカラオケに行かなかったら [→タベはカラオケに行った] (肯) きのう8時に <u>起きていたら</u> [→きのうは8時に <u>起きなかった</u>] (否) 学校に <u>遅刻し</u> なかったはずだ [→学校に <u>遅刻した</u>] (肯) 学校に <u>間に合った</u> はずだ [→学校に <u>間に合わなかった</u>] (否)	断定-反実
				低	—	彼はもう <u>帰国した</u> のかもしれない (肯) 彼はそこになかったようだ (否) 試験に合格できたかどうかわからない (不)	推量-過去
					+	明日の6時からサッカーの試合がある (肯) 年内に解散総選挙は行われない (否)	予定
	未			高	—	雨が降ったら試合が <u>中止になる</u> (肯) 欠席が多ければ単位が <u>認定されない</u> (否)	予定-仮想
					+	彼女は来年結婚するらしい (肯) 明日は雨が降らないだろう (否) あのレストランがおいしいのかどうかわからない (不)	推量
				低	—	毎日勉強すれば合格できるかもしれない (肯) どんなに練習しても優勝はできないだろう (否)	推量-仮想
					+	あの子は双子なのでよく <u>似ている</u> のは当然だ (肯) こんなに見通しのよい交差点で事故が起こらないのは当然だ (否) 先週までにレポートを提出する必要があった (不)	当為
価値判断	既				—	あんなことを <u>言うべきではなかった</u> なあ [→あんなことを <u>言った</u>] (肯) やはり私が行くべきだった [→ <u>行かなかった</u>] (否)	断定-反実
					+	手を <u>上げろ</u> (肯) <u>動くな</u> (否)	働きかけ-命令
	未	異	↑↑↑ ↓↓↓ 弱		—	合格したかったらもっと勉強しなさい (肯) 命が惜しければ <u>動くな</u> (否)	働きかけ-命令-仮想
					+	結果はすぐに <u>連絡すべきだ</u> (肯) 明日 <u>言うべきではない</u> (否)	働きかけ-当為
					—	開封したら早めに <u>食べなければならぬ</u> (肯) 他人宛の封書なら勝手に <u>開封してはならぬ</u> (否)	働きかけ-当為-仮想
					+	ここに名前を <u>書いてください</u> (肯) このいすには <u>座らないでください</u> (否)	働きかけ-依頼
					—	申し込む場合はここに名前を <u>書いてください</u> (肯) チケットを持っていない方は <u>入らないでください</u> (否)	働きかけ-依頼-仮想
					+	あなたに行ってほしいんです (肯) <u>忘れないで</u> ほしいです (否)	働きかけ-願望
					—	時間があつたらあなたに行ってほしいんです (肯) 旧モデルならもっと <u>安く</u> なつてほしい (否)	働きかけ-願望-仮想
					+	交通ルールを <u>守ろう</u> (肯) これは彼女には <u>言わない</u> ことにしよう (否)	働きかけ-勧誘
					—	もしまだでしたら昼御飯をいっしょに <u>食べ</u> ましょう (肯) 持ち主がわからないなら <u>触らない</u> ことにしよう (否)	働きかけ-勧誘-仮想
		同		高	—	週末映画を見に行こうと思っている (肯) 新しい車は <u>買わない</u> ことにする (否) 進学しようか就職しようか迷っている (不)	意志
					+	雨が降ったら家の片づけをしよう (肯) 雨が降ったら <u>買い物に行く</u> のはやめよう (否)	意志-仮想
					—	海外旅行に行きたい (肯) 彼には <u>会いたくない</u> (否)	欲求
				低	—	お金があつたら海外旅行に行きたい (肯) 雨が降ったら学校に行きたくないなあ (否)	欲求-仮想
					+	来年結婚なさるんですか? (肯) 去年スキーに <u>行きませんでしたか</u> ? (否)	問いかけ
					—	時間があつたら (肯) 雨が降らなければ (否)	仮定