日本語文章の事象に対する判断情報アノテーション

江口 萌[†1] 松吉 俊[†1] 佐尾 ちとせ[†1]
乾 健太郎[†1] 松本 裕治[†1]

情報抽出や含意認識などの応用において、個々の事象に対して、その述語と項構造を認識するだけではなく、書き手が表明している態度や真偽判断、価値判断などの情報も解析し、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。本研究では、このような情報をモダリティ情報と呼ぶ。本論文では、事象のモダリティ情報を表す次の 7 つ組のタグの体系を提案する: 態度表明者、時制、仮想、態度、真偽判断、価値判断、焦点。このタグ体系に基づいて構築中である、約 4 万事象を対象としたコーパスの構築過程と現状についても報告する。

Annotating Modality and its Associated Information to Event Mentions in Japanese Text

Megumi Eguchi,[†1] Suguru Matsuyoshi,[†1]
Chitose Sao,[†1] Kentaro Inui[†1] and Yuji Matsumoto[†1]

Many NLP tasks including information extraction and textual entailment recognition involve the task of analyzing the modality status of each event mention in a given text on top of predicate-argument structure analysis. This paper reviews the literature of modality annotation and proposes a new annotation scheme to cover all the aspects of modality which are only partly treated in previous schemes. In our scheme, each event mention is annotated with seven slots: *Source*, *Tense*, *Assumptional*, *Modality type*, *Authenticity*, *Sentiment*, and *Focus*. The paper also reports on the present results of our manual annotation of a Japanese corpus consisting of about 40,000 event mentions, showing a reasonably high ratio of inter-annotator agreement.

1. はじめに

一般に、文章に記述される情報は、単純な命題のみではなく、そこには、命題に対する**情報発信者の主観的な態度**も記述される。例えば、次の文 (1a), (2a), (3a) からは、それぞれ、(1b), (2b), (3b) のような、書き手の態度を読み取ることができる。

(1) a. この夏、ぜひとも山口県に旅行に行きたい。
b. ある命題 (「この夏、私が山口県に旅行に行くコト」) が成立することを 望んでいる

(2) a. もう遅いから、きっと彼は先に帰ったんだろう。
b. ある命題 (「彼が先に帰るコト」) が成立したであろうことを 推量している

(3) a. 廊下を走らないでください。
b. ある命題 (「あなたが廊下を走るコト」) の成立の価値を否定的なものと判断し、受け手にそれを実行しないように 働きかける

命題に対するこのような態度は、言語学において**モダリティ**[6),14)] と呼ばれ、現在も、多くの研究者によって活発に研究が続けられている。

自然言語処理において、与えられた文章から情報を抽出するにあたり、個々の事象に対して、その述語と項構造を認識するだけではなく、書き手が表明している態度や真偽判断、価値判断などの情報も解析し、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。なぜならば、文章に記述されている事象が、実際に成立した事実であるのか、それとも、成立しなかったことであるのか、もしくは、書き手がその成立を望んでいるだけであるのか、といったことを自動的に認識することは、質問応答や含意認識などの応用に必須の技術の一つであるからである。この分野において、文章を解析するための技術は、古くから研究されており、これまでに様々な解析ツールが開発されてきた。例えば、形態素解析器や構文解析器は、その最も基礎的なものであり、現在、誰もが自由に利用することができるこれらの解析器[⋆1]が存在する。しかしながら、入力文章に存在する事象を特定し、その事象に対する書き手の態度や肯否極性などを高い精度で解析してくれる解析ツールは、現在のところ、

†1 奈良先端科学技術大学院大学
Nara Institute of Science and Technology

⋆1 例えば、形態素解析器 JUMAN (http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html)、ChaSen (http://chasen-legacy.sourceforge.jp/)、MeCab (http://mecab.sourceforge.net/)、構文解析器 KNP (http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html)、CaboCha (http://chasen.org/~taku/software/cabocha/) などが利用可能である。

利用可能ではない。その主な理由[⋆1]は、次の 2 つにあると我々は考える。

(i) 入力 (おそらく、構文木) に対して、どのようなマークアップを施して出力すればよいかが明らかではない

(ii) 解析ツールを開発するための基礎となるコーパスが利用可能ではない

上記の解析ツールを実現するためには、これらの問題を解決する必要がある。

問題 (i) を解く枠組みとして、次の 2 つの方法が考えられる。

A. 「たい」や「んだろう」、「きっと」のような、モダリティを示す表現と、「ない」や「不可能です」のような、否定を表す表現にその意味を付与する

B. ある事象に関して、周辺の文脈から読み取れる情報を統合し、事象そのものに態度や真偽判断などの情報を直接付与する

情報を付与した結果を、直接、情報抽出へ応用することを考えると、事象そのものに情報が付与されているほうが都合が良い。なぜならば、一般の情報抽出において、その抽出単位は、おそらく事象であり、そこにすべての情報が集約されていることが望まれるからである。方法 A. を採用する場合、事象に対する総合的な情報を得るためには、その周辺のモダリティ表現などに付与された情報を収集し、なんらかの計算を実行する必要がある。これは、そのような特別な計算機構を、情報抽出システムの開発者が独自に用意しなければならないことを意味する。情報抽出への応用しやすさを考慮し、我々は、上記の問題 (i) に対して方法 B. を採用する。

本研究では、書き手が表明する態度や真偽判断、価値判断などの、事象に対する総合的な情報を事象の**モダリティ情報**と呼ぶ。本論文において、情報抽出や含意認識などの応用を考慮した、モダリティ情報を表すタグの体系を提案する。加えて、このタグ体系に基づいて構築を進めているコーパスについて報告する。

本論文は、以下のように構成される。まず、2 章において、モダリティと否定、および、その周辺について概説し、これらの情報のマークアップ方法に関する先行研究について述べる。次に、3 章で、我々が提案する 7 つ組のタグの体系について説明する。4 章において、このタグ体系に基づくコーパスの構築過程と現状について報告する。最後に、5 章で全体をまとめる。

⋆1 その他の理由として、日本語では、事象に対する書き手の態度や事象の肯否が多様な言語表現で記述されるということが挙げられる。

2. 関 連 研 究

2.1 モダリティと否定、および、その周辺

前章で述べたように、モダリティは、文に存在する命題に対する情報発信者の主観的な態度を表す。言語学において、用語も含めて、統一した見解は存在しないようであるが、モダリティは、おおよそ、次のように分類される[6),14)]。

真偽判断 (or 認識) のモダリティ 断定か、推量かを表す

価値判断 (or 評価) のモダリティ 必要か、許可できるか、そうでないかを表す

表現類型 (or 発話類型) のモダリティ 叙述、意志、行為要求、勧誘、疑問、感嘆のいずれかの態度を表す

丁寧さのモダリティ 普通体か、丁寧体かを表す

伝達態度 (or 対話態度) のモダリティ 聞き手の存在に対する話し手の意識のありようを表す

説明のモダリティ 文と先行文脈との関係づけを表す

「表現類型のモダリティ」は、書き手の中心的な態度を、「真偽判断のモダリティ」は、命題の真偽に対する書き手の**確信度**を、「価値判断のモダリティ」は、命題成立を書き手が望んでいるか否かを表すと見ることができる。モダリティに関連する重要な周辺項目は、**態度表明者**である。態度表明者は、文字通り、モダリティの態度を表明している人物や団体を指す。多くの場合、態度表明者は文章の書き手であるが、文章に伝聞の表現や情報の出所を表す表現が用いられた場合、これを特定することは有用である。

文献[7)] の定義によると、事態の成立を表すことを肯定といい、事態の不成立を表すことを否定という。文章に否定表現が存在する場合、どの部分が否定されているのか、すなわち、どこが否定の**焦点**であるのかを知ることは、文章の意味を正確に捉える上で重要となる。

モダリティと否定の周辺項目のうち、記述された情報の信憑性を判断するための情報として、次のような項目が有用であると思われる。

仮想性 仮想世界の話であるのかどうか

時制 本質的に真偽が定まらない未来のことであるのかどうか

アスペクト 真偽が一方から他方へ変化するアスペクト (真偽アスペクト) を持っているかどうか

また、含意認識への応用を考慮すると、**推量や疑問の焦点**を特定することは重要である。なぜならば、一般に、推量、もしくは疑問の焦点になっている部分を除いた命題は、前提とし

て成立していることが含意されるからである。

モダリティや否定に関する情報を文章にマークアップする際には、何に応用するかなどの目的を踏まえつつ、上で挙げた項目を検討することになる。

2.2 マークアップ方法に関する先行研究

我々が事象のモダリティ情報と呼ぶ情報をマークアップするためのタグ体系に関する研究は、近年、主に英語や日本語を対象として進められており、純粋な自然言語処理分野の研究[1),8),9),11),12)] ばかりでなく、生物医学分野における研究[5),13)] も存在する。前節で挙げた項目のうち、本研究と先行研究がどれを対象としているかを表 1 にまとめた。

Rubin ら[9)] は、ニュース記事を対象に、態度などを表す表現に対して、態度表明者、時制、確信度、その表現の焦点が真偽判断と価値判断のどちらに当たっているのかを、4 つ組のタグとして記述している。

TimeML[11)] は、テキストに存在する事象、時間情報、そして、事象間の関係を表すためのマークアップ言語である。この体系においては、各事象に対して、時制を表す Tense 属性、アスペクトを表す Aspect 属性、モダリティ表現を表す Modality 属性、肯否極性を表す Polarity 属性などを持つ <MAKEINSTANCE> 要素が付与される。加えて、主節の事象が示唆する、従属事象の真偽判断情報が <SLINK> 要素によって表現される。TimeML の Modality 属性には、事象の核となる述語に接続する助動詞 (must, may, should など) をそのまま記述することになっているため、日本語など、述語の後にたいてい複数の助動詞が接続する言語に対してこの体系を直接適用することは難しい。また、態度表明者に関する情報はタグ付けされない。

Prasad ら[8)] は、Penn Discourse TreeBank (PDTB) に存在する談話関係とその項に対して、次の 4 種類の項目からなる attribution という属性タグを付与している。

<Source>: 情報の発信源または判断主体、<Type>: 対象の確信度を表す動詞のクラス、<Scopal Polarity>: スコープを考慮した否定の有無、<Determinacy>: これら 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

談話関係もタグ付与の対象であるため、以下のような、否定や疑問の焦点が談話関係にある場合に、それを自然に表現することができる。

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

ここで、彼らは、**because** の <Scopal Polarity> に “neg” を記述する。

FactBank[10),12)] というコーパスの構築において、Saurílらは、文に存在する各事象の態度表明者ごとに、< 事実らしさに対する態度表明者の確信度, 肯否極性 > をマークアップする枠組みを提案している。

川添ら[1)] は、日本語のニュース記事や報告書を対象に、モダリティに関連する表現、否定表現、仮定表現、およびそれらのスコープをマークアップする枠組みを提案している。川添らは、表現のスコープが二文以上に及ぶような場合を考慮しており、次のような例において、空要素タグを用いることにより、モダリティに関連する、省略された表現を補っている。

- 市の健康推進課 **によると**、感染したのは男性。市内の料理店で食事をした後症状を訴え、感染が確認された。

ここで、下線の文にはモダリティに関連する表現が存在しないが、その前文にある「によると」のスコープに入っているとして、空要素タグを補い、情報を記述する。

Medlock ら[5)] は、生物学分野のテキストを対象に、確信度に関係する表現の有無をマークアップするタグ体系を提案し、その体系に基づくコーパスを作成している。ここで、確信度に関係する表現が現れる文全体をスコープ内としている。

BioScope[13)] は、生物医学分野のテキストを対象に、各文に現れる否定表現、確信度に関係する表現、そしてそのスコープをマークアップしたコーパスである。BioScope は態度表明者、仮想性や価値判断に関する表現はマークアップの対象としていない。

3. 提案するタグ体系

3.1 タグ付与の対象

1 章で述べたように、本研究では、文章に存在する事象をタグ付与の対象とする。ここで、**事象**とは、行為、出来事、状態の総称であり、例えば、次の文においては、「雑誌を購入するコト」、「来週から購入を中止するコト」、「(略) と思うコト」が事象であり、すべてタグ付与の対象となる。

- 来週からこの雑誌の購入を中止しようと思う。

なお、文献[14)] に従い、事象に受動と使役のヴォイスまで含める。これにより、次の文 (1)、(2) に含まれる事象は、それぞれ、「あのケーキが友達に食べられるコト」、「今度、母親が太郎をお使いに行かせるコト」とする。

(1) あのケーキは友達に食べられた。
(2) 今度、母親が太郎をお使いに行かせるそうだ。

3.2 モダリティ情報タグ

本研究では、以下の 7 つの項目に対するタグを、事象の**モダリティ情報タグ**と呼ぶ。

表 1 本研究、および、先行研究で対象とする項目

	タグ付与対象	態度表明者	時制	仮想性	肯否極性	否定の焦点	確信度	推量の焦点	真偽アスペクト	表現類型	価値判断	事象の範囲
本研究	事象	○	○	○	○	○	○	○	○	○	○	×
Rubin ら[9]	語や句	○	○	×	×	×	○	×	×	○	×	×
TimeML[11]	事象	×	○	○	○	×	○	×	×	○	×	○
Prasad ら[8]	談話関係と事象	○	×	×	○	○	○	×	×	○	×	○
FactBank[10),12]	事象	○	○	○	○	×	○	×	×	○	×	○
川添ら[1]	語や句	○	×	○	○	×	○	×	×	×	×	×
Medlock ら[5]	語や句	×	×	×	×	×	○	×	×	×	×	×
BioScope[13]	語や句	×	×	×	○	×	○	×	×	×	×	×

態度表明者 対象とする事象の成否の判断や、他者への働きかけや問いかけをしている人物、もしくは、団体

時制 態度表明時から見た、対象事象の相対的な時制

仮想 仮定された条件の有無

態度 叙述、意志、働きかけ、問いかけなどの伝達的態度

真偽判断 態度表明者による対象事象の真偽判断

価値判断 態度表明者による対象事象の価値判断

焦点 対象事象に関する否定や疑問などの焦点

以下、この節では、上に挙げた各項目について、それを採用した理由を説明し、その項目に対するタグのリストを概観する[⋆1]。

3.2.1 態度表明者

情報検索や情報抽出などの応用においては、発信された情報の信憑性を判断する手がかりの一つとして、態度を表明する人物を特定することは重要である。本研究では、Wiebe ら[15]が導入した「態度表明者の入れ子構造」を用いて、事象に対する態度表明者 (の列) を記述する。タグ付与の例を原稿末の表 5 の ID=1∼6 に示す。“wr” タグは、対象事象に対する態度表明者が、文章の書き手 (writer) であることを表す。“wr_arb” タグは、対象事象に対する態度表明者が不特定の (arbitrary) 個人や集団であると、文章の書き手が述べていることを表す。“wr_ot” タグは、対象事象に対する態度表明者が、書き手以外の特定の (other) 個人や集団であるが、文章の書き手がその名称を明記していないことを表す。“wr_*STRING*” タグは、対象事象に対する態度表明者が「*STRING*」であると、文章の書き手が述べていることを表す。

3.2.2 時　　制

事象に対する真偽判断が推量として表現されている場合、その事象の時制を特定することは重要である。なぜならば、事象の時制を知ることができれば、態度表明者が行った推量を、次のように解釈することができるからである。

(1) もしその事象が未来のコトであるならば、その推量は、いまだ真偽が定まっていないことによる断定保留を表す

(2) もしその事象が未来のコトでないならば、その推量は、事象の真偽を確認していないことによる断定保留を表す

そこで、本研究では、事象の時制が未来であるのか (“未来”)、そうでないのか (“非未来”) を、モダリティ情報の 1 項目として記述する。上記の 2 つの区別に必要であるのは、書き手が情報を発信したときに対する相対的な時制ではなく、態度表明時に対する相対的な時制であるので、我々は、「態度表明時を基準とする時制」を記述する。タグ付与の例を原稿末の表 5 の ID=1∼38 に示す。“非未来” タグは、対象事象の相対的な時制が、過去、現在、脱時間的 (順接条件節の中など、時間軸上のどこにも位置づけられない場合[7]) のいずれかであることを表す。

3.2.3 仮　　想

情報抽出においては、文章に記述される情報が事実であるのか、それとも、単なる仮想的な話であるのかを区別することが求められる。そこで、事象の仮想性をモダリティ情報の 1 項目として記述する。タグ付与の例を原稿末の表 5 の ID=8,9,10 に示す。“条件” タグは、事象が条件として仮想的に述べられていることを、“帰結” タグは、事象が仮想的な条件の

⋆1 個々のタグの詳しい説明については、次の URL で一般公開している作業基準に関する資料を参照してほしい。
http://cl.naist.jp/nltools/modality/manual.pdf

表 2 ＜ 真偽判断 ＞ に対するタグと肯否極性の関係

	成立	高確率	0	低確率	不成立	
肯定 ⟵	肯定の断定	肯定の推量	詳細不明	否定の推量	否定の断定	⟶ 否定

帰結として述べられていることを、“0” タグは、上記 2 つのいずれでもないことを表す。

3.2.4 態　　度

「表現類型のモダリティ」は、態度表明者の中心的な態度を表す重要な項目であるので、これを ＜ 態度 ＞ に記述する。「表現類型のモダリティ」内の分類と ＜ 態度 ＞ の分類はおおよそ同じであるが、＜ 態度 ＞ には、「感嘆型」に相当するタグが存在しない。これは、我々の興味が、伝達的態度に基づく事象の分類にあることによる。文の表層的な形式によって認定される「感嘆型」を設定することは、我々にとってそれほど重要ではないと考える。＜ 態度 ＞ に対するタグは、以下の 8 種類である: **叙述、意志、欲求、働きかけ-直接、働きかけ-間接、働きかけ-勧誘、許可、問いかけ**。タグ付与の例を原稿末の表 5 の ID=17∼33 に示す。

“働きかけ-直接” タグは、態度表明者が、直接、相手に対して、行為の実行、もしくは、非実行を求めることを表す。それゆえに、この “働きかけ-直接” タグが付与されている場合、行為が実行されたかどうかはともかく、態度表明者が行為の実行 (もしくは、非実行) を望んでいることが相手に伝えられたことを含意する。一方、“働きかけ-間接” タグは、“働きかけ-直接” タグとは異なり、上記のような含意を示すとは限らない。

3.2.5 真 偽 判 断

「真偽判断のモダリティ」は、事象の真偽に対する態度表明者の確信度を表すため、応用に有用であるので、これを ＜ 真偽判断 ＞ に記述する。＜ 真偽判断 ＞ では、この「真偽判断のモダリティ」の分類とともに、事象の肯否極性と一部のアスペクト情報を表現する。＜ 真偽判断 ＞ に対するタグは、以下の 9 種類である: **成立、不成立、不成立から成立、成立から不成立、高確率、低確率、低確率から高確率、高確率から低確率、0**。事象の肯否極性に関する軸に沿って、真偽の変化を持たない 5 つのタグを並べると、表 2 のようになる。真偽の変化を含意する 4 つのタグは、これらの間の遷移を表す。タグ付与の例を原稿末の表 5 の ID=1∼38 に示す。

3.2.6 価 値 判 断

＜ 価値判断 ＞ では、言語学における「価値判断のモダリティ」の根幹にある、「事象成立の望ましさ」という概念を、極性情報[⋆1]として取り扱う。言語学の先行研究[6] において、上記のカテゴリーの表現は、基本的意味の面から、「必要」、「許可・許容」、「不必要」、「不許可・非許容」の 4 つに分類され、個々の使用場面において、「当為判断」、「働きかけ」、「後悔・不満」などの意味を帯びるとされる。本研究では、これらの意味を独自の体系により＜ 態度 ＞ に記述し、態度表明者が事象の成立を望ましいと判断しているのか、それとも、望ましくないと判断しているのかを、＜ 価値判断 ＞ に記述する。このような記法をとることにより、＜ 態度 ＞、＜ 真偽判断 ＞、＜ 価値判断 ＞ に関して、表現力があり、かつ、見やすいタグ体系の構築を目指している[⋆2]。タグ付与の例を原稿末の表 5 の ID=15∼30 に示す。

3.2.7 焦　　点

事象が “不成立” であるとき、実際に否定されているのは、事象を構成する一部の要素であり、文章が、暗に、同類の別の事象が成立していることを含意していることがある。例えば、次の文では、態度表明者は、「太郎が仕事で東京に行くコト」は不成立であると判断しているが、この文からは、「太郎が東京に行くコト」が成立していることが読み取れる。

- 太郎は東京に仕事で行ったのではない。

このような含意が存在することを表現するため、本コーパスでは、否定、推量、問いかけを対象として、その焦点を記述する。なお、対象事象内に焦点が存在せず、対象事象と別の事象の間の関係を示す接続表現などに焦点がある場合も、例外的に、その焦点を記述する。＜ 焦点 ＞ に対するタグは、以下の 7 種類である。

否定 (*FOCUS*)、否定 (*FOCUS*; *EVENT*)、推量 (*FOCUS*)、推量 (*FOCUS*; *EVENT*)、問いかけ (*FOCUS*)、問いかけ (*FOCUS*; *EVENT*)、0

タグ付与の例を原稿末の表 5 の ID=31∼38 に示す。ここで、「*FOCUS*」は焦点を、「*EVENT*」は、対象事象と別の事象の間の関係を示す接続表現などに焦点がある場合に、後者の事象の核となる述語を表す。

4. モダリティ情報タグ付与コーパス

前章で説明したタグ体系に基づいて現在構築中であるコーパスについて報告する。

⋆1 語句に対して設定される評価極性 (例えば、「給料が上がる」は評価極性ポジティブ、「事故に遭う」は評価極性ネガティブ) とは異なる。例えば、「あのとき事故に遭ってよかった。」における「あのとき事故に遭うコト」という事象に対する ＜ 価値判断 ＞ は、“ポジティブ” であると判断する。

⋆2 本研究では、1 つの事象に対するモダリティ情報にのみ着目するので、「説明のモダリティ」に関する情報は、タグ体系に含めていない。他のモダリティに比べ、自然言語処理において重要度が低いと思われるので、「ていねいさのモダリティ」と「伝達態度のモダリティ」に関する情報も、タグ体系に含めていない。

4.1 構　　築

コーパス構築にあたり、本研究では、次の 4 種類のテキストを対象とした。

(a) ブログ記事 (20,000 事象/5,687 文)
(b) 数種類のトピックに基づき収集された Web 文書 (4,858 事象/4,858 文)
(c) 言明間意味的関係コーパス[3] における文リスト (14,402 事象/2,878 文)
(d) KOTONOHA プロジェクト[⋆1]および文科省科学研究費特定領域研究「日本語コーパス」プロジェクト[⋆2]が共同で構築している “現代日本語書き言葉均衡コーパス” の一部である「Yahoo!知恵袋」(19,920 事象/6,362 文)

これらのテキストを形態素解析器 Mecab と構文解析器 Cabocha で解析し、各文に含まれる述語を抽出した。本来ならば、事象の範囲を特定して、それを明確にマークアップするべきではあるが、現在のところ、これを高い精度で自動的に実現することは困難であり、そのほとんどを人手で行うとすると、かなりのコストがかかる。文章において、ほとんどすべての事象は、それぞれ、1 つの述語を核として表現されるため、その述語を事象の代表形態素 (列) として用いることができると考えられる。そこで、本コーパスでは、事象の範囲を明確にマークアップすることはせず、述語に対してタグを付与することで、その述語を核として持つ事象にそのタグを付与したと見なす。

タグ付与作業は、主に、1 人の作業者が行っている。作業者には、事象が含まれる文全体と、事象の核となる述語の位置以外に、述語を含む文節とその前後の文節における形態素の列を、表層形の列、および、基本形の列で提示する。もし誤って事象の核となる述語として抽出されてしまった補助動詞や、機能語相当表現[4] の一部などがあった場合、モダリティ情報タグ付与の対象から除外する。タグ付与にかかる時間は、1,000 事象あたりおよそ 5 時間である。論文執筆時点における、上記の (a)、(b)、(c) のコーパス内のタグの分布を表 3 に示す[⋆3]。ここでは、紙面の都合上、タグ内に自由な文字列を含むものは 1 つの欄にまとめた。構築作業は現在進行中であり、ここで示した値は変動する可能性がある。

現実の事象に対するタグ付与という観点から、我々のタグ体系を評価するため、作業者間の一致度を算出した。タグ付与済みコーパスからランダムに 300 事象を抽出し、上記の作業者とは別の作業者にタグ付与作業を依頼した。一致度の指標である κ 統計量は、はじめ 7 つの項目に対する平均で 0.4 程度であったが、作業者間でタグ付与基準を確認することに

⋆1 http://www.kokken.go.jp/kotonoha/
⋆2 http://www.tokuteicorpus.jp/
⋆3 (d) のコーパスにはまだ着手していない。

表 3 構築中のコーパスにおける、モダリティ情報タグの分布

対象テキスト		(a) ブログ記事	(b) 一般 Web 文書	(c) 村上らのコーパス[3]
タグ付け対象事象数		19,259(100%)	4,428(100%)	13,674(100%)
態度表明者	wr	19,188(100%)	4,421(100%)	13,524(99%)
	wr_arb	66(0%)	7(0%)	72(1%)
	wr_*STRING*	5(0%)	0(0%)	78(1%)
時制	非未来	17,071(89%)	3,956(89%)	11,775(86%)
	未来	2,188(11%)	472(11%)	1,899(14%)
仮想	条件	646(3%)	130(3%)	476(3%)
	帰結	56(0%)	2(0%)	101(1%)
	0	18,557(96%)	4,296(97%)	13,097(96%)
態度	叙述	18,303(95%)	4,202(95%)	13,060(96%)
	意志	394(2%)	89(2%)	244(2%)
	欲求	261(1%)	21(0%)	51(0%)
	働きかけ-直接	85(0%)	23(1%)	22(0%)
	働きかけ-間接	131(1%)	53(1%)	218(2%)
	働きかけ-勧誘	26(0%)	15(0%)	18(0%)
	許可	3(0%)	0(0%)	7(0%)
	問いかけ	56(0%)	25(1%)	54(0%)
真偽判断	成立	16,627(86%)	3,990(90%)	12,044(88%)
	不成立	988(5%)	128(3%)	641(5%)
	不成立から成立	0(0%)	0(0%)	13(0%)
	成立から不成立	0(0%)	0(0%)	10(0%)
	高確率	981(5%)	142(3%)	518(4%)
	低確率	84(0%)	33(1%)	56(0%)
	低確率から高確率	0(0%)	0(0%)	9(0%)
	高確率から低確率	0(0%)	0(0%)	9(0%)
	0	579(3%)	135(3%)	374(3%)
価値判断	ポジティブ	860(4%)	157(4%)	522(4%)
	ネガティブ	61(0%)	45(1%)	69(1%)
	0	18,338(95%)	4,226(95%)	13,083(96%)
焦点	焦点あり	30(0%)	25(0%)	68(0%)
	0	19,229(100%)	4,403(100%)	13,606(100%)

表 4 作業者間の一致度 (κ 統計量)

態度表明者	時制	仮想	態度	真偽判断	価値判断	焦点	左記の平均	7 つ組全体
0.69	0.76	0.68	0.66	0.70	0.72	0.75	**0.71**	0.58

より、現在は、0.71 という高い値となった (表 4 参照)。文献[2] によると、この値は良い一致度を示唆しているので、我々の体系は実際のタグ付与という観点からも良いものであると言える。

4.2 現在のタグ体系の問題点

コーパス構築作業において明らかになった、我々のタグ体系の問題点について述べる。

4.2.1 動詞の可能形

文脈により、“許可” タグを付与することもあるが、現在の体系では、動詞の可能形 (例えば、「使える」、「買える」、「使うことができる」、「買うことができる」) を核とする事象に対して、動詞の基本形 (例えば、「使う」、「買う」) を核とする事象と同様のモダリティ情報タグを付与している。

4.2.2 補 助 用 言

文章に補助用言「やすい」、「にくい」、「がたい」、「づらい」が存在する場合、現在の体系では、これらが意味するところを適切に表現できない。暫定的に、補助用言「やすい」が存在する場合、主に、< 時制 > に “未来” タグを、< 真偽判断 > に “高確率” タグを付与している。補助用言「にくい」、「がたい」、「づらい」が存在する場合、主に、< 時制 > に “未来” タグを、< 真偽判断 > に “低確率” タグを付与している。

5. お わ り に

本論文では、書き手が表明する態度や真偽判断、価値判断などの、事象に対する総合的な情報を表すタグの体系を提案し、このタグ体系に基づいて構築を進めているコーパスについて報告した。

我々の最終目標は、入力文章に存在する事象を特定し、その事象に対するモダリティ情報を高い精度で解析する解析ツールを実装することである。今後、構築したコーパスを学習コーパスとして用いることにより、現在試作している解析ツールの精度向上に努めたい。

タグ付与の作業基準に関する資料やモダリティ情報解析に関する最新情報を、次のサイトで公開中である。なお、構築したコーパスは、同場所において公開する予定である。

`http://cl.naist.jp/nltools/modality/`

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憑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)。

参 考 文 献

1) 川添愛, 齊藤学, 片岡喜代子, 戸次大介. 確実性判断に関わる意味的文脈アノテーション. 情報処理学会研究報告書, 2009-FI-93, 2009-NL-189, pp. 77–84, 2009.
2) J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, Vol.33, No.1, pp. 159–174, 1977.
3) 村上浩司, 増田祥子, 松吉俊, 乾健太郎, 松本裕治. 言明間の意味的関係の体系化とコーパス構築. 言語処理学会第 15 回年次大会発表論文集, pp. 602–605, 2009.
4) 国際交流基金, 財団法人日本国際教育協会（編）. 日本語能力試験出題基準【改訂版】. 凡人社, 2002.
5) Ben Medlock and Ted Briscoe. Weakly supervised learning for hedge classification in scientific literature. In *the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 992–999, 2007.
6) 日本語記述文法研究会（編）. 現代日本語文法 4. くろしお出版, 2003.
7) 日本語記述文法研究会（編）. 現代日本語文法 3. くろしお出版, 2007.
8) Rashmi Prasad, Nikhil Dinesh, Alan Lee, Aravind Joshi, and Bonnie Webber. Annotating attribution in the penn discourse treebank. In *the COLING/ACL Workshop on Sentiment and Subjectivity in Text*, pp. 31–38, 2006.
9) Victoria Rubin, Elizabeth Liddy, and Noriko Kando. *Chapter 7: Certainty Identification in Texts: Categorization Model and Manual Tagging Result*, pp. 61–74. Springer-Verlag New York, 2005.
10) Roser Saurí. *FactBank 1.0 Annotation Guidelines*. `http://www.cs.brandeis.edu/~roser/pubs/fb_annotGuidelines.pdf`, 2008.
11) Roser Saurí, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky. *TimeML Annotation Guidelines Version 1.2.1*. `http://www.timeml.org/site/publications/timeMLdocs/annguide_1.2.1.pdf`, 2006.
12) Roser Saurí and James Pustejovsky. Factbank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 2009.
13) György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *the Workshop on Current Trends in Biomedical Natural Language Processing*, pp. 38–45, 2008.
14) 益岡隆志. 日本語モダリティ探究. くろしお出版, 2007.
15) Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. In *Language Resources and Evaluation 39 issue 2-3*, pp. 165–210, 2005.

表 5 本コーパスにおけるタグ付与例

ID	文（下線が事象の核となる述語）	態度表明者	時制	仮想	態度	真偽判断	価値判断	焦点
1	九州には阿蘇山の雄大な景色が あり ますし、温泉も豊富です。	wr	非未来	0	叙述	成立	0	0
2	やっぱり騒がしいのは性に 合わ ないから、今度は温泉とか行ってのんびりする ···	wr	非未来	0	叙述	不成立	0	0
3	投薬はどうやら三半規管の機能を安定させるために 続ける のだそうだ。	wr_arb	非未来	0	叙述	成立	0	0
4	最初は脱ステの症状に 苦しん だものの、··· 今ではもうステロイド剤は使われていないそうです。	wr_ot	非未来	0	叙述	成立	0	0
5	··· 食べた後にはキシリトールガムを噛むことで虫歯予防にも なる とリーッタは言います。	wr_リーッタ	非未来	0	叙述	成立	0	0
6	早く家に 帰り たいと太郎が言っていたよ、と二郎が言った。	wr_二郎_太郎	未来	0	欲求	0	ポジティブ	0
7	··· この水には病のもと・活性酸素を消す還元力が ある のではないか。	wr	非未来	0	叙述	高確率	0	0
8	吸入ステロイド剤でも急性副腎不全が起きることを 知ら なければ、医師も診断できないことがある ···	wr	非未来	条件	叙述	不成立	0	0
9	吸入ステロイド剤でも急性副腎不全が起きることを知らなければ、医師も 診断でき ないことがある ···	wr	非未来	帰結	叙述	低確率	0	0
10	··· ほとんどの皮膚科の医師は、ステロイド剤を長期にわたって使用した経験など 無い でしょう。	wr	非未来	0	叙述	高確率	0	0
11	クローン技術やナノテクノロジーの発達で移植可能な人工臓器が 出現する のは時間の問題だと思う。	wr	未来	0	叙述	高確率	0	0
12	このような場合、副作用の少ない薬剤では十分な効果が 得 られないことが予測されるため、多少の ···	wr	未来	0	叙述	低確率	0	0
13	副腎皮質ホルモンから作られるステロイド剤の劇的な効果に医師達が驚き、競って 使い 始めたのです。	wr	非未来	0	叙述	不成立から成立	0	0
14	ステロイド剤を使用している場合は、急激に 使用 を中止するとリバウンド現象が起こり、症状が ···	wr	非未来	条件	叙述	成立から不成立	0	0
15	花子からのメールで、彼が京都に 来る ことを確信した。	wr	未来	0	叙述	低確率から高確率	0	0
16	治療は病状の 進行 をおさえるためにステロイド剤、免疫抑制剤の投与が行われています。	wr	未来	0	叙述	高確率から低確率	0	0
17	··· 現在、クローン技術を応用して臓器を 供給する 試みを否定することは困難であろう。	wr	未来	0	意志	高確率	ポジティブ	0
18	化膿を防ぐために大切なのは、入浴した時に患部を 濡ら さないように気をつけることです。	wr	未来	0	意志	低確率	ネガティブ	0
19	新宿行って、なんか食べに 行き たい。	wr	未来	0	欲求	0	ポジティブ	0
20	旅立ちと出会いの季節、いろんなことがあるけると、後悔し ないように生きたい。	wr	未来	0	欲求	0	ネガティブ	0
21	ぜひ一度、「還元水」の持つ効果を 実感し てみてください！	wr	未来	0	働きかけ-直接	0	ポジティブ	0
22	戦争で死ぬのはいつも兵士より民間人市民が多いことは 忘れ ちゃいかん。	wr	未来	0	働きかけ-直接	0	ネガティブ	0
23	日本人フランス人問わずみんなに 飲ん でもらいたい。	wr	未来	0	働きかけ-間接	0	ポジティブ	0
24	··· 何度も挑戦することの出来る根気がない方は、手をださ ない方が無難です。	wr	未来	0	働きかけ-間接	0	ネガティブ	0
25	手打ち蕎麦食べに 行き ませんか？	wr	未来	0	働きかけ-勧誘	0	ポジティブ	0
26	僕らは絶対に 近づか ないことにしよう。	wr	未来	0	働きかけ-勧誘	0	ネガティブ	0
27	クローン技術の適用についてはその危険性の観点から規制を 加える ことが許されるであろう。	wr	未来	0	許可	0	ポジティブ	0
28	メーカーとしても、犬に効果があるか分からないキシリトールを多く いれる こともないと思います。	wr	非未来	0	許可	成立	ネガティブ	0
29	あなたはその時に彼女に真実を 伝える べきだった。	wr	非未来	0	叙述	不成立	ポジティブ	0
30	彼が来ると知っていたら、行か なかったのに。	wr	非未来	0	叙述	成立	ネガティブ	0
31	それでは、キシリトールにはどんな効果が ある のでしょうか	wr	非未来	0	問いかけ	0	0	問いかけ (どんな)
32	クローン技術について考え、この技術をどう 受け入れ ていくか、考えなければならないと思います。	wr	未来	0	問いかけ	0	0	問いかけ (どう)
33	クローン技術もここまで き たか、というところでしょうか。	wr	非未来	0	問いかけ	高確率	0	0
34	太郎は、栄養を摂ったから、元気に なっ たのですか？	wr	非未来	0	叙述	成立	0	問いかけ (から; 摂っ)
35	彼は東京に仕事で 行っ たわけではない。	wr	非未来	0	叙述	成立	0	否定 (仕事で)
36	薬を飲んだから元気に なっ たわけではない。	wr	非未来	0	叙述	成立	0	否定 (から; 飲ん)
37	おそらく生後 2ヵ月前位からステロイド剤の 使用 をしてたと思われますが ···	wr	非未来	0	叙述	高確率	0	推量 (生後 2ヵ月前位から)
38	吸入ステロイド剤は更に効果的に吸入できるので、それでこんなに早く効果が 出 たのではないか。	wr	非未来	0	叙述	成立	0	推量 (ので; 吸入できる)