拡張モダリティタグ付与コーパスの設計と構築

松吉 俊[†]　佐尾 ちとせ[†]　乾 健太郎[‡,†]　松本 裕治[†]
[†]奈良先端科学技術大学院大学　[‡]東北大学
{matuyosi, chitose-s, matsu}@is.naist.jp, inui@ecei.tohoku.ac.jp

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 (1), (2), (3) からは、それぞれその次に記述したような書き手の態度を読み取ることができる。

(1) この夏、ぜひとも九州に旅行に行きたい。
→ ある命題 (「この夏、私が九州に旅行に行くコト」) が成立することを 望んでいる

(2) もう遅いから、きっと彼は先に帰ったんだろう。
→ 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) 廊下を走らないでください。
→ ある命題 (「あなたが廊下を走るコト」) が成立することを否定的に評価し、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**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。文章に表現されるモダリティを解析する技術は、情報抽出や含意認識など、自然言語処理の応用に有用ではあるが、現在のところ、高い精度でこれを実現するシステムは利用可能ではない。

我々は、モダリティとその周辺情報を整理した**拡張モダリティ**の体系を独自に設計し、この体系に基づくタグ付与コーパスの構築を開始した [16]。本論文では、このコーパスの設計方針と現状、および、構築時に直面した問題とその対応について述べる。我々は、モダリティ解析の精度向上に必要な技術や言語資源について理解を深めるため、このコーパスと最大エントロピーモデルに基づくモダリティ解析システムを実装し、その誤り分析を行った。本論文では、この誤り分析の結果についても報告する。

2 関連研究

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 [14, 5]。

真偽判断のモダリティ 断定か、推量かを表す
価値判断のモダリティ 必要か、許可できるかを表す
表現類型のモダリティ 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 先行文脈との関係づけを表す

我々の拡張モダリティは、自然言語処理において特に重要であると思われる、真偽判断、価値判断、表現類型のモダリティを含む。

モダリティとその周辺情報をマークアップするための体系、および、その解析手法に関する研究は、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 [7, 9, 6, 10, 8, 11, 2, 1] だけでなく、生物医学分野における研究 [3, 4, 13] も存在する。

マークアップ体系やコーパス構築に関する重要な先行研究は、Saurílらによる FactBank[8, 11] である。Saurílらは、事象とその時制、肯否、モダリティをマークアップする TimeML[9] の体系の上に、事象を対象として、態度表明者 (source)[15] ごとに、事実らしさに対する態度表明者の確信度と独自の肯否極性をマークアップする枠組みを提案している。モダリティに関する TimeML のマークアップは、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。

3 拡張モダリティ

3.1 事象

本研究の対象は、文章に存在するすべての事象のモダリティである。ここで、**事象**とは、行為、出来事、状態の総称である。本研究では、文献 [14] に従い、事象にヴォイスを含めるが、使役においては、ガ格が使役者の事象とガ格が被使役者の事象を分けて認識する。

表 1: 拡張モダリティの項目とラベル、および、コーパスにおける現在の分布

		Yahoo!知恵袋 (OC)	白書 (OW)	新聞 (PN)	書籍 (PB)
文数		6,404	5,835	16,433	9,869
形態素数		110,649	228,651	360,814	234,540
事象候補数		31,528(–%)	78,596(–%)	103,824(–%)	67,521(–%)
事象候補数 (タグ付与済み)		26,592(–%)	22,497(–%)	13,561(–%)	16,385(–%)
事象数 (タグ付与済み)		14,089(100%)	7,733(100%)	8,819(100%)	9,466(100%)
項目	**ラベル**				
態度表明者	wr:筆者	13,757(98%)	7,320(95%)	8,149(93%)	8,155(86%)
	wr:筆者_arb:不特定	112(1%)	88(1%)	33(0%)	86(1%)
	(その他)	220(1%)	325(4%)	637(7%)	1,225(13%)
相対時	非未来	11,972(85%)	6,214(80%)	7,726(88%)	8,164(86%)
	未来	2,117(15%)	1,519(20%)	1,093(12%)	1,302(14%)
仮想	0	12,445(88%)	7,348(95%)	8,484(96%)	8,388(88%)
	条件	1,167(8%)	290(4%)	242(3%)	724(8%)
	帰結	477(4%)	95(1%)	93(1%)	354(4%)
態度	叙述	11,146(79%)	6,440(83%)	7,923(90%)	8,236(87%)
	意志	314(2%)	754(10%)	280(3%)	394(4%)
	欲求	293(2%)	44(1%)	180(2%)	150(2%)
	働きかけ-直接	496(4%)	40(1%)	41(1%)	85(1%)
	働きかけ-間接	458(3%)	385(5%)	268(3%)	236(3%)
	働きかけ-勧誘	13(0%)	0(0%)	1(0%)	20(0%)
	許可	28(0%)	35(0%)	27(0%)	29(0%)
	問いかけ	1,341(10%)	35(0%)	99(1%)	316(3%)
真偽判断	成立	9,192(65%)	5,672(73%)	6,888(78%)	6,600(70%)
	不成立	985(7%)	188(3%)	671(8%)	919(10%)
	不成立から成立	74(1%)	18(0%)	11(0%)	58(1%)
	成立から不成立	34(0%)	7(0%)	3(0%)	31(0%)
	高確率	874(6%)	930(12%)	508(6%)	804(8%)
	低確率	143(1%)	72(1%)	88(1%)	154(2%)
	低確率から高確率	18(0%)	83(1%)	22(0%)	20(0%)
	高確率から低確率	11(0%)	18(0%)	6(0%)	4(0%)
	0	2,758(20%)	745(10%)	622(7%)	876(9%)
価値判断	0	12,337(88%)	6,458(84%)	8,014(91%)	8,465(89%)
	ポジティブ	1,462(10%)	1,196(15%)	685(8%)	818(9%)
	ネガティブ	290(2%)	79(1%)	120(1%)	183(2%)

本研究における事象の例を以下に示す。文 (4) においては、「あの本が無理であるコト」、「僕が『少しずつ学ぶ量子力学』を読めるコト」(可能態)、「『少しずつ学ぶ量子力学』を貸すコト」が事象であり、文 (5) においては、「成績が良いコト」、「母親が太郎を塾に行かせるコト」(使役態)、「太郎が塾に行くコト」が事象である。

(4) あの本が無理なら、僕でも読めた『少しずつ学ぶ量子力学』を貸してあげます。

(5) 先生によると、期末試験の成績が良くなかったので、母親が太郎を塾に行かせたそうだ。

3.2 拡張モダリティの項目とラベル

我々は、次の 6 項目からなる、事象の**拡張モダリティ**を設計した: <**態度表明者**>、<**相対時**>、<**仮想**>、<**態度**>、<**真偽判断**>、<**価値判断**>。それぞれの項目に対するラベルの一覧を表 1 に示す[1]。< 態度 >、< 真偽判断 >、< 価値判断 > の組が、2 章で述べた、表現類型、真偽判断、価値判断のモダリティにほぼ相当する。ただし、FactBank[11] と同様に、< 真偽判断 > は肯否極性の情報も含む。残り 3 つの項目は、事象の事実性をより明確に記述するために導入したものである。< 態度表明者 > は、「態度表明者の入れ子構造」[15] により、態度を表明する人物や情報源を表す。< 相対時 > は、真偽が定まっていない未来のことかどうかを態度表明時に対する相対的な時間関係で表す。< 仮想 > は、条件節の中など、仮想的な事象であるかどうかを表す。

例として、文 (4), (5) の事象のうち、次の 4 つの事象に対する拡張モダリティ“< 態度表明者 >, ···, < 価値判断 >” を示す。

「僕が『少しずつ学ぶ量子力学』を読めるコト」
→ “wr:筆者, 非未来,0, 叙述, 成立,0”

[1]個々のラベルに関する詳しい説明は、次の URL で公開している作業基準マニュアルを参照してほしい。
http://www.cl.ecei.tohoku.ac.jp/resources/modality/manual.pdf

「『少しずつ学ぶ量子力学』を貸すコト」
→ “wr:筆者, 未来, 帰結, 意志, 高確率, ポジティブ”
「成績が良いコト」
→ “wr:筆者_1:先生, 非未来,0, 叙述, 不成立,0”
「母親が太郎を塾に行かせるコト」
→ “wr:筆者_1:先生, 非未来,0, 叙述, 成立,0”

4 拡張モダリティタグ付与コーパス

4.1 設計方針

前章で説明した拡張モダリティの情報を、拡張モダリティタグとして文内の事象に付与する。本研究では、タグ付与対象のテキストとして、現代日本語書き言葉均衡コーパス (BCCWJ)[2]を利用した。BCCWJ を選択した理由は、BCCWJ は著作権処理の済んだデータであり、タグ付与結果を自由に公開でき、それを他の研究者と共有することができるからである[3]。BCCWJ 内の 4 ジャンル (Yahoo!知恵袋 (OC)、白書 (OW)、新聞 (PN)、書籍 (PB)) における文数と形態素数を表 1 の上部に示す。

文において、ほとんどすべての事象は、1 つの述語を核として表現されるので、拡張モダリティタグ付与コーパスでは、そのような事象のみを対象とする。ただし、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。このようにした理由は、現在のところ、述語が与えられた時に高い精度で事象の範囲を自動的に特定することは困難であり、その作業を人手で行うとすると、かなりのコストがかかるからである。本研究では、述語を表す品詞として、主に、動詞、形容詞、形状詞、名詞-普通名詞-サ変可能/形状詞可能を用いた。さらに、網羅性を重視して名詞述語を抽出するため、後続形態素列に基づく抽出規則を作成して用いた。これらの品詞や規則により抽出されるのは事象 候補 の述語のリストであり、その中には、事象の述語だけでなく、文 (4) の「(て) あげ」のような補助動詞や、文 (5) の「(に) よる (と)」のような複合辞の一部、名詞述語でない名詞などが含まれる。本コーパスでは、これらに対して「対象外」という補足欄に “機能表現” や “名詞” などを記述し、拡張モダリティの情報を付与しない。

文に、否定や問いかけの焦点、または、程度や頻度を表す表現が存在する場合、これらは拡張モダリティを判断する際に重要な情報であるので、それぞれ、補足欄「焦点」と「程度」にその情報を記述する。

[2] http://www.tokuteicorpus.jp/

[3] 我々が構築したコーパスは、BCCWJ との差分データに変換し、次の URL で 2011 年 3 月に公開予定である。
http://www.cl.ecei.tohoku.ac.jp/resources/modality/

4.2 現状

本コーパスにおける事象数、および、現在のラベルの分布を表 1 に示す。タグ付与作業は、主に 1 人の作業者が行っている。本論文執筆時点において、タグ付与済み事象数は 40,107 であり、このうち、OC の 14,089 事象に対しては、実装した解析システムの解析結果をフィードバックさせ、それを参照しながらのタグ見直し作業を数回行い、タグの質を向上させている。

表 1 から、それぞれの項目において、全体の 70%〜90%の事例を占めるラベルが存在することが分かるが、自然言語処理の応用においては、残りの 10%〜30%の事例に関して、そのラベルを正確に判定することが重要となる。

4.3 直面した問題とその対応

事象と見なすかどうかに関して次の 2 つの問題に直面した。それぞれ、以下で述べるように対応した[4]。

限定修飾 これまで限定修飾の事例は対象外としていたが、文 (6) のようにタグを付与すべき事例が見つかった。含意認識などの応用を考慮し、補足欄「対象外」に “限定修飾” と記述しつつ、拡張モダリティの情報も付与することにした。

(6) あなたが卒業した 小学校はどこですか?
「あなたが小学校を卒業するコト」
→ “wr:筆者, 非未来,0, 叙述, 成立,0”

一部の機能表現 前接の述語が表す事象と独立の事象であると見なすかどうか悩ましい、「〜と思う」、「〜を図る」、「〜を期待する」、「〜気がする」などの表現が存在する。補足欄「対象外」に “機能表現-事象可能” と記述しつつ、これらに対して拡張モダリティの情報も付与することにした。

5 試作した解析システムの誤り分析

本研究では、モダリティ解析の精度向上に必要な技術や言語資源について理解を深めるため、次のようなモダリティ解析システムを試作し、その誤り分析を行った。

- 入力は、文の構文解析結果と事象の核となる述語の位置。出力はその事象の拡張モダリティ
- 最大エントロピーモデルに基づく機械学習
- 構築したコーパスの OC を利用。5 分割交差検定
- 素性: 後続形態素列の表層形 1,2,3-gram と活用形、後続形態素列に存在する機能表現のクラス [12]、係ってくる文節内の形態素列の表層形 1,2,3-gram

[4] 表 1 では、これらの数は事象数に含めていない。

表 2: < 態度 >=“叙述” 関連の誤り分析結果

今後解くべき課題	事例数
(a)推量形式、感嘆形式、文末基本形などの**後続形態素列の語義曖昧性解消**	33(30%)
(b)主述語の拡張モダリティが**連体節内の述語に及ぼす影響の解明**	20(18%)
(c)並列節、条件節、目的節など、**節間の意味的関係の認識**	19(17%)
(d)素性として用いる**連用句の取捨選択**	18(16%)
(e)叙実表現やモダリティ副詞などの**手がかり表現の分類・集積**	8(7%)
(f)**省略解析**・否定や推量の**焦点特定**	7(6%)
(g)手がかり表現の**作用域を制限する要素の整理**	4(4%)
(h)文が含意する**前提の認識**	2(2%)
計	111(100%)

このシステムの < 態度 > に関する正解率は 0.90、正解ラベルとの一致率 (κ 統計量) は 0.69 であった。< 態度 >=“叙述” 関連の誤り 111 事例を分析し、今後解くべき課題ごとに整理した結果を表 2 に示す。

この表において合わせて 38%(42/111) を占める (b), (d), (g) は、文型と述語の位置に応じて現在の素性集合から不要な素性を上手く除去する枠組みを確立する課題である。言語直観が働きやすく、比較的取り組みやすいと考えられるので、今後はこれらの課題に取り組む。一方、(a), (c), (f), (h) は、意味解析に関わる難しい課題である。例えば、(a) においては、文末の「〜でしょうかね。」が推量か疑問かの判定が、(c) においては、接続助詞「て」の用法が並列か理由かの判定が求められる。

6 おわりに

本論文では、文章に存在する事象のモダリティおよびその周辺情報を適切に捉えた拡張モダリティの体系について述べ、この体系に基づいて構築した拡張モダリティタグ付与コーパスの設計と現状について報告した。また、解析システムの誤り分析を行い、その精度向上のために解くべき課題の一部を明らかにした。

今後は、コーパスの構築を続けるとともに、モダリティ解析システムの改善に取り組む予定である。

謝辞 本研究は、独立行政法人 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。本研究を遂行するにあたり多大な助力を頂きました東北大学の渡邉陽太郎助教に心より感謝いたします。

参考文献

[1] 川添愛, 齊藤学, 片岡喜代子, 崔栄殊, 戸次大介. 言語情報の確実性アノテーションのための様相表現の分類. 九州大学言語学論集, 第 31 巻, pp. 109–129, 2010.

[2] Kentaro Inui, Shuya Abe, Hiraku Morita, Megumi Eguchi, Asuka Sumida, Chitose Sao, Kazuo Hara, Koji Murakami, and Suguru Matsuyoshi. Experience mining: Building a large-scale database of personal experiences and opinions from web documents. In *the 2008 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 314–321, 2008.

[3] Marc Light, Xin Ying Qiu, and Padmini Srinivasan. The language of bioscience: Facts, speculations, and statements in between. In *Proceedings of BioLink 2004 workshop on linking biological literature, ontologies and databases*, pp. 17–24, 2004.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Roser Saurí. *FactBank 1.0 Annotation Guidelines*. `http://www.cs.brandeis.edu/~roser/pubs/fb_annotGuidelines.pdf`, 2008.

[9] 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.

[10] Roser Saurí and James Pustejovsky. Determining modality and factuality for text entailment. In *the International Conference on Semantic Computing*, pp. 509–516, 2007.

[11] Roser Saurí and James Pustejovsky. Factbank: a corpus annotated with event factuality. In *Language Resources and Evaluation*, 2009.

[12] 松吉俊, 佐藤理史. 文体と難易度を制御可能な日本語機能表現の言い換え. 自然言語処理, Vol. 15, No. 2, pp. 75–99, 2008.

[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.

[16] 松吉俊, 江口萌, 佐尾ちとせ, 村上浩司, 乾健太郎, 松本裕治. テキスト情報分析のための判断情報アノテーション. 電子情報通信学会論文誌 D, Vol. J93-D, No. 6, pp. 705–713, 2010.