

モダリティ、真偽情報、価値情報を統合した拡張モダリティ解析

江口 萌[†] 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

1. はじめに

一般に、文章には事象だけではなく、その事象に対する**情報発信者の主観的な態度 (モダリティ)** も記述される。例えば、以下の文 (1) では、下線部のようなモダリティを読み取ることができる。

- (1) ぜひとも、この薬品の使用を中止したい。
→ 事象「この薬品を使用するコト」が不成立になることを望んでいる

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 (誰が何をするのか) を認識するだけではなく、書き手が表明しているモダリティや真偽判断、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、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等を自動的に認識することは、情報抽出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。そこで、我々は、書き手が表明しているモダリティや真偽判断、価値判断などを統合した情報を、事象の**拡張モダリティ**と呼び、これを解析するシステムの開発に取り組んでいる。

拡張モダリティの解析システムを構築するにあたって、次の3つの難しい問題がある。I) 拡張モダリティに相当する情報を表す統一的な分類体系は存在しない。II) 拡張モダリティに影響する言語表現は非構成的に意味が変化する。例えば、否定辞の数により、単純に事象の真偽を判断することはできず、周辺の文脈を考慮する必要がある。

- (2) この薬品を使用しないこともない。
→ 事象「この薬品を使用するコト」の**成立**を述べ立てている。
- (3) この薬品を使用しないのではないだろうか。
→ 事象「この薬品を使用するコト」の**不成立**を推量している。

III) 拡張モダリティは、多種多様な言語表現に影響を受ける。例えば、文 (1) は、「中止する」という言語表現により、否定辞が存在しないにも関わらず、事象の成立が否定されている。

本論文では、これらの問題を考慮し、独自の拡張モダ

表 1 本研究、および、先行研究で対象とする項目

	確信度	価値判断	表現類型	肯否極性	態度表明者	時制	仮想性	真偽アスペクト	焦点
本研究	○	○	○	○	○	○	○	○	○
Rubin ら ⁹⁾	○	×	○	×	○	○	×	×	×
TimeML ¹⁰⁾	○	×	○	○	×	○	○	×	×
Prasad ら ⁸⁾	○	×	○	○	○	×	×	×	○
Sauri ら ¹¹⁾	○	×	×	○	○	×	×	×	×
FactBank ¹²⁾	○	×	○	○	○	○	○	×	×
Inui ら ²⁾	○	×	○	○	○	○	○	○	×
川添 ら ¹⁾	○	×	×	○	○	×	○	×	×
Light ら ⁵⁾	○	×	×	×	×	×	×	×	×
Medlock ら ⁶⁾	○	×	×	×	×	×	×	×	×
BioScope ¹⁵⁾	○	×	×	○	×	×	×	×	×

リティタグ体系を構築するとともに、この体系に基づく拡張モダリティ解析システムを提案する。

2. 先行研究

我々は拡張モダリティの体系を設計するに当たり、言語学におけるモダリティ、および、その周辺項目のうち情報の信憑性を判断するために有用であろう項目を整理した。表 1 に、我々が整理した項目のうち、本研究と先行研究がどれを対象としているかを示す。これらの項目の一部を対象としているが、これらを包括的に捉えた先行研究はない。

分類体系に関する重要な先行研究は、Sauri らによる FactBank¹²⁾ である。Sauri らは、事象とその時制、肯否、モダリティをタグ付与する TimeML¹⁰⁾ の体系の上に、事象を対象とし、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と独自の肯否極性を付与する枠組みを提案している。TimeML のタグ付与は、事象の核となる述語に接続する助動詞 (must, may など) を直接記述するため、日本語などの膠着言語に対してこの体系を直接適用することは難しい。

解析手法に関しては、人手による語彙的・統語的知識を用いる手法^{5),11)} や、機械学習による手法^{2),5),6)} が提案されている。例えば、Sauri ら¹¹⁾ は、項と述語間の知識、否定辞の有無、モダリティ表現を利用して、事象の入れ子の外側から、順に事象の事実性を解析するアルゴリズムを提案している。Inui ら²⁾ は、条件付確率場を学習器として用いて、事象のモダリティと時間ごとの肯否極性を解析する手法を提案している。

3. 拡張モダリティ

3.1 拡張モダリティタグ体系

表1で挙げた項目をもとに、モダリティや真偽判断、価値判断などの情報を統合した拡張モダリティタグ体系を設計した。これは、〈態度表明者、時制、仮想、態度、真偽判断、価値判断、焦点〉の7項目から構成され、文章に存在する各事象に対して付与される。タグ付与例を表2に示す。タグの詳細については、文献⁷⁾、および、脚注のURL^{*}で公開している作業基準書を参照して欲しい。

3.2 拡張モダリティタグ付与コーパス

現在、前節のタグ体系に基づくコーパスを構築中である。このコーパスの対象テキストは、下の7種類である。

- (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) 言明間意味の関係コーパス³⁾における文リスト (14,402 事象/2,878 文)
- (D_i) 現代日本語書き言葉均衡コーパス^{**}
 - (D₁) Web (14,439 事象/5,432 文)
 - (D₂) 白書 (27,121 事象/3,725 文)
 - (D₃) 新聞 (40,841 事象/11,606 文)
 - (D₄) 書籍 (34,200 事象/8,747 文)

ランダムに取り出した 300 事象に対する作業者間の一致度 (κ 統計量) は、7 項目に対する平均で 0.71 という高い一致を示した。なお、7 項目全体に対する κ 統計量は 0.58 であった。

現在、(A),(B),(C),(D₁) のコーパスについてはタグ付与が完了している。なお、(C),(D_i) のコーパスについては、今後公開を予定している。

3.3 手かかり表現辞書

コーパス作成時に、拡張モダリティに影響する動詞、形容詞が少なからず存在することが明らかになった。そこで、直前事象に与える影響を〈態度、真偽判断、価値判断〉の3つ組で記述し、解析のための手かかり表現辞書を作成した。分類語彙表に存在するすべての動詞と形容詞、形容動詞を対象として、格(「を」や「に」)、直前の事象の形式(スル形、タ形)、直後の否定辞の有無を考慮して、上の3つ組を記述した。作成した辞書は8,122 エントリーからなり、例えば「中止する」の“を:スル形; 無”に対して〈意志、不成立、ネガティブ〉が、“を; タ形; 無”に対して〈意志、成立から不成立、ネガティブ〉が記述されている。この辞書も公開を予定している。

4. 拡張モダリティ解析システム

文の構文解析の結果と事象の核となる述語の位置が与えられると、その事象に対する拡張モダリティタグを出力

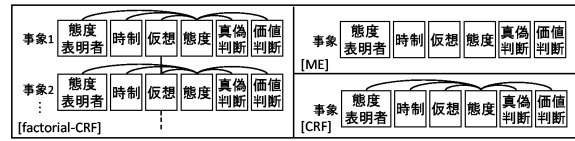


図1 グラフ構造: [factorial-CRF] は項目間・事象間の依存関係を考慮、[ME] は項目間の依存関係を考慮しない、[CRF] は項目間の依存関係のみを考慮するグラフ構造である

する解析システムを構築する。3.1 節のタグ体系は、〈態度〉を中心として項目間に強い依存関係を持っている。また、同じ文に存在する複数の事象の間には、それらのタグの間に依存関係があることが想定される。そこで、本研究では、項目間、および、事象間の依存関係を考慮できる条件付確率場⁴⁾を利用した解析システムを構築した。解析に用いた素性を表3の左側に示す。素性(a)において、係り先の文節は、南による節の分類¹⁷⁾により制限している。素性(b)は、素性(a)の3-gramよりも長い特徴的な形態素列を捉えるのに用いる。依存関係を表すグラフ構造を、図1の左の [factorial-CRF] に示す。上記の素性はエッジ素性として与え、事象間の依存関係を表現するエッジには、直前の事象の素性を与える。

〈態度表明者〉と〈焦点〉は、表2のID=7,9に示すように、その内に自由文字列を含むタグが存在するため、本実験においては、〈態度表明者〉の“wr_(文字列)”は“wr_STR”に統合し、〈焦点〉は対象から外した。

5. 実験

5.1 設定

利用した素性の有効性を評価するため、表3のような素性リストを利用する、5つのシステム [ME],[CRF],[f-CRF1],[f-CRF2],[f-CRF3]、および、素性(g)の語彙統語パターンに基づく人手規則により、決定的に出力を行なうシステム [BASE] を実装した。図1の右側のようなノード素性を持つ [ME]、エッジ素性を持つ [CRF] と比較することで、項目間、事象間の依存関係を考慮することが有用であるかを調査する。

学習・評価コーパスには、前述のコーパス (A), (B), (C), (D₁) (但し、(D₁) については一部分) の計 41,704 事象を使用し、5 分割交差検定を行なった。本実験では、各項目ごとに「最頻出のタグを除くマイクロ F 値」を算出し、評価尺度として用いた。最頻出のタグを除いた理由は、表5,6の右から2列目に示されるように、最頻出のタグが全体の約90%を占めているが、実際に重要であるのは、残りの10%を正確に捉えることであり、上記の評価尺度はこれを測るのに妥当であると考えたからである。

5.2 結果

実験結果を表4に示す。[f-CRF3] が他のシステムよりも高い値を示した。正解率に関して検定を行った結果、[f-CRF3] と [ME] の間には全ての項目において、[f-CRF3] と [CRF] の間には〈仮想〉において有意差(有意水準1%)

^{*} <http://cl.naist.jp/nltools/modality/manual.pdf>

^{**} <http://www.tokuteicorpus.jp/>

表 2 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	虫菌菌は「キシリトール」を代謝しようとして大きなエネルギーを消費するため、虫菌予防にもなります。	wr	非未来	0	叙述	成立	0	0
2	これらを達成するため、非抜歯矯正治療とフッ素やキシリトールを用いた虫菌予防に力を入れております。	wr	未来	0	意志	高確率	ポジティブ	0
3	最初の3ヶ月ほど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用されていません。	wr_ot	非未来	0	叙述	不成立	0	0
4	新潟の酒を美味しく飲むにはウコンが必須だそうです。	wr_arb	非未来	条件	叙述	成立	0	0
5	今後更に育種改良を進展させるため、... 生産性の向上を図る。	wr	未来	0	意志	高確率	ポジティブ	0
6	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
7	新宿行って、なんか食べに行きたいと太郎が言った。	wr_太郎	未来	0	欲求	0	ポジティブ	0
8	... 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
9	両親によると彼は東京に仕事で行ったわけではないらしい。	wr_両親	非未来	0	叙述	成立	0	否定(仕事で)

表 3 各システムで解析に用いた素性

	素性	GRMM を用いたシステム				
		ME	CRF	f-CRF1	f-CRF2	f-CRF3
(a)	格となる述語が存在する文節 (中心文節)、その前後の文節、中心文節に係る文節、中心文節の係り先 2 つまでの文節における形態素情報 (基本形、品詞、活用形) の 2-gram, 3-gram	○	○	○	○	○
(b)	機能語列 (核となる述語以降の形態素で、2 回目の内容語が出現するまでの形態素列)	○	○	○	○	○
(c)	中心文節の係り先 2 文節での意志動詞 ¹³⁾ の有無	○	○	○	○	○
(d)	分類語彙表 ¹⁶⁾ における「人間活動の主体」に属する形態素の有無	○	○	○	○	○
(e)	日本語機能表現辞書 ¹⁴⁾ の意味コード	○	○	×	○	○
(f)	モダリティ解析用手かがり表現辞書	○	○	×	○	○
(g)	人手により構築した語彙統計パターン (モダリティ関連 215 個、肯否極性関連 153 個、条件関連 27 個)	○	○	○	×	○

表 5 〈態度〉の各タグの分類精度 (Precision, Recall) と分布

正解タグ	システムが出力したタグ	システムが出力したタグ								計	Recall
		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		
正解タグ	叙述	38,020	123	20	10	43	0	0	115	38,331	0.992
	意志	446	462	6	1	17	2	0	4	938	0.493
	欲求	86	14	382	1	3	0	0	3	489	0.781
	働きかけ-直接	94	5	1	210	6	0	0	3	319	0.658
	働きかけ-間接	200	16	4	3	290	5	0	4	522	0.556
	働きかけ-勧誘	16	11	0	0	9	25	0	3	64	0.391
	許可	20	0	0	0	0	0	0	0	20	0.0
	問いかけ	240	5	3	0	5	0	0	768	1,021	0.752
	計	39,122	636	416	225	373	32	0	900		
Precision	0.972	0.726	0.918	0.933	0.777	0.781	0	0.853			

表 6 〈真偽判断〉の各タグの分類精度 (Precision, Recall) と分布

正解タグ	システムが出力したタグ	システムが出力したタグ								計	Recall	
		成立	高確率	0	低確率	不成立	成立から 不成立	不成立から 成立	高確率から 低確率			低確率から 高確率
正解タグ	成立	34,401	193	131	2	55	0	0	1	0	34,783	0.989
	高確率	872	1,048	55	2	11	0	0	0	1	1,989	0.527
	0	568	103	1,737	3	71	0	0	0	0	2,482	0.7
	低確率	42	16	7	60	120	0	0	0	0	245	0.245
	不成立	146	6	22	15	1,932	0	0	0	0	2,121	0.911
	成立から不成立	11	1	1	0	9	1	0	0	0	23	0.043
	不成立から成立	28	3	0	0	0	0	0	0	0	31	0.0
	高確率から低確率	12	1	0	0	1	0	0	0	0	14	0.0
	低確率から高確率	11	3	0	0	1	0	0	0	1	16	0.063
	計	36,091	1,374	1,953	82	2,200	1	0	1	2		
Precision	0.953	0.763	0.889	0.732	0.878	1.0	0	0.0	0.5			

表 4 実験結果: 最頻出のタグを除くマイクロ F 値

システム	態度表明者	時制	仮想	態度	真偽判断	価値判断
BASE	.042	.538	.312	.666	.723	.651
ME	.109	.639	.553	.662	.747	.688
CRF	.182	.636	.594	.710	.761	.683
f-CRF1	.182	.636	.594	.710	.761	.683
f-CRF2	.174	.579	.538	.662	.676	.607
f-CRF3	.188	.641	.599	.718	.763	.687
1 回目	.231	.642	.591	.717	.761	.690
2 回目	.261	.641	.589	.719	.761	.687
3 回目	.267	.644	.608	.722	.762	.689
4 回目	.285	.646	.614	.753	.763	.695

が認められた。この結果より、項目間、ならびに、事象間の依存関係を考慮することが、拡張モダリティ解析においては、有効であったと言える。また、[f-CRF3] と [f-CRF1]の間には有意差が見られなかったが、素性 (e)(f)

を追加することで、例えば、表 2 の ID=5 の事象に対して、〈態度〉を“叙述”、〈真偽判断〉を“成立”、〈価値判断〉を“0”と誤って判定していたものを正しく判定できるようになった。これは、素性 (f) の見出し語:「図る」の情報が追加されたためと考えられる。[f-CRF3] と [f-CRF2]の間については、〈時制, 態度, 真偽判断, 価値判断〉で有意差が見られ、素性 (g) が精度の向上に大きく影響していることが分かった。

5.3 考察

[f-CRF3]における、〈態度〉、および、〈真偽判断〉の詳細な解析結果を表 5 と 6 に示す。紙面の都合上、〈態度〉

に着目し、学習事例が3番目に多いにも関わらず、精度が低かった“意志”、特に最も誤り数が多かった“意志”を“叙述”とした事象(446個)の誤り分析を行なった。分析した結果、“叙述”と“意志”に共通して出現する特徴的な形態素情報を含む事象を多く誤ってしまう傾向が見られた。例えば、直後に「ため」(約14%(63/446))や、「よう」(約12%(53/446))を含む事例である。このような事例の拡張モダリティを正しく解析するには、例えば、表2のID=1の「ため」が‘理由’、ID=2の「ため」が‘目的’を述べるのに使われているのかを正しく把握することが重要である。これには、素性(c)の意志・無意志動詞辞書が重要な指標として考えられるが、現在、有効な素性として上手く利用できていないようである。これについては、事象の核となる述語と係り先の述語が意志動詞かどうかを素性とするなど、今後、素性(c)を再検討する必要がある。

コーパスにおけるタグの分布が偏っているために、出現頻度の低いタグに関する学習がうまくいっていないという問題もある。この解決策として、選択的サンプリングにより、モデルにとって有用であろう事象を選択し学習することが考えられる。次節では、選択的サンプリングの予備実験の結果について報告する。

5.4 選択的サンプリングの予備実験

モデルにとって有用と予想される事象を選択し、学習していくことが、出現頻度の低いタグの精度の向上に有効かを調査した。タグ付与作業者の経験則により、出現頻度の低いタグが付与されそうな事象を選択し、タグ付与を行い、学習データに追加していくことで、「最頻出のタグを除くマイクロF値」がどの様に変化するか観察した。今回作業対象としたコーパスは(D₂)~(D₄)である。[f-CRF3]のシステムを初期値として、新しくタグ付与したデータを500事象ずつ4回に分けて学習データに追加していった。結果を図4の下部示す。4回試行した結果、精度の向上が見られた。

6. おわりに

我々は、書き手が表明する態度や真偽判断、価値判断などの情報を統合し、これを表すタグ体系を提案し、このタグ体系に基づくコーパスを構築した。そして、項目間・依存関係を考慮できる条件付確率場を用いた解析システムを提案した。また、更なる精度の向上を目指し、拡張モダリティ解析用手法が表現辞書を構築した。

実験の結果、人手により作成した語彙・統語パターンが有効な素性という事が分かった。しかし、人手によるパターンの作成には膨大な手間・時間がかかるため、我々は、これに頼らない解析システムの構築を目指している。また、選択的サンプリングの予備実験の結果、事象を選択し学習していくことで、精度の向上が見られた。これは、語彙・統語パターンを素性としなくても、有効である学習データを効率的に追加していくことで精度が向

上していく可能性も示唆している。そこで、今後も、解析に有用な素性を検討すると共に、選択的サンプリングを続けることで、更なる精度の向上を目指している。また、この解析システムを応用である含意関係認識などに適用し、フィードバックを受けることで、タグ体系についても改良していく予定である。

タグ付与の作業基準やモダリティ情報解析に関する最新情報は、次のサイトで公開中である。なお、構築したコーパス・拡張モダリティ解析用手法が表現辞書は、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)、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「経験マイニング技術の高度化と実用化」(課題番号: 21013036, 代表: 乾健太郎)。

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, 2009-FI-93, 2009-NL-189, 2009.
- 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*, 2008.
- 3) 村上浩司, 増田祥子, 松吉俊, 乾健太郎, 松本裕治. 言明間の意味的關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) John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc of ICML*, 2001.
- 5) Marc Light, Xin Ying Qiu, and Padmini Srinivasan. The language of bioscience: Facts, speculations, and statements in between. In *Proc of BioLink 2004 WS on BioLINK*, 2004.
- 6) Ben Medlock and Ted Briscoe. Weakly supervised learning for hedge classification in scientific literature. In *the 45th Annual Meeting of the ACL*, 2007.
- 7) 江口明, 松吉俊, 佐尾ちとせ, 乾健太郎, 松本裕治. 日本語文章の事象に対する判断情報アノテーション. 情報処理学会研究報告 2009-NL-193, 2009.
- 8) Rashmi Prasad, Nikhil Dinesh, Alan Lee, Aravind Joshi, and Bonnie Webber. Annotating attribution in the Penn discourse treebank. In *the COLING/ACL WS on Sentiment and Subjectivity in Text*, 2006.
- 9) Victoria Rubin, Elizabeth Liddy, and Noriko Kando. *Chapter 7: Certainty Identification in Texts: Categorization Model and Manual Tagging Result*. Springer-Verlag New York, 2005.
- 10) Roser Saurí, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky. *TimeML Annotation Guidelines Version 1.2.1*, 2006.
- 11) Roser Saurí and James Pustejovsky. Determining modality and factuality for text entailment. In *IEEE ICSC*, 2007.
- 12) Roser Saurí and James Pustejovsky. Factbank: a corpus annotated with event factuality. In *Language Resources and Evaluation*, 2009.
- 13) 阿部修也, 乾健太郎, 松本裕治. 共起パターンによる事象間関係知識の獲得. 自然言語処理, Vol. 16, No. 5, pp. 79-100, 2009.
- 14) 松吉俊, 佐藤理史, 宇津呂武仁. 日本語機能表現辞書の編纂. 自然言語処理, 第14巻, 2007.
- 15) György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *the WS on Current Trends in BioNLP*, 2008.
- 16) 国立国語研究所. 分類語彙表. 大日本図書, 2004.
- 17) 南不二男. 現代日本語の構造. 大修館書店, 1974.