語彙知識と構成性に基づく日本語事実性解析

成田 和弥　　水野 淳太　　乾 健太郎
東北大学
{narita, junta-m, inui}@ecei.tohoku.ac.jp

1 はじめに

事実性とは，文に記述されている事象が，起こったことなのかそうでないことなのか，といった情報である[1]．

(1) a. 彼はさきほど部屋を出た。
b. もう遅いから、彼は先に帰ったんだろう。
c. 問題が発生するのを防いだ。

例えば，(1a) の事象「出る」は，実際に起こったことだと著者は判断している，と解釈できる．(1b) の事象「帰る」は,「だろう」という表現により，起こった可能性が高いことだと著者は判断している，と解釈できる．(1c) の事象「発生する」は,「防いだ」という表現により，実際には起こっていないことだと著者は判断している，と解釈できる．

事実性を解析する技術は，情報抽出や含意関係認識など，自然言語処理の応用に対して有用であるが，多岐に渡る言語表現を捉えきれておらず，いまだ十分な解析性能は実現されていない．また，内在する問題の分析・整理が不十分であることも原因の一つとなっている．そこで，高精度な日本語事実性解析器を実現するために，事実性解析を行う上で，何が問題となっているかを分析し，明らかにする必要がある．

我々は，事実性に影響を与える言語表現の存在に着目した．(1b) の事象「帰る」の事実性は,「だろう」という表現に，(1c) の事象「発生する」の事実性は,「防いだ」という表現に影響されるように，事実性に影響を与える手がかりとなる表現が多数存在している．本論文では，事実性に影響を与える語彙知識に焦点を当て，語彙知識と構成性に基づく日本語事実性解析を行うことで，どのような問題があるかを明らかにし，今後どのような研究に取り組むべきかを議論した．

2 関連研究

事実性に大きく関連する概念として，情報発信者の主観的な態度（モダリティ），および，肯定または否定があげられる．本研究における事実性は，事象の真偽に対する書き手の確信度を表した「真偽判断のモダリティ」[7] と，肯定または否定の組み合わせに相当している．

事実性およびその周辺情報を付与するための体系に関する研究としては，Sauríらによる FactBank (2009) [5] や，松吉らによる拡張モダリティタグ付与コーパス (2010) [10]，川添らによる MCN コーパス (2012) [12] などがある．Sauríらは，事実らしさに対する態度表明者の確信度と，その確信の方向を表す肯否極性の組で事実性を定義し，事象に付与した．松吉らは，＜態度＞，＜真偽判断＞等からなる拡張モダリティタグ体系を設計し，それを事象に付与したコーパスを構築した．拡張モダリティタグのうち，＜真偽判断＞が事実性に対応する．川添らは,「はず」「だろう」のようなモダリティ関連表現に対し，事実性の認識に関連する表現の意味・用法を特定し，新聞記事等のテキストにアノテーションを施したコーパスを構築している．

解析手法に関する研究としては，機械学習に基づく手法や，パターンベースの手法があげられる．江口ら [8] は，拡張モダリティタグの項目間，文内の事象間の依存関係を考慮できる，条件付確率場を用いた拡張モダリティ解析を行った．Sauríら [4] は，事実性に影響を与える手がかり表現を利用し，確信度と肯否極性を，依存構造木の根から伝搬させて解析するモデルを提案した．

3 日本語事実性解析器の構築

日本語事実性解析における問題を明らかにするため，Sauríら (2007)[4] の英語事実性解析モデルをもとに，日本語事実性解析器を構築した．今回構築した日本語事実性解析器は，構文解析結果を入力とし，各事象に対する事実性を出力する．

3.1 事実性の定義

Sauríらは，事実らしさに対する確信度を Certain (CT)・Probable (PR)・Possible (PS)・Underspecified (U) の 4 段階，その確信の方向を表す肯否極性を positive (+)・negative (−)・underspecified (u) の 3 値として扱い，これらの組み合わせによって事実性の値を定義した．例えば (1a) の事象「出る」の事実性は CT+と表される．同様に，(1b) の事象「帰る」の事実性は PR+，(1c) の事象「発生する」の事実性は CT− と表される．

英語では，PR は *probable*，PS は *possible* といった表現によって，事実性を解釈しているが，日本語では表現が多彩であるため，PR と PS の境界が曖昧で，その区別は容易でないことが予想される．そこで我々は，PR と PS を 1 つの値 PR としてまとめて扱った．また Sauríらは，事実性が明らかでないときに，CTu，Uu といった事実性を定義している[2]．我々は，簡単化のためこれ

[1] その事象が真に起こったことかはわからないため，本論文では，著者による事象の成否の判断として事実性を解釈する．また，松吉ら (2010) [10] と同様に，事象は行為，出来事，状態の総称であると考える．

[2] PRu，PSu，U+，U− は利用不可値と考えられている．

表 1: 確信度と肯否極性の組み合わせによる事実性の値

確信度 \ 肯否極性	positive (+)	negative (−)
Certain (CT)	実際に起こっている (CT+)	実際には起こっていない (CT−)
Probable (PR)	起こっている可能性が高い (PR+)	起こっている可能性が低い (PR−)
Underspecified (U)	不明 (U)	

らを区別せず，事実性が明確ではない場合をまとめて U とした．

以上 2 点を変更し，それ以外は Sauríらの定義を利用した．即ち，確信度を Certain (CT)・Probable (PR)・Underspecified (U) の 3 段階，肯否極性を positive (+)・negative (−) の 2 値として扱い，これらの組み合わせによって事実性を表す．これら組み合わせをまとめたものを表 1 に示す．

3.2 使用する語彙知識

Sauríらのモデルは，確信度と肯否極性の組で表される事実性を，factuality marker と呼ばれる，事実性に影響を与える表現を利用して解析する．例えば *not* は肯否極性を反転させる表現，*may* は確信度を下げる表現である．機能語だけでなく述語についても考えられ，例えば *know that* という表現は *that* 節の内容が成立していることを前提としているので，*know* は *that* 節内の事象を CT+とする表現と考えることができる．

日本語においても「〜ない」は肯否極性の反転,「〜だろう」は確信度の減少，というように，同等の表現が存在する．このような表現を集めた語彙知識として，日本語機能表現辞書「つつじ」（機能表現辞書）[11] およびモダリティ解析手がかり表現辞書（手がかり表現辞書）[8] を利用した．機能表現辞書は，文の構成に関わる要素である機能表現を，意味，文法的機能などに応じて網羅的に収録した辞書である．この辞書は意味クラスによって表現を分類しており，例えば「否定」の表現は肯否極性を反転,「推量」の表現は確信度を減少，といった手がかりとして利用できる．手がかり表現辞書は，述語が拡張モダリティにどのような影響を与えるかを，先行する事象の時制および肯否環境ごとに収録しているが，これは事実性とも大きな関わりをもっている．例えば (1c) の「防いだ」のような述語が収録されており，先行する事象「発生する」の肯否極性を反転させる手がかり表現として利用できる．

3.3 解析モデル

前節であげた，事実性に影響を与える手がかりとなる表現を集めた語彙知識を利用し，確信度と肯否極性を依存構造木の根から伝搬させて事実性を解析する．このモデルは，依存構造木をもとに，以下の 3 つの要素に基づいて事実性解析を行っている．

親の辞書情報 親の述語により，手がかり表現辞書をもとに事実性を更新する

機能語の辞書情報 解析したい事象に付随する機能語により，機能表現辞書をもとに事実性を更新する

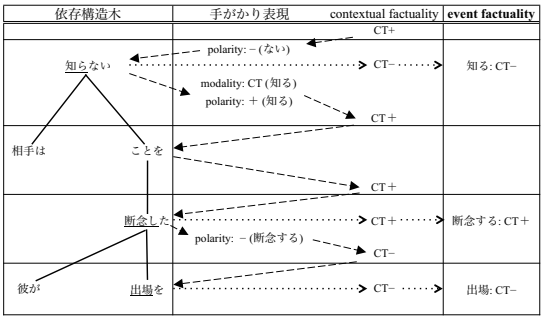

図 1: (2) に対する事実性解析アルゴリズムの動き

親の事象が持つ事実性 親まで伝搬されてきた事実性をもとに，伝搬，更新を行う

ここでは，文全体で伝搬させる事実性を contextual factuality (CF)，各事象に割り当てる事実性を event factuality (EF) と呼ぶ．入力は一文の構文解析結果，出力は各事象に対する EF であり，CF は EF を伝搬させるための変数である．

まず，多くの事象の事実性は CT+であるため，初期値として CF に CT+を割り当て，文末の文節から依存構造木の走査を開始する．各文節で行うことは，機能表現辞書を用いた更新 (Step1)・事象がその文節内に存在するかの確認 (Step2)・手がかり表現辞書による更新 (Step3) の 3 ステップである．

例として，(2) に対する解析の流れを図 1 に示す．

(2) 彼が出場を断念したことを相手は知らない。

まず CF の初期値 CT+を割り当て，依存構造木を文末から走査していく．最初の文節「知らない」に対して，Step1 では，極性を反転させる否定の機能表現「ない」が存在するので,「機能語の辞書情報」により，CF を CT−に更新する．Step2 において,「知る」は事実性を割り当てるべき事象であるため，そのときの CF (CT−) を「知る」の EF として出力する．Step3 では,「知る」という述語が手がかり表現辞書内に存在するので，それに基づいて CF を CT+に更新する．そのまま子に伝搬をしていき,「断念する」の EF を出力する際には，親である「知る」が持つ事実性（CT−），親である「知る」の辞書情報（CT− → CT+），そして「断念する」に付随する機能語の辞書情報（今回は該当する機能語は存在しない）により，解析が行われる．これを繰り返し,「知る」の事実性が CT−,「断念する」の事実性が CT+,「出場」の事実性が CT− という出力が最終的に得られる．

4 実験・問題分析

4.1 実験設定

日本語事実性解析における問題を整理・分析するために，構築した解析器をウェブ上の数千文に対して適用し，その評価を行った．

実験では，拡張モダリティタグ付与コーパス [10] の一部である，Yahoo!知恵袋 (OC) に含まれる 6,404 文に対してアルゴリズムを適用した．このコーパスは現代日本語書き言葉均衡コーパス (BCCWJ)[3]に対して，＜態度

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

表 2: ＜真偽判断＞と事実性との対応

確信度 \ 肯否極性	+	−
CT	成立 不成立から成立 (CT+)	不成立 成立から不成立 (CT−)
PR	高確率 低確率から高確率 (PR+)	低確率 高確率から低確率 (PR−)
U	0 (U)	

表 3: 文末の事象，確信度に関する Confusion Matrix

正解 \ 出力	CT	PR	U	Total
CT	**4,148**	106	576	4,830
PR	280	**91**	111	482
U	536	33	**1,451**	2,020
Total	4,964	230	2,138	7,332

表 4: 文末の事象，肯否極性に関する Confusion Matrix

正解 \ 出力	+	−	Total
+	**4,028**	149	4,177
−	41	**427**	468
Total	4,069	576	4,645

表明者＞，＜態度＞，＜真偽判断＞などの 6 項目からなる拡張モダリティを付与したものである．この中の＜真偽判断＞は，我々の事実性に相当する．表 2 に示すように，＜真偽判断＞と我々の事実性を対応付けた．

本実験で対象とする事象は，＜態度表明者＞が著者となっているものの中で，限定修飾・機能表現のように「対象外」が付与されていない 14,917 事象である．正解の形態素情報を入力し，構文解析を行った結果を，構築した事実性解析器の入力とする．尚，事実性を付与すべき事象の同定は，あらかじめ正解を与えた．

また，更新した事実性を伝搬していくため，一度誤りを生じると，伝搬先の事実性にも影響を及ぼす可能性がある．分析の目的として，そのような誤りが生じてしまうのは本意ではない．そこで，誤りの伝搬を除くため，事実性の伝搬元となる「親の事象が持つ事実性」として，コーパス中の正解ラベルを利用した．

4.2 問題分析

語彙知識と構成性に基づく事実性解析器の誤り分析を通して，日本語事実性解析において，今後どのような課題に取り組んでいくべきかを分析した．

我々は，文末の事象における事実性に着目した．現在のアルゴリズムでは，「親の辞書情報」「機能語の辞書情報」「親の事象が持つ事実性」をもとに事実性を決定しているが，文末の事象では，親からの影響がないため，「機能語の辞書情報」のみを利用している．文末の事象においては，主に機能語に関する問題が起こることが考えられ，それ以外の事象では，他の要素も考慮した複合的な問題が起こることが考えられる．そこで，文末の事象における問題と，それ以外の事象における問題を区別し，分析を行った．表 3〜6 に，それぞれにおける，確信度，肯否極性に関する Confusion Matrix を示す．

文末の事象における解析誤りを，ランダムに 430 事例確認し，どのような問題があるかを分析したところ，半分の 215 事例が機能語の影響による問題であった[4]．文

表 5: 文末以外の事象，確信度に関する Confusion Matrix

正解 \ 出力	CT	PR	U	Total
CT	**4,293**	291	1,556	6,140
PR	362	**131**	135	628
U	320	45	**452**	817
Total	4,975	467	2,143	7,585

表 6: 文末以外の事象，肯否極性に関する Confusion Matrix

正解 \ 出力	+	−	Total
+	**4,333**	306	4,639
−	108	**330**	438
Total	4,441	636	5077

末以外の事象では，ランダムに 1,483 事例を分析し，そのうち，機能語に関する問題が 436 事例（29 %），親の内容語の問題が 47 事例（3 %），事実性を伝搬する範囲（スコープ）の問題が 789 事例（53 %）であった[5]．以降では，文末の事象における機能語の問題，文末以外の事象における内容語・スコープの問題について述べる[6]．

4.2.1 文末の事象における機能語の問題

文末の事象における機能語の問題 215 事例のうち，U に関する誤りが最も多く，「正解は U ではないが，U だと誤認してしまっている」場合が 123 事例（57 %）を占めた．このうち，U に関わる機能語がもつ曖昧性によって，意味を誤認してしまう，という問題がほとんどであった．

(3) 知らないのも不思議ではないです。（正解：CT−，出力：U）

(3) は，機能語の曖昧性が問題となっている例である．最長一致により，網羅的に事実性に影響を及ぼす機能語を認識している．そのため，「では」を勧めの意味としてとらえ，「不思議である」の事実性が U と誤認された．このように，事実性解析において，機能語のもつ曖昧性が問題であることが明らかになった．機能語の曖昧性は簡単な問題ではなく，先行研究でも議論されている [9, 13, 12]．特に，川添ら (2012) [12] は，「ではない」「よう」のようなモダリティ関連表現の曖昧性解消を目的の一つとし，MCN コーパスの構築を行っている．また，事実性に影響を与える機能語が，そもそも機能表現辞書に載っていない場合もあったが，その数は機能語の問題 215 事例のうち，10 事例もなく，概ね現在の機能表現辞書でカバーできていることが明らかになった．

また，機能語が省略されることによってうまく認識できていないという問題も見られた．

(4) 知っている人は教えて。（正解：U，出力：CT+）

(4) では，「教えて」のあとに「ください」が省略されているため，事象「教える」の事実性が U だということを，表層のみから当てるのは容易ではない．この問題は，事例数としては多くないが，インターネット上のテキスト等を扱う機会が多くなるにつれ，大きな問題となることが考えられる．

4.2.2 文末以外の事象における内容語の問題

文末以外の事象における内容語の問題は，47 事例あり，このうち，38 事例は内容語が不足していることによ

[4]残りは，正解ラベルのアノテーション誤りや，そもそも事実性を判断できない事象といった問題があった．また，一部「あまり」「おそらく」などの事実性に影響を与える副詞を考慮できていない，という問題も存在した．

[5]他にも，文末の事象と同様に，「副詞の問題や，正解ラベルのアノテーション誤り，そもそも事実性を判断できない事象があった．

[6]機能語に関する問題は，文末の事象における分析と同じような傾向が見られたため，本稿では割愛した．

る問題であり，9 事例が，内容語が誤った影響を及ぼしたことによる問題であった．

(5) 正しいことを確認してください。（正解：CT+，出力：U）

(5) では，親の述語である「確認する」につけられた事実性 U を伝搬させたことにより，事象「正しい」の事実性を U と出力している．しかし，「確認する」は先行する文脈を前提する述語であるため，「正しい」の事実性は CT+となるべきである．これは，手がかり表現辞書に「確認する」が存在しなかったため，うまくいかなかった問題である．

今回分析した，文末以外の事象における誤り 1,483 事例のうち，手がかり表現辞書中のない用語が関与していたものが 422 事例あった．ところが，内容語に関する誤りはわずか 9 事例であったことから，内容語の曖昧性の問題は，ほとんどないことが明らかになった．また，分析したコーパス領域中の正解事例 2,207 件のうち，手がかり表現を使って正解できたものは，628 件であった．これに対し，手がかり表現が辞書に登録されていないことが原因となっている誤りは，わずか 38 件であった．このことから，内容語のカバレッジ不足も，大きな問題ではないことがわかった．

4.2.3 文末以外の事象におけるスコープの問題

3 節で述べた通り，このモデルは「親の辞書情報」「機能語の辞書情報」「親の事象が持つ事実性」に基づいて事実性解析を行っている．しかしながら，そもそも事実性を伝搬させるべきなのか否かという要素が重要であり，半分以上を占める問題となっていることが明らかになった．即ち，どこまで事実性を伝搬させるか，という範囲に関する問題である．

(6) 少し郊外にでると音声が聞き取れません。（正解：CT+，出力：CT−）

(6) の事象「でる」はの肯否極性は+であるが，システムは − と出力してしまった．これは，否定の機能語「ん」の影響を受けた，事象「聞き取る」の肯否極性 − を，そのまま伝搬させてしまったことによる誤りである．否定表現「ん」は「音声が聞き取れる」ことに対するものであり，「少し郊外にでる」こととは関係がない．このように，否定表現や推量表現が影響を及ぼす範囲（スコープ）の問題が重要であることが明らかになった．例えば，親の肯否極性が − であるときに，子の肯否極性を − として出力した 256 事例のうち，正解しているものが 75 事例あり，スコープで誤っていると分析されたものは 147 事例あった．このように，スコープは大きな問題となっていることがわかるが，− のまま伝搬させるべき事例も多く存在し，難しい課題であることが明らかになった．

否定表現および推量表現のスコープを同定する研究は，CoNLL-2010 や*SEM 2012 における Shared Task[2, 3] で扱われるなど，英語では近年盛んに行われている [6, 1]．例えば，BioScope (2008) [6] は，否定表現，様相表現，そして，それらのスコープをマークアップしたコーパスであり，スコープを特定する研究に広く利用されている．しかし，日本語では十分に研究されておらず，今後の重要な課題であり，この問題を解決することが，事実性解析の精度向上に大きくつながると考えられる．

5 おわりに

本論文では，情報抽出や含意関係認識などの，自然言語処理の応用に対して，有用な情報である，事実性の解析に関して述べた．特に，語彙知識と構成性に基づく日本語事実性解析を行うことによって，どのような問題があり，今後どのような研究を行うことで事実性解析につなげられるかを議論した．その結果，「機能語の曖昧性の問題」「内容語の知識の問題」「事実性を伝搬させる範囲の問題」などがあることを明らかにした．また，今回は人手により除外している，そもそも事実性を議論すべき対象ではない事象を，実際にどのように判断し，解析対象から除いていくかも大きな問題であると考えられる．

今後は，これらの課題を進め，語彙知識および言語現象を加味した，高度な事実性解析器の構築を行う．さらに，事実性の枠組みを拡張することで，高度なモダリティ解析器の構築につなげていきたいと考えている．

謝辞 本研究は，文部科学省科研費（23240018）の一環として行われた．

参考文献

[1] Emilia Apostolova, Noriko Tomuro, and Dina Demner-Fushman. Automatic extraction of lexico-syntactic patterns for detection of negation and speculation scopes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pp. 283–287, 2011.

[2] Richárd Farkas, Veronika Vincze, György Móra, János. Csirik, and György Szarvas. The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning — Shared Task*, pp. 1–12, 2010.

[3] Roser Morante and Eduardo Blanco. *SEM 2012 shared task: Resolving the scope and focus of negation. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the main conference and the shared task*, pp. 265–274, 2012.

[4] Roser Saurí and James Pustejovsky. Determining Modality and Factuality for Text Entailment. *First IEEE International Conference on Semantic Computing*, pp. 509–516, 2007.

[5] Roser Saurí and James Pustejovsky. FactBank: a corpus annotated with event factuality. *Language resources and evaluation*, Vol. 43, No. 3, pp. 227–268, 2009.

[6] György Szarvas, Veronika Vincze, Richárd Farkas, and J. Csirik. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pp. 38–45, 2008.

[7] 益岡隆志. 日本語モダリティ探求. くろしお出版, 2007.

[8] 江口萌, 松吉俊, 佐尾ちとせ, 乾健太郎, 松本裕治. 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. 言語処理学会第 16 回年次大会発表論文集, pp. 852–855, 2010.

[9] 今村賢治, 泉朋子, 菊井玄一郎, 佐藤理史. 述部機能表現の意味ラベルタガー. 言語処理学会第 17 回年次大会発表論文集, pp. 308–311, 2011.

[10] 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.

[11] 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, Vol. 14, pp. 123–146, 2007.

[12] 川添愛, 田中リベカ, 戸次大介. MCN コーパス：モダリティ関連表現の曖昧性解消のためのアノテーションと言語学的テストの利用. テキストアノテーションワークショップ・コンテスト, 2012.

[13] 鈴木敬文, 阿部佑亮, 宇津呂武仁, 松吉俊, 土屋雅稔. 大規模階層辞書と用例を用いた日本語機能表現の解析. 『現代日本語書き言葉均衡コーパス』完成記念講演会予稿集, pp. 105–110, 2011.