テキスト情報の事実性解析
Analyzing the Factuality of Textual Information

森田 啓[†] 佐尾 ちとせ[†] 松吉 俊[†] 松本 裕治[†] 乾 健太郎[‡]
Hiraku Morita Chitose Sao Suguru Matsuyoshi Yuji Matsumoto Kentaro Inui

1. はじめに

大規模なテキスト集合から，述語を核とした事象を抽出するタスクにおいて，その事実性を解析する技術は重要である．なぜならば，事象は，確定した事実だけではなく，まだ予定にすぎないことや仮定なども表わしうるからである．それゆえ，事実性解析は，特に，情報の信憑性や有効性を判断するときに，必要不可欠な技術であると言える．

このような背景により，我々のグループは，事象の事実性を捉えるタグ体系とそれに基づいて解析を行う事実性解析器を構築した [1]. 本論文では，この解析器の精度向上を目指して行っている，次の 2 つのアプローチについて報告する．

1. 事実性タグ体系の精緻化
2. 複数事象の事実性における依存関係を考慮した解析

2. テキスト情報の事実性

本研究では，次のように事象を定義する．

項構造+述語+時間情報+極性+モダリティー

ここで，

項構造 名詞句+助詞のリスト

述語 動詞，形容詞，サ変名詞，名詞-形容動詞語幹．「降り始める」など，補助動詞が付いたものも 1 つの述語と見なす

時間情報 テンスとアスペクト

極性 否定表現の有無

モダリティー 「そうだ」,「べきだ」,「と思う」,「というのは信用しかねる」など，かなり多様な言語表現によって表明される話者態度や発話意図

である．事象の例を以下に示す．

- 彼はカレーライスを食べなかったそうです．

この文では,「なかっ」が否定の極性を,「た」が過去の時制を,「そうです」が伝聞のモダリティーを表している．

本研究では，事象に対する時間情報，極性，モダリティーの 3 つの値を事象の事実性と呼び，事象が与えられたときに，これらの 3 つの値を推定することを事実性解析と呼ぶ．日本語には，モダリティーを表す言語表現が大量に存在するため，この事実性解析は容易ではない．

本研究で提案する事実性解析器は，テキストが入力されたときに，そこに含まれるすべての事象に対して，次章で説明する事実性タグの組を出力する．

[†]奈良先端科学技術大学院大学 情報科学研究科, Nara Institute of Science and Technology
[‡]情報通信研究機構 知識処理グループ, National Institute of Information and Communications Technology

3. 事実性タグ体系の精緻化

3.1 事実性タグ

本研究で用いる事実性タグは，原ら [1] のタグ体系に準拠するものであり，次のようなタグの 7 つ組で表わされる．

(瞬間, ・, ・, 推量, ・, 状態, ・)

最初の 3 つは，それぞれ，過去，現在，未来における時間情報 (+極性) の値を表し，中央の値は事象のモダリティーを，後ろの 3 つは，モダリティー表現自身に対する，過去，現在，未来の時間情報 (+極性) の値を表す．

3.2 精緻化

事実性タグの時間情報 (+極性) に関する 6 つのスロットには，次の 7 種類のタグのいずれかを記述する．

瞬間，状態，反復/継続，始，止，否定，
・(言及なし)

ここでは，原らの「反復」と「継続」を一つにまとめた．

モダリティーに関しては，原らの 16 種類のタグを統合・整理することによって得られた，次の大きく 3 種類，細かく 10 種類のタグを用いる．

事実確定 過去に起きたこと，もしくは現在の状況を表す

断定，伝聞

不確定 確信をもって述べることができないこと，もしくは未来の状況を表す

仮定，推量，疑い，質問，意志・予定，依頼・当為

無関係 事実であるか仮定であるかに関係しない

体，用

3.3 事実性タグ付きコーパス

事実性解析器の学習コーパスとして，精緻化したタグ体系に基づく事実性タグ付きコーパスを作成した．

ブログ記事から，ある商品について書かれたテキストを抽出し，そこに含まれるすべての事象の述語に対して，1 人の作業者が上で述べた事実性タグを付与した．このとき，対象の事象を含む文の前後一文を参照するとともに，誤って述語と判定された補助動詞，機能語相当表現を作業の対象から除外した．

作成したコーパスの一部を表 1 に示す．このコーパスには，異なりで 2646 文，延べ 4417 個の事象が含まれている．上記の作業者と別の作業者との判定者間一致による κ 統計量は 0.68 であった．この事実より，付与されたタグは信頼できると言える．

表 1: 事実性タグ付きコーパスの一部

事象を含む文 (下線は対象述語)	時間	モダリティー	モダリティー時間
プロフィールにもあるとおり、私はレガシィが 好き なんです。	(・, 状態, ・,	断定,	・, 状態, ・)
ヘルシア，これであと少し 安く なったらうれしいんですけどね.	(・, ・, 瞬間,	仮定,	・, 状態, ・)
アジエンス使ったら感想 聞か せてくださいね！	(・, ・, 瞬間,	依頼・当為,	・, 状態, ・)

表 2: タグ付きコーパスによる実験結果

手法	過去	現在	未来	モダリティー
ベースライン	0.58	0.54	0.81	0.66
原ら [1] の解析器	0.66	0.61	0.85	0.77
提案手法	0.75	0.69	0.84	0.73

4. 事象間依存関係を考慮した事実性解析器

4.1 提案手法

事実性解析器の精度向上を目指して，本研究では，その学習モデルとして factorial CRF[2] を用いた．このモデルは，一文内に複数の事象が存在する場合，それらのタグ間の依存関係を考慮してモデルのパラメーターを調整する．本研究では，以下に対する形態素情報と，機能表現辞書 [3] に含まれる意味分類をモデルの素性として用いた．

述語，前後の文節，係り先・係り元文節

4.2 実験

3.3 節で述べたコーパスと上記の手法を用いて事実性解析器の性能評価実験を行った．このコーパスのデータの約 99%において，モダリティーに対する時間情報は”・, 状態, ・” であったため，本実験ではこれに対する評価を行わなかった．

事実性タグ付きコーパスにおいて，ある商品名を含むデータを評価コーパス，それ以外のデータを学習コーパスとして実験を行った．この操作を 3 つの商品名に対して繰り返し，得られた評価値の平均を取った．実験結果のシステムの精度を表 2 に示す．この表には，ベースラインと原らの解析器 [1] の精度も示した．ここで，ベースラインは，コーパスにおいて最も多く出現したタグを解析器が選択し続けた時の精度である．原らの解析器に対する評価値は，彼らの論文から引用した値の平均値である．

4.3 考察

複数事象の依存関係を解析モデルに組み込んだことで，時間情報の解析精度を向上させることができた．

我々の解析器のモダリティー解析精度は，原らの解析器に比べて劣っているように見える．この主な原因は，原らが使用した学習コーパスにおいて「宣言」タグがかなり頻出していたからである．我々のコーパスにおいては，タグ体系の精緻化に基づき，これらのタグが付与されていたデータに対して複数のタグを付与し直した．これにより，我々の評価実験においては，それらのデータに対する分類が難しくなった．

5. 関連研究

事象の時間情報を扱った関連研究に，Pustejovsky らの TimeML[4] がある．これは，事象と時間情報と事象間の関係を表すマークアップ言語である．この TimeML を用いてタグ付けされたコーパスに TimeBank などがある．TimeML の枠組みにおいては，主に，事象間の時間関係の順序づけを重要視しているのに対し，我々の事実性タグ体系においては，事象がそもそも事実なのか，それともそうでないのかということに主眼を置いている．

6. おわりに

本論文では，事実性解析器の精度向上を目指し，事実性タグを事実性の観点から精緻化し，テキスト内で隣り合う事象の相互関係を解析モデルに組み込んだ．

本論文で示した手法は，我々のグループが進めているブログマイニングの検索ツールにすでに組み込まれており，今後一般公開される予定である．また，言論マップ生成システム [5] にも組み込まれる予定である．

謝辞

本研究は，独立行政法人 情報通信研究機構 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」の支援の下に実施した．

参考文献

[1] 原, 乾：“事態抽出のための事実性解析”, 情報処理学会研究報告 2008-FI-89, 2008-NL-183, pp. 75–80 (2008).

[2] C. Sutton, K. Rohanimanesh and A. McCallum: “Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data”, The Journal of Machine Learning Researchf, pp. 693–723 (2004).

[3] 松吉, 佐藤, 宇津呂：“日本語機能表現辞書の編纂”, 自然言語処理, **14**, 5, pp. 123–146 (2007).

[4] B. Boguraev and R. K. Ando: “TimeML-compliant text analysis for temporal reasoning”, Proceedings of the 19th International Joint Conference on Artificial Intelligence, pp. 997–1003 (2005).

[5] 村上, 松吉, 隅田, 森田, 佐尾, 増田, 松本, 乾：“言論マップ生成課題：言説間の類似・対立の構造を捉えるために”, 情報処理学会研究報告 2008-NL-186, pp. – (2008).