経験マイニングのための事実性解析

原一夫　東山昌彦　乾健太郎　松本裕治
奈良先端科学技術大学院大学　情報科学研究科

1 はじめに

ブログに代表される個人型情報発信メディアが爆発的に普及し，個人の行動や成功体験，トラブル，感想など，個人の経験に関する膨大な情報が Web 上に加速度的に蓄積されている．こうした情報は，うまく整理し再構成すれば，個別の状況にあった意思決定やトラブルの回避・解消に有用な「知」の宝庫に変えられる可能性がある．本稿では，商品やサービスなど，様々な事物（以下，トピック）の利用に関する個人の経験情報を広く Web 文書集合から収集し，意味的な分類を行ってデータベース化する「経験マイニング」という新しい課題を考え，その核となる事実性解析について我々の取り組みを報告する．

経験マイニングでは，Web 文書集合から抽出した経験情報を述語項構造に基づく表現形式に構造化するとともに，トピック，著者，事態の評価極性，事実性情報等のきめ細かい情報で索引付けする．これによって，例えば「あるサービスの利用に伴うトラブル」や「ある商品に関心を持ちながらまだ買っていない人」，「ある商品の利用を止めた理由」といった複雑な検索ができるようになり，個人の利用はもとより，企業のマーケティングやリスク管理，行政サービスの評価などの情報源として Web を有効活用できるようになると期待できる．

経験の分類において，事態の評価極性および事実性の情報はとくに重要である．例えば，次の例のように，これによって，例えば，次の例のように，ネガティブな出来事が事実として語られていれば，著者の経験した「トラブル」と解釈できるし，ポジティブな出来事を伝聞形で述べていれば，著者がトピックに関心を持ちながらまだ自分では利用していないことがわかる．

(1) ランプがつかない ときがある
　　negative　　事実

(2) 寝癖がつきにくくなる って友達が言ってました
　　positive　　伝聞

このように，評価極性と事実性の情報を組み合わせれば，例えば個々の事態を「利用成功」「トラブル」「トラブルの未然回避」「トラブルの解消」「満足」「不満」「安心」「心配」といったクラスに分類することができ，これまで主として評価極性（ポジティブ／ネガティブ）だけで分類してきた意見分類を，それを包含する経験分類に一般化することができる．

2 事実性解析

経験マイニングはテキストから事態情報を抽出する情報抽出の問題と見なすことができる．事態抽出の研究は，MUC（Message Understanding Conference）や ACE（Automatic Content Extraction）に代表される問題設定の中で，固有表現抽出，照応解析，抽出パターンの自動獲得等の技術を向上させる一方，述語項構造解析など，抽出する関係や事態の種類を限定せず，テキストに記述される事態を網羅的に抽出する方向にも進んでいる．述語項構造解析については，英語では PropBank や NomBank などの資源を用いた研究が盛んであり[4, 5]，日本語でも例えば Kawahara らや Iida らの研究がある[1, 2]．

その一方で，事態の事実性を同定する研究はあまり進んでいない．事態は，既に事実として起こったこと（テロ事件や企業合併結果など）や真実であると証明されたことだけでなく，まだ計画や仮説でしかない不確実なものも含む．事態の種類を限定しない情報抽出を行うとき，これらを混在させたままの解析結果を提示することは利用者にとってあまり意味がなく，抽出した事態の事実性をテキストから読み取り，解析結果に付与することが望まれる．我々は，経験マイニングという特定の応用を想定することによって，事実性解析に求められる要件を明確にしつつ，可及的に一般性の高い意味解析基盤技術を開発することをねらう．

3 事実性解析の方法

事態は，コアとなる述語（事態表現と呼ぶ）および，それを修飾する時間情報（テンス・アスペクト），極性（成立／不成立），話者態度（モダリティ）から構成されると考えられる．たとえば，

(3) 商品 A は店舗 B で３割 値引き してました

では，事態表現は状態変化を表わす「値引きする」であり，機能表現「て・ます」によって状態変化の結果残存状態というアスペクトが加えられ，「た」によって過去にその事態が成立していたという宣言（テンスとモダリティ）が付加される．別の例，

(4) 明日から飲料水 A を 飲み 始めるつもりです

では，行為を表わす「飲む」という事態表現が，「始める」による行為の開始のアスペクトと「つもり」による意志のモダリティが加えられている．

3.1 時間情報と極性

時間情報と極性（事態の成立／不成立）は，過去，現在（＝記述時刻），未来のスロットからなる３つ組で表す．それぞれのスロットには，｛▲，■，□，↑，↓，×，・｝のいずれかのラベルが入る．単発的な瞬間的事態（状態変化や行為など）の成立を▲，瞬間的事態の反復的継続の成立を■，状態等の継続的事態の成立を□，反復的事態または継続的事態の開始を↑，終了を↓，事態の否定（不成立）を×，言及なしを・で表す．なお，本研究では開始，継続，終了などのアスペクト表現（～し始める，～にハマる，～するのをあきらめる，～するのをやめる，等）は事態とみなさず，主となる事態を補助するための表現と位置づける．

3.2 話者態度

話者態度は，極性以外の心的状態を抽象する．事態成立の真偽を問うもの，すなわち，事態の事実性に関連するもの

　宣言（事実），確信，不確実，伝聞，保留，疑問，反実仮想，反語

と，それ以外

Factuality Analysis for Event Mining.
Kazuo Hara, Masahiko Higashiyama, Kentaro Inui, and Yuji Matsumoto, {kazuo-h, masahiko-h, inui, matsu}@is.naist.jp, Computational Linguistics Laboratory, Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara, 630-0192, Japan.

仮定，予定，可能，願望，意志，質問，推奨，当為をタグとして用意する．一つの事態に対して，伝聞・予定などのように複数のタグが付与されることもある．従来，話者態度を表す表現としては，「よう」「まい」「らしい」のような機能語や「～するべきだ」「～するところだった」のような複合辞が主な分析対象だったが，本研究では，「～と思う」（宣言），「～の感がある」（不確実），「～という話を聞いた」（伝聞），「～というのは信用しかねる」（疑問）など，内容語を含む表現もそれが話者態度を表すなら広く話者態度表現として解釈する．したがって，これを解析するには，従来の機能表現辞書（例えば松吉らの辞書[7]）に加え，話者態度を表しうる表現を広く識別できる資源あるいはモデルの開発が必要である．以下に，いくつか例を示す．

(5) 商品Aはまだ食べたことがない．
行為（利用）－〈×，×，・〉－宣言
(6) 画像はちょっと前からハマって飲んでる商品Aです．
行為（利用）－〈■，■，・〉－宣言
(7) 商品Aを飲んでおなかを壊した人の話を聞いてて，
出来事（ネガティブ）－〈△，・，・〉－伝聞

4 事実性解析のためのタグ付きコーパスの作成

以上のような事態の抽象化を計算機で自動化して行うために，本研究では，まずテキストに出現する事態に対して各情報（時間情報，話者態度）を人手で付与してタグ付きコーパスを作成し，それを訓練データとして用いて解析モデルを学習する方法をとる．そのための予備的な試みとして，次の手順でタグ付けコーパスを作成した．

まず，ドメインを選び（ここでは，価格や買い換え頻度など，やや性質の異なる 2 つのドメインとして，飲料水と自動車を選択），対象商品名を決定した後（飲料水 A，自動車 B），商品名をキーとしてブログ記事を収集し，形態素解析および文節区切りを施す．タグを付与する対象は，動詞および動詞化しているサ変名詞（名詞-サ変接続の直後に動詞-自立がある場合）に限定する．さらに，対象とする商品と直接関係のない表現（e.g. 飲料水 A を買って帰りました），アスペクト表現（e.g. ～し始める），モダリティ表現（e.g. ～と思う，～の話を聞いた），機能語相当表現（e.g. ～ことができる，～ことにする，～といっても，～てもらう，～てみる等）を人手によって同定し，タグ付け対象から外す．こうして，タグ付け対象を特定した上で，事態タイプ，時間情報，話者態度のタグを前後一文を見て読み取れる範囲で付与した．時間情報タグについては，言及なしかどうか定かでない場合は，？付きのタグを用いた（□？，■？，等）．

タグ付けは一人の作業者により行った．同一データに対し，2 週間の間隔を空けて 2 度タグ付けをしてもらい，判定者内一致（intra-rater agreement）による κ 統計量は 0.8 以上を得た．

5 実験

前節のコーパスを使用して，時間情報と話者態度について予測実験を行った．学習モデルは，過去，現在，未来のラベル系列からなる時間情報に対しては Hidden Markov SVM を，話者態度については multi class SVM を使用した[6]．素性は，予測対象の文節，その前後の文節，文全体を区別した上で，品詞と原型を組み合わせたものを用いた．leave-one-out による実験結果（正解率）を表 1 に示す．時間情報の欄には，過去，現在，未来が揃って正解した場合の正解率を記載した．

表 1：作成したタグ付きコーパスを用いた実験結果

ドメイン（データ数）	過去	現在	未来	時間情報	話者態度
飲料水（591）	65.8%	63.6%	88.3%	51.1%	76.5%
自動車（484）	67.1%	58.5%	81.4%	41.9%	77.9%

6 おわりに

本稿では，経験を時間情報，極性，話者態度の観点から抽象化する枠組みを提案した．さらに，ブログ記事からの経験抽出を応用例として想定し，実際にコーパスを作成し，予備的な実験を行った．

我々の目的は，例えば時相論理のような過度に複雑な意味表現を導入することなく，経験抽出/分類のような事態抽出の応用に広く有益で現実的な事実性解析の枠組みを設計し，それを実現する解析モデルを開発することである．本稿ではその最初の試みを報告したが，課題も多い．とくに，3 節で述べた枠組みでは表現できない情報が種々残っており，時間情報，極性，話者態度それぞれの精緻化が必要である．解析モデルについては，今回の実験では個々の事態の事実性解析を独立した問題として扱った．しかし，一連の文脈のなかでいくつかの事態が言及されている場合，それらの事実性の間には，例えば主節が過去であれば従属節も過去になりやすいといった依存関係があると考えられる．今後は，事態間の事実性の依存関係も含めて学習できるようなモデルを設計し，実験を進める予定である．

謝辞

本研究は，文科省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：Web 文書からの個人の経験の抽出と分類」（19024057，代表：乾健太郎），およびニフティ株式会社から支援を受けた．記して深く感謝する．

参考文献

[1] Iida, R., Inui, K. and Matsumoto, Y.: Exploiting Syntactic Patterns as Clues in Zero-Anaphora Resolution, Proc. of COLING-ACL, pp. 625-632 (2006).
[2] Kawahara, D. and Kurohashi, S.: A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis, Proc. of HLTNAACL, pp. 176-183 (2006).
[3] Matsumoto, Y.: Morphological Analysis System ChaSen: Easy-to-Use Practical Freeware for Natural Language Processing, Journal of Information Processing Society of Japan, Vol. 41, No. 11, pp. 1208-1214 (20001115).
[4] Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B. and Grishman, R.: The NomBank Project: An InterimReport, Proceedings of the HLT-NAACL Workshop on Frontiers in Corpus Annotation (2004).
[5] Palmer, M., Gildea, D. and Kingsbury, P.: The Proposition Bank: An Annotated Corpus of Semantic Roles, Computational Linguistics, Vol. 31, No. 1, pp. 71-106 (2005).
[6] Tsochantaridis, I., Joachims, T., Hofmann, T. and Altun, Y.: Large Margin Methods for Structured and Interdependent Output Variables, Journal of Machine Learning Research (JMLR), Vol. 6, pp. 1453 - 1484 (2005).
[7] 松吉俊，佐藤理史：体系的機能表現辞書に基づく日本語機能表現の言い換え，言語処理学会第 13 回年次大会発表論文集，pp. 899- 902 (2007).