経験マイニング： Web テキストからの個人の経験の抽出と分類

乾 健太郎　　原 一夫
奈良先端科学技術大学院大学 情報科学研究科
{inui,kazuo-h}@is.naist.jp

1 経験マイニングという課題

ブログに代表される個人型情報発信メディアの爆発的な普及に伴い，個人の行動，成功体験，トラブル，興味，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つつある．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．しかし，個人が Web 上に発信する経験情報は不均質で無秩序に分散しているため，現在の社会はこれを有効に活用できていない．

こうした背景から我々は，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 文書集合から抽出し，意味的な索引付けを行う新しい技術の研究に取り組んでいる．我々が「経験マイニング」と呼ぶこの課題では，図 1 に示すように，Web 上のテキストから個人の経験情報を，トピック，経験主，事態タイプ，事実性情報，事態表現といったスロットからなる構造化情報として抽出する．このうち事態タイプは，経験の核となる事態表現を入手や利用などの行為，ポジティブ/ネガティブな出来事等に分類する．また，事実性情報は，その事態が実際に起こったことなのか，可能性を述べただけなのかといった，いわゆるテンス・アスペクト・モダリティに相当する情報である．

この技術が現実的な規模で機能すれば，最終的な生成物として期待されるのは，特定の商品（車，携帯電話など）や特定の機能を持った場所（飲食店，病院，温泉など）から行政サービス（子育て支援制度，花火大会など）にいたる様々なトピックに関する膨大な数の個人の経験を集積した経験データベースである．個々の経験は，トピックや経験主，事態タイプ，事実性，経験表現等のきめ細かい情報で索引付けされ，トラブルや要望といった意味的な概念を使って効率的に閲覧できるようになる．また，とくにブログのように一連の記事の著者が特定できる場合は，著者の経験をプロファイル情報として利用することもできる．こうした情報は，著者のバックグラウンドを知り，信頼性を判定する際の手がかりとして利用することができる他，例えば「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人」といった複雑な検索を可能とし，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．

2 何が新しいか？

2.1 技術的焦点

上に述べた経験マイニングの重要なポイントは，経験を分類する基準として特定の利用シーンに特化した基準を仮定するのではなく，次のような一般性の高い意味的な情報によって個別の経験のインスタンスを索引付けする点にある．

(1) a. **トピック：** どの利用物，サービスに関する経験か
　　b. **経験主：** 経験の主体
　　c. **事態タイプ：** 経験情報の核となる事態の種類（e.g. ポジティブ/ネガティブな出来事・状態・性質，入手・利用等の行為）
　　d. **事実性：** (c) の事態の事実性に関する情報（e.g. テンス，アスペクト，極性，モダリティ情報）
　　e. **事態表現：** 経験の核となる事態の表現（典型的には述語項構造）

これらのうち，とくに重要なのは事態タイプと事実性である．例えば次の例のように，ネガティブな出来事が過去または現在の事実として述べられていれば，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，といったことがわかる．

(2) ランプがつかない（ネガティブな出来事） ときがある（事実）

(3) 寝癖がつきにくくなる（ポジティブな出来事） って友達が言ってました（伝聞）

(4) 発売当初は何度か飲んで（利用行為） いたのですが（過去に繰り返していた行為を現在はやっていない）

このように，事態タイプと事実性の情報を組み合わせることによって，トピックへの「関心」や「要望」,「トラブル」といった意味的な検索要求を表現することができ，データベース上に蓄積された経験情報を柔軟に検索することができる．重要な点は，いずれの技術も言語解析の基盤技術として，経験マイニングに限らず，言語の意味解析を必要とする多様な応用に広く適用可能なことが期待できる点である．我々のねらいは，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，意味解析に求められる要件を明確にしつつ，できる限り一般性の高い意味解析の基盤技術を開発することである．

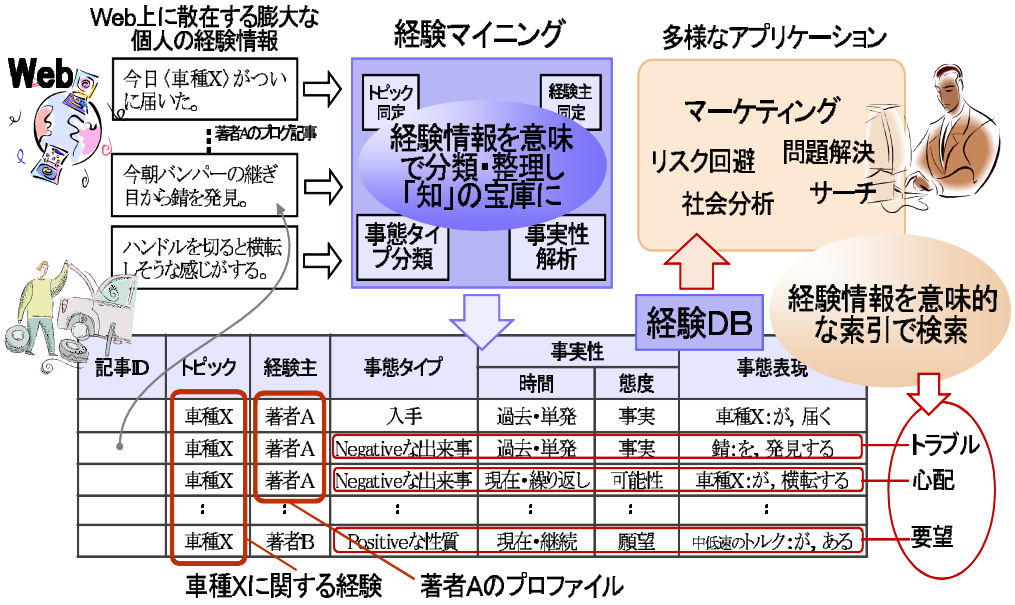

図 1: 経験マイニング

2.2 先行研究との関係

関連する研究分野としては評判分析や情報抽出が挙げられるが，我々がここで考えるタスクは従来の評判分析や情報抽出のスコープに収まらない問題をいくつか含んでいる．

まず，評判分析は，商品やサービス等に対する個人の評価を収集する点で個人の経験の一部を扱っているが，現在のところ「美味しい」や「エンジンが静か」など，明示的に評価を述べた記述から評判情報を抽出するという課題設定 [4, etc.] に留まっており，成功体験やトラブルといった個人の経験を広く収集するものではない．例えば (2) の「ランプがつかない」という事態は〈ネガティブな出来事〉と解釈できる．こうした評価極性（ポジティブかネガティブか）を暗に持つ事態（”evaluative factual”）の認識は，経験の意味的な分類には欠かせない技術と考えられるが，評価情報抽出の研究はこうした問題に焦点を当てるに至っていないのが現状である [2]．

また，我々のタスクは，米国の研究プログラム MUC[1]や ACE[2]におけるイベント抽出課題とも関連性が高い．例えば，ACE における event type および sub-type の同定は，イベントの分類軸が我々の考える事態タイプ分類とは異なるものの，同類の問題と考えることができる．ただし，我々のタスクでは次節で述べるように事実性情報に関して詳細な分類を想定するのに対し，現状の ACE では次のような粒度の解析しか行っていない．

(5) TENSE: Past, Present, Future, Unspecified
POLARITY: Positive, Negative
MODALITY: Asserted, Other

[1]Message Understanding Conference
http://www-nlpir.nist.gov/related_projects/muc/
[2]Automatic Content Extraction
http://www.nist.gov/speech/tests/ace/

GENERICITY: Generic, Specific

事実性解析については，テキスト中の事態情報の注釈言語として開発された TimeML[9] によって規定される時間解析タスクとも関連が深い．ただし，事実性に関する限り，TimeML は個別のイベントインスタンスについて次のような文法的情報を付与するに過ぎず，我々が想定するような意味的な解析を要求するものではない．

(6) POS: Adj, Noun, Verb, Prep, Other
TENSE: Future, Infinitive, Past, etc.
ASPECT: Progressive, Perfective, etc.
POLARITY: Positive, Negative
MODALITY: must, should, may, etc.

この他，事実性の解析に焦点を当てた研究は，筆者らの知る限り Medlock ら [7]，Zhou ら [13] の研究などが散見される程度であり，この分野の発展が急がれる．

以下，事態タイプ分類と事実性解析について我々の現在の試みを簡単に報告する．

3 事態タイプ分類

1 節で述べたようなアプリケーションを想定すると，経験データベースに対するユーザからの検索要求には，例えば次のようなものが考えられる．ここで「対象」とは，特定の商品やサービスなど，1 節で述べた「トピック」の対象を指す．

(7) 対象に関心を持っている人
対象を欲しいと思っているが，未入手の人
対象を入手あるいは利用した人
対象の入手・利用に関するトラブル
対象に対する満足，安心，期待の記述
対象に対する不満あるいは不安の記述
対象の継続的な利用の実績または意志の記述

対象の入手あるいは利用の中止
対象の推薦する記述またはその反対

これに対し我々は，試験的に 3 種類の商品（清涼飲料水，自動車，シャンプー）に関する記述をブログ記事から収集・分析し，上のような想定検索要求に照らしてどのような事態タイプを用意すべきかを検討してきた．その結果，現時点で少なくとも次のような観点および粒度の分類が必要であることがわかっている．

(8) a. 評価・感情： トピックに関して経験主が持つ主観的評価および感情．それぞれ評価極性を持つ
関心（目が離せない），好悪（お気に入りだ），体験による評価（美味しい，重宝する），伝聞による評価（人気だ），感情
b. 出来事： トピックの入手・利用等に伴って起こる出来事や状態
評価極性を持つ出来事（壊れる，騙される，髪に腰が出る），入手・利用の可否に関する出来事・状態（慣れる，発売される）
c. 行為： トピックに関して経験主が意図的に行う行為．評価極性なし
情報収集（資料請求する），入手（買う，入庫する），利用（食べる，試乗する），利用中止（解約する），決定（選ぶ），検討（検討する）

これらのうち，(8a) の評価・感情については，例えば小林らの評価表現辞書 [4] など，既存の資源である程度網羅できると考えられる．また (8c) の行為についても，例えば入手や利用の表現を集めた Torisawa の資源 [11] や我々のグループで開発中の事態オントロジー [3] などが利用できるであろう．

一方，(8b) 出来事のサブタイプである「評価極性を持つ出来事」は他に比べて顕著に大きなクラスタをなすもので，このタイプのインスタンスを識別する計算モデルの構築が技術的には最も重要な課題である．これについて我々はこれまでのところ，事態表現の最も典型的な形式である「名詞＋格助詞＋述語」の評価極性の分類に焦点を当て，名詞単体の評価極性の知識が鍵になることを明らかにするとともに，名詞の評価極性をコーパスから自動獲得する実験を行い，既存の方法を大幅に改善できることを確認している [1].

4 事実性解析

事実性解析の問題設定についても，前節で述べたブログ記事の分析の中で検討し，現在のところ以下に概観するような仕様でコーパス作成およびモデル学習を行っている．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．

4.1 問題設定

事態の事実性情報は，事態の時間に関する情報（テンス，アスペクト）と極性（成立／不成立），および話者態度（モダリティ）から構成されると考えられる．

時間情報と極性 時間情報と極性は，〈過去，現在，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．▲は瞬間的事態（状態変化や行為など）の単発的な成立，■は瞬間的事態の反復的継続の成立，□は状態等の継続的事態の成立，↑と↓は反復的事態または継続的事態の開始および終了，×は事態の否定（不成立）あらわす．また，当該の時間における事態の成立/不成立にコミットしていない場合は”•” でそれを表現する．

話者態度 話者態度はいわゆるムード/モダリティに概ね相当する情報である．文献 [5] などの既存の分類，および前述のブログ記事の分析を基に，現在のところ次のような分類を想定している．

(9) 宣言，確信，不確実，伝聞，保留，疑問，反実仮想，仮定，予定，可能，願望，意志，質問，推奨，当為

話者態度の分類も，事態タイプの分類や時間情報の分類と同様，どのような観点と粒度で分類すべきかについて一般的な最適解を見つけ，それを証明することは極めて難しい．ここでも，具体的な応用をいくつか想定し，それらにまたがる公約数的な要件を抽出するという経験的な方法をとらざるを得ないだろう．上記の分類は今後も修正を加えていく予定である．

話者態度についてもう 1 点重要なことは，対象とする言語形式の範囲の決め方である．従来のモダリティ研究では，「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 [8]．しかし，情報抽出のような応用では，これらの言語形式に収まらない話者態度表現を広くカバーする必要がでてくる．例えば，「〜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，内容語を含む表現の中にも話者態度を表すものが少なくない．我々の事実性解析ではこうした表現を広く話者態度表現と捉える．これには，従来の機能表現辞書（例えば松吉らの辞書 [6]）に加え，より多様な表現を識別できる資源あるいはモデルの開発が必要である．

いくつか例を示す．

(10) 商品 A はまだ食べたことがない．
行為（利用）－〈×，×，•〉－宣言

(11) 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

(12) 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

4.2 解析モデルと評価実験

3 節冒頭で述べた 3 つのドメイン（清涼飲料水，自動車，シャンプー）について表 1 に示す規模のタグ付きコーパスを人手で作成し，時間情報と話者態度の予測実験を行ったので，簡単に報告する．解析モデルとしては次の 2 種類を試みた．

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）	モデル	過去	現在	未来	話者態度	事実性判定
飲料水（1134）	SVM	49%	52%	72%	82%	81%
自動車（725）	SVM	38%	48%	74%	84%	82%
シャンプー（958）	SVM	53%	63%	80%	84%	84%
飲料水（1134）	Factorial CRF	66%	61%	90%	83%	80%
自動車（725）	Factorial CRF	75%	59%	88%	85%	84%
シャンプー（958）	Factorial CRF	68%	58%	90%	85%	84%

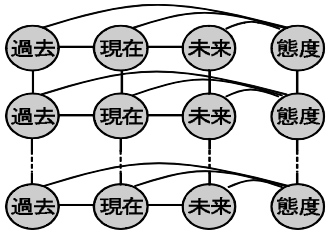

図 2: 時間情報と話者態度および複数の事態インスタンスの依存関係を表現できる Factorial CRF モデル

(a) 事態ごとに独立に解くモデル: 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[12] を，話者態度については multi class SVM を使用する（表 1 の SVM）

(b) 隣接する事態間の依存関係を考慮するモデル： 同一文に現れる複数の事態の間の依存関係を図 2 のようなグラフで表現し，Factorial CRF[10] 上にモデル化する（表 1 の Factorial CRF）

なお素性には，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を用いた．

タグ付き事例を３つのドメインで分割し，３分割交差検定を行った結果を表 1 に示す．Factorial CRF が SVM より全般的に良い結果を得ており，事態間の依存関係のモデル化が精度向上に貢献することがわかる．例えば，

(13) 店頭で見本の匂いをかいでみて，やめた．

の「かいで」の時間情報は「やめた」の時間情報に依存すると考えられるが，Factorial CRF のモデルではこうした依存関係を自然に組み込むことができる．

表 1 が示すように，我々の事実性解析タスクは，上述のような比較的単純な教師あり学習として解くだけでも一定の精度が得られる問題設定になっており，過度に複雑な非現実的設定に陥っていないことはある程度確かめられた．今後は，問題設定をさらに洗練するとともに，機能表現辞書 [6, etc.] などの既存の資源を効果的に利用し，精度の向上をはかりたい．

5 おわりに

Web 上に分散的かつ大量に語られている個人の経験の情報を時間情報，極性，話者態度の観点から索引づけする経験マイニングの枠組みを提案し，中心的な部分問題である事実タイプ分類および事実性解析について我々の取り組みを報告した．今後は，個別の要素技術の水準を引き上げるとともに，企業のマーケティングや自治体の地域モニタリング等の具体的応用を想定した経験マイニング・プロトタイプシステムの開発に取り組む予定である．

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．

参考文献

[1] 東山昌彦, 乾健太郎, 松本裕治. 述語の選択選好性に着目した名詞評価極性の獲得. 言語処理学会第 14 回年次大会予稿集, 2008.

[2] 乾孝司, 奥村学. テキストを対象とした評価情報の分析に関する研究動向. 自然言語処理, Vol. 13, No. 3, 2006.

[3] 乾健太郎. 事態オントロジー： 言語に基づく推論のためのコトに関する基本知識. 言語処理学会第 13 回年次大会ワークショップ「言語的オントロジーの構築・連携・利用」論文集, pp. 27–30, 2007.

[4] N. Kobayashi, K. Inui, and Y. Matsumoto. Opinion mining from web documents: Extraction and structurization. *Journal of the Japanese Society for Artificial Intelligence*, Vol. 22, No. 2, pp. 227–238, 2007.

[5] 益岡隆志, 田窪行則. 基礎日本語文法（改訂版）. くろしお出版, 1992.

[6] 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.

[7] B. Medlock and T. Briscoe. Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007.

[8] 森山卓郎, 仁田義雄, 工藤浩. モダリティ. 日本語の文法３. 岩波書店, 2000.

[9] J. Pustejovsky, J. Castano, R. Ingria, R.Saur ＼ i, R. Gaizauskas, A. Setzer, G. Katz. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *the Fifth International Workshop on Computational Semantics*, 2003.

[10] C. Sutton. GRMM: A graphical models toolkit. http://mallet.cs.umass.edu, 2006.

[11] 鳥澤健太郎. 対象の用途と準備を表す表現の自動獲得. 自然言語処理, Vol. 13, No. 2, 2006.

[12] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research (JMLR)*, Vol. 6, pp. 1453 – 1484, September 2005.

[13] L. Zhou, G.B. Melton, S. Parsons, and G. Hripcsak. A temporal constraint structure for extracting temporal information from clinical narrative. *Journal of Biomedical Informatics*, Vol. 39, No. 4, pp. 424–439, 2006.