みんなの経験：ブログから抽出したイベントおよびセンチメントのＤＢ化

阿部修也　江口萌　隅田飛鳥　大崎梓　乾健太郎
奈良先端科学技術大学院大学 情報科学研究科
shuya-a, megumi-e, asuka-s, osaki, inui}@is.naist.jp

1 背景

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，こうした個人が発信する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．この問題に対処する研究分野としては情報抽出と評判分析が挙げられるが，情報抽出の研究はモノの属性（e.g., 飲食店の住所や営業時間）や特定の領域の特定のイベント（e.g., 企業の人事情報）など，極めて狭い領域に対象を限定するものがほとんどで，個人の経験を広く収集し再構成するには至っていない．一方，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 [5, 8, 7] が，現在のところ「美味しい」や「エンジンが静か」など，明示的な評価，すなわちセンチメント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] に留まっており，やは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「ランプがつかないときがある」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）を暗に持つこうしたイベント（“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].

2 経験マイニング

このような背景から，図 1 に示すように，商品やサービスなど，様々な事物（トピック）の利用に関するセンチメントの情報やセンチメンを暗に含むイベントの情報を広く Web 文書集合から抽出し，意味的な索引付けを行う新しい課題を考える [3]．我々が「経験マイニング」と呼ぶこの新しい課題の核は経験抽出と経験分類からなる．経験抽出では，例えば「エリシオンが届いた」という文章断片から，「『エリシオン』という車種の車が著者のもとに『届いた』」という著者の経験が記述されていることを認識し，述語と項の組（述語項構造）の形式で抽出する（経験表現）．経験分類では，抽

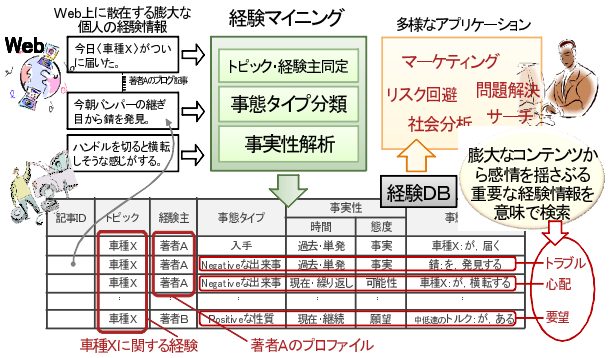

図 1: 経験マイニング

出した経験を事態のタイプで意味的に分類し，それが入手や利用などの行為なのか，ポジティブあるいはネガティブな出来事なのかといった情報を区別する．また，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事実性の観点からも解析し，これらの意味的な情報で経験を索引付けする．

2.1 基本方針

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，事態タイプや事実性のような一般性の高い意味的なタグによって経験情報を索引付けする点にある．具体的には，次の４つの軸を考える．

(a) トピック： どの利用物に関する経験か
(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. 既遂の事実，未遂の願望，伝聞）

これによって，次の例のように，ネガティブな出来事が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ことがわかる．

- ランプがつかないときがある
  〈ネガティブな出来事〉－〈既遂の事実〉
- 寝癖がつきにくくなるって友達が言ってた
  〈ポジティブな出来事〉－〈既遂の事実〉－〈伝聞〉

図 2: ワインに関連する経験情報のサマリ

このように，4 つの一般的な軸を用意してこれらを用途に応じて組み合わせることで，様々な観点からの経験情報を検索することが可能になる．

2.2 応用

この技術が現実的な規模で機能すれば，最終的に得られるのは，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，抽出元文書へのリンクとともに格納される．

これが実現すれば，Web ユーザは特定のトピックに関する他人の経験を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収集し閲覧できるようになり，企業や自治体にとってはマーケティングやリスク管理の情報源として利用することができる．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，例えば車種 X を実際に購入したユーザの記事だけをサンプルしたり，キャンペーンで伝え聞いた評判と実際に商品を利用した感想の区別して分析するなど，マーケティング等においても有用な情報になると期待される．

2.3 デモサイト「みんなの経験」

これらを体現するために，文書集合から抽出した経験情報をデータベース化し，Web ブラウザから検索できるデモサイト「みんなの経験」を開発し，公開した[1]．

[1]http://minna.naist.jp/

現在のデータベースには約 4 千万件の経験情報が格納されている[2][3]．

図 2 は，入力されたトピックに関して抽出された経験情報のサマリを表示しているスナップショットである．経験のインスタンスはそれぞれ「買う／利用するつもり」「買った／利用した」「良さそうな噂」などに分類されている．この分類は，事態タイプと事実性情報の組み合わせからなる規則で実現している．例えば，「悪かった出来事」は，〈ネガティブな出来事〉＋〈過去または現在〉＋〈断定〉の組み合わせ，あるいは〈ネガティブな出来事〉＋〈過去または現在の否定〉＋〈断定〉の組み合わせなどで規定できる．また，このサマリ表示では，「ボルドーワイン」と「ブルゴーニュワイン」についての経験の分布や主要な経験表現を比較することもできる．

3 Web からの経験獲得

本節では 2 節で説明した経験情報を Web から獲得する手法を説明する．「みんなの経験」はこの手法に基いている．

3.1 経験情報

我々の経験マイニングでは 2.1 で説明した「トピック」「経験主」「事態タイプ」「事実性」の 4 つ組からなる経験情報をブログから獲得し表 1[4]のようなデータベースを作成する．データベースの「話者態度」は著者の経験か著者以外の経験かを表わし，著者の経験であれば「ブログの URL」を著者とみなすことで 2.1 の「経験主」を表現する．また，データベースの「時間」，「肯定／否定」，「話者態度」は 2.1 の事実性に相当する．「みんなの経験」はこのデータベースを参照している．

3.2 手法

ブログとトピックが与えられたとき表 1 のような経験情報を獲得する手順を示す．

1. トピックを含む記事を発見する．「ブログの URL」，「記事の URL」，「文」，「トピック」を記録する．
2. トピックに対応する経験表現を取得する（関連性解析）．「経験表現」を記録する．
3. 経験表現の事態タイプを判定する（事態タイプ解析）．「事態タイプ」を記録する．
4. 経験表現の事実性を判定する（事実性解析）．「時間」，「肯定／否定」，「話者態度」を記録する．

[2]Wikipedia に登録されている約 50 万件の項目のうち，商品や観光地，制度やサービスなど，広く利用物と見なせる項目約 20 万件を，Wikipedia のカテゴリ情報を利用して選別し，経験抽出の対象トピックとした．

[3]2007 年 5 月から 2008 年 10 月までの 18ヶ月分のブログ記事約 2 億 7 千万記事を抽出源として利用した．ただし，これにはスパムと考えられるページが多数含まれており，とくに他の記事の一節のコピーあるいは一部書き換えたものが目立つ．そこで，経験抽出の結果を走査して類似表現が多数出現するものを削除するなど，いくつかのヒューリスティクスを実装し，スパムの可能性が高い情報を削除してある．

[4]例文中のトピックを 下線，経験表現相当を**ボールド体**で装飾している．

表 1: 経験情報データベース

ブログの URL	記事の URL	文	トピック	経験表現	事態タイプ	時間	肯定／否定	話者態度
http://.../hai	.../090110	商品Ａ を使うと寝癖が**つきにく**くなるって友達が言ってた.	商品Ａ	つきにくい	ポジティブ	過去／現在	肯定	断定－伝聞
http://.../bal	.../090110	バレーボール で一回も**勝つ**ことなく終りを迎えた.	バレーボール	勝つ	ポジティブ	過去／現在	否定	断定
http://.../atm	.../090113	月末の**混み合った**ＡＴＭ に並び、順番を待つ。	ＡＴＭ	混み合う	ネガティブ	過去／現在	肯定	断定
http://.../sfm	.../090111	ＳＦ 的な要素のある話は好きだから、**楽しめそう**。	ＳＦ	楽しめる	ポジティブ	未来	肯定	推量
http://.../tmt	.../090112	トマトジュース を夜な夜な**飲んでいます**。	トマトジュース	飲む	利用	過去／現在	肯定	断定
http://.../bkk	.../090112	バイキング で**食べ過ぎ**て …	バイキング	食べ過ぎる	利用	過去／現在	肯定	断定

3.3 事態タイプ解析

2.2 で述べたアプリケーションからの検索要求には，次のような事態を想定することで対応できると考えた．

(1) 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
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
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

このカテゴリの表現に対して〈ポジティブ〉,〈ネガティブ〉,〈利用（購入）〉の 3 種類のラベルを付与した約 6 万表現からなる辞書を作成した．この辞書は次に例を示すように名詞と述語を組み合わせた事態表現や評価極性を暗に持つ事態表現を含む．

- ポジティブ：レベルが高い，力強い，奥深い
- ネガティブ：価格が高い，暴落した，盗まれた
- 利用（購入）：行った，飲む，（ワインを）開けた

この辞書は，東山らの手法 [1] を用いた結果の一部を人手でクリーニングすることによって作成した．経験表現の事態タイプの判断はこの辞書と照合することで行なう．

3.4 事実性解析

本研究では森田らの事実性解析器 [10] で経験表現を解析した結果と係り受け関係に基く規則を適用した結果を組み合わせて事実性を推定する．森田らの解析器は述語とその周辺の文脈を入力として，事態の時間に関する情報（過去／現在，未来の 2 値）と事実極性（肯定，否定の 2 値），および話者態度から構成される事実性を出力する．話者態度は次のラベルから構成されている．

- 断定，断定－反実，推量－過去，予定，推量，当為，働き（命令，当為，依頼，願望，勧誘），意志，欲求，問いかけ，仮定

さらにこれらに直交するラベルとして〈仮想〉,〈伝聞〉がある．このラベルは佐尾らの定義 [9] に従う．このようなラベル体系として他に，Prasad ら [6] が Penn Discourse TreeBank 中の談話関係とその項に対して情報の発信源や話者態度情報を付与した attribution タグがある．

3.5 関連性解析

トピックと経験表現の関連性の判定については，現在のところ次のようなヒューリスティックを用いている．この手法の洗練は今後の課題である．

- 経験表現が 3.3 で述べた辞書に含まれている．
- トピックと経験表現が同一文内に存在し，係り受け木においてトピックと経験表現が先祖と子孫の関係になっている．
- トピックと経験表現の間に存在する文節の数は N 文節未満である[5]．
- 同一文内において任意のトピックと頻繁に共起する経験表現をそのトピックの経験表現の候補とみなし，それ以外の経験表現をそのトピックの経験表現候補と見なさない[6]．

4 評価

1ヶ月分のブログに対して我々の手法を適用して経験情報を獲得した結果を人手で評価した．このとき表 1 のラベルの中で「みんなの経験」で用いているラベルのみを評価した．

4.1 設定

1ヶ月分のブログから既知のスパムを除いて ChaSen[7]で形態素解析，CaboCha[11] で係り受け解析を行ない，Wikipedia 日本語版[8]の見出し語をトピックとして我々の手法を適用して経験情報を獲得した．ここから，1ヶ月文のブログの経験情報における頻度が 100〜1000 回となるトピック集合を作成し，この集合をトピックとする経験情報からランダムに経験情報を抽出した．

後に述べる基準 (a)〜(c) の精度を測るのに相応しくない誤りを含んだ事例を人手で除いた．例えば，3.3 で説明した検索要求の対象とならないトピックや曖昧性のあるトピックを含む事例を除いた．他に複合語の部

[5]実験では $N = 8$ とした．

[6]実験では，18ヶ月文のブログ記事において同一文内で任意のトピックと共起する経験表現の候補の合計頻度が共起した全ての経験表現の合計頻度の 3/4 程度になるように高頻度順に経験表現候補を選んだ．例えばあるトピックの経験表現の頻度が，経験表現 P_1 の頻度は 75 回で，経験表現 P_2 の頻度は 25 回であるとき，P_1 のみが候補となる．

[7]http://chasen-legacy.sourceforge.jp/

[8]http://ja.wikipedia.org/

表 2: 評価結果

評価＼条件	全事例	(b) を満たす事例
(a)	0.83 (393/472)	-
(b)	-	0.98 (386/393)
(c1)	-	0.86 (339/393)
(c2)	-	0.99 (388/393)
(a)(b)(c)	0.71 (337/472)	0.86 (337/393)

分文字列がトピックである場合は文意から部分文字列がトピックになりえない事例を除いた．

ここで残った事例のみを次の (a)〜(c) の 3 つの基準の評価を人手で実施した．

(a) 経験表現がトピックに関する経験を述べているか，すなわちトピックと経験表現の関連性の正誤．例えば，「A駅$_{\times}$ 前の喧噪から離れた川沿いの B店$_{\bigcirc}$ で**くつろいだ**.」の「くつろぐ」は「B店」に関連する経験とは解釈できるが，「A駅」に関連する経験とは解釈できない．

(b) 事態タイプ（〈ポジティブ〉, 〈ネガティブ〉, 〈利用（購入）〉, 〈事態タイプなし〉）の分類の正誤．

(c) 次の 2 つの観点からの事実性．

(c1) 経験表現の時間が〈過去／現在〉かつ話者態度が〈断定〉であるか，〈それ以外〉の 2 値分類の正誤．

(c2) 文脈における経験表現の事実極性（〈肯定〉, 〈否定〉の 2 値）の正誤．

基準 (a) を満たしていない経験情報を基準 (b) で評価することはできず，また我々の実験の目的を考慮すると基準 (a) を満たしていない経験表現を基準 (c) で評価する理由はない．そこで基準 (a) を満たした経験情報のみを基準 (b)，(c) で評価した．

4.2 結果

精度を表 2 に示す．事態タイプの精度は 0.98，事実性の精度は 0.86，0.99 と高く，我々の事態タイプ解析（3.3）と事実性解析（3.4）が経験情報獲得において有効に機能していることを示している[9]．さらに，事態タイプと事実性が同時に正解した場合の精度も 0.86 と高く，この 2 つの解析器は組み合わせた場合も有効に機能している．一方で関連性解析の精度は 0.83 とこれらの結果と比較して低く，関連性と事態タイプと事実性が同時に正解した場合の精度は 0.71 と特に低い．この結果は，経験獲得の問題において関連性解析の精度向上が重要な課題であることを示唆している．

トピックと経験表現の間に関連性がない事例を関連性解析機が誤って関連性ありと判定した事例を挙げる．

(a) 商品Aはこの世の全ての ファーストフード$_{\times}$ の中で一番**美味しい**と思う。

一般的に「ファーストフード」と「美味しい」の間には強い関係があるにも関わらず，事例 (a) のトピックと経験表現の間に関係はない．この例が示唆することはトピックと経験表現だけを情報として用いても関連性を正しく判断できない事例が存在するため，周辺の文脈を情報として用いる必要があるということである．

[9] 事実性の精度が高い理由のひとつは，粗い分類で事実性を評価したという点にある．より詳細な分類での事実性の評価は今後の課題である．

5 今後の展望

実験結果で問題となった，文書中の経験表現とトピックの関連性を判別する関連性解析を改善する．一方，実験では高い精度を得た事態タイプ解析と事実性解析であるが，「みんなの経験」に適用した結果を見るとより高い精度が必要であると思われるため，こちらも改善する．また現在はトピックと経験表現が同一文内に出現する経験のみを獲得しているが，トピックと経験表現が文の境界を越えて共起する場合も関連性解析を拡張することで獲得したい．

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）とニフティ株式会社およびアクセラテクノロジ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] 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 *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 314–321, 2008.

[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] 小林のぞみ, 乾健太郎, 松本裕治. 意見情報の抽出/構造化のタスク仕様に関する考察. 情報処理学会研究報告 NL-171, pp. 111–118, 2006.

[6] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Aravind Joshi, and Bonnie Webber. Annotating attribution in the penn discourse treebank. In *Proceedings of the COLING/ACL Workshop on Sentiment and Subjectivity in Text*, pp. 31–38, 2006.

[7] J Wiebe, T Wilson, and C Cardie. Annotating expressions of opinions and emotions in language. In *Language Resources and Evaluation*, Vol. 39, pp. 165–210, 2005.

[8] 宮崎林太郎, 森辰則. 製品レビュー文に基づく評判情報コーパスの作成とその特徴の分析. 情報処理学会研究報告 2008-NL-187, 第 15 巻, pp. 99–106, 2008.

[9] 佐尾ちとせ, 江口萌, 松吉俊, 乾健太郎. 日本語文のモダリティ・極性情報を捉えるために. 言語処理学会 第 15 回年次大会 (to appear), 2009.

[10] 森田啓, 佐尾ちとせ, 松吉俊, 松本裕治, 乾健太郎. テキスト情報の事実性解析. 第 7 回情報科学技術フォーラム（FIT2008）, 第 2 巻, pp. 259–260, 2008.

[11] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2006.