E-012

# ブログからの未来に関する表現の抽出
# Extraction of future information from blogs

坂井 俊之† 佐藤 吉秀† 川島 晴美† 奥田 英範†
Toshiyuki Sakai Yoshihide Sato Harumi Kawashima Hidenori Okuda

## 1. はじめに

ブログには，例えば予想や予定，願望などの未来に関する表現が含まれる．このうち，他者の予想は閲覧者が自身の行動を決定する上で参考となり，予定や願望はマーケッターがブロガーの行動を推測することに対して有用であると考えられる．このように，未来に関する表現には有用なものがあると言える．本稿では，時間，文末，可能性表現等を用いた，ブログ記事中の未来に関する表現を含む文の抽出手法を提案する．

## 2. 関連研究

Mani[1]らは，文に時間を割り当てる方法として，次のような方法を提案している．1 つは，yesterday や，last month などの相対的な時間表現があった場合，文章のタイムスタンプから時間の差を取る方式である．２つ目は，The action taken Thursday などのように，動詞の時制と曜日から時間を推定する方式である．これらは，文書にタイムスタンプが付いていることを前提としている．

一方，上嶋ら[2]は，タイムスタンプ付き文書群を用いて，タイムスタンプの付いていない文書の時間を推定している．方式としては，EM アルゴリズムによって文章のクラスタリングを行ない，同じクラスタ内の文書のタイムスタンプ分布から，時間の推定を行なっている．

また，野呂ら[3]は，ブログ記事中のイベントが発生した時間帯として，朝，昼，夕，夜という，４つの時間帯を推定している．例えば，「朝から自転車で郵便局に行く」という文が存在した場合，次の「郵便局の帰りに某ショップによる」という出来事は，「郵便局」が「朝」の連想語であるとし，朝の出来事だと判断する方式である．加えて，ブログ記事内の時間の流れも考慮している．

上嶋らや野呂らが過去を対象にしているのに対し，我々は，ブログ記事から未来に関する表現を文単位で抽出する．また，Mani らとは異なり，明示的な時間表現がない未来表現に対しても，抽出を行なう．提案手法としては，時間表現の他に，文末表現等も用いるという点や，上嶋らとは違い，同じ内容を表したタイムスタンプ付き文書を必要としないという点において，上記の関連研究とは異なる．

## 3. 未来に関する表現の分類

ブログ中の未来に関する表現を調査するため，まずは，ランダムに選んだブログ記事から，未来について述べている文を人手で抽出した．次に，抽出した文を，文が指し示す意味によって分類したところ，以下のような 8 分類が得られた．それぞれの分類と代表例を示す．未来記号は，競馬予想等で本命を表わす記号などに対応している．

†日本電信電話株式会社 ＮＴＴサイバーソリューション研究所 NTT Cyber Solutions Laboratories, NTT Corporation

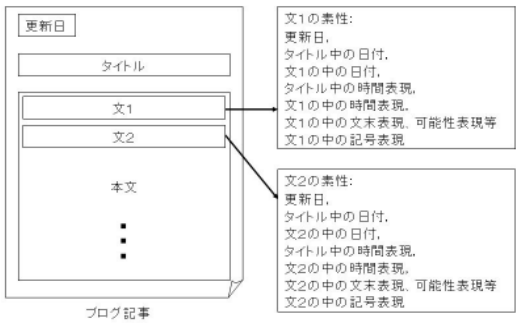


図１. 素性の種類

・未来予定：「来週から長期休暇に入る．」
・未来予想：「日本が確実に白星スタートをしそう．」
・未来願望：「でも，せめて面接までは進みたい．」
・未来心情：「あーあ年末つまんなくなるな・・・」
・未来呼掛：「インフルエンザに気をつけましょう．」
・未来疑問：「原油高は何時まで続くのでしょうか？」
・未来伝聞：「FNS では Christmas Night も歌うらしいし」
・未来記号：「◎ノワールシチー」

これらより，未来か否かの判定には，時間表現だけでなく，「～しそう」等の，事象の不確定さを表すような文末表現なども有効であることが伺える．次章では，未来に関する表現を抽出する方式について述べる．

## 4. 未来に関する表現の抽出

提案手法では，SVM[4]を用いた機械学習によって未来に関する表現を抽出する．3 章の分類時に得た知見を元に，学習に用いる素性を以下のように抽出した．

### 4.1 学習に用いる表現

学習には，以下の表現を文毎に素性として与える(図１).

・更新日（年月日，曜日）
・タイトル中の日付表現（年月日，曜日）
・本文中の処理対象文の日付表現（年月日，曜日）
・タイトル中の時間表現
・本文中の処理対象文の時間表現
・本文中の処理対象文の文末表現，可能性表現等
・本文中の処理対象文の記号表現

まず，更新日は，ブログ記事が投稿された更新日であり，年，月，日，曜日の 4 つを素性とした．

次に，タイトル中の日付表現は，更新日の各年月日の値より，タイトル中の年月日の値を別々に減算し，＋，－，0 をそれぞれ past, future, now の表記に対応させ，素性とした．例えば，更新の日付が 2008 年 6 月 6 日で，タイトル中の日付が 2007 年 6 月 13 日であれば，それぞれを減算することで，1（年）0（月）-7（日）となるため，素性としては，past（年）now（月）future（日）となる．年月日を別々の素性とするのは，必ずしもタイトル中に年月

日の全てが記入されるとは限らないためである．そして，曜日については，月曜～日曜の表記を，そのまま素性に加えた．本文中の日付表現についても更新日時からの減算を行なうが，文中の複数の日付表現に対応するため，future, past, now となった回数を素性ベクトルとした．

次に，時間表現について述べる．時間表現は，文が指し示す時間を表わす上で重要である．そこで我々は，形態素解析ツール JTAG[5]を用いてブログ記事に付与された品詞のタグから時間表現を抽出し，更にそれに，他のブログ記事から人手で抽出した時間表現を加え，拡張した．時間表現の例としては，「明日」のような現在からの相対時間，「季節」や「国民の休日」などの特定期間を示す表現，「長年」などの期間の長さを表す表現が挙げられる．そして，これらの時間表現の内，タイトル中の表現については，本文と別扱いにするため，出現の有無として，0 か 1 の 2 値を素性として加え，本文中の時間表現については，出現した回数を素性として加えた．

未来表現の判定に寄与する表現は，時間表現だけではない．例えば，「～だろう」等の文末表現，「きっと」などの可能性を示す表現，「～たら」などの仮定表現，「どうか～」，「どれだけ～」等の願望・疑問表現なども未来表現の判定に寄与する．これは未来が不確定であることに起因し，物事を断定しているか否かが，未来表現の判定に有用なためである．上記の表現は，タイトル中に存在することが少なかったため，本文中に関してのみ，出現の有無に対し，0 か 1 の値を素性として加えた．ただし，文末表現に関しては，句読点を含めずに，文末から 10 形態素以内に存在する表現のみを素性とした．

最後に，本文中の記号について述べる．これは競馬予想等で本命を表わす記号などに対応する．記号に関しては，◎，○，△，▲，×の５種類について，文中の出現回数と，ブログ記事全体の出現回数を素性として加えた．

### 4.2 素性の重み付け加算

未来表現には，必ずしも時間表現が出現するわけではない．例えば，次の例を考えてみる．

・「明日は運動会だ．どうか晴れますように．」

２つの文を見ると，どちらも未来表現であるが，第１文には時間表現が明示的に記述されているのに対して，第２文には時間表現が記述されていない．これは，第２文が，第１文の指し示す時間に関して，継続した事柄を述べているためである．このように，未来表現の判断には，周囲の文脈の考慮が必要であることが分かる．そこで，本文中の時間表現に関しては，判定対象文より前の文に出現する時間表現の素性を，文間距離に反比例した重みを付けて加え合わせた．これは，判定対象文から離れるほど，判定対象文に対する影響が小さくなると考えたためである．重み付け加算後の時間表現の素性の重み t_add(s)は以下のようになる．

$$t_{\_add}(s) = \sum_{k=1}^{s} (t(k) \times 0.5^{|s-k|}) \quad \cdots (1)$$

ここで，k，s はそれぞれ，ブログ記事中の文の位置と判定対象文の位置を示す．t(k)は，k 番目の文における，重み付け加算前の時間表現の素性の重みを表す．また，0.5 の項は，判定対象文から離れるほど，他の文からの影響が小さいことを示す．

## 5. 実験

### 5.1 実験条件

ランダムに収集したブログ記事から，未来表現の抽出を行なった．学習に用いたのは，未来表現 638 文と，未来以外の表現 1904 文である．素性に関しては，時間表現は 638 素性，文末表現等は 156 素性を用いた．また，未来表現の分類別の文数は、予定が 274 文，予想が 136 文，願望が 76 文，心情が 27 文，呼掛が 17 文，疑問が 24 文，伝聞が 21 文，記号が 63 文であった。

学習には weka の SVM を用い，重み付け加算の有無に対して 10 fold cross validation で交差検定を行なうことで，両者を比較した．カーネルは線形カーネルを用いた．

### 5.2 実験結果と考察

実験結果を表 1 に示す．重み付け加算を行なった場合は，行わない場合に比べて，適合率が 1.9%，再現率が 4.0%向上した．このことから，周囲の文脈を考慮したモデルの有効性が確認できた．

しかし，素性の有効性の確認のため，決定木学習で学習してみたところ，大きい値をもつ素性がないために判別できない文が 169 文存在した．これは，決定木における誤りの内，約半数にあたる．この原因として，重み付けが不適切である場合と，素性が不足している場合が考えられる．前者の例としては，始めにしか明示的な時間表現がない文章において，後半の文の重み付けが過剰となり，他の文の素性の影響が弱い場合が考えられる．また，後者の例としては，W 杯等のイベント名などで時間を表わしている場合が考慮されていないことが考えられる．

表 1 未来表現抽出結果

| | 適合率 | 再現率 |
|---|---|---|
| 重み付け加算なし | 74.5% | 50.9% |
| 重み付け加算あり | 76.4% | 54.9% |

## 6. まとめ

本研究では，ブログ記事に記述された未来に関する表現の抽出手法を提案した．手法としては，判定対象文中の時間表現，文末表現，可能性のモダリティ表現等を素性とし，判定対象文より前の文の時間表現素性を重み付け加算して機械学習で判定することにより，精度約７６％，再現率約５５％を達成した．今後は，重み付けの方式や，イベント名の時間推定などを行なっていきたい．

### 参考文献

[1]Inderjeet Mani, George Wilson, “Robust Temporal Processing of News”, In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000).

[2]上嶋 宏, 三浦 孝夫,塩谷 勇, “不完全なニュース集合からのタイムスタンプ推定”, データ工学ワークショップ（DEWS）, 3C-o4 (2005).

[3]野呂 太一, 乾 孝司, 高村 大也, 奥村 学, “イベントの生起時間帯判定”, 情報処理学会研究報告, NL-170 (2005)

[4]前田栄作, “痛快！サポートベクトルマシン -古くて新しいパターン認識手法”, 情報処理学会誌 ,Vol.42, No.7 (2001)

[5]Takeshi Fuchi, Shinichiro Takagi, “Japanese morphological analyzer using word co-occurrence -JTAG-”, In Proceedings of COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, volume 1, pages 409-413, Montreal, 1998