E-061

Blog におけるイベント体験情報の判別と抽出

Recognition of Event Experience Information and its Extraction from Blogs

小林 聡† 山田 剛一‡ 絹川 博之†
Satoru Kobayashi Koichi Yamada Hiroshi Kinukawa

1．はじめに

近年の Blog の普及に伴い、実世界で体験した情報を Blog を使って情報発信することが非常に増えている。

体験した実世界情報が記述してある Blog エントリの中でも、美術展や、お祭り等のイベントに参加した際の体験情報が Blog エントリに記述されている場合、その Blog エントリがイベントへの参加を検討している人にはとても有益な情報になると考えられる。

そこで本研究では、実際に参加したイベントの体験が記述されている Blog エントリかを判別し、その Blog エントリからイベントに参加するか否かを判断できる情報を抽出し、イベントへの参加を検討している人にその情報を提供することを目指している[1] 。

2．Blog エントリに現れるイベント情報

Blog エントリを分析した結果、イベントの関連情報が書かれている Blog エントリは大きく二つのタイプに分類できることが分かった。それを以下に示す

（1）実際にイベントへ行き、その感想やその場の状況等を記述した Blog エントリ
（2）イベントの紹介や宣伝等を記述した Blog エントリ

実際に観察された実世界情報を Blog から抽出するという本研究の趣旨として、今回は（1）を研究の対象とする。

3．Blog におけるイベント体験情報の判別と抽出システム

本システムの処理の流れを図 1 に示す。

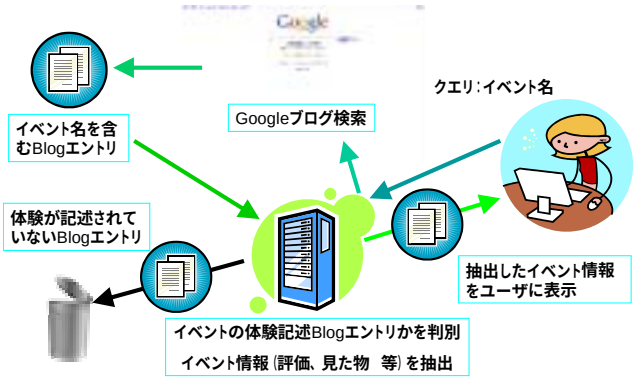

図 1. Blog におけるイベント体験情報の判別と抽出システム

（1）イベント名を検索質問として入力
（2）入力したイベント名が記述された Blog エントリ収集
（3）収集した Blog エントリにイベント体験情報が記述されているか判定
（4）イベント体験情報（評価、見たもの等）を抽出
（5）抽出したイベント情報をユーザに提供

†東京電機大学大学院情報メディア学専攻
‡東京電機大学/JST-CREST

3.1 イベント体験情報記述の有無の判定

以下のいずれかの条件を満たしている Blog エントリは、体験情報が記述されていると判定する。

（1）「～に行った」「～を見た」等の「実体験を示す語」が記述されていて、かつ、それに係っている語が検索質問のイベント名だった場合
（2）イベントの内容、雰囲気、評価等の「実際にイベントを体験しなければ記述できない表現」が記述されていて、かつ、Blog エントリのタイトルが検索質問のイベント名だった場合

3.1.1「実体験を示す語」の係り語による判定

イベントの体験情報が Blog エントリに記述されている場合、イベント名が「実体験を示す語」（例：「行ってきました」「見てきた」等）に係っている場合が多い。

例：大江戸骨董市に行ってきました。
　　ルーブル美術館展を見てきた。
　　ロートレック展を観賞しました。

そこで、「実体験を示す語」に係っている語を抽出し、抽出した語が検索質問のイベント名であった場合、その Blog エントリにイベント体験情報が記述されていると判定する。

「実体験を示す語」に係っているイベント名候補を抽出するためには、Blog エントリの本文を抽出し、抽出した本文に茶筌[2]による形態素解析と CaboCha[3]による係り受け解析を行う。この情報を基に「実体験を示す語」に係っている語を抽出する。このとき抽出したイベント名候補が検索質問のイベント名であるとは限らない。抽出したイベント名候補と検索質問のイベント名が同一のイベントを表していることが必要であるので、これを文字列の共通度により評価する。

3.1.2 文字列の共通度による評価

抽出したイベント名候補と検索質問のイベント名が同一のイベントを表しているかを、「文字列の共通度」により以下の手順で判定する。

（1）抽出したイベント名候補と検索質問のイベント名の両方に形態素解析を行う
（2）形態素単位に分割した抽出イベント名候補と検索質問のイベント名を比較し、一致した形態素の文字数の総和を求める
（3）一致した文字数の総和と検索質問全体の文字数との割合を求める
（4）（3）で求めた割合が設定した閾値（現在、閾値を 0.5 に設定）を越えるか判定
（5）閾値を超えた場合、抽出したイベント名候補が検索質問のイベント名と同一のイベントを表していると判定

3.1.3 体験に基づく記述の有無の判定

イベントの体験情報が記述されている Blog エントリには「実体験を示す語」が記述されていない場合もある。その場合には「実際にイベントを体験しなければ記述できない表現」が記述されているか判定し、イベント体験情報記述の有無判定を行う。

「実際にイベントを体験しなければ記述できない表現」は、「楽しかった」等のブロガー自身が感じたことや、イベントの状況等のブロガー自身が観察したことを表す表現が多い。これらの表現のうち確実性の高い表現の判別ルールを作り、その判別ルールに適合した数が閾値を超えた場合、Blog エントリに「実際にイベントを体験しなければ記述できない表現」が記述されていると判定する（現在、閾値を 1 に設定）。

また「実際にイベントを体験しなければ記述できない表現」が検索質問のイベントに対するものであるか判定しなければならない。そこで、検索質問のイベント名と Blog エントリのタイトルが一致しているという制約を与えている。検索質問のイベント名と Blog エントリのタイトルが一致しているかの判定には、3.1.2 で述べた文字列の共通度の評価を用いている。以上の流れを図 2 に示す。

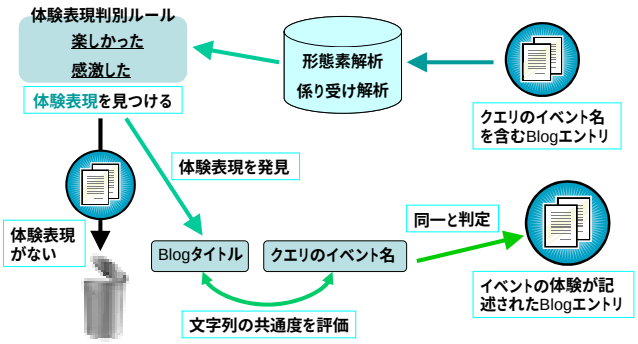

図 2.「実際にイベントを体験しなければ記述できない表現」の有無判定

本研究で用いる「実際にイベントを体験しなければ記述できない表現」判別ルール：

（1）形容詞の連用形（連用タ接続）
　　例：「面白かった」「よかった」等
（2）形容詞、形容動詞、副詞、サ変名詞＋「でした」「だった」
　　例：「きれいでした」「満足だった」等
（3）サ変名詞＋「する」の連用形＋「た」
　　例：「感動した」「堪能した」等
（4）動詞、サ変名詞＋接続助詞「て」＋補助動詞「いた」
　　例：「盛り上がっていた」「賑わっていた」等
（5）動詞（未然形）＋接続助詞「れ、られ」＋接続助詞「て」＋補助動詞「いた」
　　例：「行われていた」「売られていた」等
（6）サ変名詞＋動詞（未然形）＋接続助詞「れ、られ」＋接続助詞「て」＋補助動詞「いた」
　　例：「展示されていた」等

なお、（6）以外のサ変名詞は特定の語句（「感激」「満足」「満喫」等）のみを対象とする。

（1）（2）はイベント自体かその内容に対するブロガーによる評価、（3）はイベント自体かその内容に対してブロガーが感じたこと、（4）（5）（6）はイベントの場にいてブロガーが観察したことを表現したものである。

3.2 イベント体験情報の抽出

3.1.3 で示した判別ルールは Blog エントリにイベント体験情報が記述されているか否かを判定するためのもので、イベントに関して記述される表現すべてを網羅しているものではない。そのため、3.1.3 で対象とした表現以外のイベントに対する記述があった場合にも、それをイベント体験情報として Blog エントリから抽出する。

Blog からの経験情報抽出に関しては[4]等の研究があるが、我々は形態素情報と構文情報の両方を利用し、より正確に抽出部分の特定を行う。また、抽出対象はイベントを体験したブロガーの行動や評価だけでなく、観察したその場の状況もあわせて抽出する。

4．実験・考察

「Blog エントリにイベントの体験情報が記述されているかの有無」を判定し、その精度と再現率を求める実験を行った。

評価方法：

10 種類のイベントに対して、Google ブログ検索[5]を行い、得られた各上位 30 件の Blog エントリにより評価

精度・再現率：

精度 88.1%、再現率 84.9%

なお、精度と再現率の定義は以下の通りである。ただし、実体験に基づいて記述された Blog エントリのことを実体験記事と呼ぶことにする。

$$\text{精　度} = \frac{\text{出力のうちの実体験記事の数}}{\text{システムが実体験記事とした記事の数}}$$

$$\text{再現率} = \frac{\text{出力のうちの実体験記事の数}}{\text{全ての実体験記事の数}}$$

考察：

精度に比べ、再現率が若干低い値となっている。再現率を上げるためには、評価表現以外の「実際にイベントを体験しなければ記述できない表現」について分析を進める必要がある。

5．まとめ

本稿では Blog におけるイベント体験情報の判別と抽出システムの処理である「Blog エントリにイベントの体験情報が記述されているかの有無の判定」と「イベント体験情報の抽出」について述べた。

今後は、イベント体験情報抽出の実験評価を行う。

参考文献

[1] 小林聡, 山田剛一, 絹川博之, ”Blog からのイベント体験情報の抽出”, 電子情報通信学会 2008 年総合大会, D-5-12 (2008)
[2] 形態素解析システム 茶筌 http://chasen.naist.jp/hiki/ChaSen/
[3] 係り受け解析システム CaboCha http://chasen.org/~taku/software/cabocha/
[4] Takeshi Kurashima, Taro Tezuka and Katumi Tanaka, ”Blog Map of Experience form City Blogs”, Web Information Systems (WISE2005), pp.465-503, November 2005
[5] Google ブログ検索 http://blogsearch.google.co.jp/