ブログ記事からの特定の時間帯を表す時間表現の抽出

廣嶋 伸章　　戸田 浩之　　松浦 由美子　　片岡 良治
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
{hiroshima.nobuaki,toda.hiroyuki,matsuura.yumiko,kataoka.ryoji}@lab.ntt.co.jp

1. はじめに

情報検索サービスは、ある物事についての詳細な情報を調べたり、行動を起こしたりする上で日常生活に必要不可欠なものとなってきている。特に、携帯端末の普及により、外出先などから携帯端末を用いて検索を行う機会が増加してきている。PC を用いた検索と比較して、携帯端末を用いた検索では、外出先で今夜行われるイベントの情報を知りたいといったような時間に応じた検索に対するニーズは高いと考えられる。

時間に応じた検索を実現するための最も単純な方法として、ブログなどの文書に付与されているタイムスタンプを用いることが考えられる。しかし、タイムスタンプは文書がある時間帯に関する内容であるかどうかによらずすべての文書に付与されるため、ある時間帯に書かれた文書を検索することはできるが、ある時間帯に関する内容が書かれた文書を検索することができないという問題がある。ある時間帯に関する内容が書かれた文書のみを検索するためには、文書中に含まれる時間表現を手がかりとして文書が表す時間帯に関する情報を付与する必要がある。そこで本研究では、文書に時間帯の情報を付与するために必要な時間表現を抽出し、時間表現に対して時間帯を結びつけることを目的とする。

時間表現は、表現と時間帯との対応付けが必要かどうかによる分類および表現が一つの時間帯を表すか複数の時間帯を表すかによる分類により、大きく 4 種類に分類することができる。表現と時間帯との対応付けが不要な時間表現は、「2009 年 12 月 17 日」や「毎週木曜日」のような表現であり、これらが文書中に含まれていた場合には、いくつかの単純なパターンを用いることにより、文書に時間帯を付与することが可能である。表現と時間帯との対応付けが必要な時間表現は、「バンクーバー五輪」や「お昼ご飯」のような表現であり、これらが文書中に含まれていた場合は、表現と時間帯を対応付けた辞書を参照して文書に時間帯を付与する必要がある。このうち、「バンクーバー五輪」のようなある一つの時間帯を表す時間表現については、イベントカレンダーなどの情報から比較的容易に辞書を構築することが可能である。一方、「お昼ご飯」のように周期的に繰り返される複数の時間帯を表す時間表現については、辞書を作成するのにコストがかかると考えられる。本研究では、文書に時間帯の情報を付与するために必要な時間表現のうち、特に周期的に繰り返される複数の時間帯を表す表現を抽出することを目的とする。

2. 関連研究

時間表現の抽出および時間表現を用いた文書への時間情報の付与に関する研究はこれまでにもいくつかの手法が提案されてきている。

Setzer ら[1]や Mani ら[2]は、ニュース記事中のイベントや時間情報に対してアノテーションを行う手法を提案している。しかしながら、時間情報のアノテーションが行える時間表現は日付や曜日などに限られるため、多くの時間情報が付与できない。

小倉ら[3]は、文書中に含まれる各文を時系列化す

る手法を提案している。しかしながら、ここで用いられている時間表現も日付や「深夜」などの特定の表現に限られる。

土屋ら[4]は、単語とその単語の概念の組からなる概念ベースを用いて様々な単語に対して時間帯を付与する手法を提案している。しかしながら、単語に付与できる時間帯は「冬」などの単位であり、細かい時間帯を付与することは難しいと考えられる。

野呂ら[5]は、ブログ記事を対象としてイベントの生起時間帯を判定する手法を提案している。SVM を用いて記事中の各文がイベントを表しているかどうかを判定し、ナイーブベイズ分類器を用いてイベントを表す文が「朝」「昼」「夕方」「夜」のうちどの時間帯に属するかを判定している。このナイーブベイズ分類器において分類に寄与した素性を調べることにより時間帯ごとの時間表現に相当する語を抽出することが可能である。しかしながら、学習ベースの手法であるため訓練データを用意する必要がある。また、1 日を分類する時間帯を細かくしていくと分類が難しくなると考えられる。

3. 提案手法

提案手法では、ブログ記事がイベントの発生した時間帯に書かれていることに着目し、各表現が含まれるブログ記事の作成日時をもとにその表現がある時間帯に偏って用いられるかを判定することにより、特定の時間帯を表す時間表現の抽出を行う。以下では、提案手法の詳細について述べる。

3.1 時間表現候補の抽出

まず、「X の時間」のような時間表現パターンを持つブログ記事を取得し、記事の中から名詞句 X を抽出し、時間表現候補集合を獲得する。

3.2 ピーク時間帯の推定

各時間表現候補に対し、その表現が表している時間帯を推定する。各時間表現候補 e に対し、3.1 節で獲得した記事のうち e に関する記事を取得する。各文書の作成日時が複数の時間帯のうちどの時間帯に属するかを判定してピーク時間帯を決定する。ここで周期を A 秒、時間帯分割数を N とする。例えば、1 日のうちの特定の時を表す時間表現を抽出したい場合には $A = 86400, N = 24$ とすればよい。全ブログ記事および e に関する記事が分割された各時間帯 $p_1, p_2, \cdots, p_N$ のうちどこに属するかを求め、時間帯ごとに文書数をカウントする。全ブログ記事のうち時間帯 p_i に属する文書数を D_i、e に関する記事のうち時間帯 p_i に属する文書数を d_i とし、d_i を D_i で割った値が最大となる p_i をピーク時間帯 P とする。

また、ピーク時間帯とは別に、e に関する記事が何番目の周期に属するかを求め、各周期に属する文書数の最大値を L とする。

3.3 時間表現スコアの算出

時間表現候補ごとに求めたピーク時間帯をもとに、時間表現候補 e に対し、時間表現が特定の時間帯に属することをどの程度強く示しているかを表す時間表現スコア S を以下のようにして求める。

$$S = \frac{1}{L+\alpha}\sum_j \frac{1}{r_j}\cos 2\pi \frac{t_j - t_P}{A}$$

ここで、α は定数、$j(j = 1,2,\cdots,M)$ は e に関する記事の記事番号、r_j は j 番目の記事が属する時間帯が全ブログ記事中でどの程度を占めるかの割合、t_j は j 番目の記事の作成日時、t_P はピーク時間帯の中心を表す時間である。e に関する記事の作成日時がピーク時間帯に近ければ余弦の値は大きくなるため、ピーク時間帯に近い記事が多いほど和の値は大きくなる。同じ周期に属する記事が多い場合には周期的ではなくバースト的なものであると考えられるため、L の値が大きい場合にはスコアが小さくなるよう

表 1: 各条件において利用した値

条件	パターン	周期	時間帯分割数	定数
H	X の時間	86400	24	30
W	X の日	604800	7	30

に補正を行っている。また、e に関する記事が少ない場合にスコアが大きくなることを防ぐため、定数 α を設けている。

時間表現候補を時間表現スコアの降順にソートし、上位のものを時間表現として抽出する。

4. 実験

提案手法の有効性を検証するため、実験を行った。

実験データとして、2009 年 4 月から 10 月までに投稿されたブログ記事を作成日時とともに収集した。

実験では、1 日のうちの特定の時を表す時間表現（条件 H）および 1 週間のうちの特定の曜日を表す時間表現（条件 W）の抽出を試みた。それぞれの条件において利用した値を表 1 に示す。また、パターンを含む文書を取得する際、検索エンジンを用いて検索を行い、検索結果上位 10,000 件の文書を取得した。各時間表現候補に関する文書は最大 1,000 件取得した。

提案手法を用いて各条件のもとで抽出された時間表現上位 100 件に対し、1 人の評価者により正しく時間表現が抽出されているかの評価を行った。評価は 2 段階の基準で行った。基準 A では、抽出された時間表現が実際に表す時間帯の範囲とほぼ一致する場合にのみ正解とした。基準 B では、抽出された時間表現が実際に表す時間帯の一部であるものや、考え方によってはその時間帯であると考えられなくはないものも正解とした。例えば、条件 W において「休み」という時間表現が日曜日の時間帯を表すものとして抽出された場合、一般的には日曜日だけでなく土曜日も休みと考えられ、日曜日は実際に表す時間帯の一部であるため、基準 A では不正解となり、基準 B では正解となる。

表 2: 上位に抽出された時間表現（条件 H）

スコア	時間表現	時間帯	評価 (A)	評価 (B)
399.46	朝	7 時台	×	○
378.79	お昼	12 時台	○	○
369.16	お昼ご飯	12 時台	○	○
361.33	次	13 時台	×	×
354.96	日の出	5 時台	○	○
295.11	朝食	7 時台	×	○
291.86	お昼寝	13 時台	○	○
282.84	朝ごはん	8 時台	×	○
281.82	おやつ	15 時台	○	○
261.60	お迎え	13 時台	×	○

表 3: 上位の時間表現の正解率（条件 H）

件数	正解率（基準 A）	正解率（基準 B）
10	0.50	0.90
30	0.43	0.70
100	0.34	0.51

表 4: 上位の時間表現の正解率（条件 W）

件数	正解率（基準 A）	正解率（基準 B）
10	0.50	0.60
30	0.30	0.40
100	0.11	0.28

実験結果について示す。まず、提案手法により条件 H のもとで抽出された上位の時間表現について表 2 に示す。提案手法により適切な時間表現が抽出されていることがわかる。

次に、各条件のもとで、2 段階の基準により、10 位／30 位／100 位までの時間表現に対して正解率

を求めた。その結果を表 3、表 4 に示す。

条件 H では、上位 100 件の時間表現のうち半数以上が正解とみなせる時間表現であったことがわかる。文書に時間情報を付与するための時間表現辞書の構築を支援するツールとして用いるには十分な精度であると考えられる。しかしながら、そのようなツールとして用いるためには、表 3 で挙げたような比較的容易に思いつく時間表現だけではなく、「朝マック」のような容易には思いつかない時間表現が抽出されることが望ましいが、そのような表現は上位にはあまり抽出されなかった。提案手法では、時間表現候補に関する記事数が少ない場合に時間表現スコアが高くなるのを防ぐということを行っているのが原因であると考えられる。記事数が少ない場合でも適切な時間表現を抽出できるように改良を行っていきたい。

条件 W では、上位の結果の中にテレビ番組や雑誌の名前などの有益な時間表現が抽出されたが、上位 100 件のうち適切な時間表現と思われるものは 1/4 程度にとどまった。上位の時間表現の中には、「メンテナンス」などのように特定のブログで何度も話題に上っていると考えられる表現が見受けられたため、同一ブログサイトから抽出された記事をまとめて扱うことにより精度の向上につながるのではないかと考えられる。

いずれの条件においても、100 位以下に抽出された時間表現を見てみると、あまり有益と思われる時間表現が存在していないように思われた。これは、１つのパターンから抽出できる時間表現候補の種類がそれほど多くなく、１つのパターンだけでは不十分であることを示していると考えられる。抽出された時間表現をもとに異なるパターンを導き出し、そのパターンから別の時間表現を抽出するようなブートストラップ的な手法について検討していきたい。

5. まとめ

本稿では、ブログ記事がイベントの発生した時間帯に書かれていることに着目し、各表現が含まれるブログ記事の作成日時をもとにその表現がある時間帯に偏って用いられるかを判定することにより、特定の時間帯を表す時間表現の抽出を行う手法を提案した。実験では、提案手法により１日のうちの時を表す時間表現および１週間のうちの曜日を表す時間表現について抽出を行い、ある程度の時間表現を抽出できることを確認した。

今後は、ブートストラップ的な手法について検討し、抽出できる時間表現の種類を増やすところから始めていきたい。また、本稿では特に触れなかったが、スパムブログは同一の時間帯に同じ内容の記事が投稿される可能性が高く、提案手法に悪影響を及ぼすと考えられる。スパムブログを除去する手法についても検討していきたい。

参考文献

[1] A. Setzer, R. Gaizauskas. A Pilot Study on Annotating Temporal Relations in Text. In proceedings of the ACL-2001 Workshop on Temporal and Spatial Information Processing, pp.88-95, 2001.

[2] I. Mani, G. Wilson. Robust Temporal Processing of News. In proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, pp.69-76, 2000.

[3] 小倉牧人, 田村直良. 文間の時間制約モデルと事象の時系列化への応用に関する研究. 情報処理学会研究報告, 2000-NL-140, pp.111-118, 2000.

[4] 土屋誠司, 渡部広一, 河岡司. 連想メカニズムを用いた時間判断手法の有効性の検証. 情報処理学会研究報告, 2005-NL-168, pp.113-118, 2005.

[5] 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, 2005-NL-170, pp.7-14, 2005.