2C-2

複数ブログ記事からの関連事象を含む概要記事の自動生成

井原 健紘† 福冨 諭‡
電気通信大学 情報工学専攻

1.はじめに

日記形式のウェブサービスであるブログが多くの人に利用される世の中になった．様々な言及内容のブログが存在するが，その中でも時事問題や流行の話題に対して自らの意見を綴っている文章も少なくない．流行の事象は当然のことながら同じ時期に言及されることが多く，また，論点に類似性が見出せることも（同じ論点に対して反対の主張が書かれていることも含めて）多々ある．流行のキーワードはそれを含むブログと共に公開されており[1]，本研究ではそれらの情報を用いている．

著者らは，自動要約の技術を応用することにより，流行のキーワードの簡単な解説，および，それに対する論点を数行にまとめることができるのではないかと考えた．いわば，自動的な軽めのニュース記事の生成である．2006 年 8 月 14 日から翌年 1 月 5 日まで，実際にその思いつきを実行に移し，複数のブログを要約し，記事をブログとしてウェブに公開している[2]．

同様の試みに blogWatcher[3]があるが，blogWatcher はあくまで各ブログへのインデックスという性質が強いように見受けられる．それに対し，本研究による記事生成は，流行事象一つに対し一つの記事として成り立つことを目標に設計されている．なお，出来上がった記事および本ブログの存在に対する閲覧者の反応については 3 節「評判」において語る．

自動要約としては，既存の手法ととりわけ差別化できるような新しさはない．文章の分解・再構築の手法としては[4]などがある．

2.複数ブログ記事要約

2.1 目標

一つのキーワードが与えられたとき，以下の 3 項目を知ることができるような短い記事を生成するのが目標である．

- それがなぜ流行のキーワードであるのか
- その関連事象としては何が挙げられるか
- それにどのような言及がなされているか

Automatic Generation of Outline Article Including Related Events from Multiple Blog Articles
†Takehiro IHARA ‡Satoshi FUKUTOMI
†‡Department of Computer Science, the University of Electro-Communications

2.2 手段の概要

流行のキーワード（これを「第一キーワード」と呼ぶことにする）を含む複数のブログから 1 文単位で文を引用し，最長 7 文を用いて 1 つの記事を作成する．以降の節では文の引用基準を説明する．

2.2.1 第一文

第一キーワードを含む段落（複数の文からなる）を抜き出し，その名詞の出現回数を数え，$n(w)$とする．ただし，w は名詞であり，形態素解析には JUMAN[5]を用いた．その後，その段落の文ごとに次式にしたがってスコアを定める．

$$score = -\alpha L + \sum_{w \in W} n(w) \quad (1)$$

ここで，L は文の文字数であり，W はその文に含まれる名詞を示し，α は係数でありここでは 0.4 とした．このスコアが最も高い文を第一文として引用する．

2.2.2 第二文

第一文とほぼ同様の方法で文にスコアをつける．ただし，$n(key)$を負とする．key は第一キーワードの単語を示す．ここで，$\alpha = 0.3$とした．

また，任意の文 A と文 B の両方に存在する単語の個数を，文 A および文 B の短い方の単語の個数で割ったものを類似度として定義する．第一文との類似度が閾値(0.9)より低い文の中で最大のスコアを有する文を第二文として引用する．「それがなぜ流行のキーワードであるのか」を説明する文を引用することが，第一文と第二文の引用基準の狙いである．

2.2.3 第三文

まず，2.2.1 節で計算した$n(w)$の大きな w から$M(=7)$個抜き出し第二キーワード群とする．第二キーワード群を含む段落の文の中で，それ以前の引用文（この場合は第一文と第二文）と類似度が閾値(0.8)未満で最も高い文を第三文とする．

2.2.4 第四文

第三文で用いられた第二キーワードを含む段落から新たに$n(w)$を計算し，2.2.2 節と同様の方法で $score$ を得る．ここで，$\alpha = 0.35$ とした．それ以前の引用文との類似度が閾値(0.8)未満で，

スコアの最も高い文を第四文とした.

2.2.5 第五文と第六文

第三文および第四文とほぼ同じ選択基準で引用する．ただし，第三文および第四文に含まれる第二キーワードを持つ文は候補から除外する．

第一キーワードの関連事象を補足するのが，第三文から第六文の狙いである．また，偶数文目ではキーワードとなる単語の頻度をマイナスにしているが，これは直接キーワードが出てこない文を採用しやすくするためである．

2.2.6 第七文目

2.2.1 節と同様の方法で *score* を計算する．第一キーワードを含む段落から「形容詞（形容動詞を含む）」と「判定詞（だ・である等）」の両方を含む文のみを抜き出す．さらにそれ以前の引用文と類似度が閾値(0.8)未満でスコアの最大のものを第七文とする．事象に関する簡潔な感想を得て記事に完結感を出すことが目的である．

2.2.7 記事の棄却

以上の基準により出来上がった記事が「五文未満」もしくは「引用元が三箇所未満」となっていた場合には，記事が生成しづらいキーワードだったものと見なし，記事自体を棄却する．

2.3 引用元の明示

著作者を明示するため，各文の末尾から引用元に対してリンクを張る．気になる文が出てきたときに元の記事を見に行くこともできる．

3.評判

全ての文がもともと他人のものであるので記事の出力例は示さないが，ウェブ上にはブログの形式で公開されているので興味のある方は見ていただきたい[2]．本節ではウェブ上で得られた反応を紹介する．

公開してから一ヶ月の間に，本研究によるブログの記事を人間が書いたものと勘違いしたと思しき反応が数件見られた．コメント欄で数件，トラックバックが 1 件である．その後，著者らは誤解防止のためにブログタイトルを誤解しようのないものに変えた．なお，機械が自動引用して作った記事だと（おそらく）分かっていながら，アイドルに関する記事を自分のソーシャルブックマークに加えた閲覧者も 2 名存在した．アンテナなどのツールを用いてこのブログを定期的に見に来ている方も 20 名前後存在した．出来のよい記事に関しては人間が書いたブログに匹敵する価値を持つと見なされるようである．ただし，明らかに引用が失敗している記事も存在する．

記事ではなく，本研究によるブログそのものに対するソーシャルブックマークは 2007 年 1 月初旬の段階で約 50 件弱集まっている．「人工無能ブログ」と見たままをラベルとして貼り付ける閲覧者が多い．また，質は「中の上から上の下あたり」と見なしているように見受けられる．著作権に関して「ぎりぎりあり」と判定したコメントや，成功要因は文を生成せずに引用したことと分析したコメントなども存在した．積極的に負のコメントを残すソーシャルブックマークのコメントは今のところ見受けられない（判断不能のものはいくつか存在している）．

ブログ本体にブログへのコメントを残した閲覧者も存在し，こちらは「感心した」という意見を持つ方 2 名，「文脈のねじ曲げはやめてほしい」という意見を持つ方 1 名，意味のとれなかったコメントを残した方 1 名，趣旨とは全く関係のない部分でネガティブなコメントを残した方 1 名となっている．トラックバックを用いて分量のある言及をした方もおり，「何かが怖いが試みとしては面白い」とのことである．

また，他人の書いた文章を労力をかけずに再編集して提示するというこの要約記事自動生成指針に，著者自身が疑問を抱いたため，現在は更新を停止している．

心理実験による評価はおこなっていない．

4.まとめ

時事問題や流行の話題に対して自らの意見を綴っているブログ記事を集め，自動要約の技術を応用することにより，自動的な軽めのニュース記事の生成を試みた．2006 年 8 月 14 日から翌年 1 月 5 日まで実行に移しており，出力結果をブログとしてウェブに掲載している[2].

記事の質は悪くないと判断されているようである．

今後，このようなかたちでの記事生成（文引用）についての倫理的な観点からの議論が必要であると考えている．

参考文献

[1] http://d.hatena.ne.jp/hotkeyword
[2] http://d.hatena.ne.jp/saussure/
[3] 奥村 学，南野 朋之，藤木稔明，鈴木泰裕，"blog ページの自動収集と監視に基づくテキストマイニング，" 人工知能学会，セマンティックウェブとオントロジー研究会，SIG-SWO-A401-01，2004.
[4] 赤石 美奈，"文書群に対する物語構造の動的分解・再構成フレームワーク，" 人工知能学会論文誌，Vol.21，No.5，pp.428-438，2006.
[5]http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html