3P-8

マイクロブログに対するWikipediaを用いたタギングと要約手法

森 耕平† 疋田 輝雄†
明治大学理工学研究科†

1. はじめに

近年の情報氾濫の影響で優れた情報収集手段が求められており，その手段として Twitter[2]を代表とするマイクロブログが注目されている．マイクロブログの情報には，速報性が高い・網羅性が高いなどの長所があるが，冗長でまとまりがないという短所もある．

本稿では，マイクロブログの情報の短所を克服する手法を提案する．具体的には，（1）マイクロブログの投稿に対してタギングを行い，（2）タギングの結果ごとにまとめた投稿に対して複数文書要約を行う――ことで短所の克服を図る．

2. 関連研究

横本ら[6]は，ブログ記事集合に対して，Wikipedia[3]を用いた観点分類を行っている．本研究では，この Wikipedia による観点分類の手法を参考にしている．

坂本ら[4]は，あるトピックに関するマイクロブログの投稿をリアルタイムに収集し，そのトピックのイベントを検出する度にその要約を作成し，作成した各イベントの要約同士を組み合わせてトピックの要約を作成している．

中原ら[5]は，マイクロブログの投稿内容にマイクロクラスタリングを利用してクラスタの概念を構築し，興味対象となる投稿を出来る限り多く被覆する少数のクラスタを抽出して，投稿内容の要約を行っている．

3. 提案手法

3.1. タギング

第一に，あるトピックに関する投稿を取得し，投稿中の名詞が Wikipedia で項目名として使用されているかを確認する．使用されている場合に限りその項目名の Wikipedia 記事を取得し，当該記事の文書ベクトル$\vec{\mathrm{I}}$を式 (1)で作成する．

$$\vec{\mathrm{I}}(e) = (\mathrm{w}(r_1), \mathrm{w}(r_2), \ldots, \mathrm{w}(r_n)) \quad (1)$$

ここで，e は記事を表す．r は，記事内で使われている重要語（太字・ハイパーリンクの単語）を表す．W(r)は，r の逆文書頻度である．

第二に，マイクロブログの投稿の文書ベクトル$\vec{\mathrm{G}}$を式 (2)で作成する．

$$\vec{\mathrm{G}}(t, e) = (\mathrm{freq}(r_1), \mathrm{freq}(r_2), \ldots, \mathrm{freq}(r_n)) \quad (2)$$

ここで，t は投稿を，e は Wikipedia 記事を表す．r は，$\vec{\mathrm{I}}$の重要語と同じである．freq(r)は，投稿内での r の出現頻度である．

最後に，$\vec{\mathrm{I}}$と$\vec{\mathrm{G}}$との関連度（内積値）を求め，その値が閾値以上となる場合に，当該記事の項目名をタグとして投稿に付与する．

3.2. 複数文書要約

第一に，3.1 節で付与したタグ毎に投稿をまとめ，重要文抽出を式 (3)で行う．

$$\mathrm{score}(S) = \sum_{t \in S} \mathrm{tf}(t, S)\mathrm{w}(t) \quad (3)$$

ここで，tf(t,S)は単語 t の投稿 S での頻度，w(t)は t の TF-IDF である．

第二に，抽出した投稿の中で冗長なものを削除する．削除の方法には MMR[1]を用いる．これは式 (4)で行う．

$$\mathrm{MMR}(Q, R, S) = \mathrm{argmax}_{D_i \in R-S} \left[\lambda \times \mathrm{sim}(D_i, Q) - (1-\lambda) \max_{D_j \in S} sim(D_i, D_j) \right] \quad (4)$$

ここで，Q は投稿をまとめる際に使用したタグ，R はシステムによって検索された投稿の集合，S は既に選択された R の部分集合，D は抽出された重要な投稿の集合である．argmax とは，式の値が最大となる D_i を求めるという意味である．

4. 評価

4.1. タギングの評価

評価のため，Twitter 上にある投稿を収集した．評価対象のトピックには「原発」と「ドラフト」とを選択し，それぞれのトピックに関連する投稿を収集してタギングを行った．タギングの結果の例を**図 1** に示す．上述のタギングの結果を

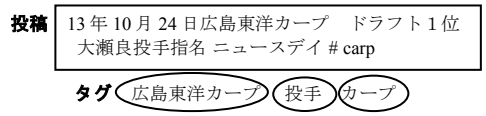

図 1 トピック「ドラフト」へのタギングの例

A Tagging and Summarizing Method for Microblogging by Using Wikipedia
†Kouhei Mori and Teruo Hikita, School of Science and Technology, Meiji University

F 値で評価したものを**表 1** に示す.

また，比較のために，Wikipedia を用いない手法でもタギングを行った．これは，投稿に形態素解析をかけて名詞を抽出し，その名詞の TF-IDF が閾値以上だった場合に，その名詞をタグとする手法である．その手法の結果を F 値で評価したものを**表 2** に示す.

表 1・表 2 より，Wikipedia を用いたタギングの方が，検出率・F 値は高くなることが分かる．一方で，再現率は，Wikipedia を用いないタギングの方が高い．この理由は，Wikipedia を用いない手法が適切・不適切に関係なく多くのタグを付与したため，付与し損ねるタグが減ったためである．また，表 1 で，「原発」よりも「ドラフト」の方が検出率が高いのは，投稿中に現れる固有名詞の数の差が影響している．「ドラフト」に出現する単語には選手名などの固有名詞が多く，また，それらが Wikipedia の項目名として使用されていたため，適切なタグが付与された．これらの結果から，本研究のタギングは Wikipedia で使用されている固有名詞が投稿中に多いほど精度が高くなることが分かる.

4.2. 複数文書要約の評価

複数文書要約の評価は，坂本らの評価指標を参考にし，我々が作成した正解要約とシステムが作成した要約とを比較して行った．具体的には，まず式 (5)と式 (6)とを求めた.

$$recall = \frac{|R \cap S|}{|R|}, \quad L = \frac{|R \cap S|}{|S|} \qquad (5), (6)$$

ここで，R は正解要約が含む内容語（名詞・動詞・形容詞）の集合，S はシステムの要約が含む内容語の集合である．続いて，$recall$ と L との調和平均を計算することで要約の精度を評価した.

評価対象として，トピック「ドラフト」に付与したタグの中で，「投手」「内野手」「ドラフト」を選択し，それぞれのタグでまとめた投稿に対し複数文書要約を行った.

要約の例を**図 2** に，評価の結果を**表 3** に示す．表 3 より，タグ毎に精度に差があることが分かる．「ドラフト」への要約の精度が特に低いのは，

表 1 Wikipedia を用いたタギングの評価

トピック	検出率	再現率	F 値
原発	0.6557	0.4996	0.5671
ドラフト	0.7289	0.4795	0.5784

表 2 Wikipedia を用いないタギングの評価

トピック	検出率	再現率	F 値
原発	0.3805	0.6089	0.4683
ドラフト	0.4247	0.4826	0.4518

ベイスターズジュニア出身の松井投手を逃したのは痛いけど、柿田投手は思ったよりずっと良い。他の指名も良いし、今回のドラフトは（少なくとも現時点では）成功だったように思える。≫続き

プロ野球ドラフト会議 柿田投手・DeNA指名「1位良かった」母校・松本工の監督ら歓喜／長野-毎日新聞

阪神のここ数年のドラフト一位選手 2013年 岩貞祐太 投手 2012年 藤浪晋太郎 投手 2011年 伊藤隼太 外野手 2010年 榎田大樹 投手 2009年 二神一人 投手 2008年 萧一傑 投手 2007年 白仁田寛和 投手

日大藤沢 金子 一輝 選手 埼玉西武ライオンズ ドラフト 4 位指名おめでとうございます。いかにも西武のショートが似合いそうな感じです。松井投手との神奈川 2013 世代対決が楽しみですね。

(以下省略)

図 2 タグ「投手」を持つ投稿への複数文書要約の例

表 3 複数文書要約の評価

タグ	正解要約との比較
投手	0.6101
内野手	0.7518
ドラフト	0.5079

タギングの際にタグ「ドラフト」の投稿への付与が上手くいかなかったためである．具体的には，Wikipedia 記事「ドラフト」は，「曖昧さ回避のためのページ」であり投稿との関連度が低くなるため，タグ「ドラフト」は投稿にあまり付与されなかった．この影響で，正解要約と大きく異なる要約結果が作成された．この結果から，本手法の要約は Wikipedia 記事の特徴で精度が左右されることが分かる.

5. 終わりに

本研究では，マイクロブログの投稿を分類・要約するために Wikipedia を使用する方法を提案し，その有効性を確認した.

今後の課題は，Wikipedia 記事の特徴で精度が左右されることへの対応である．この課題に対しては，他の Web 辞書も用いて，Wikipedia への依存度を低減することで解決を図る.

参考文献

[1] Carbonell, J. and Goldstein, J.: The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries, Proc. 21st Annual Int. ACM SIGIR Conf., pp.335–336 (1998).
[2] Twitter, https://twitter.com/.
[3] Wikipedia, http://ja.wikipedia.org/.
[4] 坂本，横山，福田，石川：マイクロブログを対象としたリアルタイムな要約生成システムの試作，第 3 回 DEIM Forum (2011).
[5] 中原，宇野，羽室：マイクロクラスタリングを用いた単語分類とトピック検知，情報処理学会アルゴリズム研究会報告 2013-AL-145(27), pp.1–8.
[6] 横本，林，牧田，宇津呂，河田，福原，神門，吉岡，中川，清田：特定トピックに関するブログ記事集合の観点分類における Wikipedia の利用，第 3 回 DEIM Forum (2011).