2N-1

マイクロブログにおける関連語の自動抽出

岩井 一晃 † 鈴木 優 ‡ 石川 佳治 ‡
† 名古屋大学 工学部 電気電子・情報工学科 情報工学コース ‡ 名古屋大学 情報基盤センター

1 はじめに

現在、Twitter に代表されるマイクロブログは多くの利用者によって利用されている．マイクロブログの利点として，日常の生活では知り合えない人の発見，交流ができるという点があげられる．

本研究ではスポーツ中継などの TV 番組に関する Twitter の投稿文におけるキーワード抽出の支援を行う．キーワード抽出の支援とは，利用者の入力したキーワードによって得られた投稿文からその分野に関連する語を抽出することである．このシステムは様々なシステムの基盤研究となり得る．このシステムを利用することにより，利用者の興味がある人物の検索を行うためのシステムの開発や，キーワード間の繋がりの発見，マイクロブログ上の情報検索の強化などのシステムに活用できると考えられる．

このような検索問合せの拡張についての研究は，Web マイニングの手法として研究されている．サーチエンジンの検索問合せの拡張として [1] のような研究が挙げられる．この検索問合せの拡張の研究は，問合せを追加することによってデータをより詳細にすることが目的である．しかし本研究ではより広くデータを取るという目的のため異なる．

また Twitter に関する研究も多数存在し，Twitter の投稿文を利用し，形態素解析器に用いられる既存の辞書に登録されていない単語を取得，辞書を改良するという研究 [2] がある．形態素解析には辞書に登録されていない単語を判別することができないという特徴がある．マイクロブログを字句解析する際，マイクロブログ特有の表現が存在するため，そのような表現を識別できるような手法をとる必要がある．そのため，本研究ではマイクロブログ特有の表現もとれるようにするために n-gram を用いた頻出単語解析を行う．

加えて，本研究では発言者の情報を用いることで，抽出した文字列が利用者の入力したキーワードに関連しているかどうかを詳しく判別する．

以上より，本研究の特徴は，形態素解析による頻出文字列の解析は，利用される辞書に依存するため，特有の表現などが多い Twitter に向いていないとして n-gram を利用した頻出単語解析を行った点，発言者の情報を用いた頻出文字列が関連語か否かの判断を行った点，そしてそれらを用いた新しい検索キーワードを加えた検索問合せの拡張により，利用者によって入力されたキーワードの分野に関する投稿文をさらに多く，得ることを可能にしたという点が挙げられる．

Development of Related Word Automatic Operation Extraction Tool on Microblog
† Kazuaki Iwai, Department of Information Engineering, School of Engineering, Nagoya University
‡ Yu Suzuki,Yoshiharu Ishikawa, Information Technology Center, Nagoya University

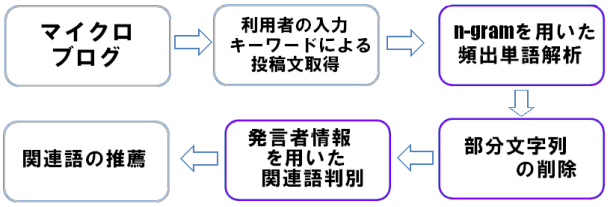

図 1: 本研究のシステムの流れ

2 関連語の抽出

本研究のシステムの流れは図 1 のようになっており，本章ではそれぞれの機能について述べる．

2.1 n-gram を用いた頻出単語解析

n-gram を用いた頻出単語解析を行う．n-gram を行う対象文章群は利用者の入力したキーワードを含む投稿文である．出現頻度の高い文字列は，利用者の入力したキーワードに高い頻度で共起する文字列であるといえる．利用者の入力したキーワードに多くの頻度で共起する単語は，利用者の入力したキーワードに関連する分野の文字列である可能性が高い．これより，頻出単語は関連語の候補となる．

2.2 部分文字列の削除

n-gram を用いた頻出単語解析により，部分文字列が頻出文字列として検出される．ところが部分文字列は検出する必要の無い頻出単語である．なぜなら，ある文字列の部分文字列であった場合，部分文字列もある文字列と同数出現するからである．

2.3 発言者の情報を用いた関連語判別

2.1 節 と 2.2 節 で述べた手法で頻出単語をとることはできるが，利用者の入力したキーワードに共起する単語が必ずしも関連する文字列であるとは限らない．例えば，ある少数の発言者が同じような関係の無い投稿文を多数投稿し，頻出単語として扱うことができると上記の二つで判別された場合，その頻出単語は関連語でないにも関わらず関連語として扱われてしまう．これを防ぐために，発言者の情報を用いて関連語であるかどうかの判別を行う．

利用者が入力したキーワードにより投稿文を取得する際，その時間，多くの人がその分野についての発言を多く行っていると仮定する．その場合，多くの発言者は利用者の入力したキーワードに関連する分野の発言を行っていることになる．よって主要な発言者群とは利用者の入力したキーワードに関連する分野を発言した発言者群と考えることができる．これより発言者群の特定を行うことによって，その発言者が利用者の入力したキーワードに関することを述べているかどうかを判別することができる．

発言者群を特定するため，発言者間の類似度, 同じ文字列を発言した人数を調べ，それらを用いたスコアリングを行う．具体的には以下の式に従う．

$$D = \sum_{i=0}^{m} A_i(M, x) \quad (1)$$

$$N = \sum_{i=0}^{m} F_i \quad (2)$$

$$S = N \times (1 - \alpha) + D \times \alpha \quad (3)$$

発言者間の類似度は，文字列毎の発言回数が重なった回数の総和である.A_i は文字列毎の発言者間の類似度である．検索の主軸となる発言者のその文字列の発言数を M とし，x は検索の主軸以外の発言者の発言数とすると M より x の値が大きい場合 M を，そうでない場合は x を発言者間の類似度の値とする．これらを発言者毎における全ての発言者間の類似度を足し合わせたものを D, 発言者の類似度とする．

また N は同じ文字列を発言した発言者数であり，各文字列の F_i の合計である．F_i は各文字列における，発言者と同じ文字列を発言した発言者の人数である．

S がスコアとなり，これは N,D を用いた式で行う．また α はパラメータである. 発言者数をより重要であると評価するため，試験的に α =0.2 としてある．

スコアが一定値未満である発言者を主要でない発言者群とする．発言者群の特定後，関連語の候補の発言者の参照する．関連語の候補の中に主要な発言者群で無い発言者が，その関連語の発言者全体の閾値以上の人数が存在した場合，関連語の候補からその文字列を外す．これにより，発言者の情報を用いない関連語判別より高い精度で関連語を抽出することができる．

3 評価実験

今回の実験は，2010 年 12 月 31 日に放映された格闘技の大会に関する投稿文を取得することを目的とし,r 利用者がキーワードとして”K-1,Dynamite”と入力したと仮定して行う．実際に取得されたデータは英文のみのものを除き 6673 ツイートであった. データは 12 月 31 日 20:53〜23:27 にデータを取得された．このデータを利用する意義として，関連語として選手名がとれる可能性が高い点，また利用者が予想できない事が起こりうる点があげられる．

上記で述べたデータを利用し，実際に本手法で関連語を抽出する．また比較対象として形態素解析を用いた関連語抽出も同じデータを用いて実験を行う．この形態素解析を用いた関連語抽出では，MeCab * を形態素解析器として利用する．

3.1 実験結果

本手法，形態素解析を用いた関連語抽出の結果はそれぞれ表 1 の左右の結果となっている．これらは出現回数が上位 10 位のものをそれぞれ示している．形態素解析の方が目的の番組に関連のある文字列を取れているように思える．しかし形態素解析で検出できているものは，本手法でも漏れなく検出することができている．反対に本手法, 第 5 位のように”自演乙”など形態素解析で判別できない文字列で番組に関係がある文字列も多数候補にあがった．

表 1: 実験結果

順位	出現回数	文字列	出現回数	文字列
1	1215	った	616	自演
2	1130	って	518	TBS
3	824	ない	449	紅白
4	616	自演	335	試合
5	601	自演乙	310	さん
6	586	てる	302	石井
7	564	して	280	MILKYHOLMES
8	557	から	274	ガキ
9	529	BS	261	青木
10	527	TB	259	渡辺

3.2 考察

本研究の手法の関連語判別の際に多く使われるである格助詞の排除を行わなかったため，今回のような結果になったと考えられる．また上位の関連語候補に”紅白”と出ているように裏番組に関するツイートも多く取得していたため，発言者群を用いた関連語判別はあまり動作しなかったと考えられる．本手法は形態素解析よりも多くの文字列を候補として取得することができているが，必要でない文字列もまた非常に多く取得してしまっている．関連語の候補に対する制約をさらに厳しくし，精度をあげることが今後の課題となる．

4 まとめと今後の課題

本研究では利用者の入力したキーワードから関連語を自動抽出する手法を提案した. これを行うことにより多くの人の投稿文を見渡せることが出来，利用者が従来のキーワード抽出より自分の興味の持てるツイートを発見しやすくなる．本研究を利用したシステムの開発により，利用者が自分の興味のある人物を探し出す支援を行うシステムの開発ができる．

今後の課題として，利用者の入力したキーワードによって得られた投稿文の解析手段の追加による関連語の精度上昇があげられる．具体的には時間ごとの取得ツイート数によってツイート自体に何らかの重みを加える，形態素解析を併用して用い格助詞などを予め取り除くなどである．

謝辞 本研究の一部は科学研究費（22300034, 21013026, 20300036）の助成による．ここに記して謝意を表します.

参考文献

[1] 大石 哲也, 峯 恒憲, 長谷川 降三, 藤田 博, 越村 三幸 ”関連単語抽出アルゴリズムを用いたクエリ拡張” DEIM Forum 2009

[2] 加藤 慶一, 秋岡 明香, 村岡 洋一, 山名 早人”ミニブログにおける注目語抽出手法の提案と注目語を用いたメディア間での話題追跡” WebDB Forum2010

*形態素解析器 ”Mecab” http://mecab.sourceforge.net/