様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月 ２０ 日現在

機関番号：２５３０１
研究種目：基盤研究(C)（一般）
研究期間：2012～2015
課題番号：２４５００２９６

研究課題名（和文）意味理解に基づくマイクロブログのチャット的メッセージの分析・構造化に関する研究

研究課題名（英文）Analyzing Microblog Articles based on Text Understanding

研究代表者
菊井　玄一郎（Genichiro, KIKUI）

岡山県立大学・情報工学部・教授

研究者番号：８０３９５０１１

交付決定額（研究期間全体）：（直接経費）　　3,900,000円

研究成果の概要（和文）：マイクロブログ（例：twitter）でどういうことが話題になっているのかをその内容に基づいて分析するための研究を行った．主な成果として，マイクロブログの記事（書込み）から「誰がどうした」，「何がどうだ」といった係り受け情報を自動抽出する手法について検討し，前段の処理である文境界推定の改良や統計的な係り受け解析における特徴量の改善などにより従来手法より１０ポイントの精度向上を達成した．また，マイクロブログの時系列データからトレンド語やトレンドになった理由などを自動抽出する処理を提案した．

研究成果の概要（英文）：This work has two major results. One result relates to parsing micro-blog sentences. Since micro-blog articles often lacks sentence boundary markers, we first developed sentence boundary detector by using machine learning technique applied to word/character sequences. Then, we developed dependency analyzer including base phrase (so-called bunsetsu) chunker. Experimental results show our method outperforms existing software by 10 points. The other result is on trend analysis for micro-blog articles. We have developed a method for choosing an article which best describes a given burst word. The method applies sentence extraction to articles within the automatically identified burst period (for the given word).

研究分野：自然言語処理

キーワード：ウエブマイニング　テキストマイニング　自然言語解析

１．研究開始当初の背景

「マイクロブログ」の普及はめざましく，代表例であるtwitterは登録ユーザ１億人以上，一日あたりの投稿数は550万件以上と言われている(2011年3月，Twitter社)．マイクロブログのメッセージ（書き込み）には，１）ニュースへのコメントや特定テーマに対する議論などのような公共的な話題を持つ**「ニュース的」なもの**と，２）これら以外の私的あるいは**インフォーマルな「チャット的」なもの**に分けられる．従来，情報抽出の主な対象は前者であり，後者は情報内容に乏しく「意味のない」ものとして除外されることが多かった．しかし，後者も，たとえば，書き手自身あるいはその周囲の状況を即応的にテキスト化したものなどは実世界の状況を検知するための情報源として有用であり，地震の発生やインフルエンザの流行の兆しを捉えるなどの研究が行われている．
これら従来研究はキーワードの有無，発信時刻，発信場所などの特徴を用いて多数のメッセージを統計的に処理して得られるものであり，個々のメッセージの表現する行為や状況を認識しているわけではない．大きな理由は，チャット的なメッセージが対話調かつインフォーマルな文体であることから，既存の技術では十分に解析できないためである．しかしながら，**今後チャット的なメッセージの持つ情報を十分に利用するためには，個々のメッセージの表す行為や状況を認識することが不可欠**である．また，これができれば，**マイクロブログの大きな特徴であるチャット的メッセージにどのような情報がどういう割合で含まれているかなどを解明することができる**．

２．研究の目的

そこで，本研究では，**マイクロブログ(twitter等)におけるチャット的なメッセージ(書き込み）を対象として，個々のメッセージが表現する行為や状況などを認識する汎用的な手法を確立する**．また，これを用いてマイクロブログのチャット的メッセージの性質を定量的に明らかにする．具体的な目標は次の２つである．
１．マイクロブログ（のチャット的メッセージ）を対象とした解析（固有表現抽出，および，係り受け解析）手法の確立．
２．マイクロブログの内容分布やその経時的変化，ユーザ属性との相関等の定量的メディア分析．

３．研究の方法
(1)研究体制:
H24,25年度は研究代表者単独で本研究遂行し，H26年度より主に内容分析のサブテーマに対して分担者が加わった．なお，データ作成と評価実験については代表者の所属する研究室の大学院生１，２名（年度による）が担当した．
(2)研究の進め方
研究開始当初は全体を以下の４つの項目に分け，まず①を行い次に②～④を並行的に行う計画であった．しかしながら②については細分類が予想を超えて困難であったこととから既存の細分類を利用することとして，他のサブテーマを優先した．
①データの収集：マイクロブログ運営会社より公開されているAPIを用いて日本語のマイクロブログテキストを網羅的に収集し，１年分のデータから１万記事（投稿単位）を収集して解析処理研究用の基礎データとした．また，バースト分析ではキーワードを指定した収集も併用した．さらにこれらのうち8000記事に対して人手でチェックした文節係り受け情報を付与した．
②固有表現抽出法の検討：話者の行動を分類するという観点から固有表現細分類を行うとともに自動分類手法を検討する
③係り受け解析法の検討：チャット的メッセージは「標準的な書き言葉から逸脱した表現」が出現し，従来の新聞等でパラメータ学習された既存手法では解析精度が低下することが知られている．本研究では解析精度低下の原因を明らかにし，統計的手法によってこれを解決することを検討する．
④マイクロブログの内容分析：マイクロブログの内容分類により内容の分析と行うこと，また，バースト分析によりトレンドの分析や予測を試みる．

４．研究成果
(1)成果の概要
本研究における主な成果は次の通りである．
１）マイクロブログにおける係り受け解析精度低下の主要因を明らかにし，改善手法を示した．
２）マイクロブログのトレンドキーワードを説明する記事を同定する方法を示した
３）マイクロブログのトレンドキーワードの早期抽出手法を示した．
４）マイクロブログの内容分類に関する新たな手法を示した．
５）マイクロブログ約8000記事に対する正解係り受けデータを作成した．
以下では主要な成果である１）および２）について詳しく説明する．

(2)マイクロブログに対する文境界推定および係り受け解析の精度向上

①　成果の概要
CRF(Conditional Random Fields)を用いた文節境界および文境界推定において効果的な推定単位および素性を明らかにした．また，マイクロブログの（文節）係り受け解析の精度を向上させる上で文節境界推定の精度向上が鍵となること，そのためには文境界の推

定精度も向上する必要があることを示した．（文献[1]）

② 問題設定

マイクロブログに対する係り受け解析の精度は必ずしも十分ではない．この原因として次の二つの問題が考えらえる．一つ目は，パラメータが，言語的性質の大きく異なる新聞でチューニングされているというモデルミスマッチの問題，二つ目は，係り受け解析の前段で行う文境界の推定が不正確であるという問題である．文境界の推定は，新聞などの統制されたテキストであれば句点や疑問符など特定の記号を手掛かりに高精度に行える．しかし，マイクロブログでは句点等が適切に出現しないため，この手法では誤りが生じる．文境界推定についてはマイクロブログと文体的に近い「話し言葉」の分野において，機械学習を用いた文分割の研究が行われているが[2]，係り受け解析への影響を含めた効果は未評価である．

そこで，本サブテーマでは文境界推定，および，係り受け解析について，既存の統計的な手法を前提に素性の調整やマイクロブログを対象としたパラメータ学習により，係り受け解析の精度が程度改善できるかについて検討した．

③提案手法の概要

提案手法の流れを図１に示す．図において網掛けした所が本研究においてチューニングを試みた部分である．

図１：係り受け解析処理の流れ

最初に形態素解析１を行う．通常，形態素解析は文境界推定の後に行われるが，本研究では文境界推定の前に tweet 全体を一つの「文」とみなして実行し，文境界推定のための素性を生成する．次に CRF による文境界推定を行う．その後，形態素解析１の結果を無視して，再度，形態素解析を行う．なお形態素解析においては研究代表者らが作成した顔文字辞書[3]や文字列規則等により非正規的な表現の書き換えや削除を行う．なお形態素解析は MeCab(https://code.google.com/p/mecab/)（辞書は IPADIC）を用いた．

④文境界推定と文節境界推定

系列ラベリングにより統計的な文境界推定を行う．その際，入力を形態素列にした場合と文字列にした場合の両方を試みた．使用した素性および素性関数の詳細は文献[1]を参照されたい． なお，入力が文字列の場合も start-end 法によって形態素情報を素性として利用している．

文節境界推定も文境界と同様に形態素単位と文字単位の双方を試みた．

⑤係り先推定

係り先の推定には CaboCha (https://code.google.com/p/cabocha/) の係り先推定部分を用いた．なお，文節境界推定の際に文字単位の系列ラベリングを行うと形態素境界でない位置に文節境界が推定されることがある，これに対処するため，MeCab の機能を用いて文節境界を強制的に形態素の切れ目とするように再度形態素解析を行った．

⑥実験結果

提案手法を Twitter からランダムに収集した 7000 記事を対象に分割数 14 の交差検定によって評価した．その結果を表１～３に示す．

表１：文末境界推定の結果

	適合率	再現率	F値
形態素単位での学習	0.913	0.832	0.870
文字単位での学習	0.906	0.800	0.849
記号区切り	0.645	0.787	0.708

表２：文節境界推定結果の結果(%)

	適合率	再現率	F 値
形態素単位の CRF	0.933	0.913	0.923
文字単位の CRF	0.937	0.922	0.930
CaboCha	0.834	0.847	0.840

表３：係り先推定の結果

文境界推定	文節境界推定	係り先推定	正解率
baseline	Cab-ORG	Cab-ORG	0.617
CRF-MBM	Cab-ORG	Cab-ORG	0.643
CRF-MBM	CRF-MBM	Cab-ORG	0.711
CRF-MBM	CRF-MBM	Cab-MBM	0.716

表１は文末境界推定の結果である．形態素単位での学習精度が一番高い結果となった．間違えた箇所は，読点などの文末境界にならない記号が句点のように扱われた箇所である．また，文末に記号やスペースがない箇所は，文境界と判別することが難しかった．

表２は文節境界推定の結果である．文境界とは異なり，文字単位での学習の方が良い結果となった．誤ったのは複数文節からなる固有表現や名詞が連続している箇所である．これらの対策としては固有表現抽出の精度向上や助詞補完などが考えられる．

表３は係り先推定の結果である．ここで，文境界，文節境界中の CRF-MBM はマイクロブログコーパスで学習した CRF であり，表１，２の結果より文境界は形態素，文節境界は文字を単位とした．Cab-ORG は CaboCha の添付モデルを利用したものである．また baseline

は記号を手掛かりにした文境界推定である．一方，係り先推定においては「Cab-MBM」がマイクロブログコーパスで学習した CaboCha による係り先推定，Cab-ORG は CaboCha の添付モデルによる結果である．
従来の手法・パラメータの組み合わせ（１行目）に比べ，新たに学習した手法・パラメータの組み合わせにより，10 ポイント(0.10)精度が向上した．とくに文節境界推定が大きく結果に影響を与えたことが分かる．実際，正解の文節境界（oracle）を用いて係り先推定を行ったところ,推定精度は 0.88 となり，文節境界の精度向上が有効であることが分かった．また，文境界推定結果でも精度の向上が見られる．文境界が正しく推定できることで，文を飛び越えるような係り関係がなくなったのではないかと考えられる．

(3)バースト依存型リランキングによるトレンド表現文の推定

① 成果の概要

マイクロブログにおけるトレンドキーワードを説明する記事を同定する方法を示した．（文献[4]）

②問題設定

流行や話題に挙がっている事柄ことをここでは「トレンド」と呼ぶ．インターネット上のトレンドを知る手掛かりの一つとして，「検索急上昇ワード」と呼ばれる単語が挙げられる．検索急上昇ワードとは,ある期間に検索サイトに入力される検索語（単語あるいは複数単語からなる文字列）のうち,それより前の期間の入力数を比べて急増したものを言う．しかし,急上昇ワードだけでは,その単語が何であるか,また,何が起きて検索数が急激に増えたのか分かるとは限らない．例えば「グリッドロック」のような,あまり聞きなれないような単語の場合,多くの人はそれが何を意味するのかも分からない．また，もし仮にこの単語が渋滞現象を表すと知っていたとしてもなぜ急上昇ワードになったのか分からないかも知れない．
そこで，本サブテーマでは検索急上昇ワードを入力すると,その急上昇ワードにより特徴づけられるトレンドがどういうものかを表現するような記事（例えばグリッドロックの例では「＜グリッドロック＞「超」渋滞現象,震災で初確認（毎日新聞） - Y!ニュース」といった記事）を抽出する手法を提案する．

③提案手法の概要

処理全体は次の３つのステップから構成される．
・STEP1:Twitter からの記事の収集
・STEP2:強くバーストしている時間帯の検出
・STEP3:重要記事の抽出（記事のランキング）

STEP1 では API を利用して与えられた検索急上昇ワードを含む記事を収集する．
STEP2 では Kleinberg[5]の手法を改良した方法により，バースト区間を推定する．
STEP3 ではテキスト自動要約で用いられる重要文抽出の手法を用いて記事の重要度を評価する.具体的には,ある記事 A の重要度 W(A)をその記事に含まれる各名詞列の重要度の総和と考え,次の式（4）のように定義する．

$$W(A) = \sum_{k \in D} S(k) \quad \cdots\cdots (4)$$

ここで,D は記事 A における名詞列の集合，S(k)は名詞列 k に対するスコアである．また，S(k)は情報検索のインデックス語抽出で用いられる tf と ridf（残差 rdf[6]）を掛けた値であ る．なお tf はバースト期間中の全記事中の頻度（token 数），ridf における df は当該キーワードが出現する記事数である．

③ 評価実験と結果

提案手法がどのくらいの精度で適切な記事を抽出できるのか評価実験を行った．実験で用いた検索急上昇ワードは「Yahoo!検索ランキング」の「急上昇ワードランキング」で公表されている検索急上昇ワードである．処理対象記事は各検索急上昇ワードの発表後 2 日後に過去 10 日分の記事から収集した（API の制約）．提案手法で抽出した記事が正解か否かは，「Yahoo!検索ランキング」の「急上昇ワードランキング」に記載されている「急上昇ワードになった理由」の内容が含まれているものを正解として人手で判断した．比較手法として,重要度計算に定数および tf 値を用いたもの，さらに，バーストを考慮せずに単語の重要度を求めたものを用いた．結果を図２に示す．

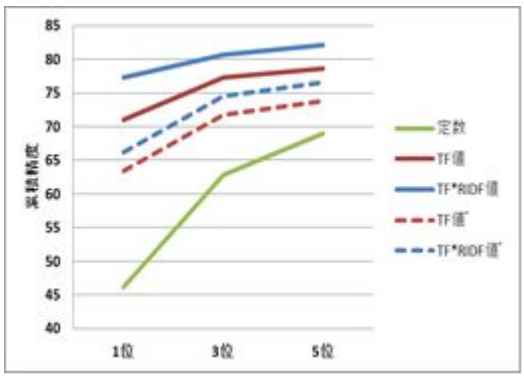

図２：急上昇ワードの要因抽出実験結果

図 2 の横軸の１位（k位）とは,記事をスコアの降順に並び替えたときの１位（k位）を表し,縦軸は各順位以内に正解記事を抽出できた急上昇ワードの相対頻度を表す．たとえば，３位の値が 70 である場合，３位以内に正解があったものが 70%存在したことを示す．破線の TF 値’,TF- RIDF 値’はバースト検出を行わずに,急上昇ワード発表以前に投稿された全記事を用いて重要度を求めた結果である．実線の TF 値，TF-RIDF 値についてはバースト検出を利用した結果である．
バースト検出と tf-ridf 値を併用することで１位正解率 77%を達成した．また 5 位までの出力で 82%の結果（急上昇ワード）に正解が含まれることが分かった．バースト検出の

導入で, 約 10%の精度の向上が見られたことから,バースト検出が有効であることが確認できた.

[1]下記学会発表①
[2]祖父江, 山本, 田村, 速水 :“ 音声認識結果の文末境界推定における識別モデルの評価 ” 言語処理学会第 15 回年次大会, pp582-585 , 2009.
[3] 渡邉, 高橋, 但馬, 菊井玄一郎, "系列ラベリングによる顔文字の自動抽出と顔文字辞書の構築", 言語処理学会第 19 回年次大会, pp.866-869, 2013.
[4]下記学会発表⑤
[5] Jon Kleinberg: Bursty and hierarchical structure in streams, In Proc. The 8th ACM SIGKDD International Conference on knowledge Discovery and Data Mining (2002)
[6] 北 研二, 津田 和彦, 獅々堀 正幹:情報検索アルゴリズム, 共立出版株式会社, pp. 43-44(2002)

５．主な発表論文等
（研究代表者, 研究分担者及び連携研究者には下線）
〔学会発表〕（計５件）
①難波悟史,門内健太, 但馬康宏,菊井玄一郎：マイクロブログに対する文境界推定および係り受け解析, 言語処理学会第 21 回年次大会, 2015.
②Yasuhiro TAJIMA, Genichiro KIKUI：Emotion estimation of comments on web news by SVM and naive Bayes based classifiers, the 2014 International Conference on Parallel and Distributed Processing Techniques and Applications PDPTA'14, 2014.
③門内健太,難波悟史, 但馬康宏,菊井玄一郎:"ランキング学習を用いたトレンド語の推定",第15回IEEE広島支部学生シンポジウム論文集, pp.451-452, Nov., 2013.
④菊井玄一郎：書き込み系メディアにおける言語表現のバーストを分析する, 電子情報通信学会北海道支部講演会（招待講演）, 2013.
⑤難波悟史, 但馬康宏, 菊井玄一郎：バースト依存型リランキングによるトレンド表現文の推定, 2013 年人工知能学会年次大会, 2013, 203-14in.

〔その他〕
ホームページ等
http://ai.cse.oka-pu.ac.jp/k/

６．研究組織
(1)研究代表者
菊井　玄一郎（KIKUI, Genichiro）
岡山県立大学　情報工学部　教授
研究者番号：80395011

(2)研究分担者
但馬　康宏（TAJIMA, Yasuhiro）
岡山県立大学　情報工学部　准教授
研究者番号： 00334467

(3)連携研究者
（　　）
研究者番号：