SNS 上での拡散を誘発する web ニュース説明文の調査と自動選択

興梠　紗和† 　木村　昭悟†† 　藤代　裕之† 　西川　仁†††

† 法政大学 社会学部　〒 194–0298 東京都町田市相原町 4342
†† 日本電信電話 (株) コミュニケーション科学基礎研究所　〒 243–0198 神奈川県厚木市森の里若宮 3-1
††† 日本電信電話 (株) メディアインテリジェンス研究所　〒 239–0847 神奈川県横須賀市光の丘 1-1
E-mail: †sawa.korogi.9k@stu.hosei.ac.jp, ††akisato@ieee.org, †††fujisiro@hosei.ac.jp,
††††nishikawa.hitoshi@lab.ntt.co.jp

あらまし　SNS の隆盛によりニュース消費に大きな変化が起きている．読者は，新聞，テレビから一方的にニュースを受け取るのではなく，SNS 上から読者の関心に合わせてニュースを選択するようになっている．この変化により，ニュースメディアは記事を SNS 上で拡散する必要に迫られているが，その方法は確立されていない．本研究では．SNS 上で多くの読者の目を引きつけるニュースの説明文を自動的に生成し配信することを目指す．この目標に向け，ニュースサイトの SNS 利用と反応を分析・抽出した拡散要因を用い，適切な説明文を候補文の中から自動的に選択する，ランキング学習に基づく手法を提案し，実際の Twitter の投稿と記事を用いた実験により，その有効性を示す．

キーワード　SNS, ニュース, 文書要約, ランキング学習

1. はじめに

SNS (ソーシャル・ネットワーキング・サービス) の隆盛により，ニュース消費の方法に大きな変化が訪れている．新聞やテレビなどマスメディア時代におけるニュース消費の方法は，報道機関から大衆＝マスに向けて発信されるニュースを消費者が一方的に受け取るだけであった．そのため，消費者がニュースに触れることができるのは，新聞では 1 日 2 回，テレビでは番組の時間帯だけであった．

しかし，インターネットの登場により，ニュースは新聞やテレビだけでなくインターネットでも得られるようになり，さらに近年では SNS とスマートフォンの台頭により，ちょっとした空き時間にも簡単にニュースを見られるように変化してきた．SNS 上には膨大なニュースが溢れ，整理して読者に提示するキュレーションサービスも登場した．これは，人々のニュース消費が追いつかなくなり，ニュース消費者がニュースを選択して消費せざるを得なくなっている状況を示している．

そのため，ニュースの配信者は，膨大なニュースの中から自らの配信するニュースを選んでもらうために，記事そのものの価値だけではなく，数多くの読者の目を引きつけ記事に誘導する様々な仕掛けを行う必要に迫られている．大多数のニュースサイトでは，SNS のアカウントを開設して情報を発信し，記事に SNS に投稿できるボタンを設置している．また，一部のニュースサイトやバイラルメディアでは，記事のタイトルや SNS 上で記事を紹介する投稿 (説明文) をより読者に訴求する構成に変更する試みもなされている．しかし，SNS 上で読者に訴求する説明文を構成する方法は未だ確立されていない．北米主要バイラルメディアの一つである「Upworthy」であっても，SNS で拡散する説明文を作るために，1 つの記事に対して見出し案を 25 本書き出した上で，編集者の経験と勘に基づいて説明文を選別しているとされる (注1)．すなわち，SNS 上で読者に訴求する説明文を構成する方法は，プロフェッショナルであっても非常に困難な問題の一つである．

そこで本研究では，SNS 上で読者に訴求するニュース記事の説明文を生成選択する指針を，データ工学の観点から明らかにすることを目指す．これにより，ニュース提供者が SNS 向けに適切な説明文を構成できるだけではなく，説明文を自動的に生成選択することも可能となり，数多くのニュースを消費者の目に触れされることに貢献できる．この目標に向け，本稿ではまず，同一記事の説明文を SNS へ複数回に投稿するソーシャルニュースサイトの説明文の構成とその反応を調査する．この調査結果に基づき，適切な説明文を候補文の中から自動的に選択する，ランキング学習に基づく手法を提案し，実際の Twitter 投稿と web ニュースを用いた実験によりその有効性を示す．

2. 関連研究

データジャーナリズム [3] の隆盛，既存ニュースメディアの SNS 活用の拡大，Flipboard や SmartNews などニュースキュレーションの拡大などが示すように，ニュースメディアと SNS がデータマイニングを架け橋としてつながろうとしている．データマイニングの主要国際会議 KDD2014 で，大手ニュースメディア技術者とデータマイニング研究者の発案により，ワークショップ NewsKDD が企画開催されたことは，この潮流を象徴するものである．この WS での発表 [2], [9], [13] を含め，ニュースと SNS をつなぐ研究は，近年多くの研究者の注目を集めている [11], [12]．Tsagkias ら [18] は，特定のニュースに言及したツイートを検索・発見するための具体的なクエリ構成法を考案した．Zhang ら [19] は，ツイートが主要ニュースに取り上げられる要因をコーパスから調査し，その結果に基づいて主要

(注1)：http:/slidesha.re/RfKQXM

ニュースに取り上げられるツイートを予測するモデルを提案した．Moniz ら [15] は，記事が Twitter で言及される回数を回帰予測することで重要ニュースを上位に推薦する手法を提案した．

しかし，これらの取り組みはいずれも，配信されたニュースを消費し，そのニュースに関してツイートを発信する，一般ユーザの視点が基本となっており，ニュース配信者と SNS との関係性に注目した研究は必ずしも十分ではない．ニュースメディアがどのようにニュースを配信すべきか，それを SNS 上でどのように実現するべきか，に関するデータ工学的なアプローチによる研究はなされておらず，本稿での報告が初めてとなる．

本稿の目標は，SNS 上で多くの読者の目を引きつけるニュースの説明文を自動的に生成し配信することにある．この目標の達成に向け，2 つの技術的課題が存在する．(1) ユーザの関心を惹起するような表現の解明，(2) そのような表現を含むニュースの説明文の自動生成．

ユーザの関心を惹起する表現の解明に関する試みとして，藤田らは，飲食店の情報が多数掲載されている web サイトから，その広告に典型的に用いられる表現を明らかにし，それを用いて広告を自動生成する手法を提案した [23]．生成された広告は実際に検索連動型広告として用いられ，一定の集客効果をもたらしたことが報告されている．検索連動型広告としての集客効果と，本稿の目指す SNS 上での読者の関心の惹起とは異なる目標ではあるが，読者の関心を惹く表現を明らかにした上でそれを利用するという点において共通している．

説明文の自動生成に関する試みとして，西川らは，文短縮を用いて読者を誘引するニュース記事の見出しを付与する方法を提案した [21]．この方法は，読者の関心を強く惹くと期待される表現があらかじめ与えられる前提で，その表現を保持したまま文を短く書き換える．しかし，この研究では読者の関心を惹起する表現を同定する課題は扱われていない．本稿は実際の Twitter 投稿を用いて読者の関心を惹起する表現を定量的に分析するものであり，西川らの研究に対して相補的な関係にある．

SNS での情報拡散については，穂積らが，mixi の「バトン」と呼ばれる日記書き込みが広がる速度とその拡散性は，友人を多く持つユーザがバトンを書くかどうかに起因することを示している [24]．拡散が急激に増加する要因について SNS の利用者に注目した研究として，榎らは，情報拡散の経路を調査し，情報拡散力の高いユーザを発見する手法を明らかにした [25]．安田は東日本大震災時のツイートを分析し，拡散されるツイートの特徴は【拡散希望】，情報源・地名等の客観的な情報を入れるとしている [22]．これらの研究は，拡散の経路やツイートからその要因を分析しているが，SNS 向けのニュース説明文を生成する方法には言及していない．

3. 調　　査

本節では，SNS において読者を誘引する要素を定性的に明らかにするために行った，2 つの調査の結果について述べる．第 1 の調査はヒアリング調査であり，SNS について深い見識を持つ編集者にヒアリングを行った．第 2 の調査はソーシャルニュースサイトの調査であり，ある記事に関して SNS に投稿される紹介文が SNS においてどのように反応されるかを分析した．

3.1 ヒアリング調査

SNS において読み手を誘引するために説明文が保持すべき性質を定性的に明らかにすることを目的に，ヒアリング調査を実施した．ヒアリングを行ったのは，ハフィントン・ポスト日本版編集者の伊藤大地氏，法政大学教授でジャーナリストの水島宏明氏，ジャーナリストの林信行氏の 3 名である．伊藤氏は，Yahoo!ニュース個人やハフィントン・ポスト日本版で記事を執筆しており，SNS における記事執筆に豊富な経験を有する．水島氏は，Twitter での RT 数 2000 以上，Facebook シェア数 1 万以上を記事を執筆しており，伊藤氏同様に SNS において読者を誘引する記事の執筆に関して豊富な経験を有する．林氏は，Twitter において 21 万人ものフォロワーを持ち，また Klout スコア (注2) が 78 であることから，SNS において読者を誘引する方法について豊富な知見を持つと期待できる．

これら 3 名に対してヒアリングを実施したところ，経験的なものではあるが，拡散の法則に共通項を見いだすことができた．読者の関心を惹く表現を用いるもしくは新しく構成する点と，読者自身との関連を想起させるような表現を利用する点である．

3.1.1 関心を惹く表現の利用と新しい表現の作成

読者の関心を惹く表現として，水島氏は，理由を示唆する「背景」「真相」「裏側」「なぜ」，驚きを意味する「衝撃の」「驚きの実態」「知られざる事実」「意外な」,「得」や「特」を連想させる「お得」「とっておき」「特大」「特報」「誰も知らない」「隠れた」などの表現が，読者の関心を惹きやすいことを指摘した．また，伊藤氏は，書き手の感情が伝わる「すごい」「驚くほど」「惹き付ける」といった表現の重要性を指摘すると共に，SNS が会話のように使われ，それ故に記事中における発言を意味するかぎ括弧とその内容が重要であると指摘した．林氏は，伊藤氏と同様に,「すごい」という表現が，多用されてはいるものの，読者の誘引のために重要な単語であることを指摘した．

また，水島氏は，既存の表現を使うのではなく，既存の表現を組み合わせ新たな表現を生成することによって，読者を惹きつけられることを指摘した．表現の組み合わせには 2 通りあり，1 つが旬なキーワードと単語を組み合わせる方法，もう 1 つは誰もがイメージできる表現と新しく登場した表現を組み合わせる方法である．後者に関しては，2007 年に新語・流行語大賞のトップテンに入賞した「ネットカフェ難民」 (注3) が例として挙げられる．誰もがイメージできる表現として「難民」，新しく登場した表現として「ネットカフェ」があり，これらを組み合わせ「ネットカフェ難民」という表現が生み出された．この表現の生みの親は水島氏である．

さらに，伊藤氏は，SNS 向けに構成される記事タイトルは，記事の要約ではなくタイトルだけで面白いと読み手に思われる必要があることを指摘した．その目的のためには，記事そのものの文脈とは無関係でも良く，単純に記事の内容を要約しただけのタイトルでは読み手を誘因することはできないと指摘した．

(注2)：http://klout.com SNS 上の影響力を数値化した指標の一つ．一般ユーザの平均スコアは 40-49 程度である．

(注3)：http://singo.jiyu.co.jp/nendo/2007.html

3.1.2 読者自身との関連を想起させる

林氏は，時間軸や場所で反応する読者に差があることを指摘した．同氏が以前,「東横線最終電車の一番後ろの車両に滑り込んだ」という内容のツイートをした際，同時刻に同じ場所にいるフォロワーから反応が多かったことから，SNS への投稿を，読者が「自分事」にできるような工夫があると良いと述べると共に，そのためには情報を絞り対象となる読者を限定することが，拡散に寄与する一つのヒントになり得ると指摘した．

3.2 ソーシャルニュースサイト調査

続いて，同一記事に関してその説明文を SNS へ複数回に投稿するソーシャルニュースサイトを対象に，説明文の構成とその SNS 上での反応を分析する．

調査対象サイトとして，本稿ではハフィントン・ポスト日本版を選択した．このサイトは，2005 年にアメリカで設立されたソーシャルニュースサイトの日本版として 2013 年より公開され，様々なジャンルのニュースを提供し，有識者・専門家・個人が記事にコメントを付けて議論できる．

本調査において使用したデータを表 1 に示す．今回の調査では，ハフィントン・ポスト日本版の公式 Twitter アカウント @HuffPostJapan のツイートの拡散状況とその特徴を調査した．このアカウントでは，7 時から 24 時の間に約 10 分後ごとにツイートが行われ，1 日の平均ツイート数は約 40 本である．各ツイートは，記事説明文とその URL から構成され，同一記事について平均 3 回ツイートされる．説明文はツイートごとに異なる場合と同一の場合とがあり，その間隔は 5-7 時間である．同一アカウントから同一記事に関する投稿が複数回行われるため，アカウントの影響を無視することができ，記事説明文やタイミングに見られる傾向から拡散に影響するヒントを発見しやすいと考えられる．以下に，ハフィントン・ポスト日本版の記事タイトル，及びこの記事について投稿されたツイートの例を示す．

記事の例

記事タイトル (注6)
旅の予習 MOVIE！いま行きたくなる，インド映画 5 選！公開中の話題作から名作までピックアップ
ツイート 1 (注7)
【新着ブログ】話題作から名作まで，いま行きたくなるインド映画 5 選 huff.to/1CZxu5d
ツイート 2 (注8)
【ブログ】旅の予習，いま行きたくなるインド映画 5 選 huff.to/1CZxu5d
ツイート 3 (注9)
【ブログ】様々な顔をもつ魅惑の国インドで生まれた傑作インド映画 5 選 huff.to/1CZxu5d

(注6)：http://huff.to/X5IjS9
(注7)：https://twitter.com/HuffPostJapan/status/510392407494762496
(注8)：https://twitter.com/HuffPostJapan/status/510477975041884160
(注9)：https://twitter.com/HuffPostJapan/status/510636514523742208

表 1　ソーシャルニュースサイト調査に使用したデータ

ツイート投稿期間	9 月 11 日から 9 月 30 日まで
ツイート収集期間	10 月 5 日から 10 月 11 日まで
収集ツイート数	858 (総投稿数の約半数)
記事数	276

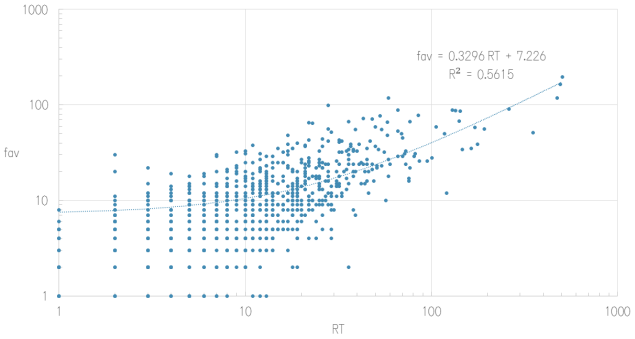

図 1　RT と fav の値の相関

3.2.1 調 査 方 法

ツイートの分析方法は以下の通りである．

（1） ハフィントン・ポスト日本版公式 Twitter アカウントから投稿されたツイートを手動で収集．その際収集したものは，ツイート本文 (説明文)，投稿日時，画像の有無，公式リツイート (以下 RT) の値，お気に入り（以下 fav）の値である．

（2） 説明文を頭の【】とその後の品詞別に分ける．

例：【新着ブログ】 / アメリカ / が / 原油価格 / で / ロシア / を / 追い詰める / 「新冷戦」 / の / 構造

（3） 【】，助詞，記号を除去する．

例：新着ブログ / アメリカ / 原油価格 / ロシア / 追い詰める / 新冷戦 / 構造

（4） 得られた単語を，以下に示す項目で記事と照合しながら調査した．(a) 最初に出てきたのは記事の何段落目か，(b) 記事本文の「」内に書かれているか，(c) 数字かどうか，(d) 単語の要素は何か (名詞 / 固有名詞 / 人名 / 動詞 / 形容詞 / 数字 / 画像 / 動画 / 否定語 / その他)，(d) 言い換え表現があるか．

（5） これらの単語を中心に，写真の有無による影響や言い換えられる単語の種類などについて分析した．

3.2.2 調 査 結 果

RT と fav の合計が最大のツイートは下記ツイート 1，最小は下記ツイート 2 であった．

ツイート 1(注10) RT: 506, fav: 197

【New】ダライ・ラマ「輪廻転生制度を廃止」発言に中国反発「絶対に認めない」

ツイート 2(注11) RT: 0 fav: 0

【新着ブログ】マンションの広さや間取り，住まい手本位の多様なプランへ

RT と fav の値の相関を図 1 に示す．図から見られるように，

(注10)：http://twitter.com/HuffPostJapan/status/509969622054219776
(注11)：http://twitter.com/HuffPostJapan/status/516897789088137218

表 2　画像の有無の拡散への影響

ツイート内容	日付	写真	RT+fav
【New!】指が届かない人用「iPhone 6 ケース」を自作（画像）(注12)	9/21	なし	3
【この発想はなかった】指が届かない人用「iPhone 6 ケース」を 3D プリンターで自作（画像）(注13)	9/21	あり	77
【コロンブスの卵】iPhone 6 で親指が届くようにする方法とは（画像）(注4)	9/22	なし	8
【この発想はなかった】指が届かない人用「iPhone 6 ケース」を 3D プリンターで自作（画像）(注14)	9/22	あり	48
【この発想はなかった】指が届かない人用「iPhone 6 ケース」を 3D プリンターで自作（画像）(注15)	9/23	あり	36

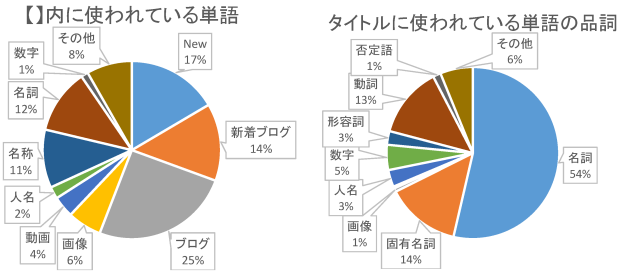

図 2　紹介文の単語分布

ツイート 1 のように RT や fav の数が大きなツイートは少なく，その合計値が 100 を超えるものは 40 本（4.6%）と限定された．逆に 50 未満は 732 本（85.3%）である．

続いて，ツイートに添付される画像に着目する．投稿されたツイートのうち，画像が添付されていたものは 285 本（33.2%）であった．該当ツイートの多くは記事の初回のツイートであり，2 回目以降のツイートでも添付されていることがある．画像が添付されているツイートは，2 回目以降でも RT や fav の値が大きいものが多く (表 2)，RT と fav の合計値上位 50 本のうち，1 本を除いてすべてに画像が添付されていた．逆に，RT や fav の合計値が 5 以下のツイートのうち，1 本を除いてすべてに画像が添付されていなかった．このことから，画像の添付が拡散に寄与する一つの要素であると考えられる．

ツイートされた記事説明文に含まれる総単語数は，【】内で 775 個，その後ろの説明文本体文で 7941 個，合計 8716 個であった．その要素を見ると，【】内では【New!】，【新着ブログ】，【ブログ】の 3 種類で約 55%を占めていた (図 2 左)．説明文では，圧倒的に名詞が多かった (グラフ 2 右)．

説明文に使われている単語を段落別に見ると，表 3 に示すように，第 1 段落に含まれる単語は，最も多いものの半分を切った．一般的な新聞では，記事の見出しを記事のリード文から構成する例が多いが，ハフィントン・ポスト日本版ではその方式

(注12)：https://twitter.com/HuffPostJapan/status/513585952913358848
(注13)：https://twitter.com/HuffPostJapan/status/513674155502604288
(注14)：https://twitter.com/HuffPostJapan/status/513885538957684736
(注15)：https://twitter.com/HuffPostJapan/status/514260403187384321

表 3　単語が記事の何段落目にあるか

段落	割合	段落	割合
1	43.3%	5	2.9%
2	9.9%	≧ 6	7.3%
3	5.8%	小見出し	2.2%
4	4.1%	なし	24.5%

を必ずしも採用していないことが見て取れる．また，記事本文に含まれていない単語も説明文の中で多数利用されており，24.5%にのぼった．このうち，記事に言い換え表現が含まれるものは 27.7%（539 個）あった．言い換えには，主に以下の 2 パターンがある．

（1） 省略型 - 例 :（記事）熟練の抜型職人→（説明文）熟練職人

（2） 置換型 - 例 :（記事）涙が止まらない→（説明文）号泣

言い換え表現に含まれる単語は名詞が多く，約 50%を占める．次に固有名詞（18.4%）が続き，動詞（16.1%），人名（6.5%），数字（3.7%），形容詞（2.3%）となっている．言い換えパターンは置換型が 60%である．もしこの観点を説明文の自動構成に導入する場合には，人間が想像し得る言い換え表現に関する辞書を構成する，もしくは単語や句の類似性を何らかの形で評価する必要がある．

また，説明文の内容における意外性も拡散に寄与すると考えられる．以下に示すツイート 3 及び 4 は，同一記事に対するツイートであり，いずれのツイートにも画像が添付されていたが，後の時刻でツイートしたツイート 4 が倍以上拡散されている．その理由として，旅雑誌の編集長が「旅をしない人が一番すごい」と書いてあることで，そのギャップに惹かれて拡散されたという仮説が考えられる．説明文の中で利用される単語のイメージの差が遠いことが，拡散の一つのヒントとなる可能性がある．

ツイート 3(注16) RT: 16 fav: 25

【新着ブログ】旅人は何かが欠けている - 創刊 10 周年の旅雑誌「Coyote」編集長に聞く，地図をつくる旅

ツイート 4(注17) RT: 43 fav: 72

【ブログ】創刊 10 周年の旅雑誌『Coyote』編集長,「旅をしない人が一番すごい」

その他の傾向としては,「初」や「発表」のような単語が入っているもの，時事的な話題，トヨタとホンダといった 2 つ以上の類似ドメインの固有名詞を比較するもの，お得・法則・ノウハウを伝えるものが拡散されている結果が見られた．

拡散するツイートに共通する一つの傾向として,「【New!】×固有名詞×名詞×動詞×画像」という構成が挙げられる． RT と fav の合計値トップ 30 ツイートに限定すると,【】内の 60%は【New!】，説明文の第 1 語の 50%は固有名詞，第 2 語の 70%は

(注16)：https://twitter.com/HuffPostJapan/status/514202687718817793
(注17)：https://twitter.com/HuffPostJapan/status/514079201209905152

名詞か固有名詞，第 3-5 語の 20%は動詞であった．これに対し，下位 30 ツイートを見てみると，第 3-5 語の動詞の割合が 6%に落ちる．また，説明文を助詞等を除いて 5 単語以内で構成することが，拡散に影響を与えると考えられる．

4. ニュース説明文の選択

前節までの調査結果を踏まえ，SNS 上で読者に訴求するニュースの説明文を自動的に生成配信する手法を考案する．この目的を達成するには，以下の 3 つの課題を解決する必要がある．

（1） ニュース記事から説明文の候補を生成する．
（2） 説明文の候補の中から最も適切な説明文を選択する．
（3） 選択した説明文を最も適切なアカウントから最も適切なタイミングで配信する．

第 1 の課題を解決する方法として，ニュース記事に文書要約 [20] や文短縮 [21] を適用し，要約文を説明文として採用する方法が考えられる．しかし，本目的への文書要約の直接的な適用は，以下の 2 つの問題がある．

（1） 文書要約そのものの技術的課題: 文書要約は，文書の中から重要文を抽出することで要約文とする抽出型要約と，文書の言語的特徴を解析して元文書にない新しい表現を構成する生成型要約がある．しかし，現在の技術水準では，生成型要約には自然性の担保に難があり，抽出型要約では人間と同等水準の高度な要約文を生成することは難しい．

（2） 要約文と説明文との性質の違い: 仮に高度な要約文を生成できたとしても，それが必ずしも SNS ユーザの目に留まる説明文になるとは限らない．第 3.1 節のヒアリング調査は，要約が本目的の説明文としては適切ではないことを示唆する．

これらのことから，本稿では，説明文候補生成については今後の課題とし，第 2・第 3 の課題，与えられた説明文候補の中から適切な説明文を選択し，それを適切なアカウントから適切なタイミングで配信する方法について検討する．

4.1 提案手法の概略

提案手法の概略は以下の通りである．

（1） 素性抽出（4.2 節）: 各説明文の候補 h から，説明文候補そのもの，説明文候補とニュース記事 a との関係，及び説明文候補の投稿情報 m に関する素性 $\boldsymbol{x} \in \mathcal{R}^N$ を抽出する．

（2） モデル学習（4.3 節）: あるニュース記事 a に言及した複数の説明文候補 $\{h_c\}_{c=1,2,C_{tr}}$ 及びそれぞれから抽出した素性 $\boldsymbol{x}_c$ と，各候補に付与された評価値 $v_c \in \mathcal{R}$ を学習データとして，説明文評価モデル $f : \mathcal{R}^N \to \mathcal{R}$ を学習する．

（3） 説明文評価: あるニュース記事 $\hat{a}$ に言及した複数の説明文候補 $\{\hat{h}_c\}_{c=1,2,C_{te}}$ からそれぞれ素性 $\hat{\boldsymbol{x}}_c$ を抽出し，それを評価モデル f に与えた際の推定評価値 $\hat{v}_c$ で説明文を評価する．

4.2 素性の抽出

説明文候補から抽出する素性は，説明文候補の善し悪しを評価する上で重要な意味を持つ．前節で説明したように，説明文候補の素性は，以下の 3 つの情報源を基にして抽出する．(a) 説明文候補そのもの h_c から抽出する素性，(b) 説明文候補 h_c とニュース記事 a との関係から抽出する素性，(c) 説明文候補の投稿情報 m_c から抽出する素性．

表 4 抽出する素性の種類 (b: 二値，i: 整数，r: 実数)

種別	素性	型	次元数
画像	Image	b	1
時刻	Time-diff	r	1
	Time-period	b	5
	Time-weekday	b	1
キーワード	Keyword-reason	b	9
	Keyword-method	b	4
	Keyword-surprise	b	5
	Keyword-special	b	4
	Keyword-publisher	i	1
	Keyword-paragraph	b	2
構造	Structure-bracket	b	1
	Structure-keyword	b	1
	Structure-propernoun	b	1
	Structure-noun	b	1
	Structure-title	r	1
	Structure-length	r	2
ユーザ	User	r	4
分散表現	Distribute-sim	r	1
	Distribute-farwords1	r	2
	Distribute-simfar	r	1
	Distribute-farwords2	r	2
	Distribute-paragraph	r	1
総計			51

本稿では，説明文を配信する SNS として Twitter を対象とし，説明文をツイートで配信することを想定して，各素性を設計する．その構成について，以下で具体的に述べる．抽出する素性すべてをまとめた表を表 4 に示す．

画像の有無 (Image)

第 3.2.2 節の表 2 に示したように，画像が拡散に寄与する一つの要素であると考えられる．画像認識分野での標準的な特徴量を用いても画像付きツイートの拡散度合いをほとんど推定できないとする先行研究 [5] に従い，ツイートの画像の有無のみを素性として採用する．具体的には，Twitter 公式ページでタイムラインに展開画像が表示される pbs.twimg.com へのリンクの有無を，素性 (Image) として採用する．

投稿時刻 (Time)

第 3.2.2 節の後半で示したように，ツイートの投稿時刻はその拡散に寄与する一つの要素である．本稿では，投稿時刻に関する以下の 3 種類の素性を採用する．

- (Time-diff) ツイート投稿時刻と記事投稿時刻との差
- (Time-period) ツイート投稿時間帯: 2-5 時，6-9 時，10-15 時, 16-19 時，20-25 時
- (Time-weekday) ツイート投稿曜日: 週末 or 平日

キーワード (Keyword)

第 3.1 節の調査結果から，SNS 熟練者が特定のキーワードを意識的に用いる可能性が示唆された．同様に，第 3.2.2 節では，特定のキーワードがユーザの関心を惹くことが示唆された．この調査結果に基づき，本稿では，以下のキーワードの有無を素性として採用する．

- (Keyword-reason) 理由を示唆: 背景，理由，秘密，なぜ，裏側，真実，知られざる，実態，とは
- (Keyword-method) 方法・法則に関連: 法則，作り方，方法，秘訣
- (Keyword-surprise) 驚きに関連: すごい，凄，驚き，素敵，びっくり
- (Keyword-special) 得や特を連想: 特別，スペシャル，とっておき，だけの

また，多くの web ニュースでは，web 検索結果で上位にランクさせる SEO 対策の一つとして，HTML の META タグにキーワードを設定している．これら META タグキーワードがツイートに含まれる数も，素性 (Keyword-publisher) として採用する．さらに，第 3.2.2 節の表 3 に示すように，記事の第一段落に含まれる単語がツイートに数多く登場する．そこで，特にツイート中の一般名詞及び固有名詞が記事第一段落に含まれるかどうかを素性 (Keyword-paragraph) として採用する．

ツイートの構造 (Structure)

第 3.2.2 節の図 2 に示したように，ツイートの先頭に括弧付きの単語もしくは単語群を表示することで，読者の注意を引き寄せる手法が用いられており，かつ括弧内に含まれる特定のキーワードが拡散に寄与することが示唆された．これらのことから，ツイートの先頭に括弧を含むかどうか (Structure-bracket)，及び先頭の括弧内に特定のキーワード (New，新着，画像) を含むかどうか (Structure-keyword) を素性として採用する．

また，第 3.2.2 節で示唆されたように，実際に拡散したツイートでは，先頭の括弧の後に固有名詞・名詞の順に形態素が続く構成が多いことから，先頭の括弧を除く最初の形態素が固有名詞かどうか (Structure-propernoun)，先頭の括弧及び助詞を除く 2 番目の形態素が名詞かどうか (Structure-noun) を素性として採用する．形態素解析には MeCab [10] を用いた．

さらに，記事タイトルとは異なる情報を提供するツイートの方がより拡散しやすい，長いツイートは一般にあまり好まれない，などの傾向を考慮し，記事タイトルと tweet との編集距離 (Structure-title)，ツイートの総単語数及び総文字数 (Structure-length) を素性として採用する．

投稿ユーザの統計情報 (User)

一般に，影響力のあるユーザによるツイートは拡散しやすいため，ツイートするユーザのフォロワー数，フォロー数，リスト被登録数，総投稿数を素性 (User) として採用する．

単語の分散表現 (Distribute)

第 3.2.2 節で示唆されたように，ツイートの内容に意外性・ギャップ・対比がある場合に拡散されやすいを考えられる．この要素を，単語の分散表現 [26] を用いて算出することを試みる．

単語分散表現は，単語 w を固定次元の実数値ベクトル $\boldsymbol{y}(w) \in \mathcal{R}^M$ で表現する枠組であり，このベクトルの中に単語の持つ言語的な性質を持たせ，類似した意味を持つ単語が類似したベクトルを持つように学習する．本稿では，単語分散表現において標準的なツールである word2vec [14] を採用し，学習用コーパスとして Wikipedia 日本語版を利用する．

この分散表現モデルを用いて，意外性・ギャップに関連する以下の 4 種類の素性を抽出する．以下，分散表現ベクトル間の類似度 $s : \mathcal{R}^M \times \mathcal{R}^M \to \mathcal{R}$ としてコサイン類似度を用い，候補ツイート h_c をそのツイートに含まれる単語の集合と見なす．

- (Distribute-sim) ツイート中で最も類似度の高い名詞対の類似度．

$$s_1(h_c) = \max_{w_1, w_2 \in g(h_c)} s(\boldsymbol{y}(w_1), \boldsymbol{y}(w_2)),$$

ただし，$g(h_c) \subset h_c$ はツイート h_c 中の名詞の集合である．

- (Distribute-farwords1) ツイート h_c の中で他の単語と最も離れている単語 $o(h_c)$ と，その単語を除いたツイート中の単語集合との類似度，及び単語を名詞に限定した場合．

$$s_{2,1}(h_c) = s\left(\boldsymbol{y}(o(h_c)), \sum_{w \in h_c - \{o(h_c)\}} \boldsymbol{y}(w)\right),$$

$$s_{2,2}(h_c) = s\left(\boldsymbol{y}(o(g(h_c))), \sum_{w \in g(h_c) - \{o(g(h_c))\}} \boldsymbol{y}(w)\right).$$

- (Distribute-simfar) ツイート中で他の単語と最も離れている単語と，最も類似度の高い名詞対との類似度．

$$s_3(h_c) = s\left(\boldsymbol{y}(o(h_c)), \sum_{w \in p(h_c)} \boldsymbol{y}(w)\right),$$

$$p(h_c) = \arg \max_{w_1, w_2 \in g(h_c)} s(\boldsymbol{y}(w_1), \boldsymbol{y}(w_2)),$$

- (Distribute-farwords2) ツイート中で他の単語と最も離れている単語との類似度が最も小さい単語との類似度，及び単語を名詞に限定した場合．

$$s_{4,1}(h_c) = \min_{w \in h_c - \{o(h_c)\}} s(\boldsymbol{y}(o(h_c)), \boldsymbol{y}(w)),$$

$$s_{4,2}(h_c) = \min_{w \in g(h_c) - \{o(g(h_c))\}} s(\boldsymbol{y}(o(h_c)), \boldsymbol{y}(w)).$$

さらに，記事の第一段落とツイートの類似度についても，単語の分散表現モデルを用いて計算し，素性 (Distributed-paragraph) として採用する．

4.3 モデル学習

ある特定のニュース記事について言及した複数の説明文候補の中から適切な候補を一つ選択する問題は，説明文候補の適切性を表現する評価値を予測する回帰と，その回帰結果としての評価値が最大になる候補を選択する最大値選択とに，分解することができる．説明文を配信する SNS として Twitter を対象とする場合，説明文の評価値としてリツイート (RT) 及びお気に入り (fav) の数を利用することができる．すなわち，本稿におけるモデル学習は，説明文候補としてのツイートの素性 $\boldsymbol{x}_c$ から RT・fav の数 $v_c \in \mathcal{R}$ を予測する回帰問題 $f : \mathcal{R}^N \to \mathcal{R}$ として定式化できる．

しかし，複数の説明文候補の中から適切な候補を選択する目的において，説明文候補からそれぞれ予測した評価値の絶対値に意味はなく，評価値の大小のみが重要となる．また，評価値

として RT や fav を採用した場合，それらの絶対値を予測する回帰モデルは，説明文候補の評価だけではなく，ニュース記事自体の評価も内包されることになるため，説明文候補の評価として望ましい定式化ではない．

そこで本稿では，説明文候補としてのツイートから RT・fav の数を直接推定する回帰問題ではなく，RT・fav の数の順位を推定するランキング学習 [4] の問題として定式化する．ランキング学習は，主に情報検索で広く用いられる学習の枠組であり，クエリに適合する web ページをページの重要性やクエリとの適合度などに基づいて順位付けする問題である．本稿では，クエリをニュース記事，web ページを説明文候補と，それぞれ読み替えることにより，ランキング学習を適用する．

5. 実　　験

5.1 実験条件

ニュース記事説明文の候補として，Twitter Firehose API を通じて取得できる日本語ツイートのうち，2014 年 8 月 14 日から 12 月 22 日までのおよそ 3ヶ月分を利用し，それらのうち，(a) ハフィントン・ポスト日本版の記事へのリンクを含み，(b) 1 回以上 RT or fav され，(c) 同一記事へのリンクを含むツイートが他に 2 つ以上あるツイートを抽出した．また，これら抽出したツイートからリンクされたハフィントン・ポスト日本版の記事を収集し，タイトル・本文・メタ情報などを抽出した．上記の手順により，総計 702 記事，2839 ツイートを収集した．

説明文候補の評価値として，ツイートの RT 数と fav 数の合計を採用する．説明文候補を選択する手法の評価尺度として，情報検索における順位付けで一般的に用いられる NDCG [6] を採用する．モデル学習では，ランキング学習の標準ツールである SVMrank [8] (RBF カーネル) を用いた．モデル学習及び評価は以下の手順で行う．まず，全データを 5 つの部分データに分割し，そのうちの任意の 4 つを用いた 4-fold 交叉検定によりハイパーパラメータを設定する．設定したハイパーパラメータと 4 つの部分データを利用してモデルパラメータを学習し，残りの 1 つで評価する．上記の手順を，評価用の部分データを変えて 5 回行い，平均の評価値を算出する．

5.2 実験結果

提案手法の主要な貢献は，目的に合わせた素性の開発にあり，実験では素性の善し悪しを NDCG を用いて評価する．提案手法で用いる 51 次元すべての素性を用いた場合に加え，素性種別ごと，およびランダムな予測と比較し，提案する素性がどの程度有効であるか，どの種別の素性がより有効であるか，調査する．また，提案手法でランキング学習を採用した妥当性を検証するため，同じ素性・同じ評価値・同じカーネルを用いた回帰学習との比較も行う．回帰学習には，SVMrank と互換性のある SVMlight [7] を用いた．

結果を図 3 に示す．この図から，提案する素性の組合せによりランダムな出力および各個別の素性種別を用いた結果よりも有意に優れた結果 (NDCG@1=0.379, NDCG@3=0.583) を示したことがわかる．また，説明文候補の評価値の順位を予測するランキング学習を用いた提案手法が，説明文候補の評価値

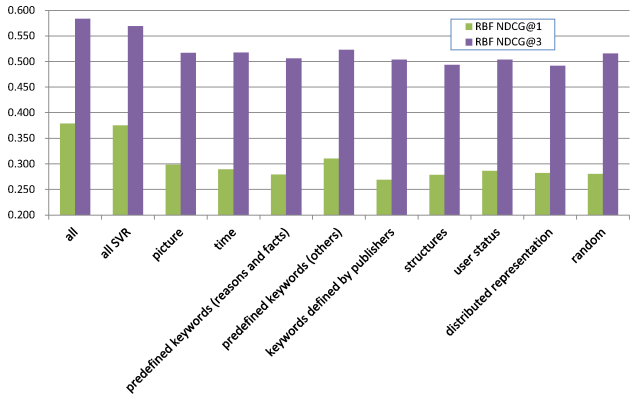

図 3　実験結果．all: 提案手法， all SVR: 提案手法の素性を用いた回帰学習，その他: 各個別素性を用いたランキング学習．

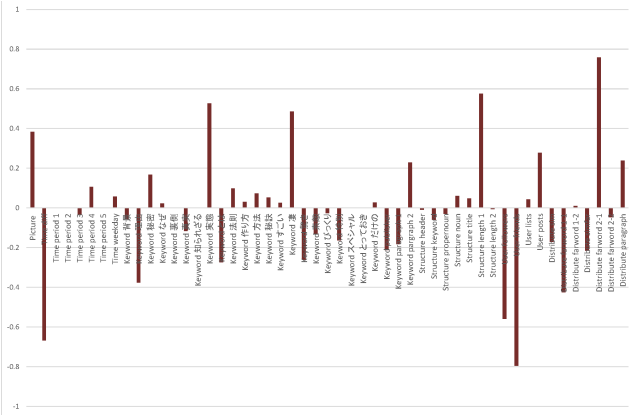

図 4　線形 SVM における各素性の係数

を直接予測する回帰学習を用いた場合に比べ，有意に良い結果となった．個別の素性種別では，画像の有無 (Image)，投稿時刻 (Time)，いくつかのキーワード，投稿ユーザの統計情報 (User)，単語の分散表現 (Distribute) が説明文候補の選択に有効であることが示唆される．

SVM のカーネルを線形カーネルとした際の係数を図 4 に示す．この図から，説明文候補の評価は，画像の有無，説明文の単語数，分散表現で最も離れている単語対の類似度，およびいくつかのキーワードと正の相関があり，ニュース記事投稿から説明文投稿までの経過時間，分散表現で最も離れている単語とそれ以外の単語との類似度，及びいくつかのキーワードと負の相関があることが，見て取れる．

ユーザに関する素性において興味深いのは，ユーザのフォロー数 User-friend およびフォロワー数 User-followers が強い負の重みを得ていることである．直観的には，ユーザのフォロワー数が多いほど RT や fav が得られ やすいように思われるが，定量的な分析の結果として得られたのは逆の結論である．

キーワードについては，予測に大きく寄与したものとそうでないものが混在している．特に大きく寄与し，予測に対して正の重みを持つものとして「実態」「凄」といったものがあり，これらの表現の利用は RT や fav を得るにあたって有効である．負の重みを持つものとしては「理由」「とは」「驚き」があった．

これら表現は，ユーザからは，使い古された，いくらか魅力に欠ける表現であるとみなされている可能性がある．

単語の分散表現において離れている 2 つの単語が Tweet に含まれていると良い結果を示すという結果は，意外性やギャップが拡散に寄与するという仮説を支持するものであるのと同時に，これまでにない単語の組み合わせを作ると良いという水島氏の指摘の有効性を示唆するものと考えることができる．

6. まとめと今後の課題

本稿では，SNS 上で多くの読者の目を引きつけるニュースの説明文を自動的に生成し配信することを目指し，ニュースサイトの SNS 投稿を多角的に分析すると共に，その分析結果を用いて，適切な説明文を候補文の中から自動的に選択する手法を提案した．実際の Twitter の投稿とニュースサイトの記事を用いた実験により，提案手法の有効性を検証した．

本稿で紹介した取り組みは，研究の最終的な目標である，SNS 上で多くの読者の目を引きつけるニュースの説明文を自動的に生成し配信するための第一歩に過ぎず，数多くの課題が残されている．その中でも代表的な課題を，以下に記載する．

より良い素性の開発 本稿で紹介した素性は，低レベルのテキスト処理・マイニング技術を用いて実装されており，改良の余地は多い．特に，投稿文字数に制限のある Twitter に説明文を投稿する場合には，文字数制約を満たすために文短縮や言い換え表現が頻繁に用いられる．照応解析や含意認識など，より高水準の自然言語処理技術の導入が必要になるであろう．また，時事性を反映した固有名詞の選択，背後にある前提知識の利用などが，より優れた候補文の選択に大きく寄与するであろう．

他の新聞社の記事への展開 本稿では，ハフィントン・ポスト日本版を対象としたが，手法の枠組そのものは他のニュースサイトにも適用可能である．しかし，素性の構成を含めた手法の詳細については，ニュースサイトの特性を考慮する必要がある．各サイトに適合した素性を検討すると共に，サイトごとに異なるモデルが必要となる．また，主要新聞社を対象して，記事数やツイート数が増加すると，モデル学習に要する計算時間が大きな課題となる．ランキング学習を採用する提案手法は，ランキング学習アルゴリズムの多くがペアワイズ学習の戦略を採用しているため，学習データ数の増大に伴って指数的に学習時間が増大する．オンライン学習 [17]，サンプリングを用いた高速化 [16] などの導入が有効であろう．

候補文の自動生成 本稿では，説明文の候補が与えられているときに適切な候補を選択する方法を主に検討し，説明文候補を生成する部分は今後の課題としていたが，候補文を生成する部分は，本研究の最終的な目的を達成する上で最も本質的な要素技術の一つである．ニュース記事の要約が記事の説明文として必ずしも適切ではないという第 3.1 節の調査結果はあるものの，文書要約 [20] や文短縮 [21] を組み合わせた要約文は有力なベースラインの一つとなり得る．また，典型的な説明文のテンプレートを用意して，そのテンプレートの単語を記事に応じて入れ替えることで候補文を生成する方法も考えられる [1]．

謝　辞

第 3.1 節のヒアリングに応じていただいた伊藤大地氏，水島宏彰氏，林信行氏，及び第 3.2 節の調査にご協力いただいた法政大学 藤代ゼミの皆様に感謝する．

文　献

[1] Alfonseca et al. Heady: News headline abstraction through event pattern clustering. In *Proc. ACL*, 2013.
[2] Chuang et al. Large-scale topical analysis of multiple online news sources with media cloud. In *Proc. NewsKDD*, 2014.
[3] Gray et al. editors. *The Data Journalism Handbook*. Oreilly & Associates Inc, 2012.
[4] He et al. A survey on learning to rank. In *Proc. ICMLC*, 2008.
[5] Ishiguro et al. Towards automatic image understanding and mining via social curation. In *Proc. ICDM*, 2012.
[6] Järvelin et al. Cumulative gain-based evaluation of IR techniques. *ACM Trans. Information Systems*, 2002.
[7] Joachims. *Learning to Classify Text Using Support Vector Machines*. Kluwer/Springer, 2002.
[8] Joachims. Training linear SVMs in linear time. In *Proc. KDD*, 2006.
[9] Kim et al. Diversity-seeking users and their influence on social news sites. In *Proc. NewsKDD*, 2014.
[10] Kudo et al. Applying conditional random fields to Japanese morphological analysis. In *Proc. EMNLP*, 2004.
[11] Lehmann et al. Finding news curators in Twitter. In *Proc. WWW*, 2013.
[12] Lehmann et al. Transient news crowds in social media. In *Proc. ICWSM*, 2013.
[13] Mclnerney et al. Discovering newsworthy tweets with a geographical topic model. In *Proc. NewsKDD*, 2014.
[14] Mikolov et al. Distributed representation of words and phrases and their compositionality. In *Proc. NIPS*, 2013.
[15] Moniz et al. Improvement of news ranking through importance prediction. In *Proc. NewsKDD*, 2014.
[16] Sculley. Large-scale learning to rank. In *Proc. NIPS Workshop*, 2009.
[17] Suhara et al. Robust online learning to rank via selective pairwise approach based on evaluation measures. **人工知能学会論文誌**, 2013.
[18] Tsagkias et al. Linking online news and social media. In *Proc. WSDM*, 2011.
[19] Zhang et al. Which tweets will be headlines? a hierarchical Bayeian model for bridging social media and traditional media. In *Proc. SNAKDD*, 2014.
[20] 奥村, 難波 **テキスト自動要約**. オーム社, 2005.
[21] 西川 他 クエリ依存文短縮と見出し生成への応用. **情処研報**, 2013.
[22] 安田 ソーシャルメディア上の情報拡散の特性 - 東日本大震災時のデマの事例とハブの役割. In **関西大学 社会学部紀要**, 2013.
[23] 藤田 他検索連動型広告の自動生成と集客効果の測定 - 飲食店ドメインを例題に. **情処論**, 2011.
[24] 穂積 他Sns における情報伝達の速度とその拡散性. In **情処全国大会**, 2008.
[25] 榎 他 ソーシャルメディアの情報拡散分析. In *DEIM* **予稿集**, 2014.
[26] 渡邉 自然言語処理分野におけるディープラーニングの現状. In *IBIS*, 2013.