SNS 上での拡散を考慮したニュース記事中重要文の自動選択

永山 孝太† 木村 昭悟‡ 藤代 裕之†

†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

E-mail †kouta.nagayama.8u@stu.hosei.ac.jp, fujisiro@hosei.ac.jp, ‡akisato@ieee.org

あらまし ソーシャルネットワーキングサービス（SNS）の普及により，ニュースメディアはニュース記事を自社サイトだけでなく SNS を利用してより多くの読者に届ける必要に迫られている．本研究では，SNS 上で拡散するニュース説明文を自動的に生成するために，記事の中から拡散に寄与する重要文を自動的に選択する仕組みを開発することを目指す． 従来の重要文選択手法は，与えられた文書を端的に表現する文を選択するものであり，拡散に寄与するかどうかについては十分に考慮されていない．本論文では，実際に拡散された SNS 投稿に用いられた内容を含む記事中の文に対してラベルを付与し，記事とラベルを教師データとして recurrent neural network (RNN) を学習することで，拡散に寄与する文を自動的に選択する手法を提案する．実際のニュース記事と Twitter の投稿を用いた実験により，提案手法が高い精度で重要文を選択できることを示す.

キーワード SNS，ニュース，文書要約，recurrent neural network

1. はじめに

SNS（ソーシャル・ネットワーキング・サービス）が登場し，ニュースを取り巻く環境は大きく変化した．新聞やテレビといったマスメディアは消費者にニュースを一方的に伝えることが出来る装置であった．夕刊や朝刊，番組枠といった時間的な制限はあったものの，ニュースのみを読者や視聴者に届けることができた．しかしながら，SNS によってニュースの伝達のあり方が全く異なるようになった．消費者はスマートフォンを用いて，いつでも，どこでも，簡単にニュースを見られるようになった．SNS にはニュースだけでなく，友人や知人の投稿から，企業の広告まで多様なコンテンツが溢れるようになった．ニュースメディアは，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．一部のニュースサイトでは，SNS 上で記事を紹介する投稿 (以降，ニュース説明文と呼ぶ) をより読者に訴求する構成に変更する試みがなされている．このニュース説明文は，従来の記事の見出しとは異なるものである．マスメディアはニュース専用の伝達装置であり，どのようなニュースであるかを読者が判断できる記事の要点をまとめた表現が求められた．一方，SNS では消費者の目を引きつけ，拡散される説明文が求められる．新聞で使っている見出しをそのまま SNS の説明文に用いているニュースサイトもあるが，朝日新聞のソーシャルメディア担当者は，従来の見出しでは拡散されないとして，SNS に投稿する際には結論となる部分を書かないなどニュース説明文の書き方を工夫していることを明らかにしている（注1）．また，米国における主要バイラルメディアの一つである Buzzfeed では，ニュース説明文と写真のパターンを最大 12 個構成して個別に配信した後，数時間後に最も拡散されたものに全てを差し替えるといった工夫をしている．また，BuzzFeed 独自のシステムにより，投稿が Twitter 上で数多くリツイート（RT）されると予測がされた場合，サイト上でツイートボタンを通常より大きくするといったことも行っている（注2）.

注1 http://web-tan.forum.impressrd.jp/e/2015/07/14/20270

注2 https://newspicks.com/news/878625/body/

このように，ニュースメディアは，SNS 上でニュース記事が拡散されるために，ニュース説明文の構成に様々な工夫を行っているが，依然として編集者が経験と勘を駆使して行っており，どのような構成が SNS 上で訴求するのか，その具体的な方法は未だ確立されていないのが現状である．

本研究では，ニュースメディアがニュースの読者に適切なニュース記事を提供するための一手段として，SNS 上でより多くの読者にニュース記事が読まれるためのニュース説明文を自動で生成することを目指す．完全に新規な説明文を自動的に構成することは非常に困難であることから，本研究では，対象のニュース記事を要約することで説明文を生成するアプローチを採用する．文書要約の技術は，近年，ニュース関連分野への適用が数多く報告されている．記事を自動で要約して配信するアプリ「Vingow」や，米国の大手新聞社 New York Times が掲載した読者の地域に応じて記事の内容を動的に書き換える geocoding 手法の適用などがその代表例である(注 3)．しかしながら，これらの事例は，与えられた文書を端的に表現する文を重要文と位置付け，抽出し，要約するという従来の要約技術の範疇にとどまるものであり，SNS 上での拡散に寄与するかどうかという観点は考慮されていない．

この課題に対して，本論文では，ニュース記事について言及した Twitter 上の投稿（ツイート）のうち，実際に SNS 上での拡散に大きく寄与したものを教師情報として利用する，教師付き学習によるニュース説明文の自動生成手法を提案する．具体的には，まず，あるニュース記事について言及したツイートのうち，最も拡散されたものに着目し，そのツイートがニュース記事のどの文から情報を取得したかを調査する．この調査により，SNS 上で拡散される説明文を構成するためには記事中のどの文を選択するべきか，すなわち拡散を考慮した重要文選択のための教師情報を構成する．この教師情報を用いて，拡散を考慮した重要文を選択するためのモデル学習を行う．モデルには，自然言語処理の様々なタスクで近年大きな成果を挙げている recurrent neural network (RNN) を採用する．このようにして学習を行ったモデルを使用して，与えられたニュース記事の中から拡散に寄与する文を予測し，この文を説明文として採用する．

注 3　http://www.businessnewsline.com/news/201511080445270000.html

2. 先行研究

本研究は，SNS 上における拡散の要因分析と，抽出型文章要約に関連する．それぞれの分野でこれまでに数多くの関連研究が行われている．

まず，SNS 上における拡散の要因分析に関する関連研究を紹介する．穂積らは，当時の有力 SNS の一つであった mixi を対象に，「バトン」と呼ばれる日記書き込みが拡がる速度とその拡散性を分析し，mixi 上で数多くの友人を持つユーザがバトンを書くことで爆発的に拡散することを明らかにした[1]．濱岡は，映画に関するツイートの分析を行い，「キャンペーン，プレゼント情報」「公開日，公開時間」などのプラスアルファの情報を付加することで RT 数が大きくなることを明らかにした[2]．また，拡散されている情報の信頼性にかかわる研究として，藤川らは，SNS 上の流言に対するユーザの反応が情報の根拠とともに書かれているかを判断するシステムを開発した[3]．斉藤らは，異常検知の考え方を用いて，SNS 上で長期間流行する話題を検出する手法を提案した[4]．上記の通り，これらの研究では，拡散経路や話題抽出，ユーザの行動分析やツイートの分類などが主に行われているが，いずれもソーシャルメディア上に投稿されたコンテンツそのもの及びその拡散経路の影響について言及しており，投稿コンテンツがリンクする記事などとの関係性は明らかにされていない．我々はこれまでに，SNS 上でのニュース説明文の拡散要因を，説明文そのものだけではなくニュース記事との関係性も利用して明らかにしてきた[5]．本研究はその流れをさらに推し進めるものである．

続いて，文書要約に関する関連研究を紹介する．文書要約は，重要文抽出と文短縮の二つの要素技術で構成されている[6]，本論文が対象とするタスクは重要文抽出である．重要文抽出では，単語の共起関係を基にした手法[7]，クラスタリングを行いトピック抽出した手法[8]，単語ベクトルを用いて単語にスコアをつけ，文のスコアを，文を構成する単語のスコアの合計として評価し，スコアの高い文を重

要文として用いる手法[9]，文の依存関係の木構造と単語の依存関係の木構造を入れ子とした入れ子依存木を用いて要約を行う手法[10] などが知られている．しかし，これらの研究は，文章を端的に表現する文を選択するものであり，説明文が SNS 上でどのように拡散されるか，という観点は含まれていない．

3. 拡散を考慮した重要文の選択

本節では，拡散を考慮した重要文の選択を教師付き学習によって実現する提案手法の具体的な処理について説明すると共に，そのための教師情報を構成する手順を記述する．

3.1 使用データ

本論文では，毎日新聞 web 版のニュース記事，及びそれらの記事に言及したツイートを解析用のデータとして利用する．具体的な手順は以下の通りである．まず，Twitter Firehose API で取得できる日本語ツイートのうち，本文中に mainichi. jp が含まれるツイートを抽出した．ツイートの収集は，2014 年 8 月 19 日から 2015 年 9 月 1 日まで継続して実施し，総計で 80 万ツイートを収集した．次に，ツイートの URL に記載されている毎日新聞の記事を，クローリングを行い取得した．ニュース記事については 2015 年 9 月 1 日時点で収集可能であったもののみを取得し，総計で 6000 記事を得た．このように収集した記事それぞれについて，最も RT 数が多かったツイートを選択し，記事とツイートの対を構成した．以降の解析では，最大 RT 数が 4 以上であった記事 1213 本のうちおよそ 150 本の記事とツイートの対を利用した．

3.2 記事の文への分割

収集した記事に重要文を選択するための教師情報を与えるに当たり，まず，収集した記事を文単位に分割した．しかし，取得した記事の中には，文末に句点が打たれていないもの，「▼」や「＝」のなどの特殊記号が存在しており，文の区切りを自動的に検出することは非常に困難である．また，記事中には鍵括弧，中括弧の中が複数文で構成されている記事も存在した．さらに，SNS は会話のように使われるため，記事中において発言を意味する鍵括弧とその中の言葉が重要であると報告がある[5]．これらの状況を考慮し，以下のルールに従って文区切りを手動で与えた． 例文中のスラッシュは文区切りを表す．

・句点や感嘆符など一般に文末で用いられる記号の直後

> 例 1） ついに、ついに、念願のメジャーデビュー！！！/支えてくださったすべての方に感謝です！/(注 4)

・括弧類の前後．

> 例 2）上村主将は/「自分たちで歴史を作ることができる」/と新興チームの魅力を話す。(注 5)

・鍵括弧の中に複数の文が存在する場合．

> 例 3）/「区間賞は狙っていたが、まさか区間新が出るとは。/前半から突っ込む自分らしいレースができた」/。(注 6)

・隅付き括弧で記者名など，著者署名が入っている場合．

> 例 4）講演会は午後 1 時から 4 時１０分。/予約不要で無料。/【垂水友里香】/(注 7)

・特殊記号が存在した場合はその前後．

> 例 5）その他の部門は次の通り。/優秀棋士賞/＝/糸谷哲郎竜王/▽/敢闘賞/＝/郷田真隆王将/▽/新人賞/＝/千田翔太五段/(注 8)

3.3 ラベリング

前節で分割した記事中の各文について，重要文であるかどうかのラベルを付与した．具体的には，各記事と対になる，最も RT 数の多いツイートに含まれる情報が記載されている文を重要文と見なし，ラベル「1」を，それ以外の文にラベル「-1」を与える．

注 4 http://mainichi.jp/mantan/news/20150617dyo00m200016000c.html
注 5 http://mainichi.jp/articles/20150328/ddg/041/050/005000c
注 6 http://mainichi.jp/graph/2015/01/02/20150102k0000m050017000c/001.html
注 7 http://mainichi.jp/articles/20150122/mog/00m/040/003000c
注 8 http://mainichi.jp/articles/20150402/ddm/012/040/109000c

表１　記事本文とラベル例（注10）

記事中の文	ラベル
秋篠宮ご夫妻の次女佳子さまは７日、奈良県橿原市の神武天皇陵を参拝された。	1
２０歳の成年皇族になった報告のため。	1
佳子さまは２００３年３月にもご夫妻と姉の眞子さまと共に参拝しているが、お一人では初めて。	1
午前１１時ごろ、参拝服の佳子さまは陵に向かってゆっくりと進み、玉串をささげて一礼した。	-1
６日は三重県伊勢市の伊勢神宮を訪れ、外宮、内宮の順に参拝した。	-1
【真鍋光之】	-1

今回の調査では，読者の興味を誘引する文が記事中に複数ある可能性があるため，ツイートに書かれている文の意味に対応する文にはすべてラベルを付与した．また，重要文が複数ある場合，その重要度は判定していない．記事に対してラベルを付けた例を表 1 に示す．

上記のラベリングを行った結果として，記事中のどの文が実際にツイートで選択されたかについての傾向を，図 2 に示す．ラベリングを行った 150 記事のうち 2/3 に当たる 100 記事が，記事の第 1 文目をツイートに引用していた．続く第 2 文は全記事の 1/3 に当たる 54 記事が引用していた．これらのことから，記事の前半から優先的に選択する傾向がうかがえる．また，ツイートに記事中の複数文を選択しているものは全体の 5/6 に当たる 123 記事あり，非常に多くのツイートが複数文を選択して構成されている．一方，興味深い点として，10 文目から 30 文目も 108 回選択されたという点が挙げられる．記事中の文数が 30 文を超えるものは 29 記事のみであったことから，この事実は，記事の前半部分のみならず，後半を含む様々な位置から文が選択されていることを示している．

3.4　文の単語への分割

文ごとにラベルを付与した記事を，形態素解析器 MeCab [11] を用いて分かち書きを行った．また，分かち書き後の全ての単語を基本形に変換した．その後分かち書きをした文の構成単語すべてに文に付与さえたラベルを付与した．

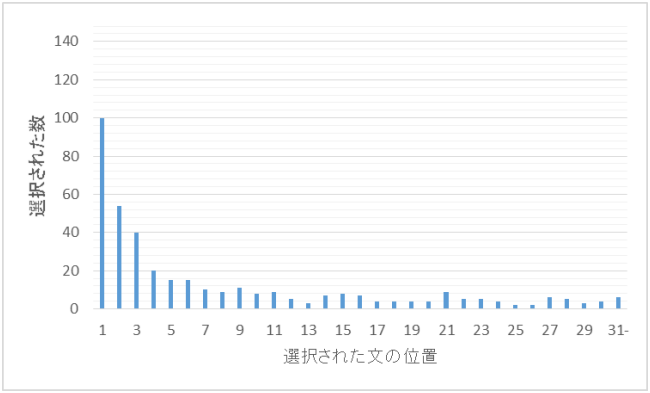

図 2　文の選択傾向

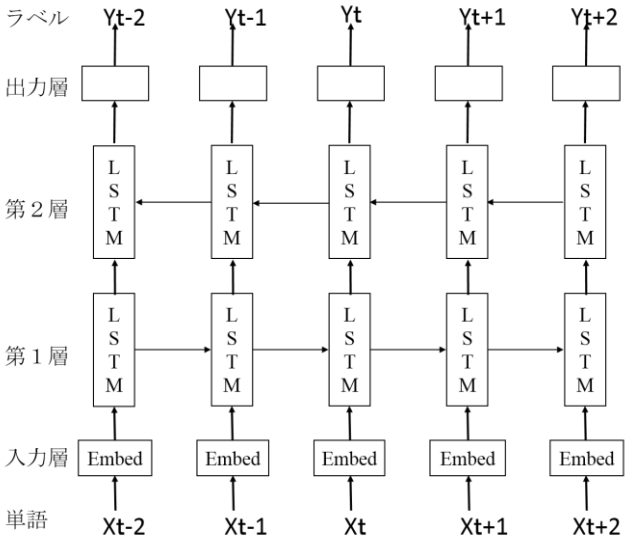

図 3　提案モデル

3.5　モデル学習

前節までで構成した教師データを用いて，記事中から重要文を選択するための recurrent neural network (RNN) モデルを学習する．特に本論文では，Long Short Term Memory（LSTM）をノードに採用した双方向ネットワークである bidirectional LSTM [12] を改変したモデルを用いる．図 3 に提案モデルを示す．

事前に用意した，各単語 X_t (t=1,2,…) にラベル Y_t が付与されたデータセットを使用し，第一層で単語 $X_1,X_2,\ldots$ を

注 9　http://twilog.org/_5newspaper/date-150307

注 10　http://mainichi.jp/feature/koushitsu/news/20150307k0000e040246000c.html

順方向に読み込み，文の先頭から対象単語までの文意を学習する．そして，第二層ではノードの連結方向を逆にし，最終単語からの文意を学習する．これにより，前後の文意を考慮した文選択を狙う. 第 3. 3 節にも示したとおり, 新聞記事は, 30 文を超える文数で構成されていることも多く, 記事の文脈によっては重要文が記事の後半から選択されることも珍しくない．しかし，一般的な RNN では，短期的な依存関係しか学習できず，長期的な記事の文脈を考慮することが困難である．そこで，各ノードに Long Short Term Memory（LSTM）を採用して，長期的な依存関係を考慮する．

LSTM の内部に持つユニットの数はすべて 600，出力は 1 及び-1 の各クラスに対応する 2 次元とした．学習においてはミニバッチを使用し，20 記事を同時に学習した．モデル学習の評価関数として，最終的に出力された数値と教師データのラベルとのソフトマックス交差エントロピーを用い確率的勾配降下法による誤差伝播により学習を行った．

3. 6 重要文選択

前節で学習したモデルを用いて，与えられた記事から重要文を選択する．学習済モデルからの出力をそのまま予測結果として用いる方法も考えられるが，本論文では，より精度の高い予測を実現するために，モデル出力の直前に位置する，第 2 層 LSTM の出力ユニット 600 次元を特徴ベクトルとして抽出し，Support Vector Machine（SVM）による分類器で最終的な予測結果を得る．RNN モデルの各ノードは単語に対応しているため，SVM に入力する特徴ベクトルは，当該単語が選択されるかどうかを判断するために用いられる．一方，本論文の目的は，記事中から文を選択することにある．そこで，各単語に対応する SVM 分類器の予測結果を当該文中の全単語で平均し，その平均値の大小により，文選択の予測結果を決定する．

4. 実験

4.1 実験条件

第 3.1-3.3 節で作成したデータセットを用いて，第 3.4-3.6 節に記載の提案手法がどの程度正確に拡散重要文を選択できているかを検証する実験を行った．SVM のカーネルは線形カーネルを採用した．データセットを 4 つの部分データに分割し，そのうちの 3 つの部分データで 3-fold 交差検定によりハイパーパラメータを決定した．このように決定したハイパーパラメータと 3 つの部分データを使い RNN モデルの学習及び SVM 分類器の学習を行い，残り 1 つの部分データで評価を行った．上記の手順を，評価に使用するデータを変えて 4 回行い，その平均値で評価する．

4.2 実験結果

今回の実験は拡散重要文を正しく選択することが目標であるため，提案モデル＋SVM 分類器で選択した文と，人手で付与した正解ラベルを比較した正解率により，提案モデルがどの程度正確に拡散重要文を選択できるかを評価する．文レベルの結果において，1 文目のみを選択する手法をベースラインとし，提案手法での選択結果と比較する．さらに実験において，今回用意したデータセットの数が非常に少なかったため，データを複製加工して増やす措置を行った．30 文以上で構成される記事を複製し，30 文以降を削除した．20 文以上，10 文以上の記事も同様に行った．

	予測: 選択	予測: 非選択	正解率
ラベル: 選択	1696	384	81.5%
ラベル: 非選択	72	14722	99.6%
	予測: 正解	予測: 不正解	正解率
合計	16418	456	97.2%

表 2　単語レベルでの予測結果

	予測: 選択	予測: 非選択	正解率
ラベル: 選択	884	256	77.5%
ラベル: 非選択	970	7764	88.8%
	予測: 正解	予測: 不正解	正解率
合計	8648	1226	87.6%

表 3　文レベルでの予測結果

	予測: 選択	予測: 非選択	正解率
ラベル: 選択	179	149	54.5%
ラベル: 非選択	90	6931	99.5%
	予測: 正解	予測: 不正解	正解率
合計	269	7080	96.3%

表 4　ベースラインの予測結果

実験結果を表 2 及び 3 に示す．表 2 は単語レベルでの予測結果，表 3 は文レベルでの予測結果をそれぞれ示す．

これらの結果から，単語レベルで 81.5%，文レベルで 77.5%と高い精度で正解重要文を選択できたことがわかる．ベースラインの 1 文目を選択する手法では 54.5%の正解率であり，提案手法が優位な結果となった．しかし，1 文目を選択する手法で 54%の正解率を得られたことは，提案手法はまだ精度向上の余地があることも示している．また，今回のタスクは，本質的に正例（選択文）と負例（非選択文）の数に非常に大きな偏りがあるため，予測結果を非選択とする方向にバイアスがかかりやすい．このことは，表 2 において，負例（非選択）に分類誤りがほとんど発生していないことからも理解できる．特に後段の SVM 分類器において，ハイパーパラメータをより精緻に選択する必要があると考えられる．さらに，今回用意したデータセットは量が非常に少なく，特に RNN モデルで過学習が生じている可能性は否定できない．今後の実験において，データセットを増やす，適切な正規化を行う，などの追加検討により，解決を試みる．また，教師付き学習の枠組で文選択を行う既存手法もいくつか知られており[13]，それらの手法との比較検討も重要な課題である．

5. まとめと今後の課題

本論文では，ニュースメディアがニュースの読者に適切なニュース記事を提供するための一手段として，SNS 上でより多くの読者にニュース記事が読まれるためのニュース説明文を自動的に生成することを目指し，対象ニュース記事から拡散に寄与すると考えられる記事中の重要文を，教師付き学習の枠組で選択する手法を提案した．実際の記事及びツイートを用いた実験により，提案手法が高い精度で拡散重要文を選択できることを示した．

本論文の提案手法は，文書要約における文選択に該当するものであり，今後，既存の文短縮手法 [14] と組み合わせる，あるいは拡散要素を考慮した新規の文短縮手法を考案することにより，説明文の自動生成を実現する．

参 考 文 献

[1] 穂積，矢吹，佐久田，”SNS における情報伝達の速度とその拡散性”，情処全国大会予稿集，2008.

[2] 濱岡，”Twitter におけるリツイート(RT)回数の規定要因”，慶応/京都連携グローバル COE ディスカッションペーパー, DP2011-028，2012

[3] 藤川，鍛治，吉永，喜連川，”マイクロブログ上の流言に対するユーザの態度の分類”，信学技報，2011.

[4] 斎藤，富岡，山西，”ソーシャルネットワークにおける長期間流行する話題の早期検出”，信学技報，2011.

[5] 興梠，木村，藤代，西川，”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”，電子情報通信学会論文誌，2016.

[6] 平尾努，鈴木潤．磯崎秀樹，”最適化問題としての文書要約”，人工知能学会論文誌，2009.

[7] 館林，原口，”文の結束性に寄与する特徴的な語を考慮した文間依存関係に基づく文書要約手法の提案”，人工知能学会全国大会予稿集，2006.

[8] 山本，斎藤，”用例利用型による文間接続関係の同定”，自然言語処理，2008.

[9] 別所，西川，牧野，松尾，”単語ベクトルを用いた文書要約の検討”，信学技報，2014.

[10] 菊池，平尾，高村，奥村，永田，”入れ子依存木の刈り込みによる単一文書要約手法”，自然言語処理，2015

[11] T. Kudo, K. Yamamoto, and Y. Matsumoto, “Applying Conditional Random Fields to Japanese Morphological Analysis,” Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.230-237, 2004.

[12] A. Graves and J. Schmidhuber，”Framewise phoneme classification with bidirectional LSTM and other neural network architectures”， Neural Networks 18:5–6, pp. 602–610, 2005.

[13] T. Hirao, H. Isozaki, E. Maeda, and Y. Matsumoto. “Extracting Import Sentences with Support Vector Machine,” In Proc. International Conference on Computational Linguistics (ACL), pp. 342--348, 2002.

[14] 西川，今村，別所，牧野，松尾，“クエリ依存文短縮と見出し生成への応用,” 情処研報，2013-NL-214(2)，2013.