SNS 上で拡散するニュース説明文の信頼性向上に向けた学習データ整理

廣瀬　友香†　　木村　昭悟‡　　藤代　裕之†

†法政大学 〒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:　†yuka.hirose.8v@stu.hosei.ac.jp, fujisiro@hosei.ac.jp, ‡akisato@ieee.org

あらまし SNS の普及により,不確実な情報やデマといった「偽ニュース」が社会問題となっており,より信頼性の高いニュースが求められるようになっている.これまで我々は,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,新聞社の記事中から拡散に寄与する重要文を recurrent neural network(RNN)を用いて自動的に選択する仕組みの開発を行った.しかし,その精度は十分ではなく,教師データを改めて検証した結果,教師データ内に一般ユーザーの投稿の混在,また記事のスタイルによる精度低下が明らかになった.この検証結果を踏まえ,本論文では,朝日新聞社のニュース説明文を教師データとして採用すると共に,手動で教師データの選別・編集を行い,重要文自動選択手法の精度向上を目指した結果,大幅な選択精度の改善を実現した.

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

1. はじめに

SNS(ソーシャル・ネットワーキング・サービス)の普及により,私たちはネット上で情報を取得し,発信する機会が増えた.しかし,SNS 上で発信される情報には,信頼性の低い不確実な情報やデマといった,「偽ニュース」が多く含まれている.

2016 年 11 月に行われたアメリカ合衆国大統領選挙の際には,フェイクニュースサイトが登場し,「ローマ法王がトランプ氏を支持した」などといった多くの「偽ニュース」が SNS 上で拡散された[1].

この状況を受け,SNS の 1 つである Facebook は,「偽ニュース」の検出を始めた.Facebook に「偽ニュース」として検出された記事を,AP や ABC といった従来のメディアなどが調査を行い,「偽ニュース」の拡散抑制に動いている[2].

一方日本でも,ネット上の情報に多くの「偽ニュース」が含まれていることが問題となっている.DeNA の医療サイト問題では,記事内容の引用元の不明さや,医療的根拠のない不確かな情報が掲載され,そのことに批判が高まり,サイトが閉鎖された[3].

このように日本国内外で,「偽ニュース」が SNS 上で拡散されるという問題が注目を集めている(注1).

さらに,カナダの BuzzFeed の創刊者兼編集者である Craig Silverman は,SNS 上では,事実よりも嘘が広がりやすい状況にあると述べている[4].

「偽ニュース」の発生と拡散が社会的な問題となったことにより,ネットや SNS で発信される情報の信頼性が疑われるようになった.そして現在,より信頼性の高い情報が求められている.

本研究では,これらの背景を受け,読者に信頼性が高いニュース記事を届けるために, SNS 上で拡散するニュース説明文を自動生成する研究を行ってきた.ニュース説明文とは,SNS 上で記事を紹介する投稿文のことを指す.適切なニュースを適切な形で届けるために,刺激的なタイトルを用いた拡散や,記事内容と異なる釣りタイトルではなく,的確に記事内容を説明し,かつ SNS 上で拡散するニュース説明文の構成が重要である.

興梠らは,社会工学的な知見とデータ工学的手法を融合させることで, この課題に初めて取り組んだ. まず, ニュース説明文の作り手であるジャーナリストらにヒアリングを実施し,その知見を踏まえ,ネットメディアであるハフィントン・ポスト日本版の記事とニュース説明文から拡散に重要な要素を明らかにした.また, ハフィントン・ポスト日本版の記事とニュース説明文を学習データとして用いて,与えられた複数の説明文候補の中から最も適切な説明文を選択する手法を提案した[5][20].

興梠らの使用データはネット限定のメディアが配信した記事とニュース説明文であったため,永山らはより信頼性の高い毎日新聞社の記事データを用い,記事中から要約してニュース説明文を生成することを目指した.実際に拡散した SNS 投稿文に用いられた内容を含む記事と,重要文であるかどうかのラベルを付与した文を教師データとした RNN によりモデルを生成し,ニュース説明文に用いる文選択を行った[6][18].

本論文では,ネットや SNS 上での不確実な情報やデマといった「偽ニュース」の拡散という社会問題を踏まえ,より信頼性の高い教師データの採用を行った.さらに,信頼性が十分ではないと思われる教師データを

(注1) 読売新聞の 2017 年 2 月 16 日付「論点スペシャル」や朝日新聞社の 2017 年 2 月 18 日付「耕論」,NHK の 2017 年 2 月 6 日,7 日放送「クローズアップ現代＋」など.

手動で選別・編集を行うことで,永山ら[6][18]の RNN に基づくニュース説明文の重要文選択の精度向上を目指す.

2. 先行研究

本研究には,SNS 上で拡散させる要因を調査する研究,文章要約についての研究が関連する.

まず,拡散に関する研究について関連研究を紹介する.日本では「偽ニュース」の拡散に関する研究として,東日本大震災をはじめとした震災時のデマが扱われてきた.安田は,東日本大震災発生直後のデマに関連するツイートを目視で分析を行い,ソーシャルメディア固有の情報拡散の特徴を明らかにし,「拡散希望」などの拡散に重要な要素を提示した[7].榎らは,ソーシャルメディア上での情報経路に着目し,情報拡散力の高いユーザーを発見した[8].白井らは,情報拡散が行われるネットワークの構造に着目し,ソーシャルメディア上でデマが拡散されやすいネットワーク構造を示した[9].

これらの研究は,「偽ニュース」の拡散を対象に研究しているものであり,信頼性の高いニュースの拡散について研究したものではない.

次に文書要約について,関連研究を紹介する.文書要約では,一般に重要文抽出と文短縮の２つで構成される場合が多い[10].舘林らは,文書の全体像の把握に向けた重要文抽出のために話題を捉え,全体からバランスよく文を抽出する手法を提案した[11].菊池らは,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の提案により,要約の精度向上を図った[12].別所らは,単語をトピックベクトルとして表現し,語及び文のスコアを対象ベクトルとの類似度として,スコアの高い語及び文を要約テキストとして出力する手法の提案を行った[13].さらに,田中らは表示領域に制限のあるデバイスでのニュース記事の要約表示を想定し,文融合及び文分割処理を取り入れた手法の提案により,限られた表示領域の中で要点をつかんだ要約生成を行うことが可能であることを示した[14].

これらの研究の目的は,文書の主要内容をできるだけ損なわずに所定の文長に短縮することにある.しかし,これらは所定の文長に短縮するにすぎず,我々は,これに加え,拡散を考慮した文書生成のために,所定の文長に短縮することを目指している.

これまでに我々は,SNS 上で拡散されるニュース説明文の自動生成をするために,「拡散」「要約」を考慮したニュース説明文の自動生成について研究を行ってきた.

本論文は,前節で紹介した我々の既存研究である興梠ら[5][20],永山ら[6][18]の流れをさらに推し進めるものである.

3.調査

本論文では,永山らの手法[6][18]のさらなる精度向上を目指し,提案モデルにより選択された文と,人手により選択した文を手動で比較し,調査を行った.本節では,その調査結果について示す.

3.1 手法概要

本節では,まず永山らの手法について概要を説明する.永山らの報告[6][18]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学習・評価用のデータとして利用した.このデータ収集のために,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 のドメイン URL が含まれるツイートを抽出した.2014 年 8 月 19 日から 2015 年 9 月 1 日まで継続してツイート収集を行い,総計 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により取得した.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 6000 記事を得た.

永山らの手法は,SNS 上の拡散に寄与する文をニュース記事中から選択する方法を提案している.以下にその詳しい手順を示す.

(1)記事の文への分割

まず収集した記事を,以下のルールに基づいて文末を決定することにより,文単位に分割した.

- 句点や感嘆符など一般に文末で用いられる記号
- 括弧類の前後
- 鍵括弧の中に複数の文が存在する場合
- 墨付括弧で記者名など,著者署名が入っている
- 特殊記号が存在した場合はその直後

(2)ラベリング

記事と対になるツイートのうち,最もリツイート(RT)数の多いものを正解ニュース説明文とみなし,その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みなすことで,分割した記事中の各文について重要であるかどうかの 2 値ラベルを付与した.

(3)文の単語への分割

形態素解析器 MeCab[15]を用いて, 記事の各文を単語単位に分かち書きを行い,分かち書きした文の構成単語全てに,文に付与されたラベルを付与した.

(4)モデル学習

上記手順により構成した,単語単位に 2 値ラベルが付与され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 RNN モデルを学習する.特に,永山らは,

音声認識において非常に有効であることが示されている bidirectional LSTM[16]を改変したモデルを用いることにより,記事のような長期的な依存関係を持つ文書への単語ラベリングを可能とした.

(5)重要文選択

(4)で学習した RNN モデルを用いて,与えられた記事から重要文を選択する.各単語に対応する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.

(6)評価手法

提案のモデルで選択した文と,人手で付与した正解ラベルを比較し,その正解率により,提案モデルの選択の正確さを評価した.

このような手法により実験を行った結果,図 1 に示すとおり, 提案モデルでは,1 文目,2 文目の選択がされやすいこと,さらに 1,2 文目以降も一定数は選択が行われていることが示された.さらに,図 2 に示す通り,ランダムに文選択を行うよりも優位に選択が行われたことが示された.

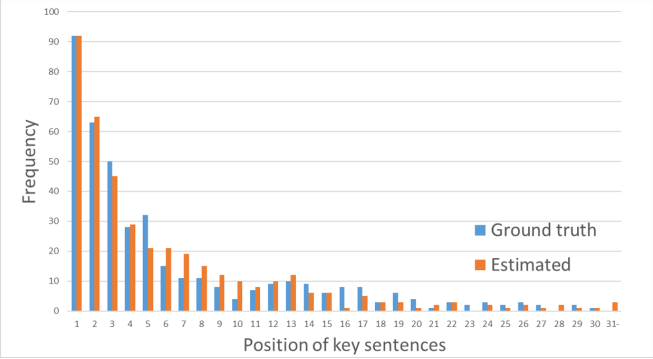

図 1:選択重要文の分布

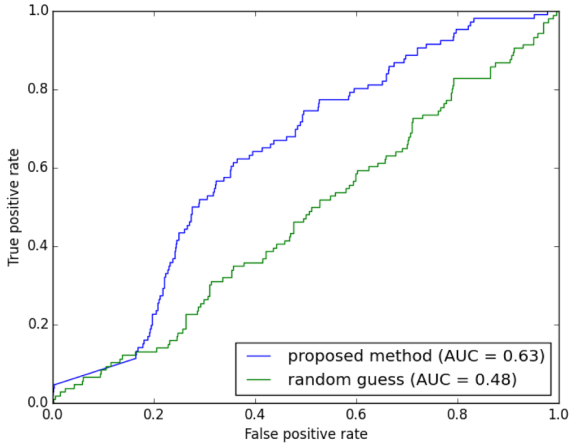

図 2:文ごとの推定

3.2 調査結果

前節の手法により選択された文章と,正解の重要文を目視で比較を行った.それにより永山らの手法の定性的な性能を検証した.その結果,(1)記事スタイルの違い,(2)使用データの信頼性,(3)文選択における区切り方,の 3 点において課題が見つかった.

なお,正解ニュース説明文であるツイートの情報が含まれる文を,モデル選択文内にて下線部で示している.

(1)記事スタイルの違い

新聞では,記事のスタイルが多数存在している.記者が記事対象から感じた印象や周囲の状況,関係者の反応などを記事にしている「雑観」や,記者の意見や感情が混ざって描かれている「コラム」,さらにはテレビ番組の内容が紹介されている「告知記事」など,様々である.

今回の永山らの使用データにも,様々なスタイルの記事が含まれていた.紙面の囲み記事であるコラム記事は,一般的な記事と違い,記者の意見や感情が混ざっているため,記事内での伝えたい箇所が分散している場合が多い(図 3).告知記事では,番組名やタイトルに言葉が偏りがちになるとともに,固有名詞が多く含まれるため,選択が難しいと考えられる(図 4).

正解ニュース説明文

世界には無数の悲劇がある。ガザの失業率は４３％で世界最悪。配給食を売って現金を得る避難民がいる。そして、ペットのえさにその配給食を買う富裕層がいる。＜発信箱:配給食を売る＝大治朋子（エルサレム支局） － 毎日新聞

モデル選択文(正解ニュース説明文との一致なし)

「人間は穀類だけでは生きられない」昨年夏、イスラエルとの戦闘で市民ら２０００人以上が死亡したパレスチナ自治区ガザ地区。避難所前で露天商を営む男性（３７）が言った。・・・「ご飯やパンしか支給されない」「収入がないから配給食を売って野菜を買っている」配給された食事を見ると、ピラフのようなものに野菜らしきかけらが混ざっている。・・・露天商の男性によると、配給食を買うのはお金に余裕のある市民で「ペットのえさにする人もいる」。世界銀行によると、ガザの失業率は４３％で世界最悪という。・・・

図 3:コラム記事の例

正解ニュース説明文

サワコの朝：脚本家・大石静直伝！“売れる”俳優の見抜き方とは！？ － 毎日新聞

モデル選択文

トーク番組「サワコの朝」（ＭＢＳ／ＴＢＳ系全国ネット、６月２７日午前７時半～８時）のゲストは脚本家の大石静。朝ドラとして初めて“婚前交渉”の描写を取り入れたＮＨＫ連続テレビ小説「ふたりっこ」や４５歳のキャリアウーマンが

１７歳年下の男性と不倫の恋に落ちていくドラマ「セカンドバージン」など刺激的な作品で社会現象を起こした。そんな大石の作品は、無名に近かった俳優がブレイクすることでも知られている。番組では、その代表ともいえる内野聖陽、佐々木蔵之介、長谷川博己を抜擢するに至った彼らの魅力に加え俳優の素質を見抜く方法を伝授。・・・

図 4：告知記事の例

(2)使用データの信頼性

永山らは,毎日新聞の記事と,それぞれの対になるツイートの収集を行った.記事自体は,信頼性が高いと考えられる毎日新聞のものであったが,ツイートについては,記事 URL をキーとして RT 数を基準に収集を行ったため,一般ユーザーが記事に対して言及しているツイートも数多く含まれていた.そのため,内容について必ずしも記事が要約されたものではなく,信頼性が担保されているとも言えないことが明らかになった.

また,図 5 に示すように,一般ユーザーの投稿には,記事中の引用ではなく,一般ユーザーの意見や感想が付加されるもしくはそれらで占められることも多い.そのため,文書要約に基づくアプローチでは正解に近い説明文を導き出すことは不可能である.

一般ユーザーの投稿

そういえば、皆さんこの記事って見られましたか？ぐんまちゃんのルーツを知る、貴重なインタビューなので是非ご覧下さい

記事

ゆるキャラグランプリ(ＧＰ)２０１４で初優勝を果たした「ぐんまちゃん」は、群馬県職員の中嶋史子さん(４７)＝前橋市＝によって今から２０年前にデザインされた。「この子と一生、一緒に生きていく。そんな思いで描きました」。中嶋さんの願いは一つ。我が子のようにかわいがるぐんまちゃんが、これからも長く県内外で親しまれる存在であり続けることだ。　中嶋さんは学生時代にデザイン事務所で働き就職後も仕事の一環でポスター製作を手がけていた。・・・

図 5：実際に収集した一般ユーザーの投稿と対の記事

(3)文選択における区切り方

永山らの文分割の方法にも課題があった.永山らの方法では,鍵括弧や「■」「◇」「＝」などの特殊記号ごとに文を分割しているため,その結果として区切られた一文が文としての体裁をなしていない場合が多く見られた.そのため,図 6,図 7 に示すように,記号が多い記事や鍵括弧の中にさらに鍵括弧が存在するような文の選択に失敗しやすい.しかしながら,図 8 に示すように,記号箇所で区切っていない文については,文選択が成功する傾向が見られた.

ニュース説明文

AKB 島崎遥香：こじはるが「おしゃれ認定」も塩対応　ファッションに無関心？ - 毎日新聞 #ぱるる #島崎遥香

モデル選択文

人気アイドルグループ「AKB48」が 29 日、国立代々木競技場第一体育館(東京都渋谷区)で行われた日本最大級のファッション＆音楽イベント「ガールズアワード 2015 SPRING／SUMMER」に登場し、ライブパフォーマンスを行った。MC でグループのメンバーの中で「おしゃれだと思う人」を聞かれた小嶋陽菜さんは「ぱるる(島崎遥香さん)」と回答。一方、島崎さんはファッションのこだわりを聞かれるも「ないです」とポツリ。この日の衣装に関しても「どうなんだろう。分からない、これ」と関心の薄そうな“塩対応”だった。・・・

図 6：記号が多く,選択がされにくい例

ニュース説明文

「自問するようになった。『おい、いつまでもノンポリでいられるのか、宝田よ』と。俳優は後から身につけた職業。生身のお前の意見はどうなんだ、人間として何を言わなきゃいけないんだ、と」《反戦の訴え、番組で遮られ》

モデル選択文

問題の背景として、籾井勝人ＮＨＫ会長の「政府が右と言っているものを左と言うわけにはいかない」といった発言や、・・・「６０歳を過ぎた頃から、自問するようになったんです。『おい、いつまでもノンポリでいられるのか、宝田よ』と。俳優は後から身につけた職業。だったら生身のお前の意見はどうなんだ、人間として何を言わなきゃいけないんだ、と。それからは、言うべきことは言ってきたつもりです。もちろん、先日のＮＨＫの番組でもね」　「物言えば唇寒し」。そんな出来事が芸能界で相次いでいる中、大俳優が自らの信念で語る言葉と、その重みに圧倒される。・・・

図 7：鍵括弧が被っているため,選択がされにくい

ニュース説明文

【陸上】陸上織田記念:男子 200 藤光謙司大会新　女子は福島千里

モデル選択文

陸上の世界選手権(８月・北京)の代表選考会を兼ねた織田幹雄記念国際大会は１８日、広島市のエディオンスタジアム広島で開幕した。男子２００メートルは藤光謙司(ゼンリン)が大会記録となる２０秒４３をマークし優勝。桐生祥秀(東洋大)は２０秒８０で３位だった。昨年の仁川アジア大会１００メートル３位の高瀬慧(けい)＝富士通＝は左足内転筋のけいれんで決勝を途中棄権した。女子２００メートルは、福島千里(北海道ハイテクＡＣ)が２３秒５４で優勝。・・・

図 8：記号で区切らなかったため,選択ができた例

3.3 調査結果まとめ

前節で述べた通り,永山らの手法で用いたデータには以下の問題が明らかになった.

(1)ニュース説明文生成にあたり,適切でない記事が多数含まれていた

(2)一般ユーザーの投稿が混ざっていたために,データの信頼性が十分でない

(3)個人の感情が含まれたSNS投稿を利用したために,記事自体の「要約」とは言えない

以上の問題点から,本論文では使用する記事を厳選する.信頼性が高いとされる朝日新聞社のweb記事及び,朝日新聞社のソーシャルエディタが作成した朝日新聞社公式TwitterアカウントのSNS投稿文(ニュース説明文)を教師データとして採用する.朝日新聞社のニュース説明文は,社内の編集者が「SNS上での拡散性」を考慮して付与した記事タイトルが用いられている.そのため,拡散性,信頼性がともに担保されており,さらに記事内容が的確に説明されていると考えられる.

4.提案手法

3.2節で述べた通り,永山らの手法で用いたデータには,コラム記事や告知記事などのニュース自体の重要性もそれほど高くなく,さらにニュース説明文を生成する際に,扱いにくい記事が含まれていたことが調査により明らかになった.そこで,本論文では使用データの見直しにより,明らかとなった永山らの手法の問題点を改善すべく,手動により教師データの収集・選別・編集を行った.これにより,既存の手法からの精度向上を目指す.

4.1 使用データ

本論文では,3.3節で述べたように,朝日新聞web版の「朝日新聞デジタル」のニュース記事,及び朝日新聞の公式Twitterアカウント@asahiのツイート(ニュース説明文)を用いた.ツイートの収集は,2016年11月3日から1月31日まで毎日13時に継続的に行った.収集する記事は,朝日新聞デジタルのトップページに上がる9記事のうち,公式Twitterアカウントのツイートと対になるもののみを収集した.

「SNS上での拡散性」を考慮して付与された,記事と対になるツイートを収集することにより,記事の内容を的確に説明しつつ,拡散性を考慮したデータを構成できる.

4.2 記事の文への分割

永山らの記事の文への分割は,記号ごとに区切りを入れていたために,一文が認識しにくくなっていた.そこで本論文では分割方法について再度検討を行った.また記事の中には,「■」「◇」などの特殊記号が存在する文章や鍵括弧や中括弧が存在する文章もあり,単純に分割するのが難しいものもあった.そのため既存手法も踏まえ,以下のように基準を設定し,手作業により文の分割を行った.

・通常の句点で一文が終わる場合

> 明治安田生命Ｊ１は３日に第２ステージ（Ｓ）最終節がある。／熾烈（しれつ）なのが・・・(注2)

・鍵括弧内に句点が含まれる場合

> ・・・／「動物や昆虫までが色っぽく、命の変容する様がエロスを放つのが父の作品の特徴。／それを描くのを楽しんでいたことが伝わる。／エロスは・・・でもあり魅力でもある」と・・・(注3)

・会話文の鍵括弧で段落や話題が変わる場合

> ・・・／検察官「金に困っていたのか」／被告「片親で６歳の子どもの世話を見ていくことと、老いていく両親の世話を１人で見るという漠然とした将来への不安がありました」／・・・(注4)

・「■」「◇」など特殊記号で見出しとして付けられている場合

> ・・・／■帽子をかぶってシェフ気分／くまモンがこの日最後に訪れたのは、リヨン郊外にある、世界的に有名なレストラン「ポール・ボキューズ」。・・・(注5)

・本文中に問いが存在する場合

> ・・・／――大ヒットですね。／時代にあった内容だったんだと思います。ドラマも、エピソードは微妙に違いますが、芯を通るものは原作と同じものを描いて下さった。・・・(注6)

4.3 ラベリング

前節で一定のルールに則って分割した記事中の各文について,重要文であるかのラベルを付与した.付与の方法としては,各記事と対のツイートに含まれる情報の入った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」を与えた.

(注2)http://digital.asahi.com/articles/ASJC15PXFJC1UTQP01S.html?iref=comtop_8_08

(注3)http://digital.asahi.com/articles/ASJC24JQWJC2UCVL017.html?iref=comtop_photo

(注4)http://digital.asahi.com/articles/ASJCJ664CJCJUTIL03F.html?iref=comtop_8_01

(注5)http://digital.asahi.com/articles/ASJCT419DJCTUEHF00C.html?iref=comtop_8_06

(注6)http://digital.asahi.com/articles/ASJDP5JJNJDPUTFL01C.html

4.4 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[15]を用いて分かち書きを行った.この際,辞書には ipadic-neologd[17]を使用した.この辞書の導入により,記事に頻出する固有名詞を含む文をより正確に分かち書きできる.それぞれの文の構成単語に文に付与されたラベルをそれぞれ付与した.

4.5 モデル学習

前節までで構成した教師データを用い,永山らのモデル[6][18]を使用した.モデル学習を行う際,入力する単語の初期値を word2vec[21]で与えた.word2vec の学習には livedoor ニュースコーパス[22]を用いた.

5. 実験

5.1 実験条件

前節までに作成したデータセットを用いて，永山らのモデル[6][18]でどの程度正確に重要文を選択できているかを検証する実験を行った.データセットを 10 個の部分データに分割し,そのうち 8 つをモデル学習,1 つをモデル検証，残り 1 つを評価に用いる，10-fold 交叉検定を行った．それ以外の実験条件は，永山らの論文に記載の内容と同一である．本論文では，分類問題で一般的に用いられる尺度の一つである Area under ROC (AUC)を評価尺度として用いた.

5.2 実験結果

実験結果を図 9 に示す．この結果から，従来の永山らの結果を上回る精度で正解重要文を選択できたことがわかる.

本論文では,永山らの結果を踏まえ,記事の文への分割を,より人が認識する一文と同意のものに近付けるように考慮した.さらに固有名詞への配慮を考え,より正確に文選択が行われるよう,新たな辞書の導入を行った.

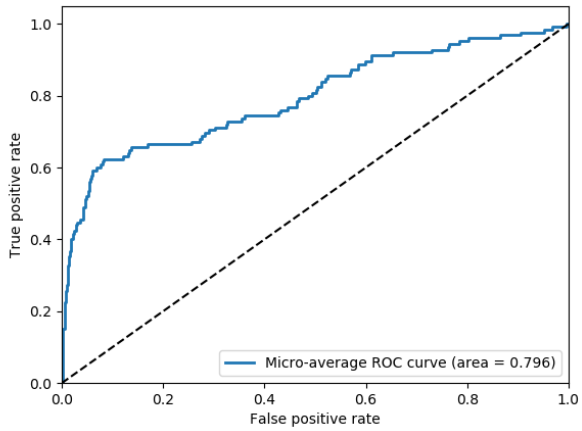

図 9：文ごとの推定

6. まとめと今後の課題について

本論文では,永山ら[6][18]の手法の精度向上を目指し,教師データの検証と,その結果をもとにモデルの改良を行った.データの検証で我々は,特に近年の不確実な情報やデマといった「偽ニュース」の発生・拡散という社会問題に着目し,教師データの信頼性を担保するように考慮した.これにより,一般ユーザーの SNS 投稿の混在や,記事スタイルによる精度の低下が明らかになった.これらの点を考慮し,手動によりデータの収集・選別・編集を行い,教師データを生成することで,永山らの結果を上回る精度で,正解重要文選択で行うことができた.

本論文は,我々の最終目標である「SNS 上で拡散するニュース説明文の自動生成」への足掛かりである.しかし,同時に多くの課題も残している.その代表的なものとしては,信頼性の高い教師データの生成,重要文の自動選択があげられる.

本論文で用いた教師データは,手動での収集・選別を行ったことにより,信頼性は担保されたが,データ量としては不十分な点があったと考えられる.そのため,収集を続け,データ量を増やしていく必要がある.または,信頼性の高いデータを自動で収集することを検討する必要があると考えられる.

また,本論文で用いた文選択手法は,文書要約に該当するが,単に重要文を選択するにすぎず,今後既存の文短縮手法[19]と組み合わせる,もしくは拡散要素を考慮した新たな文短縮手法を考案することで,ニュース説明文の自動生成を実現することができると考えられる.

参　考　文　献

[1] 『日本経済新聞電子版』 2017 年 1 月 21 日付 確認(最終閲覧日:2017 年 2 月 17 日)
http://www.nikkei.com/article/DGXMZO11928060Q7A120C1000000/

[2] 『日本経済新聞電子版』 2016 年 12 月 16 日付　確認(最終閲覧日:2017 年 1 月 16 日)
http://www.nikkei.com/article/DGKKZO010740700W6A211C1EAF000/

[3] 『日本経済新聞電子版』2016 年 12 月 1 日付 確認(最終閲覧日:2017 年 2 月 17 日)
http://www.nikkei.com/article/DGXMZO010119170Q6A131C1X13000/

[4] Craig Silverman,” LIES,DAMN LIES,AND VIRAL CONTENT”,Tow Center for Digital Journalism A Tow/Knight Report,2015.

[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] 永山, 木村, 藤代,“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”,第 8 回データ工学と情報マネジメントに関するフォーラム,2016.

[7] 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.

[8] 榎, 村上, レイモンドルディー, 小口,”ソーシャルメディア上の情報拡散分析”,第 6 回データ工学と情報マネジメントに関するフォーラム,2014.

[9] 白井 他,“Twitterにおけるデマツイートの拡散モデルの構築とデマ拡散防止モデルの推定”,人工知能学会全国大会論文集,2012.

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

[11]舘林, 原口,”文の構造と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論文集,2006.

[12]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.

[13]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.

[14]田中, 笹野, 高村, 奥村,“要約長,文長,文数制約付きニュース記事要約”,言語処理学会発表論文集,2016.

[15]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.

[16]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.

[17]https://github.com/neologd/mecab-ipadic-neologd/

[18]K.Nagayama, A.Kimura and H.Fujishiro,″Make it go viral - Generating attractive headlines for distributing news articles on social media″, In The Computation + Journalism Symposium 2016.

[19]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”, 情報研報, 2013-NL-214(2), 2013.

[20]S.Kourogi et al.″Identifying attractive news headlines for social media″, In CIKM2015.

[21]Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. ″Efficient Estimation of Word Representations in Vector Space. ″ In Proceedings of Workshop at ICLR, 2013.

[22]http://www.rondhuit.com/download.html#ldcc