DEIM Forum 2022 B34-6

リツイートの時系列データを用いた Hierarchical Attention Networks に基づく誤情報検出

神田　凌弥†　　杉山　一成†　　吉川　正俊†

† 京都大学大学院情報学研究科　〒 606–8501 京都市左京区吉田本町

E-mail: †kanda@db.soc.i.kyoto-u.ac.jp, ††{kaz.sugiyama,yoshikawa}@i.kyoto-u.ac.jp

あらまし　近年，ソーシャルメディアを発信源とした誤情報やフェイクニュースがもたらす社会への影響が深刻になっている．最近では，それらを早期に検出し，被害を未然に防ぐために，ニュースや文章の真偽を判定するための様々な手法が提案されている．本研究では，Twitter を対象とし，そのリツイートの時系列データに着目して，より効率的に文書での重要度の高い文や単語を抽出できる Hierarchical Attention Networks (HAN) に基づいて，誤情報を検出する手法を提案する．具体的には，リツイート同士の時間間隔をとり，その差をとって間隔時間加速度を定義する考え方 (RT1) と，一定時間のリツイート数を数え，既定の一定時間で割ることで速さを算出し，さらにその速さの差をとって一定時間加速度を定義する考え方 (RT2) の 2 つの手法を提案する．各ニュース記事・投稿における平均の間隔時間加速度，一定時間加速度をもとに重み付けをし，それによって得られた表現と，HAN によって得られた文書ベクトルとを組み合わせ，真偽を判定する．実験の結果，RT1 は学習中で正答率の下降が少なかったことから学習の安定化に，RT2 は HAN とほぼ同程度の精度の結果が得られたことから良い精度の検出に，それぞれ効果があることが示された．

1　はじめに

インターネットが世に広く普及し，情報技術が年を重ねるごとに発展していく現在，ツイッターやインスタグラムなどのソーシャルメディアを通じ，人々は今や自由に情報の発信と受信ができるようになった．情報発信の自由化が進んだことで，世界全体での情報の量は年々増加している．それらの情報の中には，人々の暮らしに利益をもたらすものもあれば，逆に悪影響を与えてしまうものもある．特に悪影響を受けたものについては，現実世界で大きな混乱を招くような事態に発展してしまう事例も少なからず存在する．

例えば，2021 年 1 月には，アメリカ合衆国のトランプ前大統領が，同月 6 日にワシントンで発生した，トランプ氏の支持者らによる暴動事件について，大統領選において不正があった，と主張するようなツイートが支持者らの暴動を扇動したとみなされ，自身のアカウントが一時的に凍結させられた [1]．さらに凍結の解除後，バイデン次期大統領の就任式に出席しない，などのツイートにより暴動を煽る可能性が高いとみなされ，自身のツイッターアカウントが永久凍結されることとなった [2]．すなわち，このアカウント凍結に関する一連の騒動は，ユーザが発信する情報が現実世界での悪影響を未然に防ぐための措置である．短いツイートが現実世界に大きな影響を与えてしまうことは，既に世界規模で発生しており，被害が出ている例もある．

そのような悪い影響を与え得る情報の中でも，中身が真実ではないが一見有用そうに見える情報である誤報や，意図的に情報の受け取り手を騙すような真実でない内容の情報であるフェイクニュースが，近年では大きな問題の一つとなっている．特に，2020 年はコロナウイルスが猛威を振るい，感染対策や症状，各国政府の対応，ワクチンなど，多岐にわたる話題に関する情報が多く飛び交い，同時に多くの誤報やフェイクニュースも発信されてしまった．その中には，実際に社会の混乱を招き，生活に大きな支障が出てしまうようなものも存在した．

例えば，2020 年 8 月 4 日，大阪府公館での大阪府知事の記者会見で，うがい薬を使用したうがいにより，唾液中のウイルスの陽性頻度が低下した，という研究成果が発表され，広くうがい薬を使用するように推し勧められた [3]．このニュースに影響を受け，各地のドラッグストアやスーパーではうがい薬の品切れが起こり [3]，うがい薬を製造している企業の株価が急騰する事態に発展した [4]．また，うがい薬は第三種医薬品であるため転売が禁止されているにもかかわらず，フリーマーケットサイトではうがい薬が店頭価格の五倍以上の値段で売買されていた [4]．

こうした事態を未然に防ぐため，近年では誤情報やフェイクニュースをいち早く検出する技術や，そうした情報の拡散を防ぐための技術が広く研究されている．例えば，Cui ら [5] は，ソーシャルメディアのユーザ同士で，情報が伝播する経路とエントロピーを考慮し，情報拡散の流れや拡散を防ぐユーザを推定する手法が提案されていた．このほか，ニュース記事に掲載されている写真，ニュース本文，そのニュースに付随するキーワードやトピック，ニュースに対するユーザの感情に注目してフェイクニュースを検出するような研究も存在する [6]．

本研究では，対象とするソーシャルメディアをツイッターに絞り，その機能の一つであるリツイートに関する時系列データを用いた手法を提案する．本研究では，コロナウイルスに関するニュースについてのツイッター上の投稿がリツイートされた

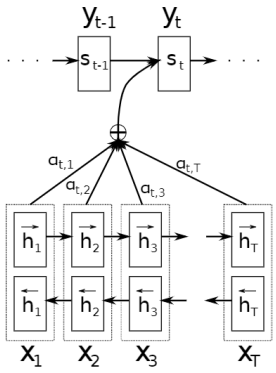

図 1　attention 機構のモデル図

時間についてのデータ，すなわちリツイートの時系列データを用いた研究を行う．一見有用に思われる誤報や，悪意を持って誤解を与えさせるようなフェイクニュースは，情報が公開・投稿されてから比較的短い時間のうちに拡散されると考えられる．つまり，誤情報やフェイクニュースは正確な情報と比較して，投稿されてからリツイートされるまでの時間が短い傾向にあると考えられる．この予測された性質を利用し，のちに説明する Hierarchical Attention Networks (階層的 attention ネットワーク，以下 HAN) モデルでニュース文についての表現の重みを学習した機構と組み合わせ，最終的にその情報の真偽を判定する真偽推定器を提案する．

2　関 連 研 究

2.1　Attention 機構

attention 機構 (図 1) は，Bahdanau ら [7] によって提案されたモデルである．[7] では主に機械翻訳モデルの新しい形として提案されている．この論文が発表される以前の機械翻訳では，エンコーダで入力をある固定長ベクトルの表現に変換し，デコーダでそのベクトル表現にマッチする翻訳先の言語に変換する，というエンコーダ・デコーダ機構が主流となっていた．しかしこの機構には，先の固定長ベクトルに変換する，というエンコーダの特性ゆえに，長い文章の翻訳には不向きとされていた．そこで attention 機構を導入することで，モデルが自動的に，すなわちソフトに，目的の単語に関連する部分を検索できるようにした．言い換えれば，文における意味理解に重要な役割を果たす単語に効率的に注目できるようになり，実際のデータセットを用いた実験では長文に対して良い結果を示し，また文章の長さに関わらず高い精度を達成した．

Zichao ら [8] が提案した Herarchical Attention Networks (HAN) は，attention 機構を階層的に利用し，効率的に文書の意味理解に重要な役割を果たす単語や文に注目するようなモデルである．文書というのは，単語が文を構成し，文が文書を構成する，というような階層構造を持っている．単語レベルの attention 機構のみでは，モデルは一つの文における単語の繋がりが理解できても，文をまたがる単語の繋がりを理解できない．そこで，文レベルに attention 機構を適用することで，文同士の繋がりも捉え，この仕組みを文書分類タスクに適用し，高い精度での分類推定を達成している．

Shu ら [9] は，dEFEND (Explainable FakE News Detection) という，ニュース記事の本文を HAN に適用したものと，その記事に対するコメントを attention 機構に適用し，二つの出力を共同 attention 機構に適用した，深層階層的 attention ネットワークモデルを提案している．このモデルでは，

（1） HAN を用いてニュース記事本文を学習し，意味的・統計的手がかりを捉える，ニュースコンテンツ符号化コンポーネント，

（2） 単語レベルの attention 機構を用いて，ユーザコメントの潜在表現を学習する，ユーザコメント符号化コンポーネント，

（3） ニュース本文とコメントとの相関関係を捉え，説明可能な上位 k 個の文・コメントを抽出する，文・コメント共同 attention コンポーネント，

というコヒーレンスプロセスを経てフレームワークが構築されている．これにより，ニュース記事本文とそれに対するコメント文から真偽を判定するのみならず，それらの文の中から，なぜその真偽の判定に至ったか，という説明ができる可能性のある部分を抽出でき，根拠のある真偽判定を実現している．

また，Vaswani ら [10] の提案する Transformer というモデルは，attention メカニズムを応用し，self-attention という新たな機構を組み込んだ，文書解釈のためのエンコーダ・デコーダモデルである．このモデルの特徴として，まず RNN (Recurrent Neural Network，回帰型ニューラルネットワーク) や
CNN (Convolutional Neural Network，畳み込みニューラルネットワーク) を用いず，しかしそれらよりも少ない計算量で計算できる．また，並列計算が可能であり，広範囲の依存関係も学習可能である．すなわち，効率的に計算が可能な上，長文であっても単語同士の関係が把握できるのである．さらに，高い解釈可能性も持ち合わせている．

その汎用性や有効性により，Transformer モデルは広く応用されており，Devlin ら [11] の提案する BERT (Bidirectional Encoder Representations from Transformers) と呼ばれる自然言語処理モデルや，Zhilin ら [12] の提案する XLNet と呼ばれる，BERT の改良版モデル など，現在でも高い精度を発揮する自然言語処理を行うためのモデルのベースとなっている．

2.2　文書以外の要素を用いての推測

誤報やフェイクニュースを検出するために，ニュース本文などの文書のみならず，他の情報を用いているような研究も存在する．

Cui ら [6] は，ニュース記事に対するユーザのコメントからユーザの感情を分析し，ニュース記事内の画像，ニュース記事本文，著者やトピックやキーワードなどのニュース記事のプロフィール情報と組み合わせてフェイクニュースを検出する，SAME という手法を提案している．ユーザの感情を何らかの検出をする手法に組み込む場合，まずユーザの表現が必要となる．

例えば，特定のトピックのニュースに対してコメントを残したり，いわゆる「いいね」を残したりするするような行動がある．しかし，このような情報は通常，高次元で疎なものであり，処理の方法に工夫が必要である．また，ユーザの感情情報の表現は疎であるのに対し，他のソースとして用いる画像の情報などは密な情報であるため，他の情報の表現と組み合わせるのが困難である．こうした問題に対処するため，[6] ではニュース記事内の画像，ニュース記事本文，著者やトピックやキーワードなどのニュース記事のプロフィール情報の各情報に異なるネットワークを使用し，異なるモダリティの表現に一貫性を強制的に持たせるため，敵対的ネットワークを使用した．そしてユーザの感情情報をモデル化し，フレームワークに組み込む手法を提案した．

また，Songtao ら [5] が提案するモデルは，ソーシャルネットワーク上のユーザ (本論文ではノードとして設定) 同士の情報の伝搬に着目している．通常，誤情報が発信されたあとなどの任意のタイミングで，その情報を広く拡散してしまう可能性のあるノードの推定は困難である．それに対し，真実を知っているノードはその誤情報を信じない (ここではこのノードを免疫ノードと呼ぶ)，という仮定のもと，真の情報を発信するノード (ここではこのノードを真ノードと呼ぶ) をうまく設定することで，効率的に免疫ノードを作っていゆく，という解決策を取ることができる．しかし，エントロピー値が情報の伝搬に影響を与えることで，真ノードの選択が正しく行えないことがある．そこで，誤情報が広まる期間中の，各ノードに連結しているノードに誤情報が伝わった時，連結しているノードが誤情報を信じる可能性のあるノード数や，情報伝達木でもって計算される二つのノード間の誤情報が伝達してしまう確率を用いて，エントロピー値を考慮した情報の伝達モデルを解析した．

2.3 ソーシャルメディアの情報を利用した健康事象の推定

本研究では，ソーシャルメディア上の情報を活かして健康事象に関する物事を推定する，という課題を扱っているが，同様な研究も多く存在する．

Wakamiya ら [13] は，ツイッターの直接情報・間接情報を利用し，地域ごとのインフルエンザ流行のピークを予測するための手法として，TRAP という手法を提案した．直接情報とは，例えば,「今日熱が出て，病院で診察受けたらインフルエンザにかかってました」というツイートのように，特定の個人がインフルエンザにかかっている，ということが具体的にわかるような情報のことを指す．一方，間接情報とは，例えば，報道局の「今日のニュース：北海道でインフルエンザが流行」というツイートのように，インフルエンザに感染した個人は特定できないが，流行しているおおよその地域や時期などが分かるような情報を指す．[13] 以前での主なソーシャルセンサーベースの予測手法では，直接情報のみが用いられていた．しかし，地域に関連したソーシャルセンサーベースの予測モデルには，問題が生じる．基本的には，人口の少ない地方の方が，人口が多い都市部よりも発信される情報の量は少ないので，直接情報のみで特定地域のインフルエンザの流行を予測することは，特に地方での流行の予測は，困難となる．そこで，

- 多くの人がインフルのシーズンの初期に間接的な情報を多く発信するが，その情報が広まるにつれ，人々は流行について認知しており，ピークの中盤や後半では間接的な情報の発信量は減少する (このように，後半で不活性化した人を [13] では「トラップセンサー」と呼ぶ).
- 情報の伝搬の度合いは，間接情報の量と相関がある．

という二つの仮定の下，間接情報を利用し，地域に依らずインフルエンザの流行を予測する，という手法を考案した．

また，Gaur ら [14] は，Reddit と呼ばれる，主に英語圏で利用されている，画像やリンクとともにコメントを投稿することができるソーシャルプラットフォーム上での，精神疾患に関する投稿から，投稿者の抱える精神疾患がどの分類のものなのかを予測するための手法として，SEDO (Semantic Encoding and Decoding Optimization) という重み行列を提案した．この研究では，以下の 4 つのステップの下で重み行列を導出した．

- DSM-5 という，精神疾患に関する分類学的・診断学的なマニュアルをもとに，他の医学情報データベースの検索の結果，同じような概念の uni-gram，bi-gram，tri-gram を各カテゴリーに格納し，その辞書を作成する．
- Reddit 内の 15 のトピック特化フォーラム内での投稿を含むコーパス上で単語埋め込みモデルを学習し，Word2Vec を用いて語彙の辞書を生成する．
- 上の 2 つのステップで生成された辞書を対応させ，Reddit の方から得られた語彙から，DSM-5 の方から得られた辞書の語彙と一致する最も頻度が高い用語と共起語を抽出する．
- エンコーディングとデコーディングの最適化により，適切な重み行列を得る．

このように，ソーシャルメディア上の情報を用いて健康事象に関する推定を行う研究は，手法や対象が多方面にわたって存在し，現在でもさまざまな研究が行われている．

3 モデル構築

3.1 Attention 機構の構造

本節では，attention 機構 [8] について説明する．まず，attention 機構への入力を作るため，文中の各単語をエンコードする．ここで、ある文 s_i に含まれる単語の集合への適用を考える．ただし，各文には T_i 個の単語が含まれており，その中の $t \in [1, T]$ 番目の単語を w_{it} と表現する．まず，t 番目の入力である単語 w_{it} と，単語埋め込みベクトル W_e，ゲート付き回帰型ユニット (Gated Reccurent Unit) [**?**] を用いて，現在の単語の入力に対する前向き・後向きの隠れ状態 $\overrightarrow{h_{it}}, \overleftarrow{h_{it}}$ は以下のように表すことができる．

$$x_{it} = W_e w_{it}, t \in [1, T] \tag{1}$$

$$\overrightarrow{h_{it}} = \overrightarrow{GRU}(x_{it}), t \in [1, T] \tag{2}$$

$$\overleftarrow{h_{it}} = \overleftarrow{GRU}(x_{it}), t \in [1, T] \tag{3}$$

ここで，x_{it} は単語埋め込みベクトルを用いて現在の入力単語をベクトル化したものである．

隠れ状態が前向き・後ろ向きの双方向をとっている理由としては，単語の意味理解に重要な他の単語が，その単語より前にある可能性と，その単語より後にある可能性があり，その両方をカバーするためである．例えば，“I have a pen and it is my treasure.” という文において，“it” が指す単語は “pen” であるため，“it” より前の部分の情報が必要となる．一方，“I run a restaurant.” という文において，“run” の意味を理解するのに，単に順方向に読むだけでは意味の特定はできないが，後半の “a restaurant” という部分から,「経営する」という意味に特定できる．つまり，“run” より後の部分の情報が必要であった，ということになる．

この双方向の隠れ状態を連結させ、$h_{it} = [\overrightarrow{h_{it}}, \overleftarrow{h_{it}}]$ と表す．

次に，式 (2)，(3) で得られた隠れ状態を attention 機構へ入力し，文ベクトルを得る．この attention 機構で行われている変換を以下に表す．

$$u_{it} = tanh(W_w h_{it} + b_w) \tag{4}$$

$$\alpha_{it} = \frac{exp(u_{it}^{\mathrm{T}} u_w)}{\sum_t exp(u_{it}^{\mathrm{T}} u_w)} \tag{5}$$

$$s_i = \sum_t \alpha_{it} hit \tag{6}$$

ここで，u_{it} は隠れ状態 h_{it} に基づいて生成されるベクトル，u_w は単語レベルのコンテキストベクトル，α は $u_i t$ と u_w により計算される，各単語の重要度を表す重みが正規化されたもの，そして s_i は文ベクトルをそれぞれ表す．

この機構により，文の意味を理解するのに重要な役割を果たす単語に大きい重みがつけられた文ベクトルが得られる．

3.2 HAN の構造

3.1 節で説明した attention 機構は，文の意味理解に重要な役割を果たす単語の推定に用いられていたが，この機構を階層的に，つまり，単語ベクトルを利用して得られた文ベクトルをさらに用いて，文書レベルでの意味理解に重要な役割を果たす文を推定する，というような仕組みになっているのが HAN である．図 2 に HAN のモデルを示す．

まず，文レベルでの attention 機構への入力のため，各文をエンコードする．ここで，ある文書 N に含まれる文の集合への適用を考える．ただし，文書 N には L 個の文章が含まれており，その中の i 番目の文の文ベクトルを s_i と表現する．3.1 節で説明した文ベクトル s_I と GRU を用いて，現在の文の入力に対する前向き・後向きの隠れ状態 $\overrightarrow{h_i}, \overleftarrow{h_i}$ は以下のように表すことができる．

$$\overrightarrow{h_i} = \overrightarrow{GRU}(s_i), i \in [1, L] \tag{7}$$

$$\overleftarrow{h_i} = \overleftarrow{GRU}(s_i), i \in [1, L] \tag{8}$$

これらの隠れ状態の前向き・後向きの必要性も，3.1 節での説明と同様の理由による．この双方向の隠れ状態を連結させ，$h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}]$ と表す．

次に，式 (11)，(12) で得られた隠れ状態を attention 機構へ入力し，文書ベクトルを得る．この attention 機構で行われている変換を以下に表す．

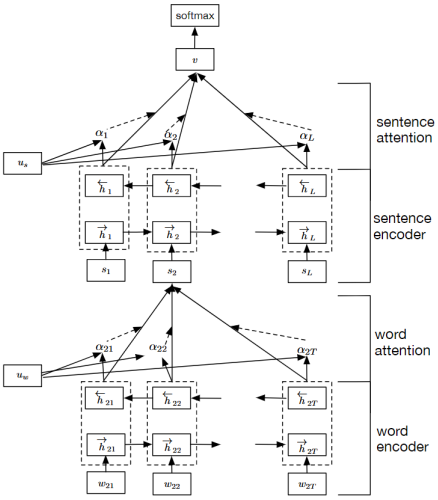

図 2　HAN のモデル図

$$u_i = tanh(W_s h_i + b_s) \tag{9}$$

$$\alpha_i = \frac{exp(u_i^{\mathrm{T}} u_s)}{\sum_i exp(u_i^{\mathrm{T}} u_s)} \tag{10}$$

$$v = \sum_i \alpha_i h_i \tag{11}$$

ここで，u_i は隠れ状態 h_i に基づいて生成されるベクトル，u_s は文レベルのコンテキストベクトル，α_i は u_i と u_s により計算される，各文の重要度を表す重みが正規化されたもの，そして v は文書ベクトルをそれぞれ表す．

文書分類は，softmax 関数を用いて行う．

$$p = softmax(W_c v + b_c) \tag{12}$$

学習時における損失は，正しいラベルでの確率の負の対数尤度をとる．

$$L = -\sum_d \log p_{dj} \tag{13}$$

ただし，j は文書 d の正しいラベルである．HAN においては，この損失が最小となるように重みを学習させ，精度を向上させている．

3.3 リツイートの時系列データ

本研究では，リツイートの時系列データについて，次の 2 つの考え方をとることとする．

(1)　ツイートが投稿されてから初めてリツイートされるまでの時間，そこから次にリツイートされるまでの時間，というように，リツイートされる時間間隔を考える．

(2)　ツイートが投稿されてから一定時間までのリツイート数，そこから次の一定時間までのリツイート数，というように，投稿から一定時間ごとのリツイート数を考える．

図 3 に，これらのモデルを示す．各モデルについて以下で説明する．ただし以下では，リツイートについての表現を，リツ

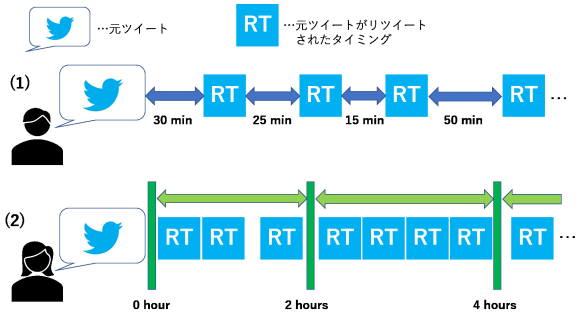

図 3　(1) リツイートされる時間間隔と，(2) 一定時間ごとのリツイート数の考え方の例

イートされた元のツイート T と，そのツイートが m 番目にリツイートされた時間 t_m を用いて，$R_m = (T, t_m)$ とあらわすこととする．

3.3.1　リツイートされる時間間隔での考え方

あるツイート T_i の上位 n 件のリツイートまでの間隔列 $B(T_i, n)$ は，

$$B(T_i, n) = \{b_{T_i1}, b_{T_i2}, \cdots, b_{T_in}\}, \tag{14}$$

$$b_{T_ik} = t_{ik} - t_{i(k-1)}, \quad 1 \leqq k \leqq n \tag{15}$$

と表される．ただし，t_{i0} はツイートが投稿された時間とする．

また、あるニュース・主張 N に対するツイート群 $\{T_1, T_2, \cdots, T_l\}$ の，上位 n までのリツイートのリツイート群 $\{B(T_1, n), B(T_2, n), \cdots, B(T_l, n)\}$ による表現行列 R_1 は，

$$R_1 = \begin{bmatrix} b_{T_11} & b_{T_12} & \dots & b_{T_1n} \\ b_{T_21} & b_{T_22} & \dots & b_{T_2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{T_l1} & b_{T_l2} & \dots & b_{T_ln} \end{bmatrix} \tag{16}$$

と表される．

ここで，b_{T_ij} はツイート T_i が $j-1$ 回目にリツイートされてから j 回目にリツイートされるまでの時間を表すが，さらにその間隔をとる．つまり，

$$ad_ij = b_{T_ij} - b_{T_i(j+1)} \tag{17}$$

を考える．これは，リツイートされる時間間隔の差を表し，この値が正に大きいほどリツイートの時間間隔は広くなり，リツイートの勢いが増している，負に大きいほどリツイートの時間間隔は狭くなり，リツイートの勢いが弱まっている，と捉えられる．また，0 に近ければ時間間隔は一定である，と考えられる．本研究では，この概念を**時間間隔加速度**，と呼ぶこととする．各々のツイートにおいてこの時間間隔加速度を求め，あるニュース・投稿全体での平均列 R_1 を求める．

$$R_1 = \{ad_1, ad_2, \cdots, ad_{n-1}\}, \tag{18}$$

$$a_j = \frac{\sum_{k=1}^{l} ad_{kj}}{l}, \quad 1 \leqq j \leqq n-1 \tag{19}$$

以下において，このモデルを RT_1 と表す．

3.3.2　一定間隔でリツイートされる回数での考え方

あるツイート T_i の，一定時間 D_j から D_{j+1} でのリツイートの回数 $C(T_i, j)$ は，

$$C(T_i, j) = m_p - m_q + 1 \tag{20}$$

$$Dj \leqq t_{m_p} \leqq t_{m_q} < D_{j+1} \tag{21}$$

と表される．また，そのツイートに対する，h 区間までのリツイートの回数列 C_{T_i} は，

$$C_{T_i} = \{C(T_i, 1), C(T_i, 2), \cdots, C(T_i, h)\} \tag{22}$$

と表せる．ここで，あるニュース・主張 N に対するツイート群 $\{T_1, T_2, \cdots, T_l\}$ の，一定時間間隔ごとのリツイートの回数群 $\{C_{T_1}, C_{T_2}, \cdots, CT_l\}$ による表現行列 R_2 は，

$$R_2 = \begin{bmatrix} C(T_1, 1) & C(T_1, 2) & \dots & C(T_1, h) \\ C(T_2, 1) & C(T_2, 2) & \dots & C(T_2, h) \\ \vdots & \vdots & \ddots & \vdots \\ C(T_l, 1) & C(T_l, 2) & \dots & C(T_l, h) \end{bmatrix} \tag{23}$$

と表すことができる．

さらに，こちらでも RT_1 のような考え方を用いる．ここで，$C(T_i, j)$ はツイート T_i が区間 j でリツイートされた回数を表すが，それを一定間隔の時間で割った値は，その区間での一定時間でのリツイートの回数が求まる．これを本研究ではリツイートの**速度**と呼ぶこととする．さらに，それらの差，すなわち，

$$ac_{ij} = \frac{C(T_i, j)}{h_j} - \frac{C(T_i, (j+1))}{h_{j+1}} \tag{24}$$

を考える．ただし，h_m は区間 m における一定時間のことを指す．これは，リツイートの速度の差を表し，この値が正に大きいほどリツイートの一定時間ごとのリツイート回数は上昇傾向にあり，負に大きいほど一定時間ごとのリツイート回数は下降傾向にある，と捉えられる．また，0 に近ければリツイートの一定時間ごとの回数は一定である，と考えられる．本研究では，この概念を**一定時間加速度**，と呼ぶこととする．各々のツイートにおいてこの一定時間加速度を求め，あるニュース・投稿全体での平均列 R_2 を求める．

$$R_2 = \{ac_1, ac_2, \cdots, ac_{h-1}\}, \tag{25}$$

$$a_j = \sum_{k=1}^{l} ac_{kj}, \quad 1 \leqq j \leqq h-1 \tag{26}$$

以下において，このモデルを RT_2 と表す．

3.4　提案モデル

本研究では，HAN をベースとしたリツイートの時系列データを用いる誤報検出器を提案する．提案モデルを図 4 に示す．

式 (22)，(29) で求めた R_{1or2} を用いて，重みをかけ合わせた値 Rp_{1or2} を求める．

$$Rp_{1or2} = W_R R_{1or2} + b_R \tag{27}$$

ここで，二つの出力について，softmax を以下のように定義する．

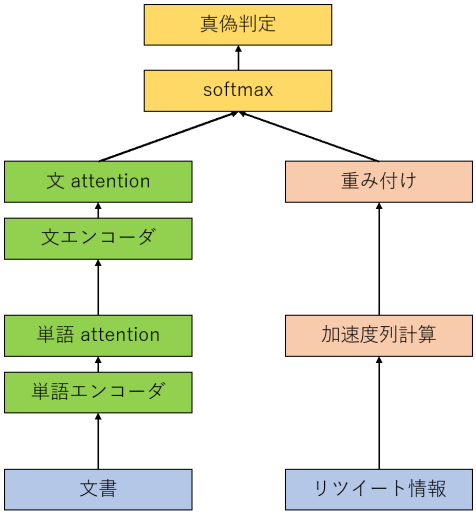

図 4　提案手法のモデル図

$$y = softmax(W_T[v, Rp_{1or2}] + b_T)$$

ただし，$y = [y_0, y_1]$ であり，y_0，y_1 はそれぞれ，ニュース・投稿 N が真である確率，偽である確率を表す．この softmax で得られた値に基づき，真偽を判定する．つまり，y_0 の値の方が大きければそのニュース・投稿は真であると判断し，y_1 の値の方が大きければ，そのニュース・投稿は偽であると判断する．

4　評 価 実 験

4.1　データセット

本研究では，CoAID (Covid-19 heAlthcare misInformation Dataset) [15] というデータセットを使用した．このデータセットは，コロナウイルスに関する健康関連の情報について，真の主張，偽の主張，真のニュース，偽のニュースの 4 種類の記事や投稿の URL，それらの情報の真偽を判定したファクトチェックサイトの URL，タイトル等の情報，各種主張・ニュースについてのツイートのツイート ID，各ツイートへのリプライのリプライ ID，がまとめられている．主張とニュースの特徴，は以下の通りである．

- 主張は，本文が 1〜2 文からなり，World Health Organization (WHO) の公式ホームページや公式ツイッター，Medical News Today (MNT) などから記事を取得している．
- ニュースは，真のニュースについては，信頼性が高いと，複数の手法でクロスチェックされている 9 つの報道発信源から取得，偽のニュースについては，PolitiFact や Health Feedback などのファクトチェックサイトから取得している．

また，これらのニュースや投稿についてのツイートは，指定期間中に投稿されたツイートについて，ニュース記事のタイトルを検索クエリにして取得している．図 5 に取得の様子を示す．

このデータセットでは,「誤情報」と「フェイクニュース」の区別はつけておらず，互換的に用いているので，本研究においても同様に，誤情報とフェイクニュースの区別は付けず，互換的に用いるものとする．本研究では，2020 年 5 月 1 日から 6 月 30 日の期間で投稿・発行されたニュース・主張の本文，及びそれらについてのツイートがリツイートされた時系列データを得るために，CoAID に記載の URL とツイート ID を利用した．リツイートの時系列データについては，CoAID に記載がなかったので，Twitter API を利用して取得した．また，ニュース記事や投稿については，CoAID に記載のニュース・投稿の数にばらつきがあるため，4 種類の情報それぞれについて，表 1 に示す数を取得し，うち 80 %を学習用に，20 %をテスト用に用いた．それぞれの選別はランダムに抽出した．記事や投稿の本文については，CoAID に記載の実験と同様に，Newspaper3k というウェブスクレイピング用のライブラリを用いて取得した．また，表 1 に記載のニュース記事や投稿は，CoAID から取得したツイートデータのうちその情報に関するツイート数が多い順に取得し，YouTube などの本文が取得できないものについては除外した．

図 5　ニュース記事のタイトルをクエリにしたツイートの取得

表 1　本研究で利用したデータセット

情報の種類	ニュース数	ツイート数
真の主張	38	1,279
偽の主張	1	27
真のニュース	100	3,686
偽のニュース	25	976
合計	164	5,968

4.2　評 価 尺 度

本研究における実験では，評価尺度として，学習を 10 エポック行った後のテストデータでの精度と学習時における学習の様子を採用する．精度は，各手法での全ニュース・投稿の真偽の判断のうち，どれくらいの正解率で真偽を判定できたかを比較する．また，学習の様子は，重み学習のなかで各手法がエポックを重ねるごとに学習データセットでの正答率がどのように変化するのかを比較する．また，本研究では，精度と学習時の正答率については 3 回の試行の平均をとることとする．

4.3　実験の概要

実験では，通常の Hierarchical Attention Networks (HAN) と提案手法を比較する．HAN は与えられたデータセットのうち，ニュースまたは投稿の本文のみを使用する．提案手法では，RT_1 では上位 11 件のリツイートについて考え，本文と，各ニュースまたは投稿につき，リツイート数の多かった上位 10 件のツイートから選んでデータとして採用した．つまり，各ニュース・投稿につき，最大で 110 リツイートまでの情報を利用する，ということである．ここで，もし各ニュース・投稿の

ツイートにおけるリツイート数が 11 に満たない場合や，ある時点でリツイートされてから次にリツイートされるまでの期間が 1 週間分より長い時間であった場合，リツイート同士の間隔として，1 週間分の時間を設定した．また，RT_2 の方では一定時間の間隔を 2 時間とし，用いたデータセットから集めたリツイート情報から一定時間ごとのリツイート数をカウントし考慮した結果，最大 8 時間までの 4 区画から得ることとした．

本研究では，単語ベクトルを表現するための埋め込み手法として，CoAID での実験と同様，Pennington ら [16] によって提案された GloVe (Global Vectors) を採用した．GloVe とは，単語埋め込み手法であり，文書全体における単語同士の共起を，重み付き最小二乗で学習することで，適切な単語のベクトル表現を得るものである．本研究では，単語埋め込みの次元数は 200 に設定した．

4.4 結果・考察

実験の結果を表 2，図 6 に示す．まず，精度については，RT_2 は HAN とほぼ同程度の精度を達成し，RT_1 は少し精度が落ち込んだ．RT_1 は特に 2 回目の試行の影響が大きく，その際には学習の精度も悪かった．原因として考えられることとしては，RT_1 で設定した，「リツイートがなかった場合に，リツイート同士の間隔を 1 週間分に設定する」，ということである．例えば，あるニュース記事・投稿のツイートの中で，リツイート数が 3 件ほどしかなかったとする．すると，残りの 8 件分はすべて 1 週間分の期間が空いてリツイートされたことになってしまう．そうすると，このツイートのリツイートの傾向としては，「とても長い期間かけて，少しずつリツイートされる」，というように認識されてしまう．さらに，それが真のニュース・主張と偽のニュース・主張に同じ割合で存在してしまった場合，そのようにリツイートの少ないツイートを持つニュースや主張の真偽については，どちらか分からなくなってしまう可能性が高い．このようなことが本研究における実験の中でもあったものと考えられる．しかし，そもそも試行回数が 3 回，と少ないので，RT_1 の精度の落ち込みは偶然結果が悪くなるような初期の重みが設定されてしまった可能性が考えられる．

また，RT_2 が良い精度を達成した原因としては，RT_1 とは異なり，そもそもリツイートが少ない場合でも，一定期間中にリツイートがない場合と同様に，リツイート回数は 0 回とカウントされるため，あるニュース記事・投稿のツイートに対するリツイート数の総数が少なかったとしても，RT_1 のときのような極端な結果にはなり得ないからだと考えられる．また，RT_1 に比べて，加速度の値に大きな差がほとんどないことも要因の一つとして考えられる．例えば，RT_1 の方では，1 回目のリツイートの間隔の時間が 1 時間，2 回目のリツイートの間隔の時間が 1 日，であった場合，その時間間隔は 24 倍も異なることとなる．さらに，今回設定したリツイートの時間間隔の最大値は 1 週間分であるので，1 時間のものとその最大値を比べると，168 倍も異なることとなる．このとき，例えば，間隔が 1 時間のものと 2 時間のもので比べる場合は，1 時間のものと 1 週間のもので比べる場合に比べて顕著な差が出現しにくくなってしまう．一方 RT_2 は，一定時間を基準としているので，例えば，投稿から 30 分後に最初のリツイートがあったとして，そこから 1 時間後に次のリツイートが，さらにそこから 6 時間後にその次のリツイートがあったとすると，時間間隔で考えると差は 1 時間と 6 時間で 6 倍の差であるが，回数で考えるとどちらも 1 回であり，大きな差の影響は受けない．さらに上限を 8 時間で設定しているので，何日か後にリツイートされてもそれは影響を受けない．これらのことが，高精度を維持できる要因であると考えられる．

表 2　各手法の 3 回の試行の精度の結果と平均

手法	1 回目	2 回目	3 回目	平均
HAN	**100.00**	96.88	**100.00**	98.96
RT_1	90.63	40.63	**100.00**	77.09
RT_2	**100.00**	96.88	96.88	97.92

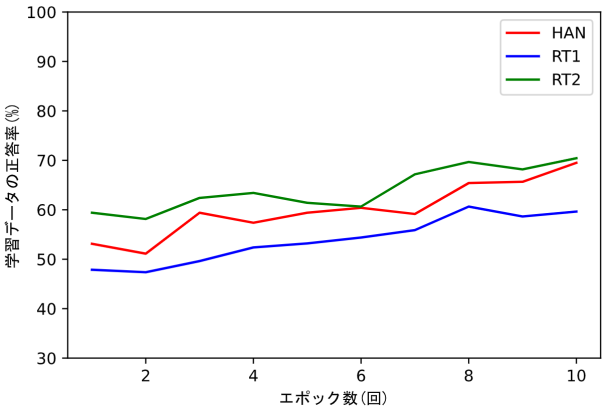

図 6　3 つの手法における 10 エポックまでの学習時の正答率

また，図 6 における学習データの正答率に対して，テストデータの精度が高いのは，データ数の少なさや，図 6 が正答率の平均をとっているため，下振れが起きると低くなってしまうことが原因だと考えられる．

次に，学習の様子について考察する．図 6 のグラフを見てみると，3 つの手法の正答率変化のおおまかな傾きは，ほぼどれも同じであることがわかる．つまり，リツイート情報を考慮することで，学習速度に大きな変化は見られない，ということがいえる．しかし，各エポックごとの正答率の変化に着目すると，HAN や RT_2 は正答率が 10 回のうち 5，6 回ほどしか上昇していないのに対し，RT_1 はほぼ全区間にわたって正答率が上昇している．特に，HAN は 4 エポック目と 7 エポック目，RT_2 は 5 エポック目と 6 エポック目に，それぞれ正答率が大きく落ち込んでいる部分が見受けられるが，RT_1 に関してはそうした中盤の正答率は常に上昇を続けており，安定している様子が観察される．しかし，やはりこれも試行回数が 3 回と少なく，さらに回数を重ねた結果，正答率が大きく変化する可能性もあり，一概に安定がとれると考えるべきではないと思われる．

また，最初のエポックにおける正答率が異なるのは，各試行ごとに初期の重みをランダムに設定しているからである．本研究における分類クラス数は 2 つ (真・偽) であるので，完全に

ランダムな状態からの正答率の期待値は 50 %である．RT_1 の 2 回目の試行の時，1 エポック目の正答率は 25 %程度と下振れしたため，他の二つと比べて平均の正答率が低くなった．

以上から，本研究において，手法 RT_1 は学習の際の安定的な重み学習に，手法 RT_2 は良い精度の維持に，それぞれ効果があるものだと考えられる．しかし，データの少なさや試行回数の少なさがゆえにさらなる検討が必要であると考えられる．

5 おわりに

本研究では，ツイッターにおけるリツイートに関する二つの時系列データの捉え方と，それらを用いた HAN ベースの誤報検出の手法を提案した．

提案手法のベースとなるモデルとしては Hierarchical Attention Networks (HAN) を用いた．HAN は，文の意味理解に重要な単語を抽出することができる attention 機構を単語レベル，文レベルと階層的に用いることで，文書の意味理解に重要な単語や文が抽出された文書表現ベクトルを得ることができる．

提案手法では，リツイートの時系列データについて，リツイートされる時間間隔ごとに捉える方法と，一定時間内でのリツイート数を数える方法の二つを考え，それぞれ，時間間隔の差をとった時間間隔加速度と，リツイート回数を一定時間の時間で割った速度の差をとった一定時間加速度を計算し，手法 RT_1 と RT_2 とした．さらに，計算して得られた加速度に重みを付け，HAN から得られた文書の表現ベクトルと組み合わせて得られた値を softmax への入力として，真偽を判別する誤報検出器を提案した．本研究では，RT_1 の方ではリツイートの時間間隔の差の最大値を 1 週間分に，RT_2 の方ではリツイートの回数をカウントする一定時間を 2 時間で 4 区間分に，それぞれ設定して実験を行った．

実験では，HAN，RT_1，RT_2 の 3 つの手法について，テストデータでの精度とエポックごとの正答率から見た学習の様子を比較し，実験により得られた結果から結果の原因や，各手法の有効性を考察した．

本研究では，今後の課題として，

- データ数を増やすこと
- モデルの構成を十分吟味すること
- 学習に用いたデータを，さらに選別すること
- 実験における設定を，十分に検討すること

が挙げられる．今後は，これまでに述べてきたデータやモデル，実験設定について再検討し，適切なモデルを調べ，慎重に吟味した上でより効果的な誤情報検出の手法を提案していく予定である．また，フォロワー・フォロウィーの関係や拡散の様子をグラフで表現して利用することや，事象の分類 (政治や健康など) による誤情報の特徴を分析・利用することも検討している．

文　献

[1] 読売新聞オンライン (1 月 7 日): トランプ氏のツイッター一時凍結,「選挙に不正」暴動あおる... 永久停止も警告.

[2] 日本経済新聞 (1 月 9 日): Twitter，トランプ大統領のアカウントを永久停止.

[3] 朝日新聞デジタル (8 月 4 日): 「うがい薬で唾液中のコロナウイルス減少」吉村知事会見.

[4] ITmedia ビジネスオンライン (8 月 6 日): イソジンは本当に効くのか? 売り切れ相次ぎ株価は急騰.

[5] Ye, S., Wang, J. and Fan, H.: Minimize Social Network Rumors Based on Rumor Path Tree, *IEEE Access*, Vol. 8, pp. 167620–167630 (2020).

[6] Cui, L., Wang, S. and Lee, D.: SAME: Sentiment-Aware Multi-Modal Embedding for Detecting Fake News, *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining(ASONAM '19)*, p. 41–48.

[7] Bahdanau, D., Cho, K. and Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate, *3rd International Conference on Learning Representations (ICLR '15)*.

[8] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. and Hovy, E.: Hierarchical Attention Networks for Document Classification, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-NLT '16)*, pp. 1480–1489.

[9] Shu, K., Cui, L., Wang, S., Lee, D. and Liu, H.: DEFEND: Explainable Fake News Detection, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*, p. 395–405.

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. and Polosukhin, I.: Attention is All you Need, *Advances in Neural Information Processing Systems*, pp. 5998–6008 (2017).

[11] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT '19)*, pp. 4171–4186.

[12] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R. and Le, Q. V.: XLNet: Generalized Autoregressive Pretraining for Language Understanding, *Advances in Neural Information Processing Systems*, pp. 5753–5763 (2019).

[13] Wakamiya, S., Kawai, Y. and Aramaki, E.: Twitter-Based Influenza Detection After Flu Peak via Tweets With Indirect Information: Text Mining Study, *JMIR Public Health Surveill* (2018).

[14] Gaur, M., Kursuncu, U., Alambo, A., Sheth, A. P., Daniulaityte, R., Thirunarayan, K. and Pathak, J.: "Let Me Tell You About Your Mental Health!": Contextualized Classification of Reddit Posts to DSM-5 for Web-based Intervention, *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*, pp. 753–762.

[15] Cui, L. and Lee, D.: CoAID: COVID-19 Healthcare Misinformation Dataset (2020).

[16] Pennington, J., Socher, R. and Manning, C.: GloVe: Global Vectors for Word Representation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP '14)*, pp. 1532–1543 (2014).

[17] Hochreiter, S. and Schmidhuber, J.: Long Short-term Memory, *Neural computation*, Vol. 9, pp. 1735–80 (1997).