2.【研究計画】※適宜概念図を用いるなどして、わかりやすく記入してください。なお、本項目は 1 頁に収めてください。様式の変更・追加は不可。

(1) 研究の位置づけ

特別研究員として取り組む研究の位置づけについて、当該分野の状況や課題等の背景、並びに本研究計画の着想に至った経緯も含めて記入してください。

当該分野の状況: フェイクニュースの自動検出

SNS の発展で情報を迅速かつ大量に取得・共有が容易になった一方、悪意により他人を騙すために作られた**フェイクニュース**も拡散されやすくなった。特に 2020 年から COVID-19 の影響による誤情報の拡散であるインフォデミックにより、メタノール飲用による死亡事故 [1] といった事象が報告された。以上から騙された人々により社会的損害が起きるため、フェイクニュース拡散の早期抑制が必要である[2]。

フェイクニュース検出へ有識者が調査する**ファクトチェック**がある。これは拡散ののち着手されるため、拡散抑制にはならない。そのため、自動でフェイクニュースを検出するべく深層学習によってファクトチェック結果をラベルとして、記事内容や利用者の反応から教師あり学習で自動検出する研究がある [3]。

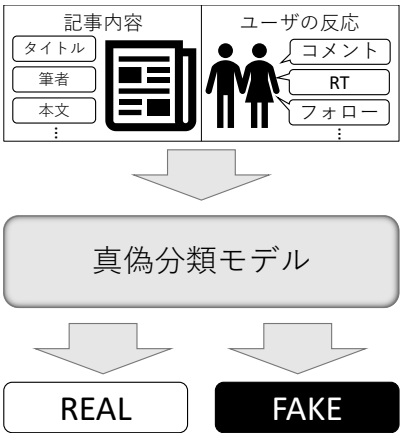

図 1: フェイクニュース自動検出の基本的な流れ

課題

フェイクニュース自動検出が抱える課題は以下の通りである。

ニュースのタイムリー性 ニュースという性質上、時間経過で扱われる情報が徐々に古くなりフェイクニュースの傾向変化への対応が難しくなる。先行研究によると学習・検証で入力するニュースの出来事を変えると検出性能が劣化する [3]。よって継続してデータセットを拡大する仕組みが必要である。

SNS プラットフォームへの依存性 SNS 上で利用者からの反応を取得する場合、その形式は取得元の SNS プラットフォームに依存する。今後主流となる SNS が変わった場合、利用者層や時代の違いにより既存のデータでは対応できない。

早期検出と正確性の両立 記事内容に加えて利用者の反応 (RT, コメント等) を扱うと検出性能が改善する [4] 一方、利用者の反応を十分に得るには時間がかかり、高い正確性と早期検出を両立できない。

日本語データセット不足 深層学習による検出は、正解ラベルとして多量のファクトチェック結果を要する。このファクトチェックが活発な地域差の影響でデータセットが**英語に集中** [5] している。もし**日本語を対象**とした場合、ラベル不足の影響により教師あり学習ができない。

本研究計画の着想に至った経緯

私は修士課程で英文フェイクニュース早期検出の研究を行った。記事に対する利用者のコメントが検出に有用とする先行研究 [7] をベースに、早期検出を想定して少ないコメントから更にコメント内容を自動生成して検出するモデルを実装した。実験にてコメントを生成した上で分類することでより多くのフェイクニュース検出を実現した (査読付き海外 IEEE 学会 発表済 [8])。

一方、研究が進むつれて社会変化の激しさを実感した。ニュースのトピックは日を追うごとに変化すると同時にフェイクニュースの内容も変わる上、ユーザの反応が現れる SNS プラットフォームも近年は新しいサービスが提供されている。この影響でこれまでの記事＋ SNS 上の反応を扱う検出形式では、データセット作成・モデル提供のみでは時代の変化に対応できない。よって記事を継続して収集する枠組み作りに併せ、ユーザの反応をSNS プラットフォームに左右されない形で得る点の重要性に着目した。

参考文献

[1] H. H-M, *et al. Critical Care* 24.1 2020: 1-3.
[2] S. Tasnim, *et al. JPMPH* 53.3 2020: 171-174.
[3] Y. Wang, *et al. KDD'18*, pp. 849-857. 2018.
[4] L. Wu & H. Liu. *WSDM '18* pp. 637-645, 2018.
[5] K. Shu,*et al. Big Data 8.3* 2020: 171-188.
[6] Y. Bang, *et al. arXiv preprint arXiv:2101.03841* 2021.
[7] K. Shu, *et al. KDD'19* 395 – 405, 2019.
[8] Y. Yuta, *et al. INES*. 2020.
[9] K. Shu, *et al. arXiv preprint arXiv:2004.01732* 2020.

【研究計画】（続き）※適宜概念図を用いるなどして、わかりやすく記入してください。なお、各事項の字数制限はありませんが、全体で２頁に収めてください。様式の変更・追加は不可。

(2) 研究目的・内容等

① 特別研究員として取り組む研究計画における研究目的、研究方法、研究内容について記入してください。
② どのような計画で、何を、どこまで明らかにしようとするのか、具体的に記入してください。
③ 研究の特色・独創的な点（先行研究等との比較、本研究の完成時に予想されるインパクト、将来の見通し等）にも触れて記入してください。
④ 研究計画が所属研究室としての研究活動の一部と位置づけられる場合は申請者が担当する部分を明らかにしてください。
⑤ 研究計画の期間中に受入研究機関と異なる研究機関（外国の研究機関等を含む。）において研究に従事することも計画している場合は、具体的に記入してください。

①研究計画における研究目的、研究方法、研究内容

研究目的

本研究では、利用者に直接反応を伺うシステムを構築することで、新たなフェイクニュース早期検出の枠組みを作る。またフェイクニュースの早期自動検出を日本語で実現・研究の促進を目指し、データセットの作成から検出モデルの実装を目的とする。

研究方法・研究内容

以下の 3 目標を目指し研究する。本研究の最終形を図 2 に記す。

目標 I 新たな検出コンセプトとして SNS 上で利用者へ反応を促すことで長期間に渡り早期検出を実現するシステムを構築する。

目標 II 日本語データセット作成に向けファクトチェック済記事とされていない記事、そして同時に SNS 上で記事に寄せられたコメントや利用者情報等を収集する。

目標III 利用者の初期反応から得られた情報より高精度な真偽分類を行うモデルを開発する。

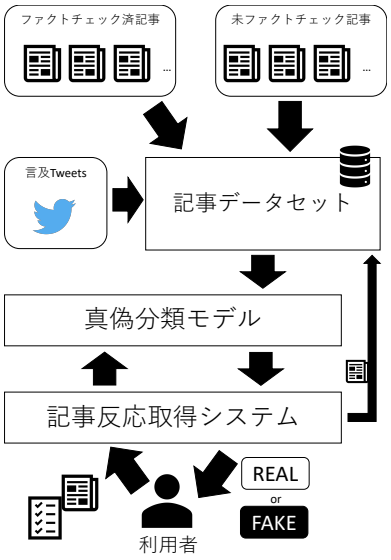

図 2: 本研究計画の完成予想図。

②どのような計画で、何を、どこまで明らかにしようとするのか

目標 I：SNS 上で利用者へ反応を促すことでさらなる早期検出を実現するシステムの構築 (採用前 - 1 年目)

データセットとモデルの構築のみでは、今後のニュース内容の変化に対応することが難しい。継続して社会の潮流の変化に対応するためには、データセットと検出方法及び利用者の反応を得る方法を工夫する必要がある。本研究では、SNS 上で利用者に対して反応を促すことで早期検出の性能向上に繋げる新しいモデルを実装する。実現形式として、フェイクの疑いが強いと判断された記事に対してキュレーターの役割であるリプライを飛ばしたり、利用者が信憑性を問い合わせたい記事に対し簡易的なチェックリスト (文体が感情的か・著者は明記されているか等) を課して回答を得たりする方法などを検討している。対象記事は英文とし、真偽分類を行うモデルは既存手法を採用する予定である。実験では SNS 利用者を対象とした主観評価実験により、フェイクニュース拡散の抑止効果を測定する。

目標 II：日本語の記事・真偽を含むデータセットを作成する (1 年目 - 2 年目)

日本語での検出を実現するためには、まずはデータセットを作成する必要がある。日本語ファクトチェック結果の取得には、特定非営利活動法人ファクトチェック・イニシアティブ (以下 FIJ) が提供する FactCheck Navi を使用する。2021 年 4 月現在で 600 超件のファクトチェック結果が公表されている。一方ファクトチェックにより正確と判断された事例はフェイクに比べて少ないため、正確なニュースとして大手新聞社やロイター通信等の記事を収集する。

目標	採用前	1年目	2年目
記事反応取得システム開発	←	—	→
日本語データセットの作成		←	— →
自動検出モデルの構築			← — →

図 3: 研究計画の年毎の流れ。

(研究目的・内容等の続き)

また目標Ⅲに向けファクトチェックが行われておらず、正解ラベルがない記事も追加する。最終的には真偽合わせてラベル付き記事を約 1,200 件、ラベルなし記事を約 5,000 件以上収集を目指す。利用者の反応として、全記事を対象に SNS 上で寄せられたコメントとして Twitter にて記事 URL を含むツイートも収集する。

目標III: 弱教師あり学習によってラベル不足を補うモデルを構築する (2 年目)

教師あり学習の課題である高精度なアノテーションのコストを補う方法として、不正確な弱いアノテーション付きデータから正確な分類を行う弱教師あり学習がある。利用者の反応を弱いアノテーションとして扱いフェイクニュースを検出する方法があり [10]、全コメントの賛否両論さやコメント者の過去の投稿、そしてフォロー関係から弱いアノテーションを付加している。これら 3 種の弱いアノテーションも併せて学習することで、推論時は利用者の反応を使わずに正確な早期検出を実現した。日本語で実現を目指す場合、日本語での検出を行う研究が少なく言語の違いによる影響が未知数であるため大幅なモデルの改変を要する可能性がある。今回は日本語での早期検出実現に向けて 3 種の弱いアノテーションに加え、投稿者のプロフィールや使用された絵文字、ハッシュタグといったコメント情報で有用なものがないか模索する。実験では、学習時には記事と利用者の反応を入力に扱い、テスト時は記事のみを入力して早期検出時の状況を再現し既存の手法と比較する。既知の手法と比較して検出性能の改善がみられたら成功とみなす一方、達成が厳しいならば目標 I の主観評価実験を日本語と提案モデルに置換して再実験し日本語での効果を測定する。

③研究の特色・独創的な点

本研究の特色

- 能動的に利用者から情報提供を得た上で、継続してデータセット拡大・モデル改善を行う点。
- 日本語を対象にフェイクニュースの自動検出を行う点。
- ファクトチェックの結果を待たず早期の検出を目指す点。

独創的な点: 先行研究との比較

先行研究は利用者の反応を利用する場合該当情報を時間が経過してから取得する受動的な形を取り、形式も SNS プラットフォーム (Twitter, Instagram, Weibo, etc.) によって微妙な差異がみられる。本研究では能動的に利用者の反応を得る枠組みを作り、今後主流 SNS が変化しても早期検出の実現が可能となる。

また深層学習でフェイクニュースを自動検出する研究対象は英文に集中し、日本語データセットがない。言語に囚われず利用者による拡散された経緯で真偽を判断する研究もあるが [12]、依然として記事の内容を考慮した研究では日本語を対象としたものがない。

独創的な点: 予想されるインパクト・将来の見通し

総務省によると SNS 利用率は 2019 年現在 69%を占める上、SNS マーケティング市場規模は 2025 年に 1 兆 1,171 億円まで成長する (出典:サイバー・バズ/デジタルインファクト調べ) と推測されている。SNS 利用者の増加によって、今後フェイクニュースは更に深刻な社会損害を起こし、謂れなき風評被害に悩む事例の増加を懸念する。本研究の完成により、これまで活発になされていなかった日本語のフェイクニュースを早期検出するモデルの開発および提供が可能となる。SNS 利用者への注意喚起に活用ができるほか、ファクトチェックの担い手への補助システムへの活用といった様々な形式で、SNS 上で騙される人を減らし社会的損害や風評被害を未然に防ぐ枠組み作りに貢献する可能性がある。

④申請者が担当する部分

本研究は所属研究室内でも萌芽的な取り組みで、環境・技術面の支援を除き申請者が全部分を担当する。データセットの生成では、正確なニュースの取得へ大手マスメディアへ協力を求める可能性がある。

⑤受入研究機関と異なる研究機関での研究従事計画

私は1 年間タリン工科大学の言語技術研究所 (Tanel Alumäe 所長) で活動予定である。当該分野は北米と欧州の研究が活発であることから、最前線の研究に従事するために必要である。

参考文献

[10] K. Shu, *et al. ECML-PKDD 2020*

[11] UNIC. `https://is.gd/UNIC_pause`. 2020

[12] T. Hamdi, *et al. ICDCIT 2020*

３．人権の保護及び法令等の遵守への対応 ※本項目は１頁に収めてください。様式の変更・追加は不可。

本欄には、「２. 研究計画」を遂行するにあたって、相手方の同意・協力を必要とする研究、個人情報の取り扱いの配慮を必要とする研究、生命倫理・安全対策に対する取組を必要とする研究など法令等に基づく手続が必要な研究が含まれている場合に、どのような対策と措置を講じるのか記入してください。例えば、個人情報を伴うアンケート調査・インタビュー調査、国内外の文化遺産の調査等、提供を受けた試料の使用、侵襲性を伴う研究、ヒト遺伝子解析研究、遺伝子組換え実験、動物実験など、研究機関内外の情報委員会や倫理委員会等における承認手続が必要となる調査・研究・実験などが対象となりますので手続の状況も具体的に記入してください。

なお、該当しない場合には、その旨記入してください。

コメント取得を予定してしている SNS は Twitter である。Twitter 社は 2020 年 3 月より学術目的で Twitter API の利用を自由化しているほか、取得したツイート ID を含む情報をデータセットとして公開することも学術目的であれば認められている [12]。

また、先行研究が提供したデータセットを使用する場合は、提供者が示しているライセンスやポリシーを遵守する。

なお、学習済みモデルの公表は平成 30 年 (2018 年) 改正著作権法第 30 条 4 号により認められている。

ただし、本研究では主観評価実験として SNS 利用者を対象としたアンケート調査を予定している。この調査により収集したデータは、個人の特定につながる情報を匿名化した上で解析を行い、解析結果の公表に際しては、匿名化を行ったデータを用い、個人情報の漏洩防止に配慮する。

参考文献

[12] Twitter 開発者ポリシーを分かりやすくアップデート, 2020 年 3 月 11 日. (最終閲覧日 2020 年 4 月 19 日) https://blog.twitter.com/developer/ja_jp/topics/tools/2020/DevPolicyUpdate.html

4．【研究遂行力の自己分析】※各事項の字数制限はありませんが、全体で２頁に収めてください。様式の変更・追加は不可。
本申請書記載の研究計画を含め、当該分野における(1)「研究に関する自身の強み」及び(2)「今後研究者として更なる発展のため必要と考えている要素」のそれぞれについて、これまで携わった研究活動における経験などを踏まえ、具体的に記入してください。

(1) 研究に関する自身の強み

主体性

自ら抱いた問題意識を出発点に研究を行う主体性

状況把握能力

貪欲な海外論文調査に裏打ちされた状況把握能力

実装能力

産学問わない活動で培ったプログラミング能力で実現される実装能力

コミュニケーション能力

プラットフォームを問わず議論を活発に行えるコミュニケーション能力

プレゼン能力

相手が小学生でも物事を分かりやすく伝えることができるプレゼン能力

成果: 国際会議における発表

(以下 1 件 査読あり・論文及び口頭発表)

1. ○ Yuta Yanagi, Ryouhei Orihara, Yuichi Sei, Yasuyuki Tahara, and Akihiko Ohsuga.
   “Fake news detection with generated comments for news articles”.
   The 24th IEEE International Conference on Intelligent Engineering Systems 2020,
   (Reykjavík, Iceland) Virtual event due to COVID-19, July 2020.

成果: 国内学会やシンポジウムにおける発表

(以下 2 件　査読なし・口頭発表)

2. ○ 柳裕太、田原康之、大須賀昭彦、清雄一
   「画像付きフェイクニュースとジョークニュースの検出・分類に向けた機械学習モデルの検討」、
   日本ソフトウェア科学会 2018 年度 MACC 研究発表会、大分、2019 年 3 月

3. ○ 柳裕太、折原良平、清雄一、田原康之、大須賀昭彦
   「フェイクニュースの早期自動検出に向けたニュース記事コメント生成モデルの提案」、
   言語理解とコミュニケーション研究会（NLC） 第 17 回テキストアナリティクス・シンポジウム、
   オンライン、2021 年 2 月

(以下 1 件　査読なし・ポスター発表)

4. ○ 柳裕太、葛西透麿、 森谷薫平、神谷岳洋、藤原徹、木村健太、榎本裕介
   「CaD428 の変異遺伝子の機能解析ツールの汎用化」、
   広尾学園高校医進・サイエンスコース研究成果報告会、東京、2015 年 3 月

(2) 今後研究者として更なる発展のため必要と考えている要素

要素 I : 学術的成果と社会問題の最前線の間にあるギャップを埋めるための発想と問題解決力

フェイクニュースの自動検出を行う研究は世界的には広く行われており、それぞれが独自の発想を追加している。この独自の発想の付加には、1 つの技術分野に絞らず他分野から得た知見がもたらす。そのため技術面では自然言語処理に限らず、利用者の拡散を考慮するためのグラフネットワークや、既知の情報を利用するためのナレッジグラフ技術など、幅広い分野の研究に論文を通して触れる必要がある。

(研究遂行力の自己分析の続き)

要素 II：多彩な分野や言語・地域圏の研究者らと活発な議論を交わす能力

要素 I の実現には、論文のみならず実際に議論を交わすことで更に深い理解を得ることが重要である。また海外で研究が活発に行われていることから、知見のアップデートも頻繁に行うことも必要である。よって分野・言語問わず多くの研究者達とプラットフォームを問わない深い議論が研究の発展をもたらすと考える。

要素III: 研究で得られた成果をどんな聞き手でも分かりやすく伝えられる表現力

新型コロナウイルス感染症蔓延の影響もあり、発表の機会や形式が大きく制限されたまま修士研究を終えた。オンライン形式での発表での経験を積めた一方、人前で発表する機会は未だにない。以上から場所を問わず誰が相手でも研究を分かりやすく伝える経験を積む必要が例年以上に必要と考える。

５．【目指す研究者像等】※各事項の字数制限はありませんが、全体で１頁に収めてください。様式の変更・追加は不可

日本学術振興会特別研究員制度は、我が国の学術研究の将来を担う創造性に富んだ研究者の養成・確保に資することを目的としています。この目的に鑑み、⑴「目指す研究者像」、⑵「目指す研究者像に向けて特別研究員の採用期間中に行う研究活動の位置づけ」を記入してください。

(1) 目指す研究者像 **※目指す研究者像に向けて身に付けるべき資質も含め記入してください。**

自分の興味のある分野を研ぎ究めると同時に社会問題を解決して人々の生活を幸せにしたい
嘘の情報に騙されて誤った風評が残り不幸になる人を 0 にしたい
ファクトチェックでは嘘は嘘であると騙された人を相手に分かりやすく説明することが重要である
自己完結のみならず成果を他人に伝えるまでが研究である

(2) 上記の「目指す研究者像」に向けて、特別研究員の採用期間中に行う研究活動の位置づけ

特別研究員の採用期間中に行う研究活動のなか、4-(2) で挙げた今後研究者としてさらなる発展のため必要と考えている要素の習得を通して、学術研究で得た知見を直接 SNS 利用者を含む日本社会に還元する研究者を目指す。その実現に向け、査読付き国際会議ないしは国際論文誌への論文発表をはじめ、国内・国際会議での口頭発表も積極的に行う。また、自然言語処理コミュニティに限らず国内ニュースメディアと積極的に連携を行い、フェイクニュースの自動検出に関連した共同研究の実現が理想である。

特別研究員として研鑽を重ねていき、現状の研究への新たな発想を追加し、実現に向けて幅広い人々と議論を重ね、得られた成果を端的に説明することが、能動的に一貫して社会課題を解決へ自ら導く研究者として大成する。その実現の大きな足がかりが本研究計画である。