主張と根拠のクラスタを用いた多様な主張を提示するニュース推薦手法の提案

指導教員 木村 昌臣　　片岡 凪

1　研究の背景と目的

ニュース記事となる出来事は, 記者によって受け取り方や伝え方が異なる．RSF(Reporters Sans Frontires) は, 政治や文化などの要因によって記者の主張が制限されていることを問題視している [1]．近年の Web ニュースには世界中の記者たちの様々な主張が見られるが, 言語の違いによる時間的コストなどから，多くの読者はそれらの主張の一部しか把握できない．

Yang らは，ニュース読者が出来事を正確かつ迅速に把握できるように階層的クラスタリングを用いて記事に対するツイートの主張をグループ化する手法を提案した [2]．この研究では COVID-19 の話題に限定した主張の文をグループ化しているが，この手法を記事に適用した場合，多くの話題について主張のまとまりが生成されることになる．これでは，COVID-19 の飲み薬やワクチンなど異なる話題に含まれる安全性に関する共通した主張がグループ化されるため，読者が興味を持っている出来事の異なる主張の収集に時間を要してしまう．

そこで本研究では，まず世界の記事を出来事の類似度よりグループ化し，同じグループの記事群から主張の類似度より再度グループ化することで，類似した出来事の異なる主張を推薦する手法を提案する．

2　提案手法

本研究では，記事の文章が出来事を述べる文と主張を述べる文に二分できると仮定する．

図 1 に提案手法の概要を示す．まず (i) 記事の文章を出来事の文と主張の文に分類し，次に (ii) 出来事の文章の類似度で記事をクラスタリングし，その後 (iii) 主張の文の類似度で主張の文をクラスタリングする．

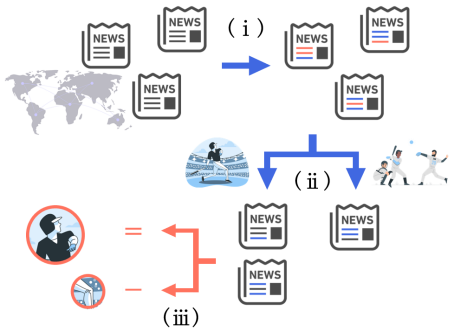

図 1: 提案手法の概要

入力する記事のデータセットには，Piyush らが収集した COVID-19 News Articles を選定した．COVID-19 についての幅広い出来事が含まれ，またインド，日本，韓国の約 8 万件の英記事が含まれるため，政治や文化の違いに由来する多くの異なる主張が分析できると考えた．また，11 カ月間という短い期間で収集されているため，より類似した出来事に関する記事が得られると期待できる．Yang らの研究と同様に話題が COVID-19 に限定されてしまっているが，限定された話題でも同じ出来事の異なる主張がグループ化される様子は観測できると考えた．

出来事と主張の文の分類には，RoBERTa を転移学習した分類器を用いる．RoBERTa は文の意味や文脈を加味した自然言語処理を行うための機械学習モデルであり，事前に学習されたモデルを利用した．

追加の学習に用いるデータセットには，Wikipedia の英文を Claim を述べるか Evidence を述べるかでラベル付けした Ruty らの IBM Debater - Claims and Evidence を利用した．2294 個の Claim は「トピックを直接サポートする一般的で簡潔な文」, 4692 組の Evidence は「トピックの文脈の中で Claim を直接サポートする文章」と定義されており，本研究では主張と Claim，出来事と Evidence を同じ定義で扱う．

クラスタリングのため，「出来事と分類した文を記事ごとに結合した文章」と「主張と分類した文」のそれぞれを，Sentence-BERT を用いてベクトルに変換した．Sentence-BERT は文章をその意味を表すベクトルに変換する機械学習モデルである．文章間の距離の算出にはベクトル間のコサイン値を使用し，クラスタ間の距離の算出には Ward 法を用いた．記事は出来事に対して主張を述べる構造を持つため，先に出来事の文章で記事をクラスタリングし，その後主張の文をクラスタリングした．

その後，読者が興味を持っている記事を 1 つ選択し，記事の文を出来事の文か主張の文かで分類し，出来事の文章を結合したものを Sentence-BERT に入力してベクトルを得る．得られたベクトルと距離が近い k 個の出来事の文のベクトルを k-NN 分類法を用いて取得し，同じ出来事を扱う記事のクラスタを同定する．同定した記事のクラスタに紐づく主張の文をクラスタごとにまとめ，記事とともに読者に提示・推薦する．

3　実験

3.1　出来事の文と主張の文の分類器の評価

分類器の COVID-19 News Articles での分類精度を確認するため，COVID-19 News Articles の 3 ヶ国の

記事から 3 件ずつ記事を抽出し，計 163 個の文を出来事か主張かで手動でラベル付けし，分類器で付与したラベルとの比較評価を行った．評価指標には，分類数が少なかった主張の文に着目した適合率と再現率を用いた．適合率は「主張だと分類した文のうち実際に主張の文であった割合」，再現率は「主張の文のうち正しく分類できた主張の文の割合」を示す．実験の結果，適合率は 1，再現率は 0.4 となった．表 1 に分類結果の例を示す．3 つ目の文は，一般の大衆にいえる筆者の主張が述べられているが，出来事に誤分類されている．

表 1: 分類器の主張の文 (C) と出来事の文 (E) の分類例

分類	COVID-19 News Articles の文
C(正)	the goal was to bolster international competitiveness.
E(正)	the government is reportedly aiming to announce guidance early next month.
E(誤)	for now there is little that families and educators can do but wait to see what the abe administration has in mind.

3.2 階層的クラスタリングの評価

クラスタリングの実験には 5000 件の記事を用いた．図 2 に主張の文の階層的クラスタリングの結果を示す．クラスタを分けるクラスタ間の距離を大きくしたときに，クラスタ数の減少が緩やかになり始めるクラスタ間の距離 0.85 を選択した．この距離は，より類似しない文章がクラスタにまとめられ始めるクラスタ間の距離に対応する．

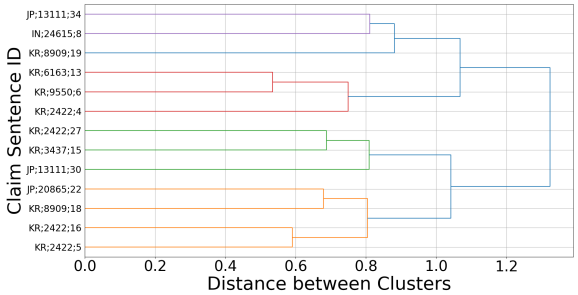

図 2: 主張の文のクラスタリング結果（抜粋）

ある 38 件の記事を持つクラスタは主張の文を 13 件持ち，うち 10 件がウイルス感染対策の出来事に対して主張を述べる文であった．表 2 にこの 13 件の主張の文をクラスタリングした結果の一部を示す．クラスタ c_1 には飲食店での感染対策に関する主張がグループ化され，クラスタ c_2 には運動時の感染対策に関する主張がグループ化されていた．また，c_1 以外のクラスタに属する感染対策に関する 7 文全てが飲食店以外の感染対策に関する主張であり，異なるクラスタに異なる主張の文が属することを確認できた．

表 2: 同じクラスタ ID と異なるクラスタ ID での文の比較

ID	主張の文
c_1	try not to eat in restaurants as much as possible.
c_1	franchise cafes and dessert shops were obligated to offer only takeout around the clock.
c_2	sports event are also obligated to keep the ceiling of 30 percent at stadiums.
c_2	outdoor exercise will be banned and wearing masks will be mandatory.

4 考察

分類の適合率が極めて大きいため，2019 年の最先端のモデルである RoBERTa が適切に転移学習できたと考える．一方で再現率は低いため，学習と分類に用いたデータセットの違いを分析する必要がある．学習に用いた文は指示語が含まれる文を出来事の文とすることが多い．表 1 で出来事と誤分類した主張の文には指示語に用いられる”there”が含まれる．しかし，誤分類した文では”there is”という構文で用いられ，指示語として”there”を用いていない．再現率の向上のため，このような構文を加味した分類手法の検討や，記事の分類に最適な学習用データセットの再検討が必要であると考える．なお，確認した 163 文のうち主張の文は 6 文しかなかったため，より信頼できる適合率と再現率の算出のためにより多くの文での評価が必要である．

飲食店の感染対策に関する文で異なるクラスタには異なる主張の文が属することを確認できたが，10 件の感染対策に関する文とその他 3 件の文を出来事として区別できていないため，クラスタ間の距離の算出方法や階層の分け方に改善が必要である．

5 まとめ

本研究では出来事の文と主張の文のクラスタを用いた多様な主張を提示するニュース推薦手法を提案した．提案手法ではより類似した出来事の異なる主張を推薦するため，記事の文を出来事の文か主張の文かで分類し，それぞれの文に基づいて 2 回の階層的クラスタリングを行った．その結果，高い適合率で分類を行い，個々のクラスタから類似した出来事の異なる主張を推薦できた．今後の課題として，クラスタ間の距離の算出方法の改善や階層の分け方の改善を行う必要がある．

参考文献

[1] RSF. 2021 World Press Freedom Index: Journalism, the vaccine against disinformation, blocked in more than 130 countries. `https://rsf.org/en/2021-world-press-freedom-index-journalism-vaccine-against-disinformation-blocked-more-130-countries` (2022 年 1 月 12 日参照).

[2] Jing Yang, Didier Vega-Oliveros, Tais Seibt, and Anderson Rocha. Scalable Fact-checking with Human-in-the-Loop. In *2021 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6, Montpellier, France, December 2021. IEEE.