記事トピックのクラスタを用いた多言語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 ニュースには世界の記者たちの様々な主張が散見されるが，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．

そこで本研究では，閲覧記事と取り扱う出来事が類似する世界の記事を抽出し，得た記事から主張が異なる文章を推薦するシステムを提案する．ここで,「出来事」は記者の解釈に依存しない事象とし,「主張」は記者が伝えるべきだと判断した出来事の解釈とする．

2　先行研究

意味や話題が類似する文章の推薦手法として，LDA(Latent Dirichlet Allocation) [2] や Sentence-BERT [3] が提案されている．LDA は記事に関連の深い単語群とそれぞれの単語の関連度を出力するが，単語群同士の関係性が得られないという点で，出来事の類似度が高い記事は得にくいと考えられる．一方，Sentence-BERT は単語同士の関係性を加味した文意の数値ベクトルを出力するため，より高い出来事の類似度が得られると考えられる．しかし，記事の全文を Sentence-BERT に入力した場合，出来事の類似度の算出に主張を述べる文も考慮することになり，より近い出来事に関する主張の違いを推薦できない．

3　提案手法

まず，英文以外の記事を DeepL API で英訳して用いる．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．そして，出来事の文のみ，主張の文のみで 2 回に分けてクラスタリングすることで，より近い出来事に関する異なる主張の文の推薦を目指す．

出来事と主張の文の分類には，Transformer を応用した分類器を用いる．Transformer は入力文から翻訳後の文を予測するような機械学習モデルであるが，Dense 層を加えたモデルのエンコード部分を用いることで分類タスクに応用できる．本研究では，入力文を数値ベクトルに変換するために，多くのニュース記事を事前学習した RoBERTa を用いる．また，分類器の学習には，英文を Evidence を述べるか Claim を述べるかでラベル付けした IBM の Debater Datasets を用いる [4]．

次に，出来事の文を Sentence-BERT を用いて数値ベクトルに変換し，それらのコサイン類似度を基に記事のクラスタリングを行う．最後に，閲覧記事が属するクラスタの主張の文を抽出し，クラスタリングを行う．

4　研究状況

分類精度の確認のため，前述の操作で出来事の文と主張の文の分類を行った．ただし，分類する日本の記事には Japanese FakeNews Dataset 中の偽物でない Wikinews を利用した [5]．学習で正解率が約 0.993 となったモデルで分類を行ったところ，表 1 の結果が得られた．

表 1: Evidence の文 (E) と Claim の文 (C) の分類結果

分類	翻訳前の入力文（一部抜粋）
E	決勝のヒットを打った 23 日の試合も 1 球だけで終わった
C	日本シリーズ進出を決めてうれしい

文 E は，記者の解釈に依存しない，誰が観測しても不変な出来事を表している．また，文 C は，出来事に対する主張を表している．一方で，文中に出来事を表す「敗れた」や「風が吹く」という言葉が含まれることより，出来事を誤分類したケースも見られた．

5　今後の予定

今後は，1 文中の出来事と主張の要素を比率で考えるなどして，推薦精度を高めるように研究を進める予定である．

[6]

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 (2021 年 7 月 19 日参照).

[2] Ming-Jie Tian, Zheng-Hao Huang, and Rong-Yi Cui. Labeled Bilingual Topic Model for Cross-Lingual Text Classification and Label Recommendation. In *2018 5th International Conference on Information Science and Control Engineering (ICISCE)*, pp. 285–289, July 2018.

[3] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. https://arxiv.org/abs/1908.10084v1 (2021 年 7 月 19 日参照).

[4] IBM Corporation. Project debater datasets. https://www.research.ibm.com/haifa/dept/vst/debating_data.shtml (2021 年 7 月 19 日参照).

[5] 坂本俊之. Japanese FakeNews Dataset. https://www.kaggle.com/tanreinama/japanese-fakenews-dataset (2021 年 7 月 19 日参照).

[6] David M Blei. Latent Dirichlet Allocation. p. 30.