7J-05

Web 議論への関連情報推薦のための検索語予測手法の精緻化と関連情報推薦エージェントの試作

木下　良輔[†1]　　白松　俊[‡2]

名古屋工業大学　工学部情報工学科[†1]　　名古屋工業大学　大学院工学研究科[‡2]

1.はじめに

本研究では，公的な社会問題に関する市民参加型議論のために，いつでもどこからでも参加可能な Web 議論を想定したファシリテーション機構の開発を目指す.

Web 議論では，議論内容に精通している人も，全く背景知識を持っていない人も参加可能であるため，持っている情報量や議論の理解度の差によって発言しづらい人が出てきてしまう．そこで，議論内容に関連した情報を参加者に推薦することで，参加者の内容理解や議論進行に役立つと考えた.

先行研究[1,3]では BERT [2]を用いて以下の 2 つのアプローチによる関連情報推薦エージェントが試作された.

(1) 議論内容から検索語を選び出し情報を集める
(2) 収集した情報から推薦する部分を抽出

しかし，有益な情報を推薦できている一方で，関連度の低い情報も推薦してしまっていた．その原因として (1) の手法の検索語の選出に問題があると考え，検索語予測の手法に注目してより実用的な関連情報推薦エージェントの設計を目指す.

2.先行研究

内藤らの研究では，検索語を決める手法としてシンプルな TF-IDF を用いて議論文脈から決定する方法と，BERT を用いた後続発言に含まれる単語の予測による検索語決定手法の 2 つを提案した．しかし，「検索結果中でより多く分類されたクラスの BERT スコア上位の候補が関連情報として有益」という誤った仮定により推薦する情報を選んでいたため，有益ではない情報も推薦してしまうという問題があった．一方安田ら [3] の研究では，ALBERT を用いて直近の議論文章に後続しやすい情報を推薦することで，参加者に有益な情報を推薦できるように改良した．しかし，関連はするが有益ではない情報も推薦してしまう問題や，リアルタイム検索ができない問題があった.

そこで本研究では，以下の 2 つを目的とする.

(1) 関連度の低い情報を選出しない関連情報推薦エージェントを設計する
(2) より有用な検索語を予測できる検索語予測の手法を検討する

なお本研究では，D-Agree 上での議論を想定し，Google Search API を用いて関連情報を検索・抽出する関連情報推薦エージェントを試作する.

3.検索語予測手法の検討

先行研究で問題点となっていた訓練データの数と質の問題は，新たな訓練データを Lancers を で収集することで改善する.

3.1 訓練データ収集

過去に行われた議論の一部分を作業者に見せ，関連情報を検索するための検索語を考えてもらうことで，検索語予測のための訓練データを収集する．作業者に見せる議論の一部分は，以下を含む.

✓対象スレッドの当該時刻における直近 2 投稿
✓対象スレッドのタイトル
✓対象スレッドを含む議論テーマのタイトル

これにより，検索語を含む確率が高い場所を考慮した検索語予測を可能にする.

D-Agree や Slack の過去の議論データから議論テーマ，スレッドタイトル，直近 2 投稿を Lancers 作業者に提示し，検索語データを 250 件の組を 2 セットで合計 500 このデータを収集した．作業者には，実際に検索に使った検索語と検索結果から最も有用そうな部分を記入させた.

3.2 検索語を含む箇所の分析

収集したデータには，検索語が含まれやすい場所に偏りがみられた．それらの割合を計算したものを以下の表 1 に示す.

表 1：検索語が含まれていた場所の割合

	テーマ	スレッド	直近の発言	一つ前の発言
割合	60.0%	70.4%	36.0%	40.8%

表から分かるように今回から情報として与えた，議論テーマ，スレッドタイトルから検索語

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.

に起用されている割合が，直近 2 投稿から起用されている割合よりも二倍近く多いことが分かった．これは議論テーマやスレッドタイトルがそもそも議論内容を要約しているからであると考えられる．また 2 つの発言は検索語にかかわる重要な要因ではあるにもかかわらず，予想していたよりも検索語になる割合が小さかった．これは「話し言葉」であることから，別の表現に言い換えられてしまっていることが原因だと考えられる．実際に，テーマ，スレッド，直近 2 投稿のどこにも含まれていない検索語の約半数の 47.6 %は言い換えの表現であったため，直近 2 投稿に含まれる検索語も十分有用であると考えた．

3.3 検索語への言い換えの分析

どのような場合に「言い換え」がなされているのかを調べた．「言い換え」が起こるパターンが上位の概念への言い換え，会話文を要約する形での言い換え，表現を名詞化する形での言い換えの 3 パターンが見られた．上位の概念への言い換えとは，言い換え前の表現が「回転ずし」であったとき，「外食産業」と言い換えることでより広義の表現にしているような場合である．会話文を要約する形での言い換えとは，会話文の中で最近の若者の選挙の投票状況について話されているときに，「若者　政治参加」のような検索語が選出されていた場合である．表現を名詞化する形での言い換えとは，二つの会話文の中で出てきたものや現象に，対応する呼称が存在するときに行われる言い換えである．例を挙げると，議会に女性を一定の比率で割り当てるべきだという旨の会話が行われているときに，その制度を表す表現である「クオータ制」が検索語になっているような場合である．検索語に「言い換え」が行われている場合，この 3 パターンのいずれかに属しているとして，言い換えが起こる場合の割合を計算したものを表 2 に示す．

表 2：言い換え表現のパターンの割合

	上位の概念	要約	名詞化
言い換えの割合	29.8%	47.4%	22.5%

表より，どのパターンも無視できない割合であるため，これらの言い換えの表現をどのように処理するか今後検討する必要があると考える．

4. GPT-3 による検索語予測

3.2 の検索語を含む箇所と 3.3 の言い換え表現の処理を行うため，本研究では GPT-3[4]による検索語予測を行う．実際の D-Agree の過去の議論データから検索語を予測した例を表 3 に示す．

表 3：検索語予測例

議論テーマ	人口減少
スレッドタイトル	年金問題
2 つ前の発言	いったい何歳まで働けるんですかね
直前の発言	そのうち80とかまでいきそうですね
予測した検索語	年金　引き上げ　対策

スレッドタイトルから「年金」が抜き出され，80 歳という具体的な年齢が「引き上げ」という語に言い換えられていることが分かる．

なお，GPT-3 はランダム性を含む生成アルゴリズムなので，同じ議論データの場合でも毎回異なる検索語が得られる．基本的には図 1 と似た検索語を予測するが，「健康保険」や「転勤」など，議論内容から少し飛躍した検索語を予測する場合もあった．そのため，適切な検索語のみを選ぶ手法を検討する必要がある．

5. まとめと今後の課題

本稿では，関連度の低い情報を選出しない関連情報推薦エージェントを設計するために，検索語予測の手法を検討してきた．その中で，新たな検索手法の導入，データの収集，データの分析などを行った．そして，検索語を含む割合や言い換え表現を処理するために GPT-3 による検索語予測を行った．よって今後は，予測によって得られた検索語による検索で集まったデータの中から，議論に投稿する情報を抽出するモデルの作成を行う．また，推薦した情報が議論参加者にどれだけ参照されたかを比較することによって，推薦した情報が役に取ったかどうかを測る議論実験を行う．

謝辞

本研究は，JST CREST (JPMJCR15E1) と NEDO (JPNP20006) の支援を受けた．

参考文献

[1] 内藤 勝太，白松 俊：Web 議論における BERT を用いた関連情報推薦エージェント，情報処理学会 第 82 回全国大会講演論文誌，5ZG-07 (2020)

[2] Devlin, J. et al : Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805(2018)

[3] 安田 隼太，白松 俊，神谷 晃：Web 議論における関連情報推薦のための発言種別の遷移確立の分析，東海支部連合大会 (2020)

[4] Brown, Tom B., et al：Language models are few-shot learners, arXiv preprint arXiv:2005.14165 (2020).

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.