5P-03

英字新聞を利用した多読支援のための学習者の嗜好に基づく記事推薦の検討

松本 大輝 † 谷本 祐次 †† 安藤 一秋 †
† 香川大学工学部 †† 香川大学大学院工学研究科

1 はじめに

多読とは，ある言語で記述された文章を大量に読むことであり，語彙力や読解速度などの向上が期待できる学習方法である．多読で学習効率を上げるには，教材の語彙が学習者の語彙レベルに適していることや，内容が学習者の嗜好に適していることが重要である．語彙の観点は，未知単語の辞書引きの増加による読解速度の低下を抑制するためである．嗜好の観点は，学習を継続的に行うために必要と考えられる．

英字新聞は，一般教養が身に付くことや，カテゴリが多く学習者の嗜好に適した記事が見つかる可能性が高いことから，教材として多読に適していると考えられる．しかし，既存の英字新聞閲覧サイトでは，各記事の語彙レベルの表示がないことや，分類先のカテゴリの粒度も大きいため，学習者が自身の語彙力や嗜好に適した記事かを判断する情報が不十分である．

本研究では，Web 上で提供されている英字新聞記事に着目し，ニュース記事の選択支援機能を有する多読支援システムの構築を目的とする．本稿では，トピックモデルを用いたカテゴリの細分類，及び学習者の嗜好に基づく記事の推薦手法について検討する．

2 提案システムの概要

本研究で提案するシステムでは，記事選択支援として下記の機能を提供する．

1. 記事カテゴリの細分類
2. 学習者の既知単語割合と嗜好に基づく記事推薦

機能 1 は，トピックモデルを用いて，記事をトピック毎に分類することで実現する．機能 2 は，既知単語割合と嗜好に適したカテゴリの予測結果に基づいて記事をランキングすることで実現する．既知単語割合の予測には，[1] で提案した手法を用いる．嗜好に適したカテゴリの予測は，学習者の閲覧済の記事に対して，機能 1 で作成したトピックモデルを適用することで実現する．具体的には，閲覧済の記事で生起確率の大きいトピックを嗜好に適したトピックとして予測する．

English News Recommendation based on Learner's Preference for Extensive Reading

† Daiki Matsumoto, Kazuaki Ando, Faculty of Engineering, Kagawa University
†† Yuji Tanimoto, Graduate School of Engineering, Kagawa University

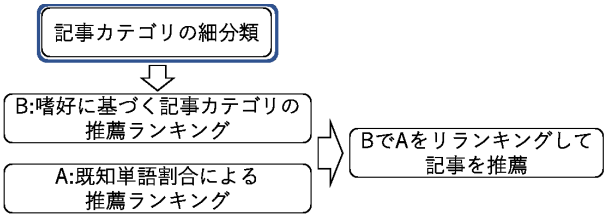

図 1: 記事推薦までの処理の流れ

提案システムでは，これらの機能を用いて，学習者の語彙力と嗜好に適した記事を推薦する．記事推薦までの処理の流れを図 1 に示す．

提案システムでは，語彙力と嗜好の観点で記事をそれぞれランキングする．次に，嗜好に基づくランキング結果を元に既知単語割合によるランキングをリランキングし，この結果を用いて記事を推薦する．本稿では，図 1 の青枠で囲まれた処理について説明する．

3 記事カテゴリの細分類

既存の英字新聞閲覧サイトでは，記事選択支援として“スポーツ”や“政治”などのカテゴリ毎に記事が分類されている．しかし，そのカテゴリの粒度が大きいため，記事をあまり絞り込めず，各カテゴリの中で学習者が記事を再度絞り込むのに時間がかかる．

そこで，本研究では，新聞社でカテゴリ分類された記事を，カテゴリ単位でトピックモデルを適用し，記事を各トピックへ細分類する．また，トピック毎に細分類された記事集合の特徴を学習者が判断しやすくするために，各トピックへ機械的にラベルを付与する．

3.1 トピックモデルを用いた細分類

トピックモデルの 1 つである LDA[2] を用いて，記事をトピック毎に細分類する．LDA は，文書が複数の潜在トピックを持つことを仮定するモデルである．そこで，文書が複数のトピックを持つ場合は，その中で生起確率が最大のトピックに文書を分類し，類似した内容毎に記事を分類できるか検証を行う．

トピック数は，HDP[3] を用いて機械的に決定する．分類対象は，2016/12/1 - 2016/12/31 間の BBC News†の”science”カテゴリの記事からランダムにサンプルした 3 日分の記事 15 件とする．分類単位を 3 日分の記事としたのは，記事数が多い場合に HDP によ

† http://www.bbc.com/news

る最適なトピック数が増加し，分類先が増えすぎないようにするためである．分類結果の一部を表1に示す．

表 1: LDA による分類結果の一部

分類先のトピック番号	分類された記事番号
2	3, 4, 10
3	2

例としてトピック2に着目すると，記事3, 10は宇宙や物理現象に関する内容であったが，記事4は食料環境についての内容であった．記事4の内容は，客観的にトピック3の記事と類似しており，トピック2，3に対する生起確率にあまり差がないことがわかった．分類先のトピック内では話題が統一されていることが望ましいため，上記のようにトピック間の生起確率に差がない場合に妥当なトピックへ分類できる手法を今後検討する必要がある．

3.2 ラベルの付与

学習者が短時間で分類された記事群の特徴を判断できるように，分類先のトピックへラベルを付与する．人手でのラベル付与は高コストであるため，ラベル候補に対して，スコアリング手法 [4] を用いてランキングを行い，トピックに適したラベルを選定し，機械的に付与する．[4] の手法は，トピックとラベルの関連度を，ラベルに対する単語の生起分布と，トピックの持つ単語の生起分布間の KL-divergence により算出する．[4] では，カテゴリ単位ではなく全カテゴリの記事のトピックモデルに対して実験している．そこで，カテゴリ単位の記事のトピックモデルに対して，適切なラベルが付与できるか確認する．

ラベル候補は，3.1 項で分類対象とした BBC の記事 15 件から抽出した全ての 2-gram とする．[4] の手法を用いてラベルをスコアリングしてランキング した結果を表 2 に示す. 表 2 では，トピック 2, 3 でスコア上位の 7 件のラベルを載せている．

表 2: 各トピックでスコアの高い上位 7 件のラベル

順位	トピック 2	トピック 3
1	prof smart	year time
2	year time	prof smart
3	space object	year array
4	space hubble	space dr
5	space time	year gathering
6	year number	space trip
7	time research	space datum

トピック 2 に着目すると，”space object”や”space hubble”のような宇宙や物理現象のラベルが存在している．トピック 2 に分類された記事は，宇宙飛行士の話や重力波といった宇宙や物理現象についての内容を多く含む．そのため，分類された記事の内容を反映したラベルを選定できていると考えられる．また，トピック 3 に着目すると，トピック 2 と類似した宇宙や物理現象のラベルが存在している．しかし，トピック 3 に分類された記事は，エネルギー環境についての内容を多く含むため，選定されたラベルは分類された記事内容を反映できていないといえる．また，トピック 2，3 のどちらにも，”year time”や”year number”など分類された記事内容を反映していないラベルがいくつか選定されている．

表 2 のように，トピック毎のラベルが類似している場合，学習者がトピック間の特徴の差異を判断するのが困難になる．これについては，集合要素間の多様性を評価できる指標 MMR を用いて，多様性の度合いが小さいラベルを排除することで改善できると考えられる．また，多様性の度合いが小さいラベルを排除することにより，分類された記事内容をより反映したラベルが選定されやすくなるため，記事内容を反映しないラベルが選定される問題も改善できると考える．

4 おわりに

本稿では，多読支援システムと，記事推薦の前処理であるトピックモデルを用いたカテゴリの細分類の手法について提案した．カテゴリの細分類では，記事内で生起確率が最大のトピックへ分類した結果，類似した話題を持つ記事集合を生成できた．しかし，一部の記事集合では話題に統一性がないことも確認した．また，ラベルの自動付与では，ある程度トピックに分類された記事内容を反映したラベルを選定することができたが，記事内容をあまり反映しないラベルも選定されていた. 今後は，細分類とラベル付与の手法の改良を行い，次のステップである記事推薦を行う．

参考文献

[1] 谷本他，”多読支援のための単語の分散表現を利用した語彙予測”，情報処理学会第 79 回全国大会講演論文集，pp.887-889, 2017.

[2] D.M.Blei, et al.,“ Latent Dirichlet Allocation ”, Journal of Machine Learning Research, Vol.3, pp.993-1022, 2003.

[3] Y.W.Teh, et al., ”Sharing clusters among related groups:Hierarchical Dirichlet processes”, Advances in neural information processing systems, pp.1385-1392, 2005.

[4] Q.Mei, et al.,“Automatic labeling of multinomial topic models ”, In SIGKDD, pp.490-493, 2007.