

読者の興味を惹くかどうかと本文との整合性を考慮した記事タイトル作成支援

古田 朋也[†] 鈴木 優^{††}

[†] 岐阜大学大学院自然科学技術研究科知能理工学専攻 〒501-1193 岐阜県岐阜市柳戸1番1

^{††} 岐阜大学工学部電気電子・情報工学科 〒501-1193 岐阜県岐阜市柳戸1番1

E-mail: †a4525067@edu.gifu-u.ac.jp, ††ysuzuki@gifu-u.ac.jp

あらまし インターネット上のニュース記事にタイトルを付ける際に考慮することとして、読者の興味を惹くかどうかと本文との整合性が取れているかどうかがある。本研究では、これらのバランスがうまく取れた良質なタイトルを付ける支援を行うことを目的としている。その一環として、記事タイトルに対して読者の興味を惹くかどうかの判定を行う。記事の本文中には、主題に関係する単語であるにもかかわらずタイトル中では使用されていない単語が存在している。その中にはタイトルに使用することによって読者の記事に対する興味を失せさせてしまう単語もあり、読者の興味を惹くタイトルかどうかに影響を与えていると考えられる。そのため、このような単語を入力として利用することによって、タイトルのみを入力とした場合と比較して読者の興味を惹くタイトルかどうかの判定精度が向上するのではないかと考えた。そこで、記事タイトルと本文中でのみ使用されている単語を入力としたニューラルネットワークモデルの分類器を構築する。また、読者の興味を惹かないタイトルのクラスタリング結果を利用することによって、興味を惹かないと判定された際の理由を提示する。上記手法を複数人の投票によって作成したデータセットにより実装し、評価実験を行った。その結果より、本文中でのみ使用されている単語が読者の興味を惹くかどうかの判定に有効に作用していること、読者の興味を惹かないタイトルをクラスタリングすることによって興味を惹かない要因に基づいた差異がクラス間に現れることを確認した。

キーワード 記事タイトル、ニュース記事、テキスト分類、機械学習、タイトル作成支援

1 はじめに

現在、インターネット上にはニュース記事が多数掲載されており、様々なタイトルが付けられている。インターネット上のニュース記事のタイトルを付ける際に考慮することとして、読者の興味を惹くかどうかと本文との整合性が取れているかどうかがある。例えば、あるショッピングサイト A で明日から 1 か月間ベッドが半額で販売されるという情報を伝える内容の記事に対しては、以下のようなタイトルが考えられる。

Title 1 「A にて特別大特価!! 対象商品は? 期間はいつまで?」

Title 2 「A にてベッド半額セール 明日から 1 か月間」

Title 1 は読者に記事を読みたいと感じさせるタイトルではあるが、誇張しすぎており、いろんな商品がセール対象であると誤認してしまう可能性がある。一方、Title 2 は整合性は完全に保たれているが、読者の知りたい情報がすべてタイトルに記載されているため、記事を読む必要性が薄れてしまう。読者の興味を惹くタイトルを付けることは、記事の閲覧数の増加が見込めるため、著者側にとってのメリットが大きいといえる。しかし、興味を惹くタイトルを付けることを追求しすぎると、誇張され過ぎたタイトルになってしまい、本文との整合性が失われてしまう。そのようなタイトルでは記事内容の誤認が発生しやすくなるため、読者側にとってはデメリットの方が大きく

なってしまう。一方、本文との整合性が取れているタイトルを付けることは、記事内容の誤認が発生しなくなるため、読者側にとってのメリットが大きいといえる。しかし、本文との整合性について追求しすぎると、タイトル中に本文中の情報の大部分が記載されることになる。そのようなタイトルでは記事本文を読む必要性が薄れてしまい閲覧数の増加が見込めなくなるため、著者側にとってはデメリットの方が大きくなってしまう。以上より、読者の興味を惹くかどうかと本文との整合性が取れているかどうかのどちらか一方について追及するのではなく、両方のバランスをうまく取ることによって著者側にとっても読者側にとってもメリットの大きいタイトルとなると考える。そこで本研究ではインターネット上のニュース記事に対して、これらのバランスがうまく取れた良質なタイトルを作成する支援を行う。読者の興味を惹くかどうかと本文との整合性が取れているかどうかについての判定をそれぞれ行い、両方のバランスを取ることを目指す。

本稿では、ニュース記事のタイトルに対して読者の興味を惹くかどうかの判定を行う。記事の本文中でのみ使用されている単語の中にはタイトルに使用することによって読者の興味を失せさせてしまったり、読者を限定してしまったりする可能性がある単語も存在している。これらの単語の存在も読者の興味を惹くかどうかに影響を及ぼしていると考えられる。しかし、タイトルのみを入力とした場合、本文中でのみ使用されている単語の情報は失われてしまい、これらの単語の及ぼす影響が判定

時に無視されることになる。そこで、タイトルと本文中でのみ使用されている単語を入力としたニューラルネットワークモデルの分類器を構築することによって読者の興味を惹くタイトルかどうかの判定を行う。本文中でのみ使用されている単語を入力として利用することによって、タイトルのみを入力とした場合と比較して読者の興味を惹くタイトルかどうかの判定精度が向上するのではないかと考えた。

また、読者の興味を惹くかどうかの判定結果が「興味を惹かない」となった場合は、その判定理由を提示する。それによって判定の説得力が増し、タイトルの改善を行う際に利用できる。読者の興味を惹かない要因は多数考えられるが、興味を惹かないタイトルをクラスタリングすることによって興味を惹かない要因に基づいた差異がクラスタ間に現れるのではないかと考えた。そこで、読者の興味を惹かないタイトル群に対してクラスタリングを行う。その後、クラスタリング結果からクラスタ内のタイトルの特徴をそれぞれ人手にて解釈する。判定理由を提示する際には、読者の興味を惹かないと判定されたタイトルが、どのクラスタに属するかを判定し、解釈したそのクラスタ内タイトルの特徴を提示する。

提案手法の評価実験として、提示手法による分類モデルの性能評価と判定理由提示手法の有効性評価を行った。タイトルのみを入力としたベースラインとタイトルと本文中でのみ使用されている単語を入力とした提案手法の分類性能を比較した。その結果、Accuracy においてはほぼ同程度の数値であったが、F 値においては提案手法の方が高い数値が得られた。また、提案手法による分類モデルはベースラインと比較して汎化性能を保った状態で学習が行われていた。これらの結果より、提案手法にて入力に利用した本文中でのみ使用されている単語が読者の興味を惹くかどうかの判定に有効に作用していることを確認した。また、興味を惹かないタイトルのクラスタリングを行い、その結果とデータセット作成時に集計した興味を惹かない要因それぞれが占める割合を照らし合わせた。その結果、興味を惹かないタイトルをクラスタリングすることによって、興味を惹かない要因に基づいた差異がクラスタ間に現れていることを確認した。

本稿では、興味を惹かないタイトルのクラスタリングと人手による各クラスタの解釈を基にした判定理由の提示手法を提案した。本手法では人手によって各クラスタの解釈を行うため、作業者の主観による影響が少なからず現れる。タイトルの改善に利用する場合、客観性の強い判定理由を提示できた方が有用性が高いと考える。そのため今後、できる限り自動で判定理由の提示を行うことができる手法を検討していく必要がある。

本論文における貢献は以下の通りである。

- 本文中でのみ使用されている単語が読者の興味を惹くタイトルかどうかの判定において有効に作用することを確認した。
- 読者の興味を惹かないタイトルをクラスタリングすることによって、興味を惹かない要因に基づいた差異がクラスタ間に現れることを確認した。

2 関連研究

タイトル作成支援に関連する研究は、既にいくつか行われている。長安ら [1] は、文書の一文概要生成手法を提案している。文書中の名詞から重要語分類を行う SVM と IDF の両方のスコアを用いて重要語を抽出している。そして、抽出した重要語をタイトルパタンに当てはめることによって文書の一文概要を生成している。金内ら [2] らは、投稿型レシピサイトにおける料理タイトルの自動生成手法を提案している。レシピのレビュー情報を分析することによってレシピの特徴や目的を抽出している。抽出したレシピの特徴や目的をタイトル文型に当てはめることによって料理タイトルを生成している。

上記の先行研究はタイトル作成の際に本文との整合性に焦点を当て、文書やレシピの特徴を正確に一文概要やタイトルで表現している。一方、読者の興味を惹くかどうかに焦点を当てた研究もいくつか行われている。西原ら [3] らは、専門外の人の興味を惹く研究発表タイトルの推定手法を提案している。タイトル中の単語の分かりやすさ、意外性、組み合わせに関する評価を Web 検索結果を利用して行い、タイトルにスコアを付与している。Stokowiec ら [4] は、動画や記事のタイトルから人気予測を行う手法を提案している。月間閲覧数の中央値を上回るものを「人気ある」として、タイトルを入力とした Bidirectional LSTM を用いた二値分類タスクとして扱っている。

このように先行研究では、読者の興味を惹くかどうかと本文との整合性が取れているかどうかのどちらか一方のみに焦点を当てている研究は多く行われている。タイトルを付与する対象やその目的によってどちらに焦点を当てるべきかは異なるが、我々が調べた限りでは両方に焦点を当てている研究は行われていない。インターネット上のニュース記事のタイトルを作成する際には、読者の興味を惹くかどうかと本文との整合性が取れているかどうかの両方を考慮すべきであると考えられる。読者の興味を惹くかどうかのみを追求していくと本文との整合性が取れていないタイトルになっていく。一方、本文との整合性が取れているかどうかのみを追求していくと読者の興味を惹くことができないタイトルになっていく。そのため、読者の興味を惹くかどうかと本文との整合性が取れているかどうかのどちらか一方を追求するのではなく、両方のバランスをうまく取ることが大切になってくる。そこで、本研究では読者の興味を惹くかどうかと本文との整合性が取れているかどうかの両方に焦点を当てる。これら両方のバランスがうまく取れた良質なタイトルを付ける支援を行うことを目的とする。

その一環として本稿では、ニュース記事のタイトルに対して読者の興味を惹くかどうかの判定手法を提案する。上記の他にも興味を惹くかどうかに焦点を当てた研究はいくつか行われている [5] [6] [7]。先行研究では、読者の興味を惹くタイトルを分析して頻出する単語や表現パターンを抽出し、それを基に作成した辞書を使用して判定を行う手法が多い。しかし、このような手法では辞書内に存在する頻出単語や表現パターンが使用されていないと興味を惹くと判定されない可能性がある。一方、

表 1 使用しないようにしている本文中単語がある記事タイトルの例

タイトル	記事の内容	本文中単語を使用しないことによる効果
「中途入社はつらいと思うとき」 2 位は「期待が大きい」、1 位は?	中途入社の男女を対象にした意識調査についての記事	1 位を伏せることにより、興味を失せさせないようにする
ロフトを寝室として使ってはいけないう納得の理由	ロフトを寝室として使ってはいけないう風水上の理由について述べた記事	風水上の理由であることを伏せることにより、読者を限定してしまうことを防ぐ

本稿の提案手法ではニューラルネットワークモデルの分類器を構築することによって判定を行う。ニューラルネットワークを利用した判定を行うことによって、教師データと完全に一致した頻出単語や表現パターンが使用されていないタイトルに対しても興味を惹くと判定を行うことができ、汎化性能が高くなると考える。

また、先行研究ではタイトルのみを入力として興味を惹くかどうかの判定を行う手法が多い。しかし、本文中でのみ使用されている単語の中にはタイトルに使用することによって読書の興味を失せさせてしまう可能性がある単語も存在している。これらの単語の存在も読者の興味を惹くかどうかに影響を及ぼしていると考えられるが、タイトルのみを入力とした場合、この情報は失われてしまう。そこで本稿の提案手法では、タイトルと本文中でのみ使用されている単語を入力としたニューラルネットワークモデルの分類器を構築することによって読者の興味を惹くかどうかの判定を行う。

3 提案手法

本研究では、読者の興味を惹くかどうかと本文との整合性を考慮した良質なタイトルを付ける支援を行うことを目的としている。その一環として本稿では、記事タイトルに対して読者の興味を惹くかどうかの判定を行う。

記事タイトルの中には、閲覧数を増やすために記事の主題に関係する本文中単語を使用しないで作成された表 1 のようなタイトルが存在している。記事の主題に関係する本文中単語の中にはタイトルに使用することによって読者の興味を失せさせてしまったり、読者を限定してしまうものもある。そのため、これらの単語は読者の興味を惹くタイトルかどうかに影響を与えていると考えられる。このような単語を入力として利用することによって、読者の興味を惹くタイトルかどうかの判定精度がタイトルのみを入力とした場合と比較して向上するのではないかと考えた。

提案手法による読者の興味を惹くタイトルかどうかの判定の概要を図 1 に示す。提案手法によるモデル構築は Step.A、モデル構築後の読者の興味を惹くかどうかの判定は Step.B の流れで行う。それぞれの詳細については、括弧内に示す節にて述べる。

- Step A-1 記事タイトルに対して興味を惹かれるかどうかのアンケートを取り、データセットを作成する。(3.1 節)
- A-2 作成したデータセットにて、ニューラルネットワークモデルの分類器の構築と判定理由提示のための

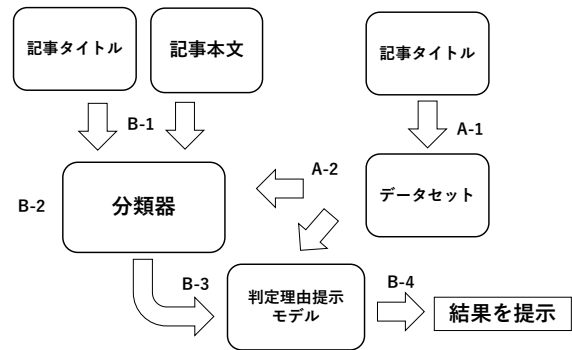


図 1 提案手法の概要

クラスタリングを行う。(3.2 節, 3.4 節)

- Step B-1 読者の興味を惹くかどうかの判定を行う記事タイトルとその記事の本文を入力する。(3.3 節)
- B-2 構築した分類器を用いて読者の興味を惹くかどうかの判定を行う。(3.3 節)
- B-3 読者の興味を惹かないと判定されたタイトルについては、興味を惹かない理由の推定を行う。(3.4 節)
- B-4 分類器による判定結果と、推定した判定理由を提示する。

3.1 データセット作成

本手法では、与えられた記事タイトルに対して読者の興味を惹くかどうかを機械学習を利用して判定する。そのため、訓練データとして読者の興味を惹く記事タイトルを大量に用意して、分類器の訓練を行う必要がある。しかし、興味を惹かれるかどうかの基準は人によって様々であるため、一人だけの意見を用いて作成したデータセットでは汎化性能が低くなってしまう。そこで、複数人の投票結果を集計してラベルを付与することによってデータセットを作成する。

記事タイトルと本文の取得には News API¹を使用した。News API にて Sports, Business, Entertainment のカテゴリの日本語記事を指定してタイトルと本文を取得した。なお、News API にて取得できる記事の本文は冒頭の 200 文字程度となっている。

取得した記事タイトルに対して読者の興味を惹くかどうかのラベルと興味を惹かない要因それぞれが占める割合を付与する。記事タイトル一つにつき、複数人に対してアンケートを取

1 : <https://newsapi.org/>

1.記事を読みたいと感じたかどうか

感じた ⇨ 質問2へ 感じなかった ⇨ 質問3へ

2.読みたいと感じた理由

興味のあるジャンルだった 興味のないジャンルだったが
内容が気になった

その他

3.読みたいと感じなかった理由

興味のないジャンルだった タイトル内の情報だけで十分、
これ以上内容が気にならなかった

タイトルから記事内容の
把握が困難だった

その他

図2 アンケートの質問と選択肢

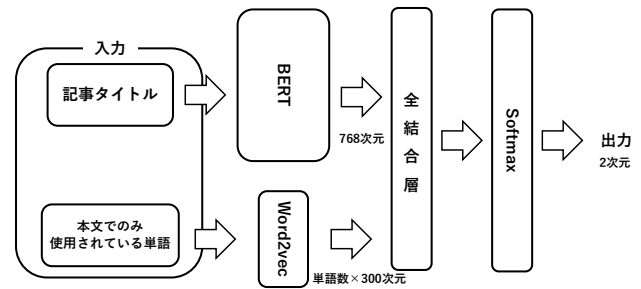


図3 構築する分類モデルの概要

りデータセットを作成していく。アンケートでは、作業者に記事タイトルを提示して、記事本文を読みたいと感じたかどうかとそうに感じた理由を質問する。アンケートにて質問した内容と選択肢を図2に示す。アンケート結果を集計して、読みたいと感じた人が半数を超えているタイトルにラベル「興味を惹く」を、読みたいと感じなかった人が半数を超えているタイトルにラベル「興味を惹かない」を付与する。また、ラベル「興味を惹かない」が付与されたデータに対しては興味を惹かない要因それぞれが占める割合も付与する。例えば、あるタイトルにおいて記事を読みたいと感じなかった人が9人、そのうちの5人が判定理由として「興味のないジャンルだった」を選択していた場合、興味を惹かない要因「興味のないジャンルだった」が占める割合は $5/9 \approx 0.55$ となる。

3.2 モデル構築

BERT [8] はあらゆる自然言語処理タスクにおいて、汎用性の高いモデルである。モデルの構造を修正せずとも転移学習することによって、様々なタスクに応用でき、高い精度を発揮している。読者の興味を惹くタイトルかどうかの判定にはタイトル内の単語の語順も少なからず影響していると考えられる。そこで、語順を考慮しているBERTを利用したニューラルネットワークモデルの分類器により、読者の興味を惹くタイトルかどうかの判定を行う。

本手法にて構築する分類モデルの概要を図3に示す。まず、入力として与えられた記事タイトルと本文中でのみ使用されている単語に対してそれぞれ処理を行う。記事タイトルに対してはBERTモデルにて特徴量抽出を行い、本文中でのみ使用されている単語に対してはWord2vecにてベクトル化を行う。そして、これらを全結合層に入力し、Softmax関数を介して2次元のベクトルで出力する。このような分類モデルを3.1節に従って作成したデータセットを用いて構築する。

本稿では、訓練済みBERTモデルとして、東北大学の乾・鈴木研究室の訓練済み日本語BERTモデル²を使用した。訓練済みモデルは日本語版Wikipediaにて、事前学習が行われて

おり、語彙数は32,000となっている。訓練時にはBERTモデルのパラメータは最終層のみ更新するように設定してファインチューニングを行った。また、Word2vecのモデルは、日本語版Wikipediaの記事全文で訓練を行ったCBOW、次元数300のモデルを利用した。このWord2vecモデルをデータセット内の記事本文でファインチューニングしてベクトル化に使用した。

3.3 モデルによる分類

構築したモデルによる分類の流れを示す。まず、与えられた記事タイトルと本文をそれぞれ形態素解析する。それらを比較して、本文中でのみ使用されている単語を抽出する。抽出した本文中でのみ使用されている単語のうち、出現回数が多い名詞を入力として使用する。本文での出現回数が多いもののほど記事の主題に関係する単語であるといえる。主題に関係する単語であるにもかかわらずタイトルに使用されていないのは、タイトルに使用すると読者の興味を惹くかどうかに影響を及ぼすためであると考えられる。そのため、本文中でのみ使用されている単語のうち、出現回数が多い名詞を使用することにした。次に、3.2節のように構築した分類モデルに記事タイトルと抽出した本文中でのみ使用されている単語を入力する。そして、出力として得られた2次元のベクトルのうち、最大値となる要素に対応したラベルを選択することによって、入力された記事タイトルが読者の興味を惹くかどうかの判定結果を決定する。

3.4 判定理由の提示

読者の興味を惹くかどうかの判定結果が「興味を惹かない」となった場合は、その判定理由を提示する。それによって判定の説得力が増し、タイトルの改善を行う際に利用できる。そこで、判定理由の提示を行うことを試みる。読者の興味を惹かないタイトルには、相応の要因があると考えられるが、その要因は多数考えられ、要因同士の組み合わせも様々である。しかし、興味を惹かない要因として同様のものが存在するタイトル同士には共通点があると考えられる。そのため、興味を惹かないタイトルをクラスタリングすることによって興味を惹かない要因に基づいた差異がクラスタ間に現れ、判定理由の提示に利用できるのではないかと考えた。

² : <https://github.com/cl-tohoku/bert-japanese>

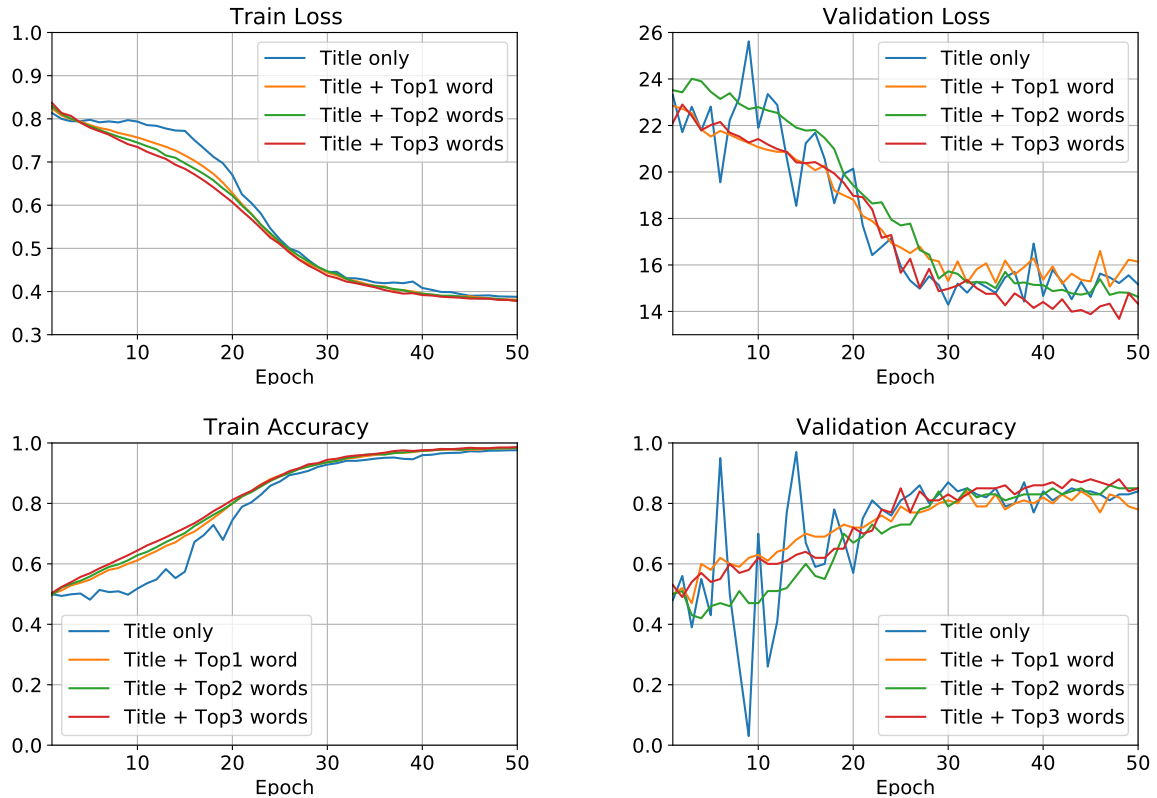


図 4 Leave-one-out 交差検証における学習曲線

判定理由の提示について示す。3.1 節にて作成したデータセット内のラベル「興味を惹かない」が付与された記事タイトル群に対してクラスタリングを行う。まず、3.2 節にてファインチューニングした BERT モデルにて記事タイトルの特徴量を抽出する。そして、抽出した 768 次元のベクトルを使用して興味を惹かないタイトルのクラスタリングを行う。次に、データセット作成時に付与した興味を惹かない要因それぞれが占める割合を参考にしつつ、それぞれのクラス内の特徴を手で解釈して決定する。それによって、教師あり学習の分類では事前に発見できなかった興味を惹かない要因同士の組み合わせや比率などが発見できるのではないかと考えた。判定理由を提示する際には、読者の興味を惹かないと判定されたタイトルが属するクラスを判定し、属するクラスの特徴を提示するという手順を踏む。

4 評価実験

提案手法について、以下二つの評価実験を行った。

実験 1 提案手法による分類モデルの性能評価

目的 1 入力に本文中でのみ使用されている単語を利用することによって分類性能が変動するかどうかを確認する。

目的 2 本文中でのみ使用されている単語の入力数によって分類性能が変動するかどうかを確認する。

実験 2 判定理由の提示手法の有効性の評価

目的 1 読者の興味を惹かないタイトルをクラスタリングすることによって興味を惹かない要因に基づいた差異がクラス間に現れるかどうかを確認する。

News API にて取得した記事タイトル 100 件に対して、3.1 節にて示したアンケートを取り、データセットを作成した。本稿では、タイトル一つにつき 11 人に対してアンケートを実施した。アンケートを集計した結果、興味を惹くタイトルが 14 件、興味を惹かないタイトルが 86 件となった。

4.1 実験 1

本節では、提案手法による分類モデルの性能評価を評価値に従って行う。本文中でのみ使用されている単語を入力として利用することによる影響の確認と、本文中でのみ使用されている単語の入力数による分類性能の変動具合の確認を目的とする。

4.1.1 実験内容

タイトルのみを入力としたベースラインとタイトルと本文中でのみ使用されている単語を入力とした提案手法の分類性能を比較する。提案手法においては、本文中でのみ使用されている単語として出現回数上位 1 つを用いた場合、2 つを用いた場合、3 つを用いた場合の分類性能をそれぞれ算出する。作成した前述のデータセットを用いて、上記の条件それぞれに対して Leave-one-out 交差検証を実行する。その際、ラベル「興味を惹く」が付与されたデータを複製することによって、訓練用データ内のラベル間の偏りを解消した。分類モデルの学習はすべて 50 エポックにて実行し、それぞれの学習曲線は図 4 のよ

うになった。

4.1.2 結果・考察

Leave-one-out 交差検証のそれぞれの結果を表 2 に示す。Recall, Precision, F 値については、それぞれ興味を惹く、興味を惹かないを正例とした場合の数値を示している。

Accuracy に注目した場合、ベースラインでは 0.84、提案手法では上位 2 つを用いた場合と上位 3 つを用いた場合にて 0.85 という数値が出ている。一方、興味を惹くを正例とした F 値に注目した場合、ベースラインでは 0.20、提案手法では上位 3 つを用いた場合にて 0.40 という結果が出ている。この結果から、Accuracy においてはベースラインと提案手法の間に大きな差は見られないが、F 値においては提案手法の方が高い数値が出ており、提案手法による分類モデルの方が高い分類性能を発揮していることを確認した。また、図 4 の学習曲線を確認してみると、ベースラインと比較して提案手法の方が安定した曲線になっていた。この結果より、提案手法による分類モデルはベースラインと比較して汎化性能を保った状態で学習ができていると考えられる。以上より、提案手法にて入力に利用した本文中でのみ使用されている単語が読者の興味を惹くかどうかの判定に有効に作用していると考ええる。

提案手法において、本文中でのみ使用されている単語の入力数の違いによる分類性能の変動具合の確認をする。上位 1 つを用いた場合は Accuracy、興味を惹くを正例とした F 値のどちらにおいても最も劣る結果となった。上位 2 つを用いた場合と上位 3 つを用いた場合を比較すると、Accuracy は同程度の値が出ていた。しかし、興味を惹くを正例とした F 値においては、上位 3 つを用いた場合の方が高い値となった。以上の結果より、上位 3 つを用いた場合が最も優れた分類性能を発揮していることが確認できた。

本実験にて、提案手法による分類モデルがベースラインよりも高い分類性能を発揮することを確認した。しかし、本稿ではタイトルの件数が 100 件と小規模なデータセットでの実験と

なっているため、データセットの総数を増やした際にも提案手法による分類モデルが高い分類性能を発揮するかどうかを確認する必要がある。

また、本稿で使用したデータセットは News API にて取得した記事タイトルのみを用いて作成している。そのため、記事内容は同じであるがタイトルが異なっているというデータ対が少ない。そこで、タイトルの付け方のみを変えたデータを作成し、追加することによってデータセットを拡張を試みる。それによって、判定時にタイトルの表現方法や本文中でのみ使用されている単語の影響をより考慮できるようになり、有用なタイトル作成支援につながるのではないかと考える。

4.2 実験 2

本節では、判定理由の提示手法の有効性の評価を行う。興味を惹かないタイトルのクラスタリングを行い、結果を確認・分析することによって提案手法が判定理由の提示に有効かどうかを評価する。

4.2.1 実験内容

4.1.2 節の結果から、入力に出現回数上位 3 つの本文中でのみ使用されている単語を利用した提案手法によるモデルを作成した。作成したデータセットすべてを訓練データとして用いて、50 エポックの訓練を実行することにより分類モデルを構築した。その際、ラベル「興味を惹く」が付与されたデータを複製することによって、データセット内のラベル間の偏りを解消した。構築した分類モデルを用いて 3.4 節に従って、データセット内の興味を惹かないタイトル 86 件に対してクラスタリングを行った。本稿では、クラスタリング手法としてクラスタ数を 5 に設定した k -means 法を、 k -means 法を適用する際の距離関数としてコサイン距離を採用して実験を行った。そして、興味を惹かない要因に基づいた差異がクラスタ間に現れているかどうかをクラスタリング結果とデータセット作成時に集計した興味を惹かない要因それぞれが占める割合を照らし合わせることによって確認する。

4.2.2 結果・考察

k -means 法による興味を惹かないタイトルのクラスタリング結果を図 5 に示す。ここでは主成分分析にて 2 次元に次元削減を行い、クラスタリング結果をプロットしている。また、興味を惹かないタイトルに付与されている興味を惹かない要因それぞれが占める割合のクラスタ内平均を計算した。興味を惹かない各要因が占める割合のクラスタ内平均を表 3 に示す。

表 3 の結果より、読者の興味を惹かない要因の内訳にクラスタごとの特徴が現れていることがわかる。Cluster0 において

表 2 Leave-one-out 交差検証の結果

	評価値		
	興味を惹く	興味を惹かない	
Title only	Accuracy	0.840	
	Recall	0.142	0.953
	Precision	0.333	0.872
	F 値	0.200	0.911
Title + Top1 word	Accuracy	0.780	
	Recall	0.142	0.883
	Precision	0.166	0.863
	F 値	0.153	0.873
Title + Top2 words	Accuracy	0.850	
	Recall	0.142	0.965
	Precision	0.400	0.873
	F 値	0.210	0.917
Title + Top3 words	Accuracy	0.850	
	Recall	0.357	0.930
	Precision	0.454	0.898
	F 値	0.400	0.914

表 3 興味を惹かない要因の内訳 (クラスタ内平均)

	興味がない	気にならない	記事内容把握困難	その他
Cluster0	84.4%	15.6%	0.0%	0.0%
Cluster1	64.0%	30.3%	5.7%	0.0%
Cluster2	79.7%	14.5%	5.8%	0.0%
Cluster3	80.0%	14.7%	4.2%	1.1%
Cluster4	78.2%	8.7%	10.9%	2.2%

表 4 タイトルを付ける際に意識したほうが良いとされる心理効果の例

心理効果	概要
ツァイガルニク効果	未完了のことは完了されたことよりも緊張感が持続しやすく、記憶に残りやすい心理現象。すべての情報をタイトルに記載するのではなく、意図的に情報を欠落させることによって、興味を惹きつける。
カリギュラ効果	禁止されるほどやってみたくなる心理現象。「絶対に〇〇しないでください」のように行動を制限するような表現を含めることによって興味を惹きつける。
ハロー効果	ある対象を評価をするときに顕著な特徴に引きずられて他の特徴についての評価が歪められる心理現象。認知バイアスの一種。「〇〇氏推薦!!」のように有名人などの良いイメージを利用して興味を惹きつける。
プロスペクト理論 (損失回避)	得を求めるよりも損を避ける人間の心理傾向。「ハイリスク・ハイリターン」よりも「ローリスク・ローリターン」を選択する。不安を煽る感じの表現を含めることによって記事を読む必要性を強く感じさせる。

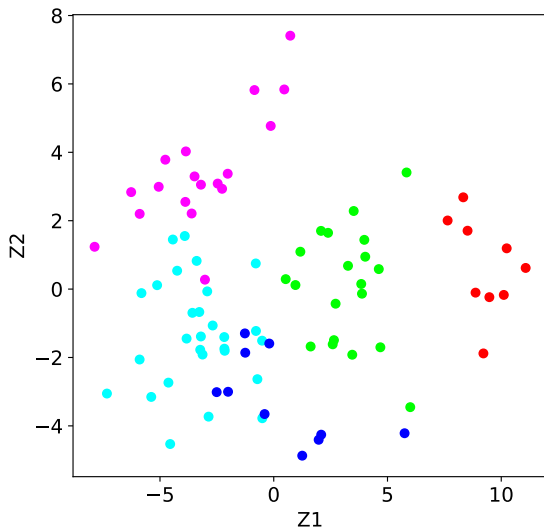


図 5 k-means による興味を惹かないタイトルのクラスタリング結果

は「タイトルから記事内容の把握が困難」の項目が唯一 0.0% となっており、「興味のないジャンル」の項目が他のクラスタよりも高い割合を占めていた。この結果より Cluster0 に分類されたタイトルは、記事内容が把握しやすいタイトルである一方、記事のジャンルが読者の興味の有無に左右されやすい傾向があると考えられる。Cluster1 においては「タイトルのみで十分で記事内容は気にならない」の項目が 30.3% となっており、他のクラスタのおよそ 2 倍の割合を占めていた。この結果より Cluster1 に分類されたタイトルは、タイトル内に情報を詰め込み過ぎて、記事本文を読む必要性を感じなくさせてしまう傾向があると考えられる。Cluster4 においては「タイトルから記事内容の把握が困難」の項目が 10.9% となっており、他のクラスタのおよそ 2 倍の割合を占めていた。この結果より Cluster4 に分類されたタイトルは、記事内容を把握するための最低限の情報がタイトル内に不足しており、記事内容が読者に伝わらない傾向があると考えられる。

上記のように、クラスタリング結果とデータセット作成時に集計した興味を惹かない要因それぞれが占める割合を照らし合わせてみることによって、読者の興味のあるジャンルかどうかの影響が大きいタイトルが分類されたクラスタ、記事本文を読む必要性を感じなくさせてしまうタイトルが分類されたクラス

タ、記事内容が読者に伝わりにくいタイトルが分類されたクラスタなどに分けられていることを確認した。以上より、興味を惹かないタイトルをクラスタリングすることによって、興味を惹かない要因に基づいた差異がクラスタ間に現れることを確認した。

本実験では、作成したデータセット内の興味を惹かないタイトル 86 件に対してクラスタリングを行うことによって、興味を惹かない要因に基づいた差異がクラスタ間に現れることを確認した。しかし、興味を惹かないタイトルのサンプル数が増加した際にも同様に差異が現れるかどうかは不明である。そのため、データセットの総数を増やした場合の実験も実施する必要がある。本手法では、読者の興味を惹くかどうかの判定を行う分類モデルを用いて記事タイトルの特徴量の抽出を行う。そして、抽出した特徴量を用いて興味を惹かないタイトルのクラスタリングを行った。そのため、特徴量抽出に用いる分類モデルの精度がこのクラスタリングの結果にも影響していると考えられる。データセットの総数を増やすことによって記事タイトルのバリエーションも増えるため、分類モデルの汎化性能の向上が期待できる。それに伴い、クラスタリングを行った際のクラスタ間における興味を惹かない要因に基づいた差異がより顕著に現れるようになるのではないかと考える。

データセット総数を増やすことに加えて、データセット作成時に質問する判定理由の選択肢や投票方法について見直しを検討するべきであると考え。本稿で作成したデータセットでは興味を惹かない要因として、興味のあるジャンルかどうか、記事内容が気になるかどうか、記事内容がタイトルから把握できるかどうかの三つの要因しか考えていない。タイトルを付ける際に意識したほうが良いとされる表 4 のような心理効果がある。これらの心理効果を考慮して興味を惹かない要因を設定し直すことによって、タイトルの改善を行う際に修正の方向性を示しやすくなるのではないかと考えた。また、本稿では判定理由を質問する際に三つの要因のうち、どれか一つを選択する方式をとった。この方式では、複数の選択肢に当てはまった場合でも一つの要因しか投票に反映されない。そこで各要因について、それぞれ二択の質問形式をとることによってデータを作成する。これによって、各要因についての良し悪しを各タイトルに付与できるようになり、その良し悪しに基づいた判定理由の提示が可能になるのではないかと考える。以上のように、データセット作成時に質問する判定理由の選択肢や投票方法につい

て見直しを行うことによって、より有用性が高い判定理由の提示につながると考える。

また、本稿で提案した判定理由の提示手法では各クラスタの興味を惹かない要因の解釈を人手で行う必要がある。人手による解釈を行うことによって、作業者の主観による影響が少なからず現れる。読者の興味を惹くタイトルへの改善を行うためには、客観性の強い判定理由を提示できた方が好ましい。そのため今後、できる限り自動で判定理由の提示を行うことができる手法を検討する必要がある。

5 おわりに

本研究では読者の興味を惹くかどうかと本文との整合性を考慮した良質なタイトルを付ける支援を行うことを目的としている。その一環として本稿では、記事タイトルに対して読者の興味を惹くかどうかの判定手法を提案した。記事タイトルの中には、閲覧数を増やすために記事の主題に関係する本文中単語を使用しないで作成されたタイトルが存在している。記事の主題に関係する本文中単語の中にはタイトルに使用することによって読者の興味を失せさせてしまったり、読者を限定してしまうものもあり、読者の興味を惹くタイトルかどうかの判定に影響を与えていると考えた。そこで、本文中でのみ使用されている単語を入力として利用することによって読者の興味を惹くタイトルかどうかの判定を行うことにした。訓練データとして読者の興味を惹くタイトルが必要となるが、興味を惹かれるかどうかの基準は人によって様々であるため、一人だけの意見を用いて作成したデータセットでは汎化性能が低くなってしまう。そこで、複数人の投票結果を集計してラベルを付与することによってデータセットを作成した。読者の興味を惹くタイトルかどうかの判定にはタイトル内の単語の語順も少なからず影響していると考え、作成したデータセットによって BERT を利用したニューラルネットワークモデルの分類器を構築した。また、興味を惹かないタイトルのクラスタリングによって興味を惹かない要因に基づいた分類を行い、読者の興味を惹かないと判定された場合の判定理由の提示に利用する。そして、上記手法についての評価実験を行った。

タイトルのみを入力としたベースラインとタイトルと本文中でのみ使用されている単語を入力とした提案手法の分類性能を比較した。その結果、Accuracy においてはベースラインと提案手法のどちらも同程度の数値であったが、F 値においては提案手法の方が高い数値が得られた。また、提案手法による分類モデルはベースラインと比較して汎化性能を保った状態での学習が行われていた。これらの結果より、提案手法にて入力に利用した本文中でのみ使用されている単語が読者の興味を惹くかどうかの判定に有効に作用していることが明らかとなった。また、興味を惹かないタイトルをクラスタリングすることによって、興味を惹かない要因に基づいた差異がクラスタ間に現れることを確認した。

しかし、本稿では小規模なデータセットを使用して実験を行っている。そのため、データセットの規模を大きくした場合

にも同様の結果を得ることができるかどうかを検証する必要がある。それに伴い、データセットの拡張や作成方法の見直しについても検討していくことによって、より有用性の高いタイトル作成支援を行うことができると考える。また、本稿で提案した判定理由の提示手法では各クラスタの興味を惹かない要因の解釈を人手で行う必要がある。人手による解釈を行うことによって、作業者の主観による影響が少なからず現れる。読者の興味を惹くタイトルへの改善を行うためには、客観性の強い判定理由を提示できた方が好ましい。そのため今後、できる限り自動で判定理由の提示を行うことができる手法を検討していく必要がある。

謝辞 本研究の一部は JSPS 科研費 18H03342, 19H04221, 19H04218 の助成を受けたものです。

文 献

- [1] 長安義夫, 山本和英. タイトルパタンによる文書の一文概要生成. 言語処理学会第 13 回年次大会, pp. 684–687, 2007.
- [2] 金内萌, 難波英嗣, 角谷和俊. 投稿型レシピサイトにおけるレビュー情報に基づく料理タイトル自動生成. 第 8 回データ工学と情報マネジメントに関するフォーラム (DEIM 2016) 予稿集. 福岡, pp. 02–29, 2016.
- [3] 西原陽子, 砂山渡ほか. 専門分野の素人に選択される研究発表タイトルの推定. 情報処理学会論文誌, Vol. 51, No. 8, pp. 1478–1486, 2010.
- [4] Wojciech Stokowiec, Tomasz Trzciński, Krzysztof Wołk, Krzysztof Marasek, and Przemysław Rokita. Shallow reading with deep learning: Predicting popularity of online content using only its title. In *International Symposium on Methodologies for Intelligent Systems*, pp. 136–145. Springer, 2017.
- [5] 辻一明, 寶珍輝尚, 野宮浩揮ほか. 新着ツイート群からの興味をひくツイートの抽出に関する考察. 平成 23 年度 情報処理学会関西支部 支部大会 講演論文集, Vol. 2011, , 2011.
- [6] 千田恭子, 篠原靖志, 奥村学. タイトルの文型が読者の関心に及ぼす影響の分析. 自然言語処理, Vol. 12, No. 2, pp. 87–107, 2005.
- [7] 岩佐由美, 藤井千枝子. テキストマイニングで見た難病に対する関心とニーズ. 医療情報学, Vol. 37, No. 3, pp. 135–145, 2017.
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.