ニュース記事の読み方の判断支援に関する研究

前川　丈幸† 　馬　　強††

† 京都大学工学部情報学科　〒606–8501 京都府京都市左京区吉田本町
†† 京都大学大学院情報学研究科　〒606–8501 京都府京都市左京区吉田本町
E-mail: †maekawa.takeyuki.28s@st.kyoto-u.ac.jp, ††qiang@i.kyoto-u.ac.jp

あらまし　ニュース記事は政治・経済や地域社会などの重要な情報源であり，隙間時間で手早く読んだり，じっくり時間をかけて知識を得たりするなど様々な読み方が存在する．本研究では記事の内容を分析して，ユーザに対して効率的な読み方を促す手法を提案する．提案手法では分散表現や BERT(Bidirectional Encoder Representations from Transformers) を利用して，記事を読み通すのにかかる時間 (読了時間) と記事の地域性を分析し，アイコンで可視化して，ユーザの記事の読み方の判断支援を図る．本研究では読了時間と地域性の推定を行うための分類器を構築した．また，ニュースサイトのアクセスログを分析してデータセットを自動構築し，分類器の学習と評価を行った．

キーワード　ニュース，アイコン，可視化，推薦，意思決定支援

1　はじめに

ニュース記事は社会の様子や政治・経済の動向を知るための重要な情報源であり，人々の生活に根付いている．インターネット上の記事はニュースサイトやニュースアプリケーションにて閲覧でき，時間を問わず利用できる点やその手軽さにより利用者は増加している．

ニュースサイトやニュースアプリケーションでは，他媒体と異なり，情報が絶えず更新されている．ユーザは膨大な量の記事の中から読むべき記事・読みたい記事を選択する必要があり，ユーザにとって大きな負担となっている．その一因として記事選択時の情報が少ないことが挙げられる．図 1 はニュースサイト Adresseavisen [1] のスクリーンショットの一部である．図中下段は提案手法による，効率的な読み方を促すためのアイコンを表示している．[1]

多くのニュースサイトやアプリケーションでは図 1 のように，ニュース記事がタイトルと代表画像の二つの情報が示された状態で掲載されている．タイトルはその記事の大まかな内容を表し，代表画像は記事内の写真あるいは記事に関係した写真である．これらは内容に関するものであるため，ユーザは記事の概略については把握することができる．しかし，例えば，記事を読み終えるのにかかる時間は表記されておらず，タイトルや代表画像から推測することも困難である．記事を読み終えるのにかかる時間が表示されていないため，時間が足りなくて記事を読み通すことができなかった，あるいは流し読みで終わってしまい内容の理解が不十分になってしまった，といったことが起こりうる．

インターネット上のニュース記事の三つの読み方を考える．

一つ目は通勤・通学の途中をはじめ，移動時間などのいわゆる隙間時間で読む場合である．ニュースを読むのに割ける時間は少なく，一人で読むことが多い．そのため短時間で一日の

図 1　ニュースサイトの例，下段は提案手法を適用した場合のイメージ

ニュースを確認し情報を得ることを目的として，緊急性や速報性の高い記事をはじめ，簡潔で理解しやすい記事が好まれて読まれると推測される．また知るべき情報を含むニュースが優先されるため，ローカルニュースより全国ニュースが多く読まれると考えられる．

二つ目は学校・職場などの自宅ではない場所での休憩時間で読む場合である．ニュースを読むのに割ける時間は都度変化するが，長いことは少ないと考えられる．この場合ニュースは一人で読むか，もしくは周りに情報を共有する人がいることがある．そのため，時間の長さに合わせて，短いあるいは内容理解が比較的容易である記事を中心に，長いあるいは専門性が高く知識を必要とする記事もある程度読まれると推測される．ま

1：図 1 中に表示したアイコンは分類器の推定結果によるものではない．

た周りにニュースに関して会話ができる人がいると，ローカルニュースを種に話をすることもあり，ローカルニュースと全国ニュースの両方が読まれると考えられる．

三つ目は休日など時間に余裕があるときに自宅で読む場合である．ニュースを読むのに割ける時間は長く，家族など情報を共有する人がいることが多い．そのため，長いあるいは専門性が高く知識を必要とする記事を含め，幅広い種類の記事が読まれる．また，ローカルニュースと全国ニュースが両方とも読まれ，ローカルニュースが話の種となることもあると考えられる．

以上の三つの読み方を考えると，ユーザがニュース記事を選択する際に，短時間で読めるかどうかと，ローカルニュースであるかどうかを示すことが重要であるが，これら二つはタイトルと代表画像など既存のシステム・サービスが提供している情報から読み取ることは難しい．タイトルと代表画像から読み取ることができないが，ユーザが記事を選択する際に有用となる情報を表示することでユーザに対し効率的なニュース記事の読み方を促すことができるのではないかと考えた．

本研究では記事を読み終わるのにかかる時間を「読了時間」，ローカルニュースか全国ニュースであるかをニュースの「地域性」とそれぞれ表現する．本研究ではユーザの記事選択とその読み方の判断を支援するため，ニュース記事の内容からその記事の読了時間と地域性を推定し，アイコンとして表示する手法を提案する．本研究では記事の読了時間と地域性の推定を分類問題として定式化し，単語の分散表現と BERT(Bidirectional Encoder Representations from Transformers) を用いて分類器を構築する．

さらにニュースサイトのアクセスログから，ユーザがあるページにとどまった時間と，アクセスしたユーザの地域の情報を取り出し，学習用のデータセットの構築を行う．読了時間のクラス設定として読書速度を用いることが考えられる．しかし読了時間と読書速度のピアソンの積率相関係数を求めたところ，相関があまりないことが分かった．そこで読了時間はあるページにとどまった時間の分布が正規分布に従うと仮定して，各クラスが全体の 20%を占めるように境界値を決定し，クラスを設定する．地域性はアクセスしたユーザの地方の種類の数が境界値を越えるか否かでクラスを設定する．

本研究での主な貢献は以下のとおりである．

（1） ニュース記事の内容を分析して，その記事の読了時間と地域性を推定するモデルを構築している．さらにユーザのニュース記事の読み方の判断を支援するためのアイコン表示法を提案している．(3 節)

（2） ニュースサイト Adressavisen [1] の記事とアクセスログを分析し読了時間と地域性の傾向を明らかにした．また，その結果に基づいて正解ラベルを設定し，データセットを構築している．(4 節)

（3） 提案する BERT ベースの分類器と SVM の精度を比較する実験を行なった．読了時間の分類では分散表現を用いる手法が有効であるとわかった．地域性の分類では分散表現を用いない手法が有効であるとわかった．(5 節)

2 関 連 研 究

2.1 ニュース記事を読むユーザへの支援に関する研究

ニュース記事を読むユーザへの支援に関する研究は，記事の内容の理解支援を目的とするものが多い．ニュース記事の内容の因果関係を，原因と結果をノードとするネットワークを構築し表現する手法を提案する研究 [2]，ニュース記事を理解のしやすい表現に変換する手法を提案する研究 [3]，ニュース記事に対する Twitter 上の反応に含まれる，特徴的な単語を可視化することで，記事の論点をわかりやすくする手法を提案する研究 [4] などが挙げられる．

田中祥太郎ら [5] は記事の内容理解に必要と考えられる背景知識を Web 上のリソースから取得する手法を提案している．ニュース記事に含まれる人物，組織，出来事，場所，建造物といったエンティティに着目し，記事からエンティティを抽出したうえで，重要度の高いエンティティに関する知識を Web 上から取得し，ユーザに提示することで理解支援をはかるものである．

西川ら [6] は同じテーマの複数の記事に対し，記事間の関係性を可視化することで，ユーザにテーマへの理解を深める手法を提案している．トピックの階層関係を取得し，それらをノードの階層構造によって表現している．またキーワード抽出も行い，トピックの階層的関係とキーワードの二つの要素からニュースのテーマに対する理解を深めやすくするというものである．

これらの研究が記事内容の理解支援に関するものである一方，本研究は読了時間と地域性という記事内容とは異なるが，必要と考えられる情報を表示することで，記事選択の際のユーザへの支援を図るものである．

2.2 ローカルニュースの抽出に関する研究

ローカルニュースの抽出に関する研究は以下のものが挙げられる．

大倉 [7] は記事の地域性を判断する際に，ユーザのクリック実績のデータを利用し，RNN を用いて学習する手法を提案している．日本をグリッドに分割し，各グリッドでのニュース記事の配信実績とクリック実績からクリック率を求め，クリック率の平均との乖離から記事の各グリッドへの適合率を求めている．求めた適合率の分布の偏りから記事の地域性を判定している．

長城ら [8] はニュース記事をツイートした Twitter ユーザの居住地域に着目している．Twitter ユーザの居住地域の分布の偏りが大きいものをローカルニュースとし，ローカルニュースの記事の特徴語を抽出して，ナイーブベイズ分類器で学習する手法を提案している．

いずれの研究も記事に反応したユーザの分布の偏りからニュースの地域性を判定している．一方本研究ではアクセスログを分析しユーザの地域分布に基づいてラベル設定し，分類器を構築して内容から直接推定している．

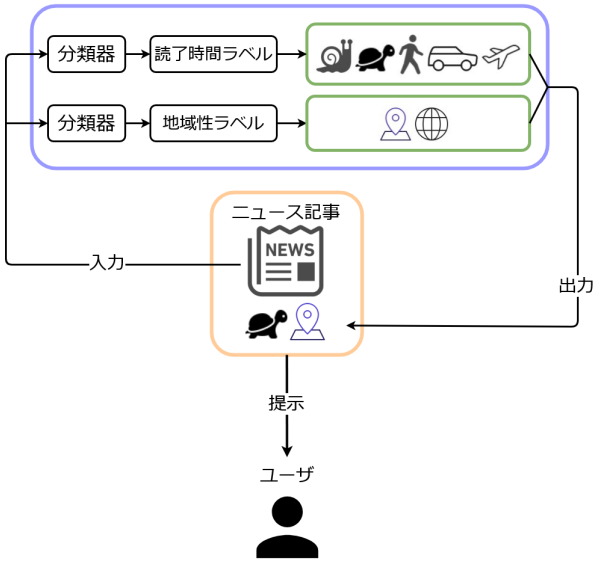

図 2　全　体　図

3　提 案 手 法

提案手法の概要を図 2 に示す．提案手法は以下の 3 ステップに分かれる．

（1）　ニュース記事の本文を分類器に入力する．

（2）　分類器が読了時間と地域性それぞれについてラベルを推定する．そして各ラベルをアイコンに変換して出力する．

（3）　アイコンを記事に付与し，ユーザに提示する．

3.1　分類器の構成と学習

分類器は BERT を用いて構成する．分散表現の利用の有無で二種類の分類器を構築しており，構成図をそれぞれ図 3，4 に示す．図 3，4 は共に図の下部から上部へ向かって処理が進むことを表している．図 3 は分散表現を利用しない分類器であり，記事本文のテキストをトークナイズして得られた単語の系列を処理する．本論文では t-BERT と表現する．図 4 は分散表現を利用する分類器であり，記事本文をトークナイズして，それぞれの単語を Word2Vec を用いてベクトル表現を得てから処理する．本論文では e-BERT と表現する．

BERT に入力できる列の最大長を L とする．t-BERT ではまず入力のテキストをトークナイザーを用いて単語に分割する．先頭 $L-2$ 単語を切り出し，先頭に [CLS]，最後尾に [SEP] という特殊なトークンを挿入し，全トークンを辞書に基づき ID に変換する．変換された ID の列を BERT に入力する．読了時間を推定する分類器の場合，BERT の出力のうち，[CLS] トークンの埋め込み表現にあたるベクトルを，5 クラスに分類する全結合層に入力し，その出力からラベルを決定する．地域性を推定する分類器の場合は 2 クラスに分類する全結合層を用いる．

e-BERT の処理の流れは t-BERT と概ね似ている．相違点は二つあり，一つはトークナイズされたテキストを Word2Vec を用いて分散表現に変換するところである．未知語は分散表現に変換せず，変換できた単語ベクトルのうち先頭 L 個を取り出し，行列とする．Word2Vec の出力と BERT の入力の次元が合わないので全結合層を通して，行列を変換し BERT に入力する．もう一つの相違点は BERT の出力の取り出し方である．入力の先頭単語の埋め込み表現にあたるベクトルを取り出し，ラベルを推定する全結合層への入力とする．

学習・検証のために各記事に対して読了時間と地域性それぞれに正解ラベルを用意する．ラベル設定の方法は 4 節で詳しく説明する．

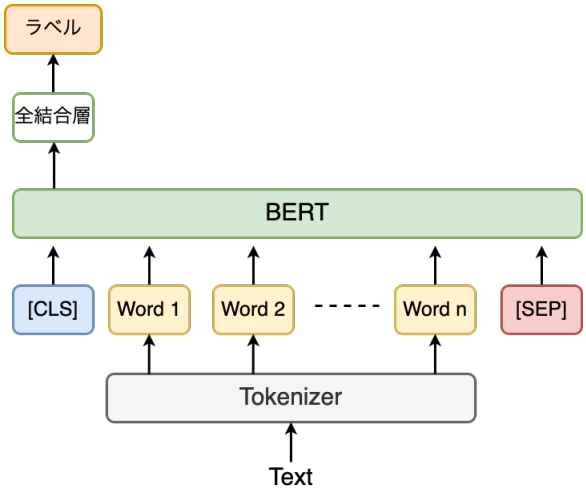

図 3　t-BERT の構成図

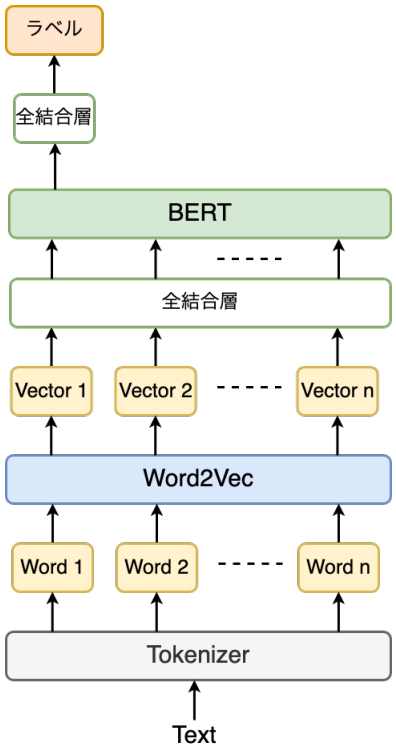

図 4　e-BERT の構成図

3.2　アイコン表示によるユーザへの支援

一目見て情報を取得できるという視認性の観点から，ラベルの可視化にあたってテキストではなくアイコンを用いる．アイコンは図 2 の右上に例示したように二種類がある．上段が読了

時間，下段が地域性のアイコンである．読了時間のアイコンは左から順に読了時間が長い，比較的長い，平均的，比較的短い，短いにそれぞれ対応する．地域性のアイコンは左から順にローカルニュースと全国ニュースにそれぞれ対応する．読了時間の左から二番目のアイコンは icons8 [9] より引用している．

読了時間のアイコンはユーザがニュース記事を読み通して情報を得るのにかかる時間の目安となる．読了時間が短いと示されている場合，ニュース記事を隙間時間，例えば通勤・通学の道中や待ち時間等で読むことができるという指標になる．読了時間が長いと示されている場合，ある程度まとまった時間がなければ読み通せないという指標になる．ユーザは読了時間のアイコンから選択した記事を今読むべきか，あるいはほかの余裕がある時間で読むのかを判断することができる．その結果効率的なニュース記事の読み方を実践することができ，ユーザにとって有益となる．

地域性のアイコンはニュースがローカルニュースか全国ニュースであるかを示す．全国ニュースは全ユーザが読むべき情報が含まれていることが多い．一方ローカルニュースはある特定の地域に関する話題について書かれており，その地域に関係のあるユーザにとっては有益なものである可能性があるが，そのほかのユーザにとってはそうではないと考えられる．ローカルニュースがどの地域に関して書かれているかは，タイトルから読み取れる場合もあり，アイコンがあることによって記事がローカルニュースであることに説得力を持たせることができる．ユーザは地域性のアイコンから選択した記事が，求めるニュースの地域性であることを確かめられ，確証をもって記事を読むことができる．

読了時間と地域性の二つのアイコンが表示された時，それらが組み合わさることによる，ニュースの読み方への影響を考える．以下の四個のケースでの読み方を考える．

読了時間：短い，地域性：全国

全国ニュースで読了時間が短い場合，緊急性や話題性の高いニュースや内容が簡潔にまとめられた記事など，全ユーザにとって読むべきものであることが考えられる．とりわけ速報性の高いニュースであるならば，短時間で読んで情報を入手する必要がある．そのため通勤・通学途中やちょっとした休憩時間など，いわゆる隙間時間で手早く読むといった読み方が想定される．

読了時間：長い，地域性：全国

全国ニュースで読了時間が長い場合，社会の動向が詳しく書かれていたり，専門性の高いニュースであることが考えられる．前者の場合，全ユーザにとって必要な情報である可能性があり，後者は対象となるユーザが限定されるが読むべき記事である可能性がある．読了時間が長いため，時間に余裕があるときにじっくりとニュース記事と向き合うという読み方が考えられる．例えば休日自宅で過ごしているときや比較的長い休憩時間などで読まれることを想定する．

読了時間：短い，地域性：ローカル

ローカルニュースで読了時間が短い場合，地域のニュースの情報を手早く手に入れることができ，ニュースのカバーする地域とユーザとの関連に関係なく有益である．ニュースのカバーする地域と関連があるユーザは，地域の情報を短時間で手に入れられるため，会話の種とすることで，ほかのユーザへ紹介することができる．ニュースのカバーする地域と関連のないユーザであっても，読了時間の短さから手軽に読むことができる．そのためユーザがその地域に興味関心を持つきっかけになることが考えられる．読み方として，例えば自宅ではない場所での空き時間などに読むことを想定する．

読了時間：長い，地域性：ローカル

ローカルニュースで読了時間が長い場合，ある地域に関する話題が比較的詳しく書かれている可能性がある．この場合記事を読むと，ある地域に関する知識を深めることができる．例えば休日家族で地元についての理解を深めるために読むという読み方ができる．

4　データセットの構築

データセットの分析と読了時間・地域性推定の学習のための正解ラベル設定を行い，学習用データセットを構築する．一般に記事のアクセスログは入手することが難しい．また配信される記事をユーザがアクセスする前にその記事の読了時間と地域性を推定する必要があり，提案手法は記事のアクセスログが存在しない場合に利用されることが多いと考えられる．そのため，アクセスログを用いない内容ベースの手法が必要である．本研究ではアクセスログからラベルを自動設定し，内容ベースの分類器を学習する．

4.1　データセット

使用するデータセットは Adressa News Dataset [10] である．このデータセットには，ノルウェーのニュースサイト Adressavisen [1] の記事とそのメタデータに加えて，2017 年 1 月 1 日から 3 月 31 日までの 90 日間のアクセスログが含まれている．記事とそのメタデータ，アクセスログの二つの JSON ファイル群から構成されている．

表 1 にデータセットの内容のうち本研究に関連するものを示す．記事本文はデータセット中の記事とそのメタデータを格納する JSON ファイルからキーを fields，field，body の順で指定することで得られる．active time はユーザがある記事のページに滞在した時間で，アクセスログの JSON ファイルからキーを activeTime で指定することで得られる．region はユーザがアクセスした端末の IP アドレスから割り出された地域で，日本の都道府県に相当する．アクセスログの JSON ファイルからキーを region で指定することで得られる．

4.2　読了時間のラベル設定

読了時間のラベルを設定するために，アクセスログ中の active time というデータを用いる．図 5 はデータセット中のある記事の active time の分布を示したものである．図 5 の分布の左端の度数が最も高くなっている．active time が短いデータは記事をクリックしてページを開いたものの，目当てのものと違いすぐに閲覧をやめた，最初の数行だけを読んだ，あるいは全体を

表 1　データセットの内容 ([10] より抜粋し改変)

データ区分	キー名	内容
記事	fields > field > body	記事本文
アクセスログ	activeTime	ユーザがある記事のページにとどまった時間 (秒)
	region	ユーザのアクセスした端末のIP アドレスから判定される region日本の都道府県に相当する

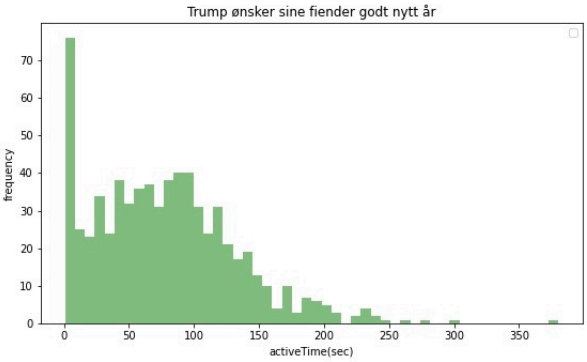

図 5　ある記事の active time の分布

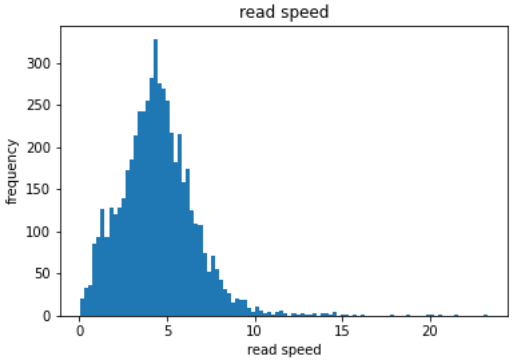

図 6　読書速度の分布

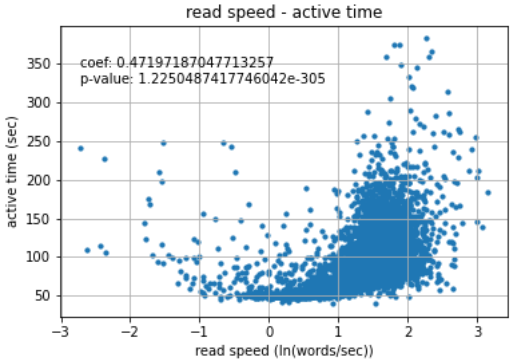

図 7　読書速度と読了時間の関係

流し読みしたなどのログである可能性が高い．これは記事を読み通したとは考えにくいため，分類器の学習に使うデータからは除外する．ゆえに閾値を設定し，その値を超えるものの平均値を記事の active time の値として使用することにする．また記事の中には active time のデータ数が極端に少なく，記事を読み通したとは考えられないような，active time が数秒程度であるものばかりが含まれているものがある．したがって active time のデータを 100 以上持っている記事のみを学習データとして用いる．

選択した記事の active time の代表値として，上位 60%のデータの平均値と，30 秒以上のデータの平均値を採用する．それぞれ上位 60%のログのデータ，30 秒以上のログのデータは記事を読み通したものであるという仮定に基づく．以下本論文ではそれぞれ「上位 60%」,「30 秒以上」と表現する．ここで設定した active time の代表値を用いて，読了時間の正解ラベルの設定を行う．

読了時間のラベル設定について以下の三つを比較する．

- 読了時間を分刻みで基準とする単純な設定
- 読書速度の分布を正規分布として捉えて利用する設定
- 読了時間の分布を正規分布として捉えて利用する設定

読了時間は 5 クラスのラベルを付与する．それぞれ，読了時間が短い，比較的短い，平均的，比較的長い，長いという意味を持つ．

読了時間を分刻みで基準とする単純な設定

読了時間を分刻みで基準とする単純なラベル設定を考える．読了時間 1 分未満，1 分以上 2 分未満，2 分以上 3 分未満，3 分以上 4 分未満，4 分以上の 5 クラスとする．この設定はデータセットによらないため，例えばほぼすべての記事が 5 分以上読了にかかる場合，極端な偏りが生じる．クラス間で下図の偏りがあるとユーザの判断支援にならないため，ラベル設定の基準として相応しくない．

読書速度の分布を正規分布として捉えて利用する設定

読書速度を利用してラベル設定を行うことを考える．読書速度とは一定時間にどれほどユーザが記事を読み進めるかを測った速度のことである．ここでは記事の長さ (単語数) をその記事の active time の代表値で割った値のことを指し，単位は words/sec である．

図 6 はデータセット中の記事の読書速度の分布を示したものである．この分布では記事の長さを 30 秒以上の active time の代表値で割った値を用いている．この分布が正規分布に従うという仮定のもとでラベル設定を行うことができると考える．しかし読書速度と読了時間の関係を考える必要がある．

読書速度が大きい記事は，読みやすい，すなわち文章が平易であると推測される．読了時間はユーザが記事を読み通すのにかかる時間であるため，読了時間は記事の読みやすさと分量の二つの要素からなっていると考える．ゆえに読書速度が大きく分量が多い記事と，読書速度が小さく分量が少ない記事は，読了時間が同程度であると推測する．

図 7 に読書速度と読了時間の関係を示す．図 7 では横軸の active time は見やすくするため自然対数をとっている．30 秒以上の active time の代表値を用いた読書速度と，読了時間のピアソンの積率相関係数を r とする．$r \simeq 0.47$ であるため両者に相関はあまりないといえる．読書速度と読了時間は単純な比

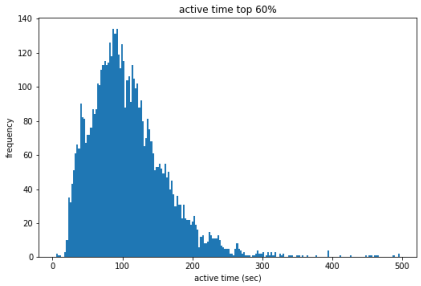

図 8　active time の代表値の分布 (上位 60%)

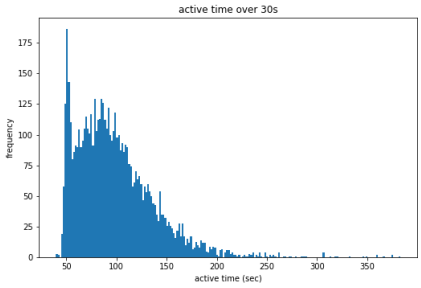

図 9　active time の代表値の分布 (30 秒以上)

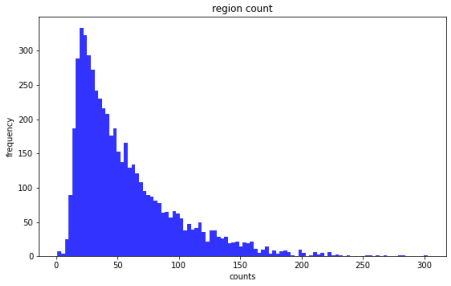

図 10　region の分布

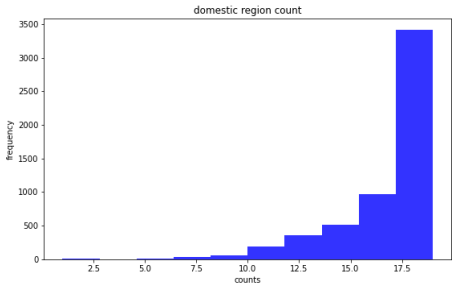

図 11　region の分布 (国内のみ)

例関係であるとはいえないことがわかった．よってこの読書速度を利用した設定は用いない．

読了時間の分布を正規分布として捉えて利用する設定

active time の代表値の分布を図 8 と 9 に示す．順に上位 60%，30 秒以上の場合を表す．この分布が正規分布 $N(\mu, \sigma^2)$ に従うという仮定のもとでラベル設定をすることを考える．

30 秒以上の分布において，60 秒あたりの度数が最も高くなっている．これは上位 60%の分布では現れなかったもので，データの大半が 30 秒以下である記事のものと考えられる．

ラベル設定においてはクラス間の偏りをなくすために，各クラスが全体の 20%をそれぞれ占めるようにする．標準正規分布表を参考に基準を $\mu \pm 0.26\sigma$ と $\mu \pm 0.85\sigma$ の四つとする．これを上位 60%と 30 秒以上の二つの active time の分布に対してそれぞれ適用し，ラベル設定を行う．

各クラス間でデータ数の偏りがある場合，多くの記事でアイコンが同じものになり，ユーザへ与える情報の意味が薄れてしまう．本研究の目的はユーザの判断支援であるため，クラス間のデータ数の偏りがあまり発生しない方法を選択する必要がある．以上の三つの設定を考えたうえで，読了時間の分布を正規分布として捉えて利用する設定を，読了時間のラベル設定とする．

4.3　地域性のラベル設定

地域性のラベル設定を行う．本研究では地域性は 2 クラスのラベルを付与し，ローカルニュース，全国ニュースをそれぞれ表す．

データセットのアクセスログのうち region のデータを用いる．ノルウェーにはデータセットの記事が配信された 2017 年当時 19 の region が存在していた．各記事の region のデータに含まれている region の種類を数えて代表値とする．

図 10 は region の代表値の分布を示したものである．横軸は各記事にアクセスのあった region の種類の数を表し，縦軸はその度数である．図 10 が示すように，ノルウェー国外からのアクセスも多く，19 より大きな値が多く存在する．よって国外からのアクセスを考慮して設定を行う必要がある．また読了時間のラベル設定とは異なり，クラス間のデータ数の偏りを考慮する必要はない．以下の二つの方法でクラス分類する．

- 国内の region のみを考慮する設定
- 全ての region を考慮する設定

国内の region のみを考慮する設定

記事の region のデータのうち，国内の region のみに着目し，その種類を数えて，記事の region の値とする．図 11 はこの場合の region の値の分布である．19 に近いデータが多いことがわかる．region の数 15 をしきい値として 15 より大きい場合を全国ニュース，15 以下をローカルニュースとする．これは 4 つ以上の region からのアクセスがない記事はローカルニュースであるという仮定に基づく．以下本論文ではこのラベル設定方法を「国内のみ」手法と表現する．

全ての region を考慮する設定

国内外を問わず記事の region のデータを使用するが，国内の region に二倍の重みをつけて種類を数える．図 12 がこの場合の region の値の分布である．重み付けをして数えた値を記事の region の値とし，region の数 50 をしきい値として 2 クラスに分ける．50 以上を全国ニュース，50 未満をローカルニュースとする．以下本論文ではこのラベル設定方法を「重みづけ」手法と表現する．

5　実　　験

ニュース記事本文から記事の読了時間と地域性を推定する分

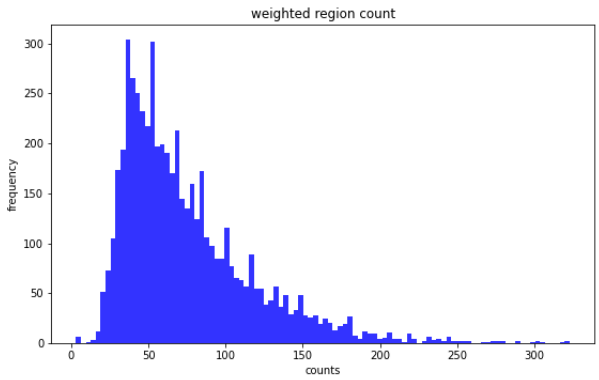

図 12　region の分布 (重み付け)

類器の学習を行い，その精度を確かめる実験を行う．4 節で述べた方法を用いて，データセットの各記事に対して読了時間と地域性それぞれにラベルを付与する．

5.1　設　　定

読了時間は 5 クラスに分類し，地域性は 2 クラスに分類する．分類器の精度は以下の三つの手法で比較する．

e-BERT　本研究で提案する分散表現に変換したテキストを入力とする BERT ベースの手法

t-BERT　本研究で提案するテキストを入力とする BERT ベースの手法

SVM　分散表現に変換したテキストを特徴量とする SVM

読了時間の分類精度の評価は，分類器のテストデータに対する予測精度によって行う．地域性はクラス間のデータ数の偏りが大きく，全国ニュースが全体の大きな割合を占めるため，クラス間のデータ数がほぼ等しくなるように，データセットからランダムに抽出したものを用いて学習・評価を行う．地域性の分類精度は推定結果の ROC 曲線の AUC により評価する．

5.2　実験結果と考察

5.2.1　読 了 時 間

読了時間の分類精度を表 2 に示す．active time の代表値の設定に関わらず，BERT ベースの提案手法は SVM より高い精度を示した．また BERT ベースの手法同士では e-BERT の精度が高いことがわかった．読了時間の分類においては BERT による分類器の構築が有用であると考えられる．さらに単語の分散表現を用いた入力の変換も有用であると考えられる．active time の代表値の設定による精度の差は 30 秒以上の設定が，上位 60%の設定よりわずかに上回ったため，前者の設定がより有用であると推察される．

読了時間のラベル設定における問題点は，記事を真に読み通した読者の読了時間がアクセスログから推測することが難しいということである．本研究では上位 60%と 30 秒以上の active time のデータの平均値を代表値としているが，どちらにもデメリットがある．上位 60%という基準は，読書の六割が記事を読み通しているという仮定のもとで設定したが，それが全記事にあてはまるわけではない．実際，上位 60%の active time の分布では 30 秒以下のデータが存在する．最小値は 4.6 秒で，これ

表 2　active time の分類精度

手法	ラベル設定	精度 (テストデータ)	精度 (学習データ)
e-BERT	上位 60%	**51.08%**	58.38%
t-BERT	上位 60%	47.35%	61.91%
SVM	上位 60%	37.80%	43.50%
e-BERT	30 秒以上	**53.97%**	60.98%
t-BERT	30 秒以上	49.10%	73.81%
SVM	30 秒以上	36.63%	41.34%

表 3　region の分類精度

手法	ラベル設定	AUC(テストデータ)	AUC(学習データ)
e-BERT	国内のみ	0.593	0.709
t-BERT	国内のみ	**0.622**	0.759
SVM	国内のみ	0.575	0.935
e-BERT	重み付け	0.613	0.647
t-BERT	重み付け	**0.652**	0.772
SVM	重み付け	0.575	0.901

はほぼすべてのログが 10 秒程度であることが原因である．また 30 秒以上という基準は，30 秒に満たない時間記事を読んだユーザが記事を読み通したとは考えにくいという仮定に基づいている．しかし記事を読み通しても 30 秒かからないものが存在することも考えられる．したがって明らかに記事を読み通したとは考えられないデータを排除したうえで，分布を考慮して，ある割合の上位データの代表値を，読了時間として採用する手法を検討する必要がある．

また精度に影響を与えていると考えられるのは，クラス幅が小さいことである．上位 60%は $\sigma \simeq 54.6$ 秒，30 秒以上は $\sigma \simeq 39.3$ 秒 であり，クラスの幅は両端のクラスを除きそれぞれ 30 秒前後となる．十数秒程度の差をもつ記事同士はあまり記事内容に差を見出すことができないが，分類問題として扱っているため，境界付近の値をもつデータが隣のクラスに誤分類されることが多い．クラスの数を減らす，あるいはデータセットを大規模なものにする必要がある．

5.2.2　地　域　性

地域性の分類精度を表 3 に示す．地域性の分類精度は BERT ベースの手法が SVM を AUC で上回っている．また t-BERT が e-BERT を AUC で上回っている．地域性の分類においてはより単純な構造である t-BERT が分類器として優れていると考える．

一方で地域性の正解ラベルの設定には改善の余地があると考える．本研究では記事にアクセスがあった region の種類を二通りの方法で数えて正解ラベルを設定している．region の種類だけを数えたことは不十分であると考える．記事のログの中での region の偏りを考慮していなかったため，例えばオスロ (ノルウェーの首都で最大の都市) からのアクセスが 1 件と 100 件の場合では，その記事のオスロからの注目度が異なることが予想されるが，本研究ではどちらも等しく region の値が 1 可算されるだけである．人口の分布を考慮し，各 region のログの数からどの region に分布が偏っているかを考える必要がある．

6 おわりに

6.1 まとめ

本研究ではユーザの記事の読み方の判断支援のために，記事の読了時間と地域性を分析して，アイコンによって表示する手法を提案している．アイコンによりユーザにニュースの効率的な読み方を促す．

アイコン表示の為にニュース記事本文から読了時間と地域性を推定する分類器を構成し学習させた．分類器は BERT をベースとして分散表現の利用の有無による二種類を構築した．分散表現を用いるものを e-BERT，用いないものを t-BERT とした．

分類器の学習のために，ニュースサイトのアクセスログを利用して，各記事に対して読了時間と地域性のそれぞれに正解ラベルを設定した．読了時間にはユーザが記事のページにとどまった時間 (active time) を用いた．active time の分布は数秒程度のデータが最も多い．これは記事を読み通したものとは考えられず，学習に用いるのは不適当とした．したがって active time の代表値を基準を設けたうえで，基準以上の値の平均値とすることにした．読了時間のラベル設定では，アイコン表示によるユーザへの支援を適切に行うために，クラス間で偏りを少なくすることが求められる．そこで読了時間の分布が正規分布に従うと仮定して，各クラスが全体の 20%を占めるようにクラスの境界値を設定した．地域性のラベル設定では国外からのアクセスを考慮にいれて行った．国内からのアクセスログの region の種類と，国内からのアクセスに重みをつけて数える二種類の方法を採用した．

実験では構築したデータセットを用いて，e-BERT と t-BERT，SVM の三つの手法で分類精度を比較した．実験の結果，読了時間の分類は，e-BERT，t-BERT，SVM の順で高い精度を記録した．読了時間の分類は提案手法の BERT ベースで，分散表現を用いる手法が有効であることを示すことができた．地域性の分類は t-BERT，e-BERT，SVM の順で高い精度を示し，BERT ベースの提案手法が有用であることがわかった．

6.2 今後の課題

今後の課題として以下の三点を挙げる．

一つ目はラベル設定の改善である．読了時間は active time の代表値を，ある基準値以上のデータのうち上位数割を分布を考慮して取り出し，それらの平均値とする手法で改善を試みる．地域性は各記事のアクセスログの region の分布と，実際の人口分布から，どの地域から関心を集めているのかという情報を取り出して改善を試みる．

二つ目はアイコン表示の個人化である．本研究では全ユーザに対して同じ記事に同じアイコンを表示することを提案している．しかし読了時間と地域性は本来個々のユーザにとって異なるものである．読了時間は同じ記事であっても個人差があり，かつ同じユーザでも記事のトピックに対する知識の量の違いによって，異なる読了時間を示すと考えられる．地域性はユーザと関係のある地域・興味関心のある地域のニュースが好まれて読まれると考えられるが，提案手法ではそれが考慮されていない．したがって読了時間は記事の読みやすさと記事のトピックに対するユーザの理解度の違いを考慮して，ユーザごとに適切な読了時間を推定することを検討していく予定である．地域性は記事の関係のある地域を推定したうえで，ユーザの興味のある地域を推定し，ローカルニュースを二つに分けて，ユーザと関連度が高い・低いという表示の仕方を検討していく予定である．

三つ目はユーザ実験である．アイコンを実際に表示することで，ユーザへニュース記事の読み方の判断支援を行うことができるのかという検証を本研究では行えなかった．提案手法に改良を加えて分類器の精度を向上させ，そのうえでユーザ実験を実施する．

謝　辞

本研究の一部は科研費（19H04116）と総務省 SCOPE（201607008）による．

文　献

[1] Adressavisen. https://www.adressa.no/, 2021 年 1 月 27 日閲覧.
[2] 石井裕志, 馬強, 吉川正俊. 因果関係ネットワークの構築によるニュースの理解支援. 第 1 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 Forum), 2009.
[3] 田中英輝, 美野秀弥. ニュースのためのやさしい日本語とその外国人日本語学習者への効果. 情報処理学会論文誌, Vol. 57, No. 10, pp. 2284–2297, 2016.
[4] 池田将, 牛尼剛聡. Twitter の反応を用いたニュース全体像の理解支援のための可視化手法. 研究報告データベースシステム (DBS), Vol. 2019, No. 5, pp. 1–6, 2019.
[5] 田中祥太郎, ヤトフトアダム, 田中克己. ニュース記事の理解支援のための背景知識抽出と補完 (データ工学). 電子情報通信学会技術研究報告= IEICE technical report: 信学技報, Vol. 114, No. 173, pp. 95–100, 2014.
[6] 西川奈都月, 盛山将広, 内藤峻, 松下光範. 初学者を対象としたニュース記事中のトピックの関係性に基づく可視化インタフェースの提案. *SIG-AM*, Vol. 15, No. 10, pp. 62–67, 2017.
[7] 大倉俊平. 閲覧実績を用いたニュース記事の地域性抽出. 人工知能学会全国大会論文集 第 31 回全国大会 (2017), p. 3G21. 一般社団法人 人工知能学会, 2017.
[8] 長城沙樹, 山口祐人, 北川博之, 天笠俊之. Twitter 上の社会・地理・内容的特徴を用いたローカルニュースの抽出. 第 78 回全国大会講演論文集, Vol. 2016, No. 1, pp. 535–536, 2016.
[9] icons8. https://icons8.jp/, 2021 年 1 月 27 日閲覧.
[10] Jon Atle Gulla, Lemei Zhang, Peng Liu, Özlem Özgöbek, and Xiaomeng Su. The adressa dataset for news recommendation. In *Proceedings of the international conference on web intelligence*, pp. 1042–1048, 2017.