1ZF-04

小学校での NIE 教材に適した Web ニュース記事の判定手法の検討

関　伸也†　　安藤　一秋‡

香川大学大学院工学研究科†　　香川大学創造工学部‡

1.　はじめに

NIE（Newspaper in Education）は，新聞を教材として活用する取り組みのことであり，小学校や中学校を中心に，幅広い教育機関で実施されている．各地域の NIE 推進協議会は，NIE 実践報告書[1]を毎年発行している．小学校での実践報告によると，地元や周辺地域に関する記事を用い，地域学習として NIE を実践している例が多い．また，NIE を継続的に行うことで，児童の読解力・語彙力や社会への興味関心が向上したと述べられている．さらに，物事を多角的にとらえる力の育成に効果があることも確認されている．

小学校での NIE では，各新聞社が発行している紙媒体の新聞や Web ニュースを使用している．しかし，これらの記事には，児童が学習していない漢字や，理解することが困難な表現などが使われており，児童にとって難解な文章であることが多い．そのため，日々発行される膨大な記事の中から，児童自身が自分の興味や課題に適した記事を探し出すことが困難という課題がある．また，教師にとっては，NIE に適した記事を探すことや，記事に関する教材研究によって，教材準備の負担が増加するといった課題がある．したがって，記事の推薦や重要語の提示，難解な言葉の言い換えなど，児童に対する支援と，NIE に活用しやすい記事や関連・補足資料の検索・選択など，教師に対する支援が必要になる．そのため，新聞記事に関する読解支援[2]や児童に対する記事推薦[3]などの研究が行われている．

本研究では，NIE を実践する教師向けの支援に注目し，NIE に活用しやすい Web ニュース記事を地域学習の教材集として推薦するシステムの構築を目的とする．NIE 教材としての価値が高い記事を推薦することで，教師が記事を探す負荷を軽減し，NIE を継続的に実践するための支援が可能と考える．

本稿では，NIE 教材に適した記事を SVM（Support Vector Machine）で判定する手法について検討する．

Examination of a Method for Determining Web News Suitable for NIE in Elementary Schools
Shinya Seki†
Kazuaki Ando‡
† Graduate School of Engineering, Kagawa University
‡ Faculty of Engineering and Design, Kagawa University

2.　NIE に適した記事

SVM は教師あり機械学習手法であるため，NIE に適した記事を判定するためには，実際に NIE に適した記事を用いて SVM の学習を行う必要がある．本研究では，学習に用いる記事として，NIE ワークシートの記事を用いる．

NIE ワークシートとは，新聞社が独自に作成し，Web 上で公開している NIE に対する補助教材である．このワークシートに採用された記事は，NIE の経験が豊富な教師の意見を基に選定されていることから，NIE 教材として十分な質の記事であると考えられる．

3.　SVM を用いた記事判定手法

SVM はベクトル表現を入力とし，二値の判定結果を出力する機械学習モデルである．そのため，事前に記事から特徴量を抽出し，記事ごとのベクトルを生成する必要がある．

我々の先行研究[4]では，記事が NIE に適しているかどうかを判断する基準として，記事内容と記事本文の読解難易度の 2 つの観点に注目した．記事内容は記事本文の BoW（Bag of Words）と形態素のカテゴリ情報から，読解難易度はリーダビリティスコア[5]と学習済みの漢字割合から特徴量を抽出し，記事ベクトルの素性とした．

本稿では，先行研究で用いた素性に加えて，新たに以下の 4 つの素性を提案する．

① 日本語教育語彙表[6]の語彙レベル

日本語教育語彙表には約 18,000 語の日本語教育用の語彙が収録されており，初級前半から上級後半の 6 段階の語彙難易度が付与されている．記事本文に出現する形態素の難易度を確認し，6 段階＋不明の 7 次元を素性に利用する．なお，素性値は各難易度の出現割合とする．

② 日本語語彙大系[7]の体言カテゴリ情報

日本語語彙大系から取得できるカテゴリ情報の内，体言のみを素性に用いる．日本語語彙大系には，2,840 種類の体言カテゴリが存在する．2,840 次元を素性に利用し，素性値として記事ごとの体言カテゴリの出現割合を用いる．なお，先行研究で用いた Juman++[8]の形態素カテゴリ情報から生成した素性と組み合わせて用いる．

③ 単語分散表現 W2V (Word2Vec)
先行研究で用いた BoW の代替として，単語分散表現を用いる．2018 年 11 月 26 日時点の Wikipedia の日本語ページ全文のデータを用いて分散表現を学習し，300 次元の単語ベクトルを生成する．そして，記事本文を 300 次元の素性で表現し，素性値には記事本文に含まれる全単語の W2V ベクトルの平均を用いる．

④ Juman++の名詞形態素の意味ドメイン情報
先行研究で使用した Juman++の形態素カテゴリ情報と同様，解析結果に含まれる名詞形態素の意味ドメイン 13 種を素性とし，その出現割合を素性値として用いる．

以上の素性を組み合わせて記事ベクトルを生成し，SVM での学習・判定に利用する．

4. 実験

十分割交差検証により，各素性の有効性と，記事判定に最適な素性の組み合わせを検討する．なお，日々膨大なニュース記事が発行されることから，本稿では適合率を重視して有効性を判定する．

実験には，先行研究と同様，神戸新聞社の NIE ワークシートの 122 記事を正例データとして，Web ニュースサイトから収集した 122 記事を負例データとして用いる．

BoW あるいは W2V を基本素性とし，以下の 5 つの素性を追加する形で，最適な素性の組み合わせを確認する．

R：リーダビリティスコア
K：習得済み漢字割合
C：Juman++と日本語語彙大系のカテゴリ情報
G：日本語教育語彙表の語彙レベル
D：Juman++のドメイン情報

基本素性に 1 つの素性のみを加えた場合の判定結果を表 2 と表 3 に示す．また，基本素性に対して素性を組み合わせた結果の内，適合率が高い上位 5 件を表 4 と表 5 に示す．

4 つの表より，W2V を基本素性とし，リーダビリティスコア（R）と漢字割合（K），ドメイン情報（D）を追加素性として用いた場合，最も高い適合率（86.6%）で判定できることを確認した．

個々の素性の有効性に関して考察する．カテゴリ情報（C）の素性は，形態素の意味をとらえることができる W2V に対して効果が薄く，形態素の出現頻度で生成される BoW に対しては効果を発揮することを確認した．このことから，意味に関する素性は，記事本文の内容をとらえるうえで有効に働くと考えられる．日本語教育語彙表の語彙レベル（G）の素性は，BoW と W2V の両方で判定性能が向上していることから，記事本文の難易度をとらえるうえで有効な素性であると考えられる．

表2：BoW+1素性の判定結果

素性	適合率	再現率	F値
BoW	0.815	0.786	0.795
BoW+R	0.817	0.794	0.800
BoW+K	0.809	0.801	0.801
BoW+C	0.858	0.882	0.866
BoW+G	0.817	0.801	0.804
BoW+D	0.812	0.826	0.815

表3：W2V＋1素性の判定結果

素性	適合率	再現率	F値
W2V	0.853	0.864	0.855
W2V+R	0.851	0.847	0.844
W2V+K	0.847	0.864	0.851
W2V+C	0.848	0.801	0.822
W2V+G	0.861	0.836	0.844
W2V+D	0.850	0.856	0.851

表4：BoW上位5件の判定結果

素性	適合率	再現率	F値
BoW+C	0.858	0.882	0.866
BoW+CG	0.858	0.882	0.866
BoW+RCG	0.858	0.882	0.866
BoW+RC	0.853	0.882	0.864
BoW+KCG	0.852	0.882	0.863

表5：W2V上位5件の判定結果

素性	適合率	再現率	F値
W2V+RKD	0.866	0.864	0.860
W2V+RD	0.865	0.864	0.860
W2V+G	0.861	0.836	0.844
W2V+RKCGD	0.860	0.830	0.843
W2V+RGD	0.856	0.835	0.842

5. おわりに

本稿では，小学校の教師に NIE 教材として利用できる Web ニュース記事を推薦することを目的に，NIE に適した記事を判定する手法を検討した．

我々の先行研究で用いた記事内容と読解難易度に関する素性に，4 つの素性を新たに追加し，素性の組み合わせを 10 分割交差検証で評価した．実験の結果，単語分散表現にリーダビリティスコアと習得済み漢字割合そしてドメイン情報を素性とした際，86.6%の適合率で NIE に適した記事を判定できることを確認した．また，組み合わせごとの精度の変化から，新規素性の有効性について確認した．

今後は，教科書内容を基にした素性を検討し，判定した記事の推薦システムを設計・開発する．

謝辞

本研究の一部は JSPS 科研費 16K00478 の助成を受けて実施した

参考文献

[1] NIE 実践報告書，http://nie.jp/report/panflet, 2017 年 7 月 26 日確認
[2] 河村，安藤，“小学生を対象とした Web ニュース読解支援システムのための重要語抽出手法の検討”，JSAI2017 大会論文集，1J1-5，2017.
[3] S. Tanaka, K. Ando, “Web News Recommendation for Elementary School Children using Degree of SNS Users’ Attention and Popular Search Queries among Children”, ACIS International Journal of Computer & Information Science, Vol.17, No.1, pp.17-23, 2016.
[4] 関，安藤，”小学校教師に対する Web ニュース推薦のための NIE 教材に適した記事判定法の検討”，JSISE2018 大会論文集，A4-4，2018.
[5] 李，“日本語教育のための文書難易度に関する研究”，早稲田日本語教育学，第21号，pp.1-16，2016.
[6] 日本語教育語彙表，http://jhlee.sakura.ne.jp/JEV.html，2018 年 6 月 8 日確認
[7] 池原他，日本語語彙大系 CD-ROM 版，岩波書店，1999.
[8] Juman++, http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN++, 2018 年 6 月 8 日確認